TOBB EKONOMİ VE TEKNOLOJİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

VERİ MADENCİLİĞİ TEKNİKLERİ KULLANARAK BİR İLAÇ SINIFLANDIRMA ÇATISI GERÇEKLEŞTİRİMİ

DOKTORA TEZİ

Aytun ONAY

Bilgisayar Mühendisliği Anabilim Dalı

Tez Danışmanı: Doç. Dr. Osman ABUL

AĞUSTOS 2017

Fen Bilimleri Enstitüsü Onayı

……..
Prof. Dr. Osman EROĞUL
Müdür

Bu tezin Doktora derecesinin tüm gereksininlerini sağladığını onaylarım.

……
Doç. Dr. Oğuz ERGİN
Anabilimdalı Başkanı

TOBB ETÜ, Fen Bilimleri Enstitüsü'nün 091111017 numaralı Doktora Öğrencisi **Aytun Onay**'ın ilgili yönetmeliklerin belirlediği gerekli tüm şartları yerine getirdikten sonra hazırladığı **"VERİ MADENCİLİĞİ TEKNİKLERİ KULLANARAK BİR İLAÇ SINIFLANDIRMA ÇATISI GERÇEKLEŞTİRİMİ"** başlıklı tezi **09.08.2017** tarihinde aşağıda imzaları olan jüri tarafından kabul edilmistir.

Tez Danışmanı :	**Doç. Dr. Osman ABUL** TOBB Ekonomi ve Teknoloji Üniversitesi	.....
Jüri Üyeleri :	**Prof. Dr. Hasan OĞUL (Başkan)** Başkent Üniversitesi	.....
	Doç. Dr. Pınar KARAGÖZ Orta Doğu Teknik Üniversitesi	.....
	Yrd. Doç. Dr. Ersin Emre ÖREN TOBB Ekonomi ve Teknoloji Üniversitesi	.....
	Yrd. Doç. Dr. Mehmet TAN TOBB Ekonomi ve Teknoloji Üniversitesi	.....

TEZ BİLDİRİMİ

Tez içindeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edilerek sunulduğunu, alıntı yapılan kaynaklara eksiksiz atıf yapıldığını, referansların tam olarak belirtildiğini ve ayrıca bu tezin TOBB ETÜ Fen Bilimleri Enstitüsü tez yazım kurallarına uygun olarak hazırlandığını bildiririm.

.

Aytun ONAY

ÖZET

Doktora Tezi

VERİ MADENCİLİĞİ TEKNİKLERİ KULLANARAK BİR İLAÇ SINIFLANDIRMA ÇATISI GERÇEKLEŞTİRİMİ

Aytun ONAY

TOBB Ekonomi ve Teknoloji Üniveritesi
Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Danışman: Doç. Dr. Osman ABUL

Tarih: Ağustos 2017

Aday ilaç moleküllerinin makine öğrenmesi metotlarını kullanarak sanal olarak taranması ilaçların ters yan etkilerinden korunmak amacıyla ilaç endüstrisinde kilit bir rol oynar. Hesaplamalı sınıflandırma metotları onaylanmış ilaçları geri çekilenlerden ayırabilir. Çalışmamızda ilaçlar üzerinde üç farklı uygulamaya odaklandık. Onaylanmış ilaçları geri çekilen ilaçlardan ayırmak amacıyla farklı makine öğrenmesi stratejileri kullandık. Öncelikle çalışmada yer alan her bir ilaç molekülü için sınıflandırma ve öznitelik seçimi problemlerinde kullanılmak üzere ToxPrint Kemotip, global moleküler, boyut ve şekil olmak üzere 760 moleküler tanımlayıcı hesaplandı. İlk uygulamada 400'den fazla sinir sistemi ve farklı hastalık gruplarına ait ilaçları onaylanmış ve geri çekilen kategorilerine ayırmak için SVM ve topluluk metotları ilaç veri setleri üzerine uygulandı. Test setleri için doğruluk oranı 0.74 ile 0.89 elde edildi. Burada ilaç veri setleri üzerinde uygulanan özellik seçimi metotları sınıflandırma performansını arttırdı. Sinir sistemi ilaçları için bir model oluşturmada the number of total chemotypes, bond CN_amine_aliphatic_ generic, XlogP, aspheric: Cor3D:ori1ve Bonds tanımlayıcıları etkin özellikler olarak belirlendi. Bunun yanında ilaç veri setlerine gSpan algoritması uygulayarak geri çekilen sinir sistemi ilaçlarının minimum % 60'ında bulunan fragmanlar belirlendi.

Çalışma spesifik bir hastalığa ait ilaçlardan oluşan veri setlerinde geri çekilen ilaçları onaylanmış olanlardan ayırmada yapılan ilk çalışmadır. Çalışmanın diğer bölümünde farklı hastalık gruplarına ait 558 ilaç hiyerarşik çoklu etiket sınıflaması ile Clus-HMC-Ens algoritması kullanılıp 3 temel seviyede sınıflandırıldı. Birinci seviye bütün ilaçları, ikinci seviye ise 3 gruptan oluşan ilaçları içermektedir. Bunlardan ilki onaylanmış sinir sistemi ilaçları, ikincisi farklı hastalık gruplarına ait onaylanmış ilaçları ve sonuncu grup ise piyasadan geri çekilen ilaçları içermektedir. Son seviye ise sinir sistemi ilaçlarının Anatomik Terapötik Kimyasal sınıflamasına göre beş gruptan ilaç içermektedir. Bu uygulamada ilaçları hiyerarşik olarak sınıflandırmada geliştirilen modeller için seçilen parametreler FTest, w_0, k, sınıflandırma eşiği, m-estimate modelin tahmin performansını arttırdı.

Çalışmanın son kısmında 1200'den fazla onaylanmış/geri çekilen ilaç çalışıldı. Sınıflandırma modellerinde etkin olan moleküler tanımlayıcılar tezde önerilen etkin öznitelik seçme stratejisi ile belirlendi. Bunlardan ToxPrint kemotiplerden olanlar ilaç molekülleri için bir dizi kurallar belirlemede kullanıldı. İlaç veri setlerinde sadece onaylanmış/geri çekilen ilaçlarda bulunan/bulunmayan kemotipler analiz edildi. bond:NN_hydrazine_alkyl_HH2 yalnızca geri çekilen ilaçların yapısında, bond:P=O_phosphorus_oxo,bond:PC_phosphorus_organo_generic,group:carbohydrate_aldohexose, group:carbohydrate_aldopentose, group:carbohydrate_hexopyranose_fructose, group:carbohydrate_hexopyranose_glucose vb. kemotipleri yalnızca onaylanmış ilaçların yapısında gözlendi. Dengesiz ilaç veri seti üzerinde sınıflandırıcı topluluk tasarımı için bir model önerildi. İlaçları onaylanmış ve geri çekilen sınıflarına ayırmada test seti için doğruluk oranları 0.80 elde edildi. Çalışmada elde edilen model ilaç aday moleküllerini elemek için ilaç tasarım evrelerinde basit bir filtre olarak kullanılabilirler.

Anahtar Kelimeler: Makine öğrenmesi, Destek vektör makineleri, İlaç keşfi, ToxPrint kemotipler, Onaylanmış ve geri çekilen ilaçlar, Hiyerarşik çoklu etiket sınıflaması, Öznitelik seçimi.

ABSTRACT

Doctor of Philosophy

FORMATION OF A DRUG CLASSIFICATION FRAMEWORK VIA DATA MINING TECHNIQUES

Aytun ONAY

TOBB University of Economics and Technology
Institute of Natural and Applied Sciences
Computer Engineering Science Programme

Supervisor: Assoc. Prof. Dr. Osman ABUL

Date: August 2017

Virtual screening of candidate drug molecules via machine learning methods plays a key role in pharmaceutical industry to prevent adverse effects of the drugs. Computational classification methods can distinguish approved drugs from withdrawn ones. In this study, we focused on 3 various applications on drugs. We studied with different machine learning strategies to distinguish approved and withdrawn drugs. To begin with, 760 molecular descriptors such as ToxPrint Chemotype, global molecular, size and shape were calculated to study classification and feature selection problems for each drug molecule in this study. In first application, SVM and ensemble methods were applied on drug data sets to categorize more than 400 drugs belonging to nervous system and various disease groups as approved or withdrawn. Accuracy rates were found between 0.74 and 0.89 for data sets. Here, feature selection methods which were applied on drug data sets increased classification performance values. The number of total chemotypes, bond CN_amine_aliphatic_ generic, XlogP, aspheric: Cor3D:ori1ve Bonds descriptors were found as more significant descriptors to form model for nervous system drugs. Moreover, the fragmans located in minimum 60 % of nervous system withdrawn drugs were determined via application of gSpan algorithms on drug data sets. This is

the first report that describes distinction of withdrawn and approved drugs pertaining to the spesific disease on the data sets. In the second part of study, 558 drugs with various disease groups were classified in 3 basic levels with hierarchical multi-label classification via Clus-HMC-Ens algorithms. While first level includes all drugs, second level consists of 3 groups of drugs. These are approved nerveous system drugs, approved drugs of various disease groups and withdrawn drugs. Last level has drugs of 5 different groups according to Anatomic Therapeutic Chemical classification of nerveous system drugs. In this application, some paremeters were selected for classification of drugs hierarchically. Selected paremeters such as FTest, w_0, k, classification treshold, m-estimate increased estimation performance of model.

In last part of study, more than 1200 approved and withdrawn drugs were studied. Molecular identifiers that are effective in classification models have been identified by an effective feature selection strategy proposed in the thesis. ToxPrint chemotypes, effective descriptors, were used for determination of a number of rules in drug molecules. Available/unavailable chemotypes were analysed in approved/withdrawn drugs on drug data sets. While chemotypes such as bond:NN_hydrazine_alkyl_HH2 only presented in withdrawn drugs, ones such as bond:P=O_phosphorus_oxo,bond:PC_phosphorus_organo_generic,group:carbohydrate_aldohexose, group:carbohydrate_aldopentose, group:carbohydrate_hexopyranose_fructose, group:carbohydrate_hexopyranose_glucose etc. just examined in approved drugs. A model for classifier ensemble design was proposed on the unbalanced drug data set. Accuracy of 0.80 was obtained for the test set in order to classify the drugs as approved and withdrawn. Developed model in this study can be used as a simple filter in drug modelling to eleminate drug candidate molecules.

Key words: Machine learning, Support vector machines, Drug discovery, ToxPrint chemotypes, Approved and withdrawn drugs, Hierarchical multi-label classification, Feature selection.

TEŞEKKÜR

Doktorada geçirdiğim yıllar süresince beni yönlendiren, bilgilendiren ve Tez çalışmalarım sırasında tecrübelerinden yararlandığım Danışman hocam Doç. Dr. Osman Abul'a, TOBB Ekonomi ve Teknoloji Üniversitesi Bilgisayar Mühendisliği Bölümü öğretim üyelerine, tezin yazımı sırasında yardımlarını esirgemeyen ve biyotıp konusunda bilgisinden, tecrübelerinden yararlandığım Yrd. Doç. Dr. Melih Onay'a, doktora süresince özellikle algoritma konusunda ve her konuda desteğini gördüğüm arkadaşım Araş. Gör. Uğur Şahin'e, B-11 labında geçirdiğimiz yıllar boyunca tüm bölüm arkadaşlarıma ve manevi desteğiyle her zaman yanımızda olan bölüm sekreterimiz ve arkadaşımız Merve Bağcı'ya teşekkürü bir borç bilirim. Hayatımın her aşamasında olduğu gibi bu aşamasında da maddi ve manevi destekleri ile yanımda olan sevgili eşim ve doğduğu günden beri doktoramın her aşamasında evimizi aydınlatan canım oğlum Rüzgar Çınar'a teşekkürler. Hayatımın ışığı oğlum Rüzgar Çınar, eşim Melih'e…

İÇİNDEKİLER

ŞEKİL LİSTESİ

Sayfa

ÇİZELGE LİSTESİ

Sayfa

KISALTMALAR

ADME : Emilim, Dağılım, Metabolizma ve Atılım
ADs : Onaylanmış İlaçlar
ANN : Yapay Sinir Ağları
AR : Doğruluk Oranı
ATC : Anatomik Terapötik Kimyasal Sınıflama
AUC : Eğri Altındaki Alan
AUPRC : Precision-Recall Eğrisinin Altında Kalan Alan
AUROC : ROC Eğrisinin Altında Kalan Alan
AW : Onaylanmış ve Geri Çekilen İlaçlardan Oluşan Veri Seti
CNS : Merkezi Sinir Sistemi
CSS : Ki-kare İstatistik Değeri
D : Boyut
DAG : Yönlü Düz Ağaçlar
DFS : Derinlik Öncelikli Arama
DS : Veri Seti
DT : Karar Ağaçları
EM : Topluluk Metotları
F-1 Score : F-1 Skor
FAW : Sınıflandırmada Etkin Öznitelik Seti
FDA : Amerikan Gıda ve İlaç Dairesi
FS : Özellik Seti
GA : Genetik Algoritma
gSpan : Çizge Tabanlı Alt Yapı Örüntüsü Madenciliği
HMC : Hiyerarşik Çoklu Etiket Sınıflaması
HOMO : Dolu Olan En Yüksek Enerjili Moleküler Orbital
K-NN : K-En Yakın Komşular
LBVS : Ligand Tabanlı Sanal Tarama
LUMO : Boş Olan En Düşük Enerjili Moleküler Orbital
MCC : Matthews Korelasyon Katsayısı
NPV : Negatif Öngörme Değeri
NS : Sinir Sistemi
NSADs : Onaylanmış Sinir Sistemi İlaçları
NSWDs : Geri Çekilen Sinir Sistemi İlaçları
PCT : Tahmini Kümelenme Ağacı
PNS : Periferik Sinir Sistemi
PPV : Pozitif Öngörme Değeri
PR : Precision-Recall
QSAR : Kantitatif Yapı Aktivite İlişkisi
RBF : Radyal Tabanlı Fonksiyon
Ro5 : Lipinski'nin Beş Kuralı
SAR : Yapısal Aktivite İlişki Analizi
SBVS : Yapı Temelli Sanal Tarama
SDF : Yapı Veri Formatı

SE **:** Duyarlılık
SP **:** Özgüllük
SVM **:** Destek Vektör Makineleri
VS **:** Sanal Tarama
WDs **:** Geri çekilen ilaçlar

SEMBOL LİSTESİ

Bu çalışmada kullanılmış olan simgeler açıklamaları ile birlikte aşağıda sunulmuştur.

Simgeler	**Açıklama**
k	Hiyerarşik Sınıflandırmada Topluluk Metodunda Kullanılan Ağaçların Sayısı
P	Hiyerarşik Sınıflandırmada Pozitif Eğitim Örneklerinin Sayısı
p	Hiyerarşik Sınıflandırmada Tüm Eğitim Seti İçerisindeki Pozitif Örneklerin Oranı
t	Hiyerarşik Sınıflandırma Eşiği
T	Hiyerarşik Sınıflandırmada Toplam Eğitim Örneklerinin Sayısı
w_0	Hiyerarşik Sınıflandırmada Sezgisel Karar Ağacında Farklı Sınıfların Ağırlıkları
ΔG	Serbest Bağlanma Enerjisi
ΔH	Entalpi Değişimi
ΔS	Entropi Değişikliği

RESİM LİSTESİ

Sayfa

1. GİRİŞ

İlaçlar genel olarak hedef bir proteine bağlanıp, bağlandığı proteinin davranışını değiştiren küçük kimyasal moleküller olarak tanımlanabilir (Wang ve diğ., 2015). İlaçlar canlı hücre üzerine tesir ettiğinde bir hastalığın iyileştirilmesini, semptomlarının azaltılmasını ve hastalıklardan korunmayı mümkün kılar. Kimyasal bir molekülün ilaç olarak kullanılabilmesi için başta hedef proteine bağlanması gerekir. Bunun dışında yan etki, toksisite gibi özelliklerden yoksun ve yerel etki gösterip hedef hastalıklar için etkin olmalıdır.

Günümüzde yeni bir ilacın geliştirilmesi (ligand tasarımı) yüksek maliyet ve bunun yanında fazla zaman yükü getirmektedir (Evens, 2007). Bilgisayar destekli veri madenciliği yöntemleri ile sofistike yazılımlar ilaç olarak düşünülen aday molekülleri erken safhalarda incelemeyi sağlar. İlaç olması muhtemel olmayan moleküller belirlenebilir ve elenebilir. Sonuç olarak, molekülün çeşitli özellikleri hakkında önceden edinilen bilgi, bize zamandan ve maliyetten kazanmamızı sağlar. Bu nedenle ilaç veri tabanlarında (örneğin DRUGBANK) saklanan onaylanmış ilaçlar kadar geri çekilen ilaçlarda yeni ilaç tasarımları için büyük önem taşır. DRUGBANK ilaç veri bankasında yaklaşık 1800 onaylanmış ve 220 geri çekilen ilaç bulunmaktadır. Kimyasal bileşik sınıflandırması problemlerinde ilaç veri tabanlarındaki onaylanmış ve geri çekilen ilaçların kimyasal yapı ve moleküler özellikleri kullanılarak bunlar için sınıflandırma modelleri geliştirilebilir.

İlaçların geri çekilmesi genellikle ciddi yan etkiler ve ölümler gibi güvenlik sorunları ile ilgilidir. Bu etkiler büyük oranda karaciğer, kardiyovasküler sistem veya daha bir çok organda ortaya çıkabilir. Ters ilaç reaksiyonları, ligand ve reseptör etkileşimi nedeniyle birincil veya istenmeyen hedef organda görülebilir. İlaçların geri çekilmesi ile ilgili bir diğer problem de ilacın etkinliğinin olmamasıdır. İlaç tasarımı problemlerinde öncelikle ilaç toksisitesinin nedenleri iyi bir şekilde ortaya konmalıdır. Bunun için, ilaç metabolizmasına katılan ilacın yapısı, fiziko-kimyasal özellikleri ve sinyal yollarının nasıl çalıştığını açıklamak gerekir. Moleküler seviyedeki ilaç-hedef etkileşimlerinde genetik değişiklikler, örnek olarak tek

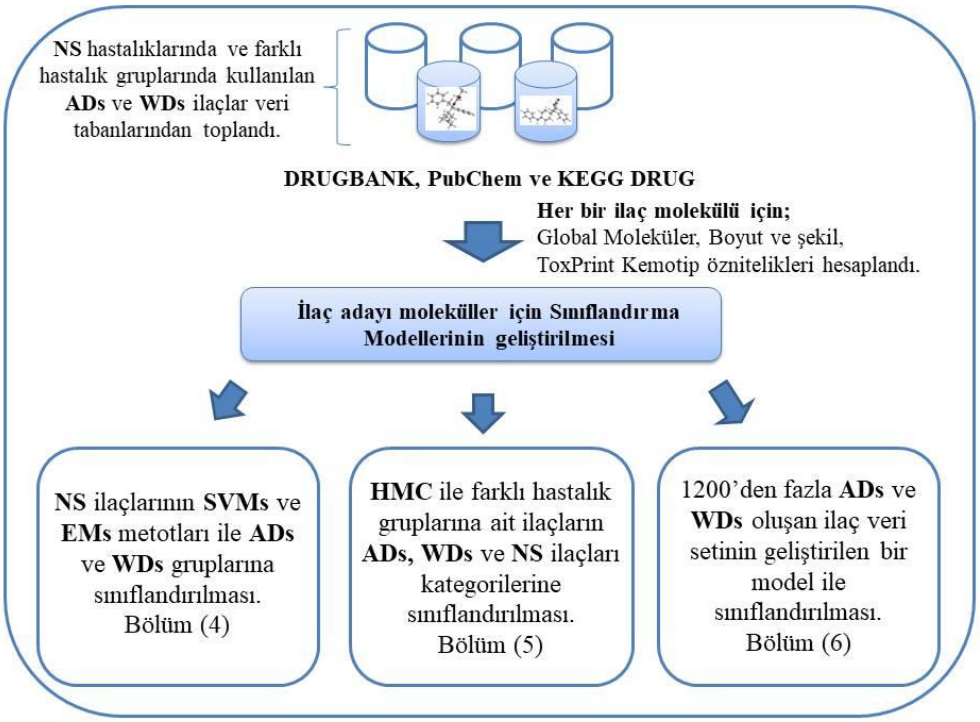
nükleotid polimorfizmleri (SNP'ler), metabolize enzim (CYP P450'ler) aktiviteleri ters ilaç etkileşimine neden olabilir. 1969 ve 2002 yılları arasında ilaçların yan etkisi sebebi ile 2.3 milyon vaka ortaya çıkmış fakat 6000 tane ilaçtan yalnızca 75 tanesi piyasadan geri çekilmiştir (McNaughton ve diğ., 2014). Yapılan bir diğer araştırmaya göre, 1950 ve 2013 yılları arasında ölüm vakaları nedeni ile 95 ilaç piyasadan geri çekilmiştir (Onakpoya ve diğ., 2015). Başta avrupa ülkeleri ve ABD hızlı bir şekilde bu ilaçları raflardan kaldırırken, dünya geneline bakıldığında malesef hala kullanılmaktadırlar. Geri çekilen ilaçların insan sağlığı üzerindeki yan etkileri açıkça ortadadır. Çalışmamız literatürde geri çekilen ilaçlarla ilgili yapılan çalışmalardaki açığı kapatmak ve bu ilaçların önceden belirlenmesine yönelik modeller geliştirmek açısından büyük önem taşımaktadır.

İlaç keşfi külfetli ve pahalı bir süreçtir. Milyonlarca ilaç alternatifi içerisinden sadece çok az bir kısmı (~%10) insanlar üzerinde test edilmektedir (Silverman ve diğ., 2004). İlaç keşfinde hesaplamalı teknikler kullanılmadan önce tek bir ilacın onaylanmış ilaç olarak piyasaya sürülmesinin maliyeti yaklaşık $2.000.000.000'dı. İlaç tasarımlarında yüksek maliyeti düşürmek ve zamandan kazanmak hesaplamalı ilaç tekniklerinin önemini arttırmaktadır. Tezde yapılan çalışmalar laboratuvar testlerinden önce ilaç adayı molekülleri arasından yanlış pozitiflerin sayısını (geri çekilen ilaçlar ve ilaç olması muhtemel olmayan moleküller) DRUGBANK, PubChem ve KEGG gibi ilaç veri bankaları üzerinde veri madenciliği yöntemleri kullanımı ile azaltmaktır.

1.1 Önerilen Çatı

Çalışmada ilaç veri tabanlarında bulunan onaylanmış ve geri çekilen ilaçlar referans alınarak yeni ilaç adayı kimyasal bileşikler için sınıflandırma modelleri geliştirildi. İlaç veri setleri DRUGBANK, PubChem ve KEGG veri tabanlarından toplandı ve ilaç moleküllerine ilişkin global moleküler, boyut, şekil ve ToxPrint özellikleri hesaplandı. Yapılan ilk çalışmada spesifik bir hastalığa ait olan ilaçlar için bir model oluşturuldu. Sinir sistemine ait çok sayıda onaylanmış ve geri çekilen ilaç molekülleri destek vektör makineleri (support vector machines, SVMs) ve güçlendirilmiş karar ağaçları gibi topluluk yöntemleri (ensemble methods, EMs) kullanılarak kategorilerine göre sınıflandırıldı. Buna ek olarak onaylamış ve geri çekilen sinir sistemi ilaçları üzerinde sık alt çizge madenciliği uygulandı. Burada

onaylanmış/geri çekilen ilaçlarda bulunan/bulunmayan fragmanlar her iki kategori içinde belirlendi ve ayrıntılı olarak incelendi. Bu fragmanlar yeni ilaçların tasarımı için oldukça önem taşımaktadır. Çalışmanın diğer kısmında oluşturulan model aday ilaç moleküllerini onaylanmış, geri çekilen ve sinir sistemi ilacı olma durumu hakkında öngörüde bulunur. Farklı hastalık gruplarına ait 558 ilaç hiyerarşik çoklu etiket sınıflaması (HMC) kullanılıp üç temel seviyede sınıflandırıldı. Son olarak farklı hastalık gruplarına ait 1200'den fazla ilaç aday ilaç molekülleri için bir model oluşturmak amacıyla önerilen bir modelle onaylanmış ve geri çekilen kategorilerine göre sınıflandırıldı. Dengesiz ilaç veri setinde sınıflandırmada en etkin olan moleküler tanımlayıcıları belirlemek için etkin öznitelik stratejisi önerildi. Bu yöntemle belirlenen etkin özniteliklerin veri setlerindeki ilaç molekülleri için ayrıntılı analizi yapıldı. Onaylanmış/geri çekilen ilaç moleküllerinde bulunan/bulunmayan ToxPrint öznitelikleri belirlendi. Bir molekülün ToxPrint özelliklerine ilişkin öznitelikler ilaçların beklenmedik ters yan etkilerine karşı önceden bilgi sahibi olmamızı sağlar. Aşağıda Şekil (1.1)'de yapılan çalışmaları özetleyen bir çatı yer almaktadır. NS, sinir sistemini; ADs, onaylanmış ilaçları; WDs, geri çekilen ilaçları temsil etmektedir.



Şekil 1.1: Kimyasal bileşikleri sınıflandırmak için yapılan çalışmaları özetleyen bir çatı.

1.2 Tez Dokümanına Genel Bakış Ve Literatüre Katkı

Tez başlıca 7 bölüm içermektedir. Bölümlerde kimyasal bileşikleri sınıflandırma problemlerine katkı sağlayacak çalışmalar ve bu alanda makine öğrenmesi metotlarıyla geliştirilen modeller ve yaklaşımlar anlatılmaktadır. İlk bölümde kimyasal bileşiklerin sınıflandırılması için önerilen çatı ve ilaç tasarımı uygulamalarına yönelik literatürdeki çalışmaların özet halinde bir analizi sunulmaktadır. İkinci kısımda yeni bir ilaç geliştirmenin önemine vurgu yapılmış, ilaçlardan beklenen özellikler, ilaç tasarımının evreleri ve çıkan maliyetler ile dünyada ilaç sanayi gibi konularda bilgi verilmiştir. Buna ek olarak, ilaç tasarım sürecinde makine öğrenmesi yöntemlerinin kullanımı, ilaç verilerini elde etmede kullanılan metotlar ve ilaç verilerinin özellikleri ile ilgili açıklamalar yer almaktadır.

Tezin 3. Bölümünde kimyasal bileşikleri sınıflandırmak amacıyla veri madenciliği tabanlı geliştirilen sınıflandırma modelleri bir tasarım çatısı altında toplanmıştır. Önerilen bu modellerde kullanılan makine öğrenmesi metotları, ilaç verilerinin özellikleri, ilaç veri kümelerinin formatları, moleküllerin sayısal verilere dönüştürülmesi işlemleri, ilaç moleküllerini ifade etmede kullanılan özellikler ayrıntılı bir şekilde anlatılmıştır.

Tezin 4. bölümünde sinir sistemi (nervous system, NS) hastalıklarının tedavisinde kullanılan çok sayıda geri çekilen ve onaylanmış ilaçların ToxPrint özellikleri hesaplanıp makine öğrenmesi metotlarıyla sınıflandırılması için geliştirilen yöntemler ayrıntılı bir şekilde anlatılmıştır. Tez boyunca çalışamalarda kullanılan tüm geri çekilen ve onaylanmış ilaç moleküllerinin ToxPrint kemotip analizi sınıflandırma çalışmalarında kullanılmak üzere yapılmıştır. Bir kemotip bağlantı için kodlanan yapısal bir fragman olarak tanımlanır ve gerektiğinde atomların, bağların, fragmanların hatta bir bütün olarak ele alındığında molekülün fizikokimyasal ve elektronik özelliklerini tanımlar. Çalışmada her ilaç molekülü için 760 tane tanımlayıcı kullanıldı, her bir veri setinde tanımlayıcılar (features) için sınıflandırmadan önce boyut azaltma (dimension reduction) uygulandı. İlaçları sınıflandırmada daha etkin olan tanımlayıcılar belirlendi. Ayrıca NS ilaçlarının onaylanmış/geri çekilen durumlarını tahmin etmek için geliştirilen modeller aday ilaç moleküllerini test etmek amacıyla bu konuda çalışan araştırmacılara ve kullanıcılara verildi.

Tezin 5. bölümünde farklı hastalık gruplarına ait 558 ilaç Clus-HMC algoritması kullanılarak hiyerarşik olarak üç temel seviyede sınıflandırıldı. Birinci seviyede bütün ilaçlar (All drugs), ikinci seviyede ise 3 grup yer almaktadır. Bunlardan ilki onaylanmış NS ilaçlarını içermektedir (NSADs). İkinci grup ise diğer hastalık gruplarına ait onaylanmış ilaçları (The other ADs) son grup ise piyasadan geri çekilen ilaçları kapsamaktadır (WDs). Son seviyede ise toplam 5 grup yer almaktadır bunlar onaylanmış NS ilaçlarının Anatomik Terapötik Kimyasal (ATC) sınıflamasına göre N02, N03, N04, N05, N06 gruplarından ilaçları içermektedir. Bu sınıflandırma NS ilaçlarını diğer ilaçlardan ayırt etmemizi sağlarken bunun yanında ATC sınıflamasına göre ilacın hangi hastalık grubuna dahil olduğunuda belirlemektedir. Çalışmada her ilaç molekülü için 760 tane tanımlayıcı kullanılmıştır.

Tezin 6. bölümünde farklı hastalık gruplarına ait 1200'den fazla onaylanmış ve geri çekilen ilaç çalışılmıştır. Onaylanmış ilaçların sayısının geri çekilen ilaç sayısından fazla olması nedeniyle dengesiz ilaç veri seti üzerinde tezde önerilen etkin öznitelik seçme stratejisi uygulanmış ve sınıflandırma problemlerinde etkin rol oynayan öznitelikler belirlenmiştir. Sonrasında kimyasal bileşiklerin onaylanmış/geri çekilen durumlarının belirlenmesi amacıyla sınıflandırıcı topluluk tasarımı için bir modelönerilmiştir. Çalışmada ayrıca etkin öznitelik stratejisi ile belirlenen ilaç moleküllerine ait tanımlayıcıların ayrıntılı analizi veri setindeki ilaçlar için verilmiştir. Tezin 3.cü ve 5.ci bölümlerinde yine onaylanmış ve geri çekilen ilaç moleküllerine sık alt çizge madenciliği uygulanıp moleküllerin yapısal özellikleri belirlendi. Geri çekilen ve onaylanmış ilaç moleküllerinin % 60'ında bulunan fragmanlar ve bu fragmanlardan yalnızca geri çekilen ilaçların yapısında bulunan ayırt edici fragmanlar kimyasal bileşiklerin sınıflandırılması problemlerine katkı sağlamak amacıyla belirlendi.

Tez boyunca yapılan çalışmalarda veri madenciliği tabanlı kimyasal bileşikleri sınıflandırma problemlerinde kullanılmak üzere modeller geliştirilmiştir. Bu modeller ilaç veri bankalarında bulunan onaylanmış ve geri çekilen ilaçların özellikleri kullanılarak elde edilmiştir. Literatürde kimyasal bileşiklerin bir çok farklı açıdan sınıflandırılması problemleri üzerinde yapılan çok sayıda çalışma yer almaktadır. Bunların bir kısmı tez boyunca anlatılmıştır ancak bileşiklerin onaylanmış ve geri çekilen ilaçlar olarak sınıflandırılması problemi literatürde ilktir. Buna ek olarak çalışmada dengesiz veri setleri için etkin öznitelik seçme stratejisi

önerilmiştir. Bu özellik seçme stratejisi ile belirlenen etkin özniteliklerle geliştirilen sınıflandırma modellerinin performansı doğruluk oranı ve seçilen öznitelik sayısının düşük olması göz önüne alındığında oldukça iyi olduğu gözlenmiştir. Kimyasal bir bileşiğin ToxPrint özelliklerine ilişkin öznitelikler ilaçların beklenmedik ters yan etkilerine karşı öngörüde bulunmamızı sağlar. Bu nedenle belirlenen etkin özniteliklerin 1200'den fazla ilaç üzerindeki analizi tezde ayrıca verilmiştir. Bu sonuçlar aday ilaç molekülleri için bir rol model oluşturur. Tezde ayrıca ilaç veri seti için geliştirilen modelfarklı hastalık türlerine ait bir kimyasal bileşiğin (ilaç adayı) sınıfı hakkında bize öngörüde bulunur. Tezde geliştirilen tüm modeller araştırmacılara DVD_Ek'ler kısmında kendi ilaç adayı moleküllerini test etmeleri amacıyla verilmiştir. Buda literatüre sağlanan bir diğer katkıdır.

Son bölümde tez boyunca elde edilen sonuçlar ve geliştirilen modeller kısa bir şekilde özetlendi ve çalışmanın önemine vurgu yapıldı.

1.3 Literatürde İlgili Çalışmalar

İlaç veri bankalarından sinir sistemi ve farklı hastalık gruplarından ilaçlar kimyasal bileşiklerin sınıflandırılması problemlerinde kullanılmak üzere modeller geliştirilmesi amacıyla toplandı. Çalışmalarımızda sınıflandırma modelleri geliştirilirken destek vektör makineleri (SVMs), topluluk metotlarından güçlendirilmiş (boosted trees) ve torbalanmış (bagged trees) karar ağaçları algoritmaları kullanıldı. Buna ek olarak dengesiz veri setleri için etkin öznitelik seçme stratejisi geliştirildi ve yine sınıflandırıcı topluluk tasarımı için geliştirilen modelilaç veri seti üzerinde uygulandı. Ayrıca çok sayıda ilaç farklı hiyerarşik yapılarda elde edilen modeller kullanılarak hiyerarşik çoklu etiket sınıflaması ile üç temel düzeyde sınıflandırıldı elde edilen sonuçlar değerlendirildi. Aşağıda bileşiklerin sınıflandırılması ve ilaç tasarım problemleri üzerine literatürde veri madenciliği yöntemleri ve çeşitli yaklaşımlar kullanılarak yapılan çalışmaların kısa bir özeti yer almaktadır.

Clark ve Pickett (2000) çalışmalarında ilaç molekülüne benzerlik tahmini için yapılan hesaplamalı yöntemleri ayrıntılı bir şekilde anlatmışlardır. Bu alanda genetik algoritma tabanlı ve sinir ağına dayalı yaklaşımlara örnekler verilmiştir. Genetik algoritma tabanlı yaklaşımlar ilaçlara ait bir takım özellikler ağırlık, moleküler özellikler, bileşiklerde topolojik şekil tanımlayıcılar (World Drug Index'ten (WDI))

hesaplamak için kullanılırlar. Bileşiklerin biyolojik aktivite profilleri ilaçları ilaç benzeri ve ilaç benzeri olmayan bileşikler olmak üzere iki gruba ayırmamızda kullanılır. Bu çalışmayı içeren bir yöntem bir filtre olarak bileşikler üzerinde yüksek verimli tarama amacıyla GlaxoWellcome'da kullanılmaktadır. Bayes sinir ağları kapsamlı tıbbi kimya bilgi sistemlerini kullanarak ilaç benzeri molekülleri tanımlamaktadır. Çoğu ilaç için tercih edilen uygulama yolu oral yoldan vucuda alınmasıdır. Araştırmacılar bu nedenle bağırsak absorpsiyonunu destekleyen fiziko kimyasal özellikleri tasvir etmeye çalışmışlardır ve bunun için hesaplama yöntemleri geliştirmişlerdir. WDI'dan 2245 ilacın analizinden Pfizer'da Lipinski ve çalışma arkadaşları tarafından geliştirilen yöntemlerden en iyi bilinen beş kuralı (rule of five) dır. David ve Stephen (2000) bunun yanında Caco-2 hücre geçirgenliğini, insan fraksiyonel absorpsiyonunu ve insan etkin geçirgenliğini öngörmeye ilişkin hesaplamalı yöntemlerden de çalışmalarında ayrıntılı olarak yer vermişlerdir.

Garcia-Serna ve diğ. (2015) yaptıkları çalışma ile prediktif ilaç güvenliği için entegre sistem yaklaşımlarının geliştirilmesine katkıda bulunmuşlardır. Bu tür yaklaşımlar kimyasal parçaların potansiyel olarak toksisitelerinin tanımlanması, istatistiksel olarak denetlenemeyen büyük ve çeşitli güvenlik olayları için mekanik bir bakış açısı yakalama imkanı sunarlar. Onlar çalışmalarında biyoaktif küçük moleküllerin güvenlik riskini belirlerken kimyasal ve biyolojik tehlikelerin birlikte tanımlanmasının daha doğru olduğunu savunmuşlardır. Onlar belirli bir ilacın kullanımı ile hastada ilaç ters etkisinin gözlenmesi arasındaki ilişkiyi kimyasal yapıları ve güvenlik olaylarını birbirine bağlayan önemli bir bilgi parçası olarak tanımlamışlardır. Bu veriler manuel olarak ve pazarlama sonrası ters etki olay bildirim sistemleri vasıtasıyla toplanarak ve bibliyografik kaynaklardan web arama günlüğü verileri yoluyla tespit edilebilir. Çalışmalarında ayrıca ilaçların ters etkilerine maruz kalmanın bir sonucu olarak meydana gelen istenmeyen ilaç olaylarını yakından izlemek amacıyla ilaç firmaları ve hatta hastaların güvenlikle ilgili verileri depolayabileceği bilgi sistemlerinin oluşturulması anlatılmaktadır. Bu kaynakların en büyüğü 1969 yılından bu yana veri topluyor ve şu anda yılda neredeyse bir milyon rapor aldığı WHO, FDA ve Health Canada organizations tarafından desteklenen the Adverse Events Reporting System (AERS)'dir. Ayrıca verilen bilgiye göre 1.332 ilaç ve 10.097 yan etki arasında 438.801 önemli açıklama içeren yeni bir ilaç ters etkisi veritabanında oluşturulduğu anlatılmıştır.

Huang ve diğ. (2011) yaptıkları çalışmada bir ilacın ters etki reaksiyonlarını doğru bir şekilde tahmin etmek için pratik bir hesaplama çerçevesi geliştirmişlerdir. Klinik gözlem verilerini, ilaç hedefi verilerini, protein-protein etkileşimi (PPI) ağları ve gen ontoloji (GO) açıklamaları ile birleştirip ilaçların geri çekilmesinin ana nedenlerinden biri olan kardiyotoksisiteyi kullanmışlardır. Geliştirdikleri siliko model tatmin edici bir kardiyotoksisite ters ilaç etkisi öngörme performansı elde etmiştir. Onlar çalışmalarında toksisite ve istenmeyen yan etkiler üzerine yapılan araştırmaların ilaç güvenliği ve etkinliğinin artırılmasına katkısının büyük olacağını anlatmışlardır. Buna ek olarak sistem farmakolojisinden bahsetmişlerdir bu yaklaşımın yeni olduğunu klinik gözlem ve moleküler biyolojiden elde edilen verileri birleştirdiğinden bahsetmişlerdir.

Jónsdóttir ve diğ. (2005) ilaç ve ilaç adayları ile ilgili özelliklere sahip veri tabanları hakkında 2-boyutlu ve 3-boyutlu yapısal bilgiler içeren en önemli veri tabanlarına genel bir bakış sunmaktadırlar. Doğru tahmin edici siliko modeller geliştirmek için deneysel verilere erişim ve bu verilerin seçilmesi ve kullanılması için sayısal yöntemler geliştirmenin önemini anlatmışlardır. Onlar potansiyel ilaçlar olarak kimyasal bileşiklerin uygunluğunun sınıflandırılmasının yanı sıra fiziko-kimyasal ve ADMET özelliklerini tahmin etmeyle ilgili birçok ilginç prediktif yöntemin son yıllarda önerildiğine vurgu yapmışlardır. Çalışmalarında ayrıca ilaç moleküllerinin üç ana veritabanı koleksiyonu olan the Comprehensive Medical Chemistry database (CMC), MDL Drug Data Report (MDDR) and the Derwent Word Drug Index (WDI) veri tabanlarından bahsetmişlerdir. CMC günümüzde 8473 ilaç bileşiği içerir ve her yıl Amerika Birleşik Devletleri Onaylı Adlar (USAN) listesinde ilk kez tanımlanan bileşiklerle güncellenmektedir.

Lipinski (2004) çalışmasında ilaç benzeri yapısal özellikleri, ilaç benzeri ve ilaç benzeri olmayanların özelliklerinin karşılaştırılması ve ilaç benzeri özelliklerin klinik ile nasıl ilişkili olduğunu ayrıntılı bir şekilde anlatmıştır. Merkezi sinir sistemi (MSS) ilaçlarının fiziko kimyasal özellikleri ve MSS ilaçlarının MSS kan-beyin taşıyıcı yakınlığına ilişkin özellikler kısaca gözden geçirilmiştir. Ayrıca oral olmayan ilaçların özellikleri ile ilgili yeni literatür gözden geçirilmiş ve kurşun benzeri bileşiklerin özelliklerinin ilaç benzeri bileşiklerin özelliklerinden nasıl farklılaştığı tartışılmıştır. Orijinal RO5 oral olarak aktif olan bileşiklerle ilişkilidir ve dört basit fizikokimyasal parametre aralığı tanımlar ve Faz II klinik statüye ulaşmış oral yoldan

aktif ilaçların % 90'ına eşlik eder. Çalışmada anlatılan bu fiziko-kimyasal parametreler kabul edilebilir sulu çözünürlük ve bağırsak geçirgenliği ile ilişkilidir ve oral biyoyararlanımda ilk adımları içerir. İlaç benzeri fizikokimyasal özelliklerden Rotatable bond count 10'dan fazla ise sıçan oral biyoyararlanımının azaldığı gözlenmiştir. Buda ilaç benzeri moleküller için yaygın olarak kullanılan bir filtredir. Çalışma ayrıca ilaçların kabul edilebilir reseptör etkileşimlerini başarmak için yeterli işlevsellik içermesinden bahsetmektedir. Düşük işlevsellik ilaç benzeri maddeleri ilaç benzeri olmayan bileşiklerden ayıran basit bir filtre olarak kullanılabilir. Çalışmasında ayrıca bileşiklerle ilgili olarak klinik çalışmalarda ilerledikçe özelliklerinde istikrarlı bir değişiklik olduğuna vurgu yapılmıştır. Örnek olarak molekül ağırlığı (MWT), log P ve polar yüzey alanı (PSA), pazarlanmış ilaçlar için bulunan yaklaşık 340'lık bir MWT ile azaldığı verilmiştir. Çalışmada bileşiklerin merkezi sinir sistemi (MSS) aktif veya pasif olarak iki yöntemden biri ile sınıflandırıldığı anlatılmıştır. Bu yöntemler sırasıyla bileşiğin deneysel olarak beyin penetrasyonuna ilişkin kanıt sergilemesine ya da bileşiğin MSS-aktif veri setinde bulunmasına ilişkindir. MSS aktivitesi veya hareketsizliği ile ilgili parametreler genellikle (1) fiziko-kimyasal özellikler veya (2) MSS taşıyıcı afinitesi (çoğunlukla P-glikoprotein (PGP) salınım taşıyıcısı) ile ilgili özelliklerdir. MSS ilaçlarının fizikokimyasal özelliklerine bakılacak olursa polar yüzey alanı 60-70'den az olanlar MSS-aktif belirleme eğilimindedir. İki kural kümesi MSS aktivitesini öngörür bunlar sırasıyla yapılan çalışmada şu şekilde verilmiştir, bir molekülde azot ve oksijen atomu sayısı beşten daha az veya eşit ise beyine girme şansı yüksektir ve LogP-(N+O) pozitif ise bileşik aktiftir.

Pauwels ve diğ.(2011) ilaç adayı moleküllerin büyük moleküler veri bankalarında uygulanabilen kimyasal yapılarına dayanan potansiyel yan etkilerini öngörmek için yeni bir yöntem geliştirmişlerdir. Çalışmalarında seyrek kanonik korelasyon analizi (SCCA) kullanarak kimyasal altyapıların (veya kimyasal parçaların) ve yan etkilerin ilişkili kümelerini belirlemişlerdir. Böylelikle DrugBank'da saklanan pek çok karakterize edilmemiş ilaç molekülü için kapsamlı bir yan etki tahmini yapılmıştır. Bu tahminleri bir yan etki kümesine sahip olma ihtimali olan ilaçlar tarafından paylaşılan kimyasal altyapı seti tarafından oluşturulan ilişkili toplulukların eşzamanlı olarak çıkarılmasıyla gerçekleştirmişlerdir. Deneylerinde 888 onaylanmış ilacın kimyasal yapılarından SIDER veri tabanındaki 1385 yan etkiyi tahmin ederek

önerilen yöntemin kullanışlılığını göstermişlerdir. Amaçları ilaç keşfi sürecinin başlangıcında potansiyel yan etkilerin klinik evrelere ulaşmadan önce tahmin edilmesidir. Çalışmada ilaçlar, sistem biyolojisi açısından metabolik yollar ve sinyal iletim pathway gibi çeşitli moleküler etkileşimlerden oluşan biyolojik sistemlere pertürbasyon uyandıran ve gözlenen yan etkilere yol açan moleküller olarak tanımlanmıştır. Vücudun ilaca verdiği yanıt yalnızca hedefiyle etkileşiminden dolayı beklenen olumlu etkileri değil aynı zamanda hedef dışı etkileşimlerin toplam etkisini de etkiler. Aslında, bir ilacın hedefi için güçlü bir afinitesi olsa bile, sıklıkla değişen yakınlıklara sahip diğer protein ceplerine bağlanır ve potansiyel yan etkilere neden olur. Bu kavram çalışmada toksik bileşiklerden etkilenen yollarla ve toksik olmayan bileşiklerden etkilenen yolları karşılaştırarak, ilaç yan etkileri ile biyolojik yollar arasında bağlantı kurularak anlatılmıştır. Son zamanlarda ilaç yan etkilerini öngörmek için çeşitli hesaplama yöntemleri önerilmiş ve yöntemler sırasıyla çalışmada pathway-dayalı yaklaşımlar ve kimyasal yapı-temelli yaklaşımlar olarak kategorize edilmiştir. Pathway-dayalı yaklaşımların ilkesi, ilaç yan etkilerini bozulan biyolojik pathway veya alt pathway ilişkilendirmektir. Çünkü bu pathways ilaç tarafından hedeflenen proteinleri kapsar. Kimyasal yapı temelli yaklaşımların ilkesi ise ilaç yan etkilerini kimyasal yapılarıyla ilişkilendirmektir.

Stelle ve diğ. (2011) yaptıkları çalışmayla protein yapı tahmini ve protein katlanması alanlarına katkıda bulundular. Çeşitli veri madenciliği tekniklerini protein dizisindeki motifleri keşfetmek için kullandılar. Yapılan bu çalışmada, proteinlerin spesifik olan ikincil yapılarını hidrofobiklik modelleri çıkarmak için arşivlediler ve bir metot tanımladılar. Çalışmada lokal bir veritabanı tasarladılar ve 20.000 proteini Protein Veri Bankasından (http://www.pdb.org) çıkardılar. Verileri depolamak için veritabanı yönetim sistemi PostgreSQL kullandılar. Veritabanı proteinleri dört katlanma sınıfına ayırdı. Daha sonra Apriori algoritması hidrofobiklik modelleri tanımlamak için uygulandı. Yapılan bu çalışmada ilaç iletimi yaklaşımlarına veri madenciliğinin katkısı oldukça büyüktür. İlaç sektörünün tüm alanlarında veri madenciliği kilit bir rol oynar.

Elayaraja ve diğ. (2012) DNA veri kümesinde bulunan biyolojik sıralar üzerinde veri madenciliği metotlarını uygulayarak tekrarlı örüntüler ve potansiyel motifler çıkardılar. Bu kısa dizileri (motif veya işaret) bulma moleküler biyolojide ve bilgisayar bilimlerinde önemli bir problemdir. Ayrıca bu motiflerin belirlenmesi bilgi

tabanlı ilaç tasarımında, adli DNA analizinde, tarımsal biyoteknolojide önem taşıyan uygulamalardan biridir. Çalışmada bölgesel protein dizi motiflerini tahmin etmek amacıyla kümeleme algoritmalarından K Means ve Rough K Means algoritmaları kullandılar ve elde ettikleri sonuçları karşılaştırdılar. Kayan pencere tekniğiyle protein dizilerinden on ardışık kalıntı ürettiler. Bu tekrar eden on sıra segmentinin K Means ve Rough K Means algoritmaları kullanıldığında farklı gruplarda kümelendiği sonucuna ulaştılar.

İlaç tasarımına veri madenciliği ile bir başka yaklaşımda Ekins ve diğ. (2006) tarafından yapıldı. İlaç keşfi ve bilgisayarlı modellemeler için kaynak veriler elde etmek amacıyla ücretsiz ulaşılabilecek veri tabanları (PubChem) önerdiler. Yine ilaç keşfinde kullanılan pathways/network analiz algoritmaları için veritabanlarının kullanımını arttırdılar ve veri madenciliği için network üzerinde veri tabanlarını araştırmada kullanılacak verimli bir 1D metot sağladılar. Langdon ve diğ.(2004) veri madenciliği ile ilaç keşfinde, fareler üzerinde ilaç etken maddelerinin doz ve değerinin hedef dokuya zarar vermediği, biyoyararlanım ölçümlerini ve bunun yanında QSAR modellerini tahmin ve yorumlamada genetik programlamayı kullandılar ve bu sonuçları insan üzerinde genelleştirdiler. Bunun sonucunda genetik programlamayla yeni ilaç tedavileri için insan biyoyararlanım ölçümlerini az sayıda oldukça kompleks biyolojik etkileşimler için otomatik olarak tahmin edebilen ve yorumlayabilen bir model oluşturdular. Yeni makine öğrenmesi metotları geliştiren ve bunları kimyasal bileşikleri sınıflandırma problemlerine uygulayan Amasyalı (2008) çalışmasında sınıflandırma problemleri için Cline adı altında bir algoritma ailesi tasarladı. Geliştirilen algoritmalar karar ağacı oluşturma algoritmaları olup yapılan denemelerde geliştirilen bu algoritmaların ilaç veri kümelerinde mevcut algoritmalarla yarışabilecek performansta olduğu görüldü. Veri madenciliği yöntemlerini kullanarak yapılan bir başka çalışmada ise Baloğlu (2006) DNA sıralarındaki tekrarlı örüntülerin ve potansiyel motiflerin çıkarılması için yukarıdan-aşağı veri madenciliği ve genetik algoritma tabanlı hibrit bir çözüm yöntemi geliştirdi. Bu amaçla birinci adımda genetik algoritma kullanıp aday motiflerin bir populasyonunu oluştururken ikinci adımda veri madenciliği yöntemi yukarıdan-aşağı haliyle kullanarak aday motiflerin uygunluğunun değerlendirmesini yaptılar.

Hendlich ve diğ. (2003) ise yapı temelli tasarım süreçlerinin birbirini izleyen veritabanı sorgu araçlarından nasıl faydalanabileceğini tanımladılar. Tüm

çalışmalarını kendilerine ait bir veritabanı sistemi olan Relibase ile gerçekleştirdiler. Ayrıca etkileşen moleküler parçalar arasında tercih edilmiş geometrik modelleri araştırmak için Relibase uygulamaları gösterdiler. Relibase kompleks protein-ligand 3-boyutlu yapısal bilgileri araştırmak için tasarlandı. Buna ek olarak, çok sayıda sorgu türünün birleşimine ve esnek bir biçimlendirme içerisinde yapı temelli ilaç tasarımına izin verir. Sorguların çoğu birkaç dakika içerisinde gerçekleştirilir. İlaç tasarımı üzerine yapılan bir diğer çalışmada Burbidge ve diğ. (2001) aittir. Onlar support vector machine (SVM) sınıflandırma algoritmasının potansiyelini yapısal aktivite ilişki analizi (SAR) için kanıtladılar. SAR ilaç tasarım sürecinde ilaç tasarım şirketleri tarafından kullanılan bir tekniktir. Yapılan çalışmalarda kuantum teori ve sayısal yapı aktivite ilişki modeli (QSAR) arasında temel bir bağlantı kurulmaya çalışılır. Diğer bir değişle, kuantum benzerlik ölçülerek bir model geliştirilmeye çalışılır. Bu konuda Carbo-Dorca ve diğ. (2000) kuantum kimyası ve QSAR arasında yoğunluk fonksiyonlarını (DF) kullanıp bir ilişki kurmaya çalıştılar. QSAR ilaç tasarım sürecinde önemli bir rol oynamaktadır. Ab initio ve B3LYP / 6-31 G(d, p) ve 6-311G(d, p) temel setlerini kullanarak Jayaprakasha ve diğ. (2011) krotonaldehidin enerjisini, geometrik parametrelerini ve titreşimsel dalga sayısını hesapladılar. Molekül içerisindeki yük transferini gösterime sokmak için krotonaldehidin HOMO ve LUMO enerji seviyelerini hesapladılar ve bantlar arası enerji farkını belirlediler. HOMO ve LUMO enerjilerinin belirlenmesi kuantum kimyasal hesaplamalarda önemli parametrelerdir. HOMO temel olarak bir elektron verici orbital olarak davranırken, LUMO büyük ölçüde elektron alıcı bir orbital olarak davranır. Daidzein, genistein, formononetin, biochanin A ve bu moleküllerin radikallerinin elektronik ve yapısal özellikleri yoğunluk fonksiyonel teori (DFT), B3LYP/6-31+G(d, p), B3LYP/6-31++G(d, p) metotları kullanılarak Zhang ve diğ. (2010) tarafından incelendi. Moleküllerin hesaplanan antioksidan aktivite değerleri sırasıyla, genistein > daidzein > biochanina A > formononetin şeklinde elde edildi. HOMO, LUMO ve Mulliken spin yoğunluğuda moleküller için ayrıca hesaplandı. Moleküler yapı ve termodinamik bakış açısından, izoflavonoidlerin B-halkasının aktif merkez olduğu ve hidrojen atom transferinin antioksidan aktivitesinde temel mekanizma olduğu ortaya konuldu.

Sınıflandırma problemlerinde moleküler tanımlayıcıların en iyi kombinasyonunu seçerek (özellik seçimi) SVM'lerin performansını arttırmak mümkündür. Fourches

ve diğ. (2010) çalışmalarında model geliştirmeye başlamadan önce veri analizinde veri kürasyonunun önemini vurguladılar. Standart bir kimyasal veri kürasyon stratejisi, QSAR analizi, sanal tarama, kümeleme vb. gibi başarılı modelleme çalışmalarını mümkün kılar. Cao ve diğ. (2012) HDAC8 inhibitör ve inhibitör olmayanları (drug and non-drug) erken safhalarda gözlemlemek ve olmayanları filtrelemek amacıyla SVM sınıflandırma metodunu çalışmalarında kullandılar. Verileri sınıflandırmadan önce ADRIANA.Code programını kullanarak veri setindeki tüm bileşikler için 23 moleküler tanımlayıcı, küresel moleküler özellikler, yüzey özellikleri ve 2D ve 3D özellikler hesapladılar. Yapılan çalışmada test seti üzerinde elde edilen en iyi modelde doğruluk oranı % 75'e ulaştı. Bununla birlikte, HDon, HAcc, NRotBond ve XlogP gibi global moleküler tanımlayıcıların HDAC8 inhibitörlerini ve inhibitör olmayanları sınıflandırmada daha etkili faktörler oldukları belirlendi. Korkmaz ve diğ. (2014) çalışmalarında üç ayrı özellik seçimi yöntemi kullanarak gerçek bir ilaç tasarım problemi üzerinde SVM modellerini oluşturdular. Amaçları aktif molekülleri aktif olmayan moleküllerden ayırmaktı. Elde ettikleri modellerde test seti üzerinde % 76 ile % 81 arasında doğruluk oranına ulaştılar. Çalışmalarında 34 moleküler tanımlayıcı kullandılar ve HBDC (Hydrogen Bond Donor Count) ve PSA (Polar Surface Area) tanımlayıcılarının sınıflandırmada daha etkili olduklarını ortaya koydular. Diğer önemli tahmin edicileri, RBC (Rotable Bond Count), logP, WI (Wiener Index) ve BI (Balaban Endeksi) olarak belirlediler.

Ghorbanzad'e ve Fatemi (2012) 326 merkezi sinir sistemi (MSS) ilacından oluşan bir veri setini kan-beyin bariyerlerine nüfuz etmelerine göre aktif ve inaktif MSS ajanları olarak sınıflandırmak için doğrusal ve kuadratik diskriminant analizi (linear and quadratic discriminant analysis) ve en küçük kareler destek vektör makinesi (least squares support vector machine, LS-SVM) sınıflandırma algoritmalarını MSS tasarım problemine uyguladı. Kan-beyin bariyerini geçen MSS ilaçları aktif olarak adlandırılır. Veri setinde 166 MSS aktif ilaç ve 160 MSS inaktif ilaç bulunmaktadır. Yapılan çalışmada geliştirilen LS-SVM modeli, eğitim setinde (% 96.5) ve test setinde (% 92.9) doğruluk oranına ulaştı. Çalışmada çok sayıda tanımlayıcı veri setindeki ilaçları sınıflandırmak için hesaplandı bunlardan sadece dokuz tanesi ilaçların kan-beyin bariyerinden geçmesine ilişkindir.

Zhang ve diğ. (2011) çalışmalarında ilaçların nöbet yükümlülüğünü erken safhalarda belirlemek amacıyla SVM tabanlı tahmin edici bir model geliştirdi. Terapötik

dozlarda nöbet uyandıran ajanlar (pozitifler) ve hiçbir nöbet riski taşımayan ajanlar (negatifler) de dahil olmak üzere 680 tane bileşik SVM modelini eğitmek için kullanıldı. Bağımsız test seti 175 bileşiği içermektedir. Elde edilen modelin doğruluk oranı % 86.9'dur. Çalışmada nöbet yükümlülüğüne sahip bileşiklerin tahmini için, moleküler elektronik özellikleri, hidrojen bağlama özelliği, moleküler aromatik fonksiyonlar, lipofiliklik, moleküler polar yüzey alanı ve moleküler yapısal bilgi içeren 18 moleküler tanımlayıcı kullanıldı. Yukarıdaki çalışmaların sonuçları, bileşik özelliklerini tanımlayan bu tanımlayıcıların etkili bir şekilde yeni bir ilacın keşfi için kullanılabileceğini ortaya koydu. Geliştirilen modeller ilaç keşfi sürecinde basit filtreler olarak kullanılabilirler.

Klekota ve Roth (2008) çalışmalarında çoklu bileşik kütüphaneleri için biyoaktiviteyi tanımladılar ve bunların üç boyutlu ayırt edilebilen alt yapılarını tespit ettiler. Biyolojik aktivite ile ilgili alt yapıları belirlemek için Chembridge Diverse Set E kütüphanesinde 4860 altyapı kümesinden alınan altyapıların varlığına veya yokluğuna göre oldukça ayırt edilebilen alt yapıları karar ağaçlarını kullanarak birbirinden ayırdılar. Sonuçlar farmasötik keşif için önemli etkileri nedeniyle ayırt edilen yapıların yeni kimyasal kütüphaneleri tasarlamak için kullanılabileceğini göstermektedir. Embrechts ve diğ. (2003) çalışmalarında evrimsel algoritmalara (EA) dayanan üç farklı özellik seçimi (feature selection) yaklaşımını QSAR problemleri için ele aldılar. Potansiyel ilaç etkinliğine sahip moleküllerin bir molekül kitaplığından verimli bir şekilde kantitatif yapı etki ilişkisi (quantitative structure-activity relationship, QSAR) modellerine dayanarak sanal olarak taranması onların biyolojik aktivitelerinin tahminde önemli bir rol oynar. Bu yöntemler, bir öğrenme modeli için bir genetik algoritma (GA), GA-ölçekli regresyon kümelemesi ve korelasyon matrisinden GA tabanlı özellik seçimi ile ortak öznitelik çıkarımı üzerine kurulmuştur. Çalışmalarında QSAR'daki özellik seçimi için ortak GA tabanlı yöntemle birlikte özellik seçimi için iki yeni yaklaşım önerdiler. Buna ek olarak GA tabanlı öznitelik seçme yöntemlerini duyarlılık analizi ile birleştiren bir hibrid özellik seçme yöntemide gösterdiler. QSAR'nın amacı, tanımlayıcı özelliklere dayalı moleküllerin biyoaktivitesini öngörmektir. Temel varsayım, biyolojik aktivitedeki değişiklikler, ölçülen veya hesaplanan moleküler özelliklerdeki karakteristiklerle ilişkilendirilebilmesidir. 2D, elektrotopolojik, 3D ve transfer edilebilir atom eşdeğerli (TAE) tanımlayıcılar da dahil olmak üzere, QSAR araştırmalarında geleneksel olarak

çeşitli tanımlayıcı türleri kullanılır. Mamitsuka (2003) verilen bir kimyasal bileşik ile ilaç etkinliği arasındaki ilişkileri hesaplamanın önemi vurgulanmıştır. Çalışmalarında kimyasal bileşiklerin ilaç etkinliği ile ilgili veri kümesinde, her satır bir kimyasal bileşime karşılık gelir ve sütunlar, bileşiğin tanımlayıcılarıdır ve bileşimin aktivitesini belirten bir etikete karşılık gelir. Son zamanlarda, tanımlayıcıların boyutunun oldukça büyümesi nedeniyle bazı ilaç verilerinin sütun sayısı (nitelikler veya özellikler) verilen bileşik grubundan daha ayrıntılı bilgi elde etmek için yüz binlere hatta milyona ulaştı bu nedenle çalışmada 4 farklı özellik seçimi yöntemi yaklaşık 140.000 özellik içeren Thrombin veri seti üzerinde denendi. Sonuçlar iki sınıflandırıcı (SVM ve C4.5) ile sınıflandırılarak değerlendirildi. Sonuçlar değerlendirilirken metotların doğru özellikleri seçebilmesi, zaman etkinliği açısından performansları ve verilerdeki gürültü seviyeleri göz önüne alındığında çalışabilme ölçütleri göz önüne alındı.

Vens ve diğ. (2008) çalışmalarında hiyerarşik çoklu-etiket sınıflandırması (hierarchical multi-label classification, HMC) için karar ağaçlarının indüksiyonuna yönelik çeşitli yaklaşımları ve işlevsel genomiklerde kullanımlarının ampirik bir çalışmasını sundular. HMC örneklerin aynı anda birden çok sınıfa dahil olabileceği ve bu sınıfların hiyerarşide düzenlendiği bir sınıflandırma varyantıdır. Uygulama doğrultusunda sınıfların hiyerarşisi her sınıfın en fazla bir ebeveyni (ağaç yapısı) veya sınıfların birden çok ebeveyni (DAG yapısı) olacak şekilde oluşturuldu. Çalışmada üç yaklaşım, tek-etiketli sınıflandırma (single-label classification, SC), hiyerarşik tek-etiketli sınıflandırma (hierarchical single-label classification, HSC) ve HMC metorları 24 maya veri setinde karşılaştırıldı. Çalışmada MIPS'in FunCat (ağaç yapısı) ve Gen Ontolojisi (DAG yapısı) sınıflandırma şemaları olarak kullanıldı. Doğruluk oranı, model boyutu ve indüksiyon süresi göz önüne alındığında HMC'nin daha iyi performans sergilediği gözlemlendi. Schietgat ve diğ. (2010) genom dizilerindeki açık okuma çerçevelerini (open reading frames, ORFs) belirleme ve bunlara biyolojik işlevler atama amacıyla karar ağacı tabanlı modeller geliştirdiler. S. cerevisiae, A. thaliana ve M. musculus, biyolojide iyi incelenmiş organizmalardır ve genomlarının dizilişi yıllar önce tamamlanmıştır. Bu genomların ORF'lerine otomatik olarak biyolojik fonksiyonlar atan metotların geliştirilmesi amacıyla hiyerarşik çoklu-etiket karar ağaçlarını (HMC) öğrenmek için bir algoritma tanımladılar. Bu tür ağaçların tek ağaçlardan daha doğru oldukları ve son teknoloji

istatistiksel öğrenme ve fonksiyonel bağlantı yöntemleriyle rekabetçi olduklarını gösterdiler. Bunlar aynı zamanda bir ORF'nin tüm işlevlerini, belirli bir gen işlevleri hiyerarşisine (FunCat veya GO gibi) riayet ederek öngörebilir. Sınıflandırma probleminde HMC'nin yaygın makine öğrenmesi metorlarından farkı (i) tek bir genin birden fazla fonksiyonu olabilir (ii) bu fonksiyonlar hiyerarşik olarak organize edilebilir. Bu algoritma ile elde edilen yeni sonuçlar, önceden açıklanan yöntemlere göre daha iyi öngörme performansı sergilemektedir.

2. İLAÇ TASARIMI VE VERİ MADENCİLİĞİ

İlaç keşfi tarihinde afyonun geçmişi ilaç keşfinde en erken ilerlemeyi gösterir. İnsanlar afyon ve onun doğal aktif bileşeni morfini yaklaşık 5000 yıl boyunca etken kimyasal maddesini ve gerçekte nasıl etki ettiğini bilmeden kullandılar. 1815 yılında F. W. Serturner, morfini afyon özünden izole etti (Drews, 2000). 1848'de bir afyon alkaloid olan papaverin izole edildi, ancak bir mide ve bağırsak gevşetici olarak kullanımı 1917'ye kadar keşfedilmedi (Sneader, 2005).

1870'lerde, Paul Ehrlich çalışmasında biyolojik dokular için boyaların seçici afinitesi parazitteki yapısal olarak benzer fakat aynı olmayan kemoreseptörlerden kaynaklandığını ve parazit ve konak dokularda görülen bu farklılıkların terapötik kullanım için istismar edilebileceği hipotezini ortaya koydu. Yapılan bu çalışma kemoterapinin doğumunu ve boya şirketlerinden ilaç endüstrisinin oluşumunun başlangıcı oldu (Drews, 2000).

1910'da Ehrlich ve öğrencisi olan Sahachiro geniş bir sistematik araştırmada bulundu. Onların 606.cı kez hazırladıkları Arsenik içeren bileşikleri (Salvarsan) Sifiliz spiroket bakterilerini öldürdü ancak konakları öldüremedi. Bu noktaya kadar bulaşıcı hastalıklardan korunmanın tek yolu bu aşıydı. Ehrlich 1915'te öldükten sonra doktorlar sihirli kurşun adını verdikleri daha az toksit içeren ilaçları aramaya devam ettiler. 1928'de İskoç bilimadamı Alexander Fleming, penisilinin antibiyotik özelliklerine sahip olduğunu keşfetti. On yıl sonra, Avustralyalı bilim adamı Howard Walter Florey, Alman bilim adamı Ernst Chain ve İngiliz biyokimyacı Norman Heatley bir ilaç olarak penisilini geliştirdiler. 1960'larda çeşitli semptomlar için kullanılan bir ilaç sınıfı olan β-blokerlerin keşfi modern çağın ilaç keşfinin başlangıcını temsil eder. β-blokerler aynı zamanda kardiyak aritmilerin yönetimi, kardiyoproteksiyon, kalp krizinden sonra ve hipertansiyon için kullanılır (Vogel ve diğ., 2013).

Bu ilaçlar, ilk mekanizmaya dayalı reseptör alt türüne özgü ilaçlardı. O zamandan beri giderek artan sayıda tasarlanan ilaçlar, örnek olarak HIV proteaz inhibitörleri, rekombinant protein ilaçları, büyüme hormonları ve terapötik antikorlar, hastaların

tedavilerinin bir parçası haline geldi. 2001'de Gleevec bir kanser tedavisi için FDA'nın onayını aldı ve bu durum Time Magazine'e "sihirli kurşun" olarak kapak oldu. Hızla bölünen hücrelerin tümünü spesifik olarak inhibe etmeden ve öldürmek yerine, belirli bir kanser hücresinin bir enzim özelliğini spesifik olarak inhibe ederek etki gösteren yeni bir sınıf ajanlara ilk örnek olmuştur.

İlaç keşfi öyküsü boyunca, şans ve kazalar önemli rol oynamıştır. Modern ilaç keşfi son 50 yılda çarpıcı bir şekilde değişmesine rağmen bu günümüzde de hala böyledir. Sildenafil (Viagra) bilim adamları tarafından ilk olarak boğaz ağrısı için Pfizer'da sentezlenmiştir. Ancak klinik çalışmalarda boğaz ağrısı (anjina) üzerinde etkili olamayınca ve bazı beklenmedik yan etkilerinin ortaya çıkmasından dolayı, 1998'de erektil disfonksiyon için tedavide kullanılan ve ilk ağızdan tedavi için uygulanan FDA'nın onaylı ilacına yönelindi. Bazı olayları öngörme ve çalışmalarda tökezlendiğinde dahi olayların faydalı yönde gelişimini sağlamak ve değerli şeyleri bulmak bir yetenektir. Louis Pasteur'un gözlemlediği gibi şans hazır bir zihni destekler. Dolayısıyla bundan sonraki bölümlerde rasyonel ilaç keşfi ve geliştirilmesi süreci anlatılmaktadır.

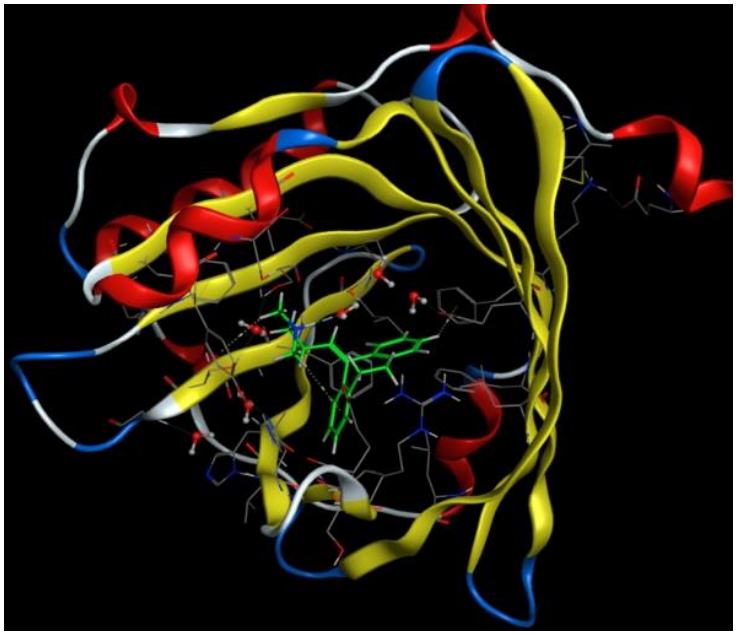
Dünyada ülkeler ilaç sanayisindeki konumlarına göre dört gruba ayrılmaktadır. İlaç araştırma ve geliştirmeye dayalı çok gelişmiş ilaç endüstrisine sahip 1.Grup ülkeler arasında ABD, İngiltere, İsviçre, Japonya, Hollanda ve Almanya bulunmaktadır. Türkiye ise 3.Grup olan 13 ülke ile birlikte, mamul ilaç ve etkin madde üreten ülkeler arasında yer almaktadır.

2.1 İlaç Tasarımı

Küçük moleküller (ilaç yapımında kullanılabilecek aday moleküller) bir hastalığa neden olan spesifik bir proteine bağlanarak onun etkisini modifiye ederler. Yani ilaçlar hastalığa neden olan bir proteinin etkisini ortadan kaldırmak için kullanılırlar. Kanser ilaçlarının etkisizliği, AIDS gibi tedavi edilemeyen hastalıklar ve yeni hastalıkların ortaya çıkması çağdaş ilaç tasarımına ihtiyaç duyulduğunu göstermektedir.

İlaç tasarımında yaygın olarak kullanılan iki yaklaşım vardır. Bunlardan ilki ligand tabanlı ikincisi protein tabanlıdır. Ligand tabanlı yaklaşımda hedef proteine bind olduğu bilinen moleküllerden bir model oluşturulur ve bu modele merak edilen ligandın ne denli yakın olduğu skorlanır. Bu yaklaşımda modelin, minimum içermesi

gereken motifler yer alır ve yapılan aslında ligandın bu motiflere sahip olup olmadığının belirlenmesidir. Motif tanımı uygulamadan uygulamaya farklılık göstermektedir, örneğin bazıları sadece yapısal (üç boyutlu molekül yapısı) bazıları ise biyokimyasal özellikleri kullanır. Protein tabanlı yaklaşımda ise hedef proteinin üç boyutlu yapısına bind olabilecek moleküller adaylar arasından alınır ve skorlanır. Skorlama fonksiyonu daha önceki bilinen binding örneklerinden elde edilmiş verilerden elde edilen veri madenciliği modelleri ile yapılabildiği gibi, doğrudan moleküler dinamik simülasyonu ile de belirlenebilir (Diniz ve diğ., 2010). Fakat moleküler dinamik çok yüksek hesaplama gücünü gerektirir. Dolayısıyla, moleküler dinamiğin kullanılabilmesi için bile adayların yine veri madenciliği teknikleri ile indirgenmesi çok önemli olmaktadır. Dikkat edilmesi gereken önemli bir nokta yaklaşım ne olursa olsun veri madenciliğinin ilaç tasarım sürecinde önemli bir role sahip olduğudur. Ligand tabanlı yaklaşımda biyokimyasal özelliklerin kısıt tabanlı veri madenciliği için kullanımı arama uzayının çok büyük olması nedeniyle hızlı sonuca ulaşmayı sağlayacaktır ve gereksiz çıktıları eleyerek etkinliğede katkıda bulunacaktır. Örnek olarak hidrofobik olduğu bilinen yüzeylerin arama uzayından çıkarılması verilebilir. Resim (2.1) İnsan alfa 1-asit glikoproteini ve ona bind olmuş amitriptilin kompleksini göstermektedir.



Resim 2.1: İnsan alfa 1-asit glikoproteini ve ona bind olmuş amitriptilin kompleksi [www.wwpdb.org].

2.1.1 İlaç tasarım ilkeleri

Başarılı bir ilacın tasarlanması için aynı anda pek çok şartın karşılanması gerekir. Öncelikle, molekül uygun hedefe odaklanmalıdır yani tasarlanan ilaç önemli metabolik görevleri olan moleküllere değil yalnızca hedefine bağlanmalıdır. Çünkü ilaç hayati görevleri yerine getiren enzimlere bağlanırsa ilacın yan etkileri ortaya çıkar. İlaç molekülleri çok büyük olmamalı, zar geçirgenliğine sahip olmalı, emilmeye ve kullanıldıktan sonra vücuttan atılmaya uygun olmalıdır. Dahası, ilacın toksisite seviyeleri en aza indirilmelidir (Freire, 2005; Yusof ve Segall, 2013). İlaç benzeri bir molekülün bu özellikleri, ADME (emilim, dağılım, metabolizma ve atılım) özellikleri olarak adlandırılır (Hou ve diğ., 2007; Yusof ve diğ., 2014). Lipinski (2000) çalışmasında dört özelliğin parametreleri için eşik değerler yayınladı. 1997'de ilaç adaylarına başlangıç filtresi olarak kullanılabilecek moleküllerin emilimini ve geçirgenliğine işaret ederek ilaç literatürü tarafından kabul edilen dört önemli kural yayınlamıştır. Lipinski'nin "beş kuralı" na göre, ilaç benzeri bir molekülün özellikleri şöyle olmalıdır: (i) toplam hidrojen bağ vericisi sayısı 5'ten büyük olmamalıdır (ii) toplam hidrojen bağ alıcısı 10'dan büyük olmamalıdır (iii) moleküler ağırlık 500 g/mol altında olmalıdır (iv) log P değeri 5'ten küçük olmalıdır. Lipinski'nin bu özelliklerini taşıyan bir molekül bir ilaç adayı olarak görülmesi için gerekli koşulları taşımış olur ancak başarılı bir ilaç olarak kabul edilebilmesi için bu koşulların yeterli olmadığını belirtmek gerekir (Lipinski ve diğ., 2001). İlk iki kural molekülün atom türleri hakkındadır, üçüncü kural molekülün çok büyük olmasına izin vermez, son kural ise molekülün çözünürlüğü hakkındadır.

İlaç moleküllerine bakıldığında oral biyoyararlanımı etkileyen daha birçok özellik vardır. Moleküler esneklik (molecular flexibility) veya molekülün yüzey alanının polaritesi biyoyararlanımı etkileyen diğer faktörlerden biridir (Veber ve diğ., 2002).

Hesaplamalı ilaç tasarımında büyük önem taşıyan bir diğer özellik protein ve ligandın bağlanma afinitesidir. Afinite bağlanma süreci sebebiyle entalpi değişimi (ΔH) ve entropi değişikliği (ΔS) ile doğrudan ilişkilidir ve doğrudan Gibbs serbest bağlanma enerjisi (ΔG) ile bağlantılıdır, burada Gibbs serbest bağlanma enerjisi (ΔG) Eşitlik (2.1) ile verilir.

$$\Delta G = \Delta H - T\Delta S \tag{2.1}$$

Burada T işlemin gerçekleştiği sıcaklığı ifade eder. Hedeflenen bir bağlanma süreci için afinitenin en üst düzeye çıkarılması ve ΔG 'nin ise minimize edilmesi beklenir. Bağlanma afinitesinde entalpi değişiminin ağırlığı, entropi değişiminden daha fazladır (Freire, 2005). Çoğu durumda, minimize edilmiş bir entalpi, maksimize edilmiş entropiye tercih edilir. Maksimuma çıkartılmış entropi daha fazla esneklik ve kendiliğindenlik anlamına gelir, ancak bu değer çok yüksek olmamalıdır çünkü kompleks bir dereceye kadar kararlı olmalıdır, aksi takdirde ilacın etkisi istenilen düzeyde olmayacaktır. Aslında rasyonel ilaç tasarımında, moleküllerin formülasyon kolaylığı, elde edilebilirliği, kararlılığı ve molekülün kristalliği gibi daha pratik özellikler değerlendirilmeye başlanmıştır (Veber ve diğ., 2002).

2.1.2 Moleküler etkileşimler

Yukarıda açıklandığı gibi başarılı bir ilacın hedefiyle arasında güçlü bir afinitesi vardır. Afiniteyi arttırmanın bir yoluda, ligand ile hedefi arasındaki kovalent olmayan bağların sayısını maksimize etmektir. Teorik olarak, ligand molekülü ne kadar çok atoma sahipse, kararlılık o kadar artmaktadır. Bununla birlikte ADME özellikleri, büyük bir molekülün iyi bir ilaç adayı olamayacağını söyler. Bu nedenle, afiniteyi en üst düzeye çıkarırken Lipinski kısıtlamalarını ihlal etmemek kimyasal bileşikler için bir model geliştirirken önem taşımaktadır.

Üç farklı türde kovalent olmayan etkileşim vardır. Bunlar sırasıyla iyonik, hidrojen ve van der Waals etkileşimleridir. İyonik bağlar, zıt yüklü atomların elektron transferiyle oluşur. Hidrojen bağları ise bir hidrojen donörü (veya bir N, F veya I atomuna bağlı bir hidrojen atomu) ile doğada elektronegatif olan bir hidrojen akseptörü arasındaki kutuplaşma ile oluşur. Van der Waals etkileşimi, yeterince yakın olan her atom çift arasında meydana gelir. Bu yakınlık her atom tipi için farklıdır ve buna van der Waals mesafesi denir (Jeffrey, 1997).

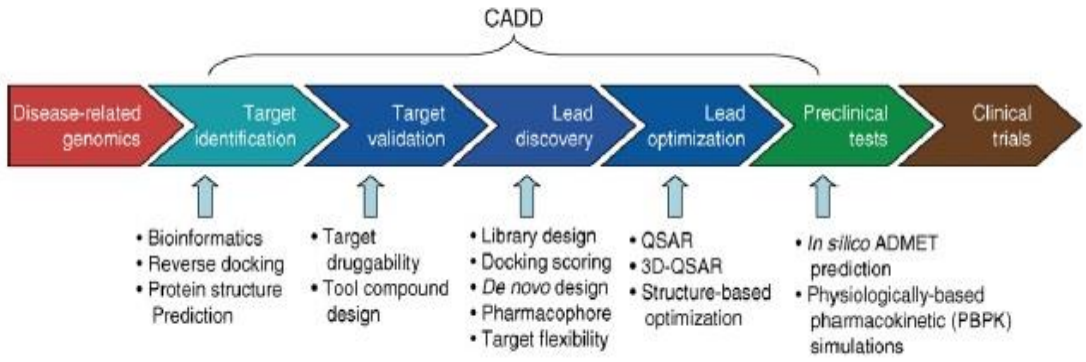
Bir bağın gücü, etkileşimi kırmak için gereken enerjiyle ölçülür, bu enerji ne kadar çok olursa, bağ da o kadar güçlü olur. Kovalent olmayan bağların enerjileri atomdan atoma farklılık gösterir, ancak ortalama atom için bunlar arasındaki en kuvvetli cazibe, su içinde yaklaşık 3 kcal/mol olan bir iyonik bağ ile oluşturulur. Hidrojen bağları bu enerjinin yaklaşık üçte birini taşır ve van der Waals etkileşimleri oldukça zayıftır, yaklaşık iyonik bağların onda bir enerjisine sahiptir (Alberts, 1998).

Kompleksin kararlılığının, taşıdığı kovalent olmayan bağların toplam gücüne bağlı olduğu söylenebilir.

2.1.3 İlaç tasarımında klinik çalışmaların türleri

İnsanlarda kullanılmak üzere, yeni bir ilaç veya biyolojik bir ürün pazarlamak ve satmak için Amerikan Gıda ve İlaç İdaresi (Food and Drug Administration, FDA) veya ABD dışındaki eşdeğer ajanslardan onay almak gerekir. Bu nedenle bu ilaçlar üzerinde bir dizi klinik çalışmalar yapılır. Bu klinik çalışmalar dört aşamada gerçekleşmektedir. Her fazın hasta türleri, hedefleri, dahil edilme/hariç tutulma kriterleri, tasarım özellikleri ve beklenen sonuçlar için belirli ve farklı gereksinimleri vardır. Bu klinik çalışmalar için geçen zaman "Klinik gelişme" olarak adlandırılır ve yaklaşık 5 yıl sürer. Bu süreçte bir ilacın onaylanmasında gerekli bilginin edinilmesine, başarılı bir şekilde pazarlanmasına ve mümkün olan en kısa sürede yeterli güvenlik bilgilerinin verilip piyasaya sürülmesi sağlanır.

İlaç tasarımı evrelerine göre, tasarlanmış yeni molekül tek ve çoklu doz toksisitesinin değerlendirilmesi için az sayıda sağlıklı gönüllü kişi (20-100) üzerinde denenir (faz 1). Faz 1 için geçen süre ~1.5 yıldır. Doz aşımı ve fazla toksisite ilişkisini tanımlamak için daha fazla sayıda gönüllü hasta (100-300) ilacı kullanır (faz 2). Faz 2 için geçen süre ~2 yıldır. Hastalığa maruz kalan birkaç bin gönüllü hasta (1000-3000) ilacın etkin doz aralığını ve bu dozların neden olduğu yan etkileri belirlemek için farklı dozlarda ilacı kullanır (faz 3). Faz 3 için geçen süre ~ 2.5-5 yıl arasındadır. Faz 3 çalışmaları başarılı yeni bir ilaç uygulaması için yeterli ve hedeflenen güvenli bilgiyi sağlamalıdır. İlaç onayını almadan önce, ürünü pazarlarken ve pazarlamadan sonra yapılacak çalışmalar belirlenir. Bunlar farmakoekonomik ve farmakogenetik çalışmalar, ilaç etkileşimi ve karşılaştırıcı çalışmalar gibi basit klinik araştırmalardır (faz 4). Bunlar ilacın temel olarak insanın yaşam kalitesini nasıl etkilediğini araştırır. Bu çalışmalar, ürün onayı sonrasında klinik araştırmalar yoluyla daha fazla bilgiye ihtiyaç duyulduğu ve hastalarda nasıl kullanılacağına dair anlayışımızı genişletir (Evens, 2007). Resim (2.2) ilaç tasarım boru hattına karşılık bilgisayar destekli ilaç tasarımı araçlarını göstermektedir (Drug discovery pipeline vs. computer-aided drug design, CADD).



Resim 2.2: İlaç keşfi boru hattına karşılık bilgisayar destekli ilaç tasarımı (CADD) araçları [Tang ve diğ.,(2006)].

2.1.4 İlaçların marketlerden geri çekilmesi

Ters ilaç reaksiyonları terapötik (veya birincil) hedefin modülasyonundan sonra ortaya çıkan birincil etkiler olarak tanımlanır. Birincil hedef birden çok organda ifade edilir ve aynı anda hedeflenir, böylece hedef dokuda terapötik etkiye ve diğer dokularda istenmeyen etkilere neden olur. İlaca bağlı toksik etkilere yol açan mekanizmaların belirlenmesinde bu etkiler hakkında hücresel ve biyokimyasal seviyede daha net bilgi edinmek oldukça önemlidir. Bu toksikolojik bilgi, ilacı mekanik olarak etkilerini inceleyebilecek ve ilaçların ters ilaç reaksiyonlarına neden olma eğilimini profilleyebilecek uygun bir in vitro deneyler paneli geliştirmek için kullanılabilir. Ters ilaç reaksiyonlarının çoğunlukla doza bağımlıdır. En sık ilişkili hedef organlar karaciğer, kardiyovasküler ve merkezi sinir sistemleridir. Hepatoselüler ve kolestatik ilaç kaynaklı karaciğer hasarı, karaciğer yetmezliği ve hepatik nekroz, karaciğer ile ilişkili tipik ilaç reaksiyonlarının ortak kalıplarıdır. Farklı hasta tepkileriyle ilişkili faktörler, tek nükleotid polimorfizmleri (SNPler) ve mutasyonlar gibi genetik öznitelikleri, cinsiyet, yaş ve birlikte tedavi gibi genetik olmayan nitelikleri içerir. İlaca bağlı olaylar, organlar üzerindeki doğrudan aktiviteden (örneğin kardiyovasküler sistemlerde), ilaçların aktif metabolitlerinin biyolojik taşıyıcılarla etkileşimlerinden reaktivitesine kadar değişen çeşitli etkilerin sonucudur (Siramshetty ve diğ., 2016).

Piyasaya sürülen onaylanmış bir ilaç onayı sonrasında ve pazarlamasından yıllar sonra bile piyasadan geri çekilebilir (Siramsheety ve diğ., 2015). Bu konudaki bir başarısızlık organizasyon üzerinde büyük maddi kayıplara sebep olabilir ve hisse senedi fiyatları üzerinde feci etkiler yaratabilir. Buda ilaç firmalarını doğrudan

küçülmeye götürür. Bir ilacın geri çekilmesindeki en büyük etken ciddi ve beklenmeyen ters etkileridir (Fliri ve diğ., 2005). Bu ciddi yan etkiler genellikle seyrek görülür. Tüm klinik çalışmalarda birkaç yüz ile birkaç bin hastada yalnızca çok az sayıda vaka ortaya çıkar ve bunların yalnızca araştırılan yeni ilaçla ilişkisi olmaz. İlaç şirketleri ilaç üzerinde Ar-Ge çalışmaları için yüz milyonlarca dolar harcamasının yanında birde ürünü pazarlamak için milyonlarca dolar harcar. 1982 ile 2002 yılları arasında 20 yıllık bir süre boyunca piyasadan 21 ilaç kaldırılmıştır (Evens, 2007). Bu ilaçların çeşitli organ sistemleri üzerindeki ters etkilerine bakıldığında çoğunluğun karaciğer (4) ve kalp üzerinde (9) hakim olduğu gözlemlenmiştir. Yapılan bir diğer çalışmada, 2002 yılından 2011 yılına kadar güvenlik nedenleriyle Avrupa Birliği çapında (the European Union, EU) piyasadan çekilen 19 ilaç incelenmiştir (McNaughton ve diğ., 2014). Geri çekilen ilaçlar hakkındaki raporların % 95'ini (18/19), vaka kontrol çalışmaları (4/19), kohort çalışmaları (4/19), randomize kontrollü çalışmalar (RCTs) (12/19) oluşturmaktadır ve meta-analiz raporları (5/19) geri çekimlerin % 63'ünde rol oynamıştır. Kardiyovasküler olaylar ya da bozukluklar, çekilme için ana sebep olmuştur (9/19), bunu karaciğer rahatsızlıkları (4/19) ve nörolojik ya da psikiyatrik bozukluklar (4/19) izlemiştir.

Birincil hedeflerindeki aktiviteleri nedeniyle ters ilaç reaksiyonlara neden olan iyi bilinen bir ilaç sınıfı antiaritmik ilaçlardır ve bunların faydaları tedavi edilen endikasyon olan aritmilerin şiddetlenmesinden dolayı birkaç durumda engellenmiştir. Bu etki esasen kardiyak aksiyon potansiyellerinin düzenlenmesi ile ilişkili olan bir potasyum iyon kanalının (insan Ether-a-go-go-ilişkili gen, hERG) alfa alt biriminin modülasyonundan kaynaklanmaktadır. Sonuç olarak, hERG kanalını inhibe eden ilaçların piyasadan çekilmesine yol açtı, örnek olarak şiddetli aritmiler ve ölüm nedeniyle antihistaminik ilaç terfenadinin geri çekilmesi (Siramshetty ve diğ., 2016).

İlaç güvenliği bilimi CDER'in (CDER Drug Safety Priorities) ilaç etkileri ve ürün kalitesi ile ilgili güvenlik konularını daha iyi tanımlamasına ve yönetmesine yardımcı olan araçlar, teknikler ve veri kaynaklarının geliştirilmesini içerir. FDA'nın düzenleyici bilim portföyünün ayrılmaz bir bileşenidir. Hücresel ve moleküler seviyedeki hastalık süreçlerine ilişkin derin yeni bilgiler, dağıtılmış sayısal ağların, veri madenciliğinin, özel olarak hazırlanmış yazılım platformlarının ve mobil cihaz uygulamalarının teknolojik olarak gelişmiş ortamlarında çalıştığı için ilaç güvenliği

bilimi ters ilaç reaksiyonlarıyla ilgili olayları tahmin etmek için daha güçlü bir potansiyele sahiptir. Yeni onaylanmış bir ilaç piyasaya girdikten sonra pazarlama sonrası deneyim ve klinik araştırmalar sırasında veya onay öncesi inceleme sırasında saptanamayan ters ilaç reaksiyonları olayları ortaya çıkarabilir. FDA pazarlama sonrası ilaç gözetimi amacıyla "pasif" bir sistem yani FAERS (FDA Adverse Event Reporting System) ve "aktif" bir sistem olan "Sentinel Sistemi" olarak bilinen iki önemli sistemi destekler.

İlaç güvenliğinde ilaçların temin edilmesinden saklanmasına, bir yere order edilmesinden sonra transferine, hastalar üzerinde uygulanmasından uygulama sonrası gözetim altında denetlenmesine kadar olayın her aşamasında güvenli olmasını sağlamak amacıyla bir dizi düzenleme yapılmıştır. İlaç güvenliği uygulamalarını yerine getirmek aynı zamanda çalışan güvenliği içinde önemli ve zorunlu bir süreçtir. Medikasyon hatalarını gruplandıracak olursak, alerji bilgilerinin olmaması, atlanan ve uygun olmayan dozlar, veriliş yolundaki uygunsuzluk, dubligasyon, ilaç-ilaç etkileşimleri, ilaç ve yiyecek etkileşimleri, ilaç seçimindeki hatalar, tedavi süresinin yanlış belirlenmesi, hastanın boy ve kilosunun yanlış belirlenmesi, formülasyondaki yanlışlıklar, ilaç verilişlerinin uygun olmayan aralıklarda yapılması, ilaç verilmesinin unutulması, kısıtlanan ilaç kullanımı olarak gruplandırılabilir. Medikasyon hatalarının sonuçları incelenecek olursa ilaç uygulamalarına bağlı olarak gelişen ters ilaç reaksiyonlarının %27-50'si önlenebilirdir. İlaçlar hatalı kullanıldıklarında hasta üzerinde kalıcı yan etkilere neden olabilirler. İlaçların vucuda alındıktan sonra etki etme yolları ilacın uygulanma yöntemi ile ilişkilidir. İlacın hasta üzerindeki etkileri hastaya, ilacın uygulandığı doza ve yola, ilacın metabolizmasına bağlı olarak değişir. İnsülin, morphine, heparin, warfarin, enoxaparin, potassium chloride, furosemide ağır medikasyon hatalarına sıklıkla neden olan ilaçlardan bazılarıdır.

2.2 Veri Madenciliği

Veri madenciliği (data mining) geçmişteki verilerden yola çıkıp matematiksel ve istatistiksel metotlar kullanarak yeni veriler için uygun modeller geliştirmektir. Veri madenciliğinde ağırlıklı olarak makine öğrenmesi teknikleri kullanılır. Bilgisayar teknolojilerindeki ilerlemeler sayesinde çok sayıda verinin analizi artık yapılabilmektedir. Amaç her türlü bilginin sayısal ortama kaydedilebilmesi sayesinde

biriken bu verilerden gerçek bilgiye ulaşmaktır. Örneğin bir mağaza hergün yaptığı satışları ve müşterilerine ilişkin verileri sayısal ortamda topladıkça büyük miktarda veri elde etmektedir. Kurumlar bu devasa verileri saklarken aynı zamanda bunlardan kazanç sağlamak amacıyla karar destek sistemleri geliştirilebilir. Aynı şekilde kullanılan kredi kartlarından, süper marketlerdeki kasalardan, e-ticaret uygulamalarından hergün milyonlarca veri ilgili merkezlerde toplanmaktadır. Bir mağaza, müşterileri hakkında bir portföy oluşturmak ister, benzer şekilde bankalar kredi kartı kullanıcılarının gerçek kimliklerini test etmek ve sahte kullanıcıları asıl sahiplerinden ayırt etmek ister. Benzer problemlerden yola çıkarak verileri çözümlemek ve gerekli bilgiye erişebilmek veri madenciliği (data mining) kavramınıda beraberinde getirmektedir. Veri madenciliği modellerini, sınıflandırma, kümeleme, eğri uydurma (regresyon), özellik seçimi/çıkarımı, ve ilişki belirleme olarak gruplandırabiliriz. Sınıflama veri tabanlarındaki gizli örüntüleri belirlemede kullanılır. Verilerin bir kısmı eğitim amaçlı kullanılırken bunların hangi sınıflara ait olduğu verilir ve yeni gelen verinin hangi sınıfa katıldığının belirlenmesi istenir (Özkan, 2008). Kümeleme ise verilerden birbirine benzer olanların gruplandırılması işlemidir. Burada bir verinin sınıfı/etiketi önceden bilinmez. Geçmiş verilerin sınıfları sürekli sayılardan oluşuyor ise yeni veriye ait model oluşturulurken eğri uydurma kullanılır. Örnek olarak bir hisse senedinin değeri ile ilgili bir model oluşturulurken bu değer sürekli bir sayı olduğundan eğri uydurma ile değer tahmini yapılır. Veri birçok özelliğe sahip iken bu özelliklerden hangilerinin verinin kümesini, sınıfını veya değerini belirlediği bilinmediği durumlarda tüm özellik kümesinden bir alt küme seçilir (özellik seçimi) veya veri için yeni özellikler elde etmek amacı ile bu özelliklerin birleşiminden farklı özellikler elde edilir (özellik çıkarımı) (Amasyalı, 2008). İlişki belirleme ise veritabanında bulunan verilerin birbiriyle olan ilişkilerini ortaya koyar. İlişkiler belirlenirken kayıtlar incelenerek hangi olayların aynı anda gerçekleştiği ortaya konur ve birliktelik kuralları (association rules) elde edilir. Örnek olarak pazar sepet analizleri verilebilir.

İlaç (ligand) ve protein bankalarında depolanan yüksek miktardaki verinin analizi bize veri madenciliği tabanlı ilaç tasarımını ve uygulamalarını mümkün kılmaktadır. Günümüzde veri madenciliği tabanlı ilaç tasarımı anahtar öneme sahiptir. Başarılı bir ilaç tasarımı hem ligand hemde proteinlerin yapısal ve biyokimyasal özelliklerinin çok iyi bilinmesini gerektirir. Bu amaç için çok sayıda yöntem ve yaklaşım

önerilmiştir. Bunlar, kullandıkları veriler ve yöntemler açısından farklılık arzetmektedir. Aşağıda makine öğrenmesi metotlarına dayalı literatürde yer alan kimyasal bileşiklerin sınıflandırılması ve ilaç tasarımı ve uygulamalarına yönelik çalışmaların özet halinde bir analizi sunulmaktadır.

2.3 Veri Madenciliği Yöntemiyle İlaç Tasarımı ve Uygulamaları

Veri madenciliği gelişmiş arama teknikleri ve algoritmaları kullanarak büyük veritabanlarında varolan örüntüleri, bağıntıları ve verilere ilişkin bilinmeyenleri keşfeder (Liao ve diğ., 2012; Witten ve Frank, 2005). Veri madenciliği sayesinde moleküler tanımlayıcılardan oluşan bir dizi ile biyolojik olarak anahtar özellikleri (etki, emilim, dağılım, metabolizma ve atılım, ADMET) ilişkilendiren bir model oluşturulabilir (Shen ve diğ., 2010). Elde edilen model, yeni bileşiklerin anahtar özelliklerinin değerlerini öngörmek ve yapı-aktivite ilişkilerini (structure–activity relations, SARs) belirlemek amacıyla kullanılabilir. Veri madenciliği modelleri, doğrusal tekniklerden türetilen basit, parametrik denklemlerden veya doğrusal olmayan tekniklerden türetilen karmaşık, doğrusal olmayan modellere kadar değişmektedir (Geppert ve diğ., 2010; Weaver, 2004). Veri madenciliğinde sanal tarama (virtual screening, VS) kemoinformatik spektrumda büyük önem taşır çünkü yapılan çalışmalar hedef proteine güçlü bağlanma afinitesi yüksek yeni olasılıkların büyük veritabanlarında araştırılmasını sağlar (Chen ve diğ., 2007). Sanal tarama yöntemleri, mevcut yapısal ve biyoaktivite verilerinin miktarına bağlı olarak yapı temelli (structure-based virtual screening, SBVS) ve ligand tabanlı (ligand-based virtual screening, LBVS) yaklaşımlar olarak sınıflandırılabilir. Reseptörün 3D yapısı biliniyorsa, yapı-temelli sanal tarama yöntemi yüksek verimli moleküler kilitlenme (docking) için kullanılabilir (Lavecchia ve Di Giovanni, 2013), ancak reseptör hakkındaki bilgilerin az olduğu durumlarda ligand tabanlı sanal tarama yöntemleri yaygın olarak kullanılır (Geppert ve diğ., 2010). Ligand tabanlı sanal tarama metotları özellikle reseptör için küçük 3D yapısı mevcutsa ilaç keşfinin başlangıcında önemli rol oynar. LBVS yaklaşımlarını genel olarak benzerlik araştırması ve bileşik sınıflandırma teknikleri olarak ayırabiliriz. Benzerlik araştırmaları moleküler grafikler (2D) veya moleküllerin 3D yapıları (Willett, 2005), 3 boyutlu farmakofor modeller (Mason ve diğ., 2001), basitleştirilmiş moleküler çizge gösterimleri (Gillet ve diğ., 2003) veya moleküler şekil sorgulamaları

(Hawkins ve diğ., 2007) ile yapılır . Türetilen sonuçlar moleküllerin parmak izlerini belirler. Sistem benzerlik metriklerini kullanarak veritabanı bileşiklerini ikili gruplar halinde karşılaştırır ve referans moleküllere azalan moleküler benzerlik sırasına göre bir bileşik sıralaması üretir. Bu sıralamadan aday bileşikler seçilir.

Makine öğrenmesi yaklaşımları bileşik sınıflandırmada yaygın olarak kullanılır (Mitchell, 2014). Bunlara örnek olarak destek vektör makineleri (support vector machine, SVM), karar ağaçları (decision trees, DT), k-en yakın komşular (k- nearest neighbors, k-NN), naive Bayesian metotlar ve yapay sinir ağları (artificial neural networks, ANN) verilebilir. Bu yaklaşımlar LBVS 'de oldukça popülerdir. Bütün bu tekniklerin amacı eğitim kümelerinden türetilen modeller üzerinde bileşik sınıf etiketlerini (aktif veya pasif) tahmin etmek ve aktivite olasılıklarına göre veritabanı bileşiklerinin sıralamasını elde etmektir (Bajorath, 2001). Buna ek olarak bu yöntemlerle bileşiklerin hedef-odaklı bileşik kütüphaneleri için seçimide mümkündür (Schnur ve diğ., 2004). İlaç keşfinde makine öğrenmesinin ilk uygulaması olan alt yapısal analiz (substructural analysis, SSA), Cramer ve diğ. (1974) tarafından biyolojik tarama verisinin otomatik olarak analizi için gerçekleştirildi. Makine öğrenimi artık artan veri koleksiyonlarının kullanılabilirliği ve yeni araçların gelişimiyle bilgisayar biliminde aktif bir araştırma alanıdır (Hand ve diğ., 2001). Birlikte ele alındığında, bilgisayar destekli ilaç keşfinde makine öğrenme yöntemleri geniş bir yelpazede yer almaktadır. Bu nedenle bu alanda yapılan çalışmalar yeni bir ilaç keşfi için önem taşımaktadır.

Gözetimli makina öğrenmesi algoritmalarından (supervised machine-learning algorithms) SVM'ler bileşikleri sınıflama ve regresyon temelli özellik değeri tahminlerine olanak sağlar. SVM'ler genel olarak aktivite tahminlerinde kullanılır. Ayrıca ilaçları, ilaç olmayanlardan ayırt etmek (Zernov ve diğ., 2003), bileşikler arasında spesifik bir aktiviteye sahip olmayanları belirlemek (Warmuth, 2003), ilacın sentetik erişilebilirliğini ve suda çözünebilirliği belirlemek gibi uygulamalarda kullanılır. Öncelikle bileşik kütüphanelerinin iz düşümü büyük boyutlu bir özellik uzayına dönüştürülür. Burada moleküller tanımlayıcı vektörlerle temsil edilir.

LBVS'de bir SVM sınıflandırması tarafından elde edilen skorlar, veritabanındaki bileşiklerin azalan aktivite olasılıklarına göre sıralanmasında başarıyla kullanılmıştır. Bir aday bileşik ile hiperplane arasındaki işaretli uzaklık böyle bir sıralamada kullanılabilir (Jorissen ve Gilson, 2005).

DT'ler spesifik moleküler özellikler ve tanımlayıcı değerleri aktivite ile ilişkilendirmeyi sağlayan kurallar içerir. DT yaklaşımlarının uygulandığı bazı problemler, kombinatoryal kütüphanelerin tasarımı, bir aday molekülün ilaç benzeri (drug-likeness) olmasının ve biyolojik aktivitelerin öngörülmesi ayrıca bileşikler için tanımlayıcı veriler üretilmesidir. Bu yöntem verilen bir veritabanı içerisindeki kimyasal bileşiklerin aktivite durumunu ortaya koyan alt yapıların belirlenmesinin yanında kimyasal bileşiklerin ilaç veya ilaç olmama durumlarına göre sınıflandırılmasını da sağlar (Scheneider ve diğ., 2008). DT'ler aynı zamanda bileşiklerin ADME/Tox özelliklerini öngörmek içinde kullanılırlar. Bu özellikler ilaçların emilimi, dağılımı, çözünürlüğü veya geçirgenliğine ilişkin özelliklerdir (Lamanna ve diğ., 2008; Wang ve diğ., 2015). Bu yaklaşımlar P-glikoprotein (de Cerqueira Lima ve diğ., 2006) veya kan-beyin bariyerinin geçirgenliği (Mente ve diğ., 2005) ile metabolik stabilite'nin belirlenmesinde de önemli rol oynar (Sakiyama ve diğ., 2008).

Naïve Bayesian sınıflandırıcıları genellikle kemoinformatikte kullanılır ve diğer sınıflandırıcılarla karşılaştırıldığında fizikokimyasal özelliklerin tahmininden ziyade biyolojik tahminler için kullanılır. Bileşiklerin toksisite durumlarının tahmini (von Korff ve Sander, 2006), fosfolipidoz mekanizması (Lowe ve diğ., 2012), hedef proteinin belirlenmesi ve ilaç benzeri moleküllerin biyoaktivitelerine göre sınıflandırılması problemlerinde de kullanılır.

K-NN algoritması en basit makine-öğrenme algoritmalarından biridir. Bir molekülün sınıfını (Kauffman and Jurs, 2001), özelliklerini (Konovalov ve diğ., 2007) veya rankını (Votano ve diğ., 2004) özellik uzayındaki en yakın eğitim örneklerini temel alarak tahmin eder. Ayrıca regresyon uygulamaları içinde kullanılır.

ANN'ler esnek hesaplamada oldukça popüler ve derinlemesine çalışılan tekniklerdir (Patel ve Chaudhari, 2005). Tıbbi kimyada ANN'ler bileşik sınıflaması, QSAR çalışmaları (Gleeson ve diğ., 2006), bileşiklerin birincil sanal taramalarında, potansiyel ilaç hedeflerinin tanımlanması ve biyopolimerlerin yapısal ve fonksiyonel özelliklerinin lokalizasyonu için kullanılırlar (Patel ve Goyal, 2007). ANN'ler teknikleri, robotik, model tanımlama, psikoloji, fizik, bilgisayar bilimleri, biyoloji ve diğer alanlarda da kullanılmaktadır (Fogel, 2008).

3. VERİ MADENCİLİĞİ TABANLI İLAÇ SINIFLANDIRMA ÇATISI

3.1 Çalışmalarda Kullanılan Yöntemler Ve Yaklaşımlar

Tezde ilaç veri setleri üzerinde yapılan çalışmalar üç bölümden oluşur. Bunlar kullandıkları ilaç verileri ve yöntemler açısından farklılık gösterir. Yapılan çalışmaların hepsinde ilaç moleküllerine ilişkin global moleküler, boyut, şekil ve ToxPrint özellikleri hesaplanıp elde edilen değerler sınıflandırma problemlerinde kullanılmıştır. İlaç moleküllerine ait SDF dosyaları KEGG, PubChem ve DRUGBANK veri bankalarından toplanmıştır. İlk bölümde, ilaçların onaylanmış ve geri çekilen kategorilerine göre sınıflandırılması için destek vektör makineleri (support vector machine, SVMs) ve güçlendirilmiş karar ağaçları (boosted and bagged trees) gibi topluluk yöntemleri (ensemble methods, EM) kullanılmıştır. Ayrıca, aday ilaç moleküllerinin risk ve güvenlik değerlendirmelerinin belirlenmesi için 700'ün üzerinde önceden tanımlanmış kemotip içeren CORINA Symphony programı ilaç moleküllerin Toxprint kemotiplerini tanımlamak için kullanılmıştır (Sharif ve diğ., 2015). Buna ek olarak, NS ilaçları üzerinde sık alt çizge madenciliği metotlarından gSpan algoritması kullanılarak geri çekilen ve onaylanmış ilaç moleküllerinin % 60'ında bulunan fragmanlar ve bu fragmanlardan yalnızca geri çekilen ilaçların yapısında bulunan ayırt edici fragmanlar ilaç tasarımlarına katkı sağlamak amacıyla belirlenmiştir.

NS ilaçlarının marketlerden geri çekilme oranı diğer hastalık gruplarına oranla daha yüksektir. Çünkü sinir sistemi için alınan ilaçlar periferik etkilere sahiptir yani beyine ulaşan ilaçlar aynı zamanda çok sayıda organı da etkiler. İkinci bölümde oluşturulan model aday ilaç molekülleri için moleküllerin onaylanmış, geri çekilen veya sinir sistemi ilacı olma durumu hakkında bize önceden bilgi verir. Çalışamada farklı hastalık gruplarına ait 558 ilaç Clus-HMC algoritması kullanılarak hiyerarşik olarak üç temel seviyede sınıflandırıldı. İlk seviye bütün ilaçları içermektedir (All Drugs). İkinci seviyede ise 3 grup yer almaktadır. Bunlardan ilki onaylanmış NS ilaçlarını içermektedir (NSADs). İkinci grup ise diğer hastalık gruplarına ait onaylanmış ilaçları (The other ADs) son grup ise piyasadan geri çekilen ilaçları

kapsamaktadır (WDs). Son seviyede ise toplam 5 grup yer almaktadır bunlar onaylanmış NS ilaçlarının Anatomik Terapötik Kimyasal (ATC) sınıflamasına göre N02, N03, N04, N05, N06 gruplarından ilaçları içermektedir. Model onaylanmış NS ilaçlarını diğer onaylanmış ve geri çekilen ilaçlardan ayırt etmemizi sağlarken bunun yanında ATC sınıflamasına göre aday molekülün hangi sinir sistemi hastalık grubu ilacına dahil olduğu hakkındada önceden bilgi verir.

Son olarak farklı hastalık gruplarına ait 1200'den fazla ilaç aday molekülleri test etmek amacıyla oluşturulan modellerde kullanılmak üzere ilaç bankasından toplanmıştır. Sınıflandırmada etkin rol oynayan moleküler tanımlayıcılar tezde önerilen etkin öznitelik seçme stratejisi ile belirlenmiştir. Geliştirdiğimiz etkin öznitelik seçme stratejisi ile ortaya çıkan modelin doğruluğu arttırılmıştır. Çalışmada ayrıca özellik seçimi/çıkarma metotları ile öne çıkan tanımlayıcılar (features) belirlenmiş ve bu tanımlayıcılar ilaç moleküllerinin yapı tabanlı analizleri için ayrıntılı olarak incelenmiştir. Buna ek olarak dengesiz veri setlerini sınıflandırmak amacıyla bir model önerilmiş ve ilaç veri setlerine uygulanmıştır.Yapılan çalışmalara ilişkin büyük resim Şekil (3.1)'de gösterilmiştir.

Bu bölümde tezde ilaç tasarım problemlerinde kullanılan makine öğrenmesi metotları, geliştirilen yaklaşımlar, kullanılan programlar yanında çalışılan ilaç veri kümelerinin formatları, moleküllerin sayısal verilere dönüştürülmesi işlemi ve ilaç moleküllerini tanımlamada kullanılan özellikler detaylı bir şekilde anlatılacaktır.

3.2 İlaç Veri Setleri İçin Kullanılan Formatlar

İlaç uygulamaları için yapılan çalışmalarda moleküllerin gösterimi için SDF dosya formatı (Structure Data Format) ve SMILES dosya formatı (Simplified Molecule Input Line Entry System) kullanıldı. SDF dosya formatı moleküllere ilişkin atom sayısı, bağlar, içerdikleri atom listeleri, chiral etiket ayarları, bağlantı tablosu (connection table, Ctab) sürümü, atom topluluğunun yapısal ilişkilerini ve özelliklerini açıklayan bilgileri içerir. Atomlar tamamen veya kısmen bağlarla bağlanabilir. Bu gibi koleksiyonlar, örneğin, molekülleri, moleküler parçaları, alt yapıları, ikame edici grupları, polimerleri, alaşımları, formülasyonları, karışımları ve birbirine bağlı olmayan atomları tanımlayabilir (bir atom da bağlantısız bir parça olabilir). SDF dosya formatına ilişkin bir örnek Resim (3.1)'de verilmiştir.

DRUGBANK, KEGG DRUG, PubChem

1.Aşama: İlaç moleküllerine ait SDF Dosyalarının ilaç veri bankalarından toplanması.

2.Aşama: İlaç moleküllerinin risk ve güvenlik değerlendirmelerinin belirlenmesi.

760 moleküler tanımlayıcı (öznitelik) her bir ilaç molekülü için CORINA Symphony Programı ile hesaplandı. Bunlar global moleküler, boyut ve şekil, ToxPrint Kemotip özniteliklerdir. Ayrıca moleküllerin ToxPrint Kemotip analizlerinin yapılması.

3.Aşama: Moleküler tanımlayıcılar (öznitelikler) için boyutsal küçültme.

--- Tezde geliştirilen etkin öznitelik seçme stratejisi

ile sınıflandırmada en etkin öznitelikler belirlendi ve elde edilen özniteliklerden oluşan altküme seti sınıflandırma performansını arttırdı.

4.Aşama: Onaylanmış/geri çekilen ilaç moleküllerinde bulunan/bulunmayan ToxPrint Kemotipler belirlendi ve analiz edildi.

5.Aşama:

5a. Sinir sistemi ilaçlarının onaylanmış ve geri çekilen kategorilere sınıflandırılması.

---SVMs (L SVM, MG SVM, CG SVM)

---EMs (BS T, BG T)

5b. Farklı hastalık gruplarına ait 558 ilacın hiyerarşik olarak üç temel seviyede sınıflandırılması.

---Clus-HMC algoritması

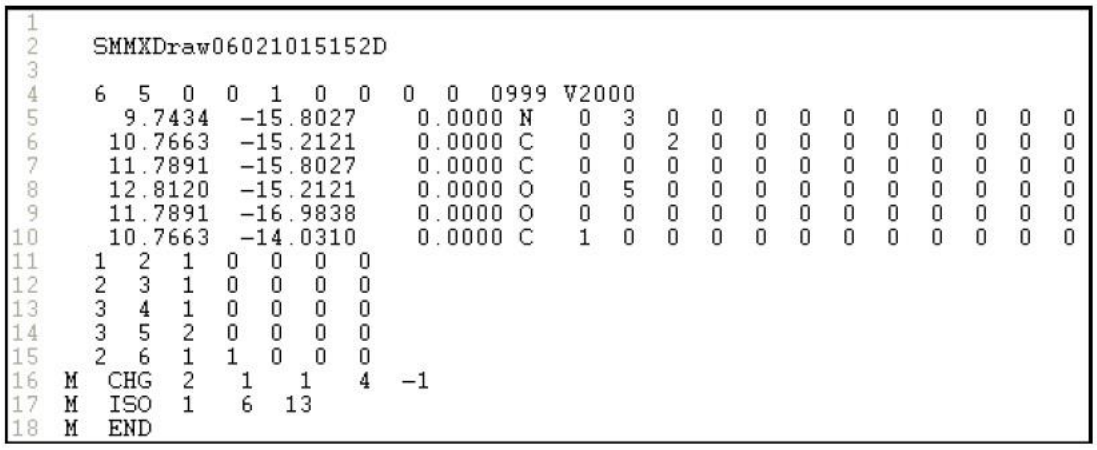
5c. Farklı hastalık gruplarına ait 1200'den fazla ilacın onaylanmış ve geri çekilen kategorilerine ayrılması.

--- Dengesiz ilaç veri seti üzerinde sınıflandırıcı topluluk tasarımı için modelönerildi.

6.Aşama: Onaylanmış ve geri çekilen ilaç veri setlerine sık alt çizge madenciliğinin uygulanması. Çalışmada onaylanmış ve geri çekilen ilaçlar için ayırt edici fragmanların belirlenmesi.

---gSpan

Şekil 3.1: Yapılan çalışmalara ilişkin büyük resim.



```
1
2     SMMXDraw06021015152D
3
4     6  5  0  0  1  0  0  0  0  0999 V2000
5       9.7434  -15.8027    0.0000 N   0  3  0  0  0  0  0  0  0  0  0  0
6      10.7663  -15.2121    0.0000 C   0  0  2  0  0  0  0  0  0  0  0  0
7      11.7891  -15.8027    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
8      12.8120  -15.2121    0.0000 O   0  5  0  0  0  0  0  0  0  0  0  0
9      11.7891  -16.9838    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
10     10.7663  -14.0310    0.0000 C   1  0  0  0  0  0  0  0  0  0  0  0
11    1  2  1  0  0  0  0
12    2  3  1  0  0  0  0
13    3  4  1  0  0  0  0
14    3  5  2  0  0  0  0
15    2  6  1  1  0  0  0
16  M  CHG  2   1   1   4  -1
17  M  ISO  1   6  13
18  M  END
```

Resim 3.1: Alanine'nin V2000 formatındaki örnek bir SDF dosyası [kaynak: Symyx CTfile Formats from www.symyx.com].

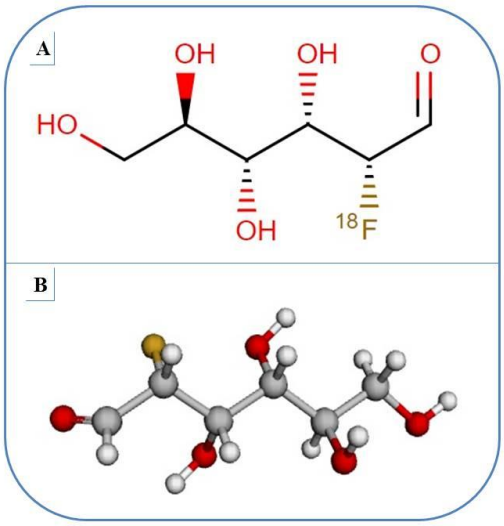
Resim 3.1'deki satır numaralarına göre 2.ci ve 3.cü satırda yorum, 4.cü satırda atom sayısı, bağlar, atom listeleri, chiral etiket ayarları ve Ctab versiyonuna ilişkin bilgiler yer almaktadır. 5-10'a kadar olan satırlarda spesifik atom sembolleri, her atom için kütle farkını, yükü, stereokimyayı, her bir atomla ilişkili hidrojenler verilir. 11-15'e kadar olan satırlarda arada bir bağla bağlı iki atom, bağ türü, her bağ için herhangi bir bağ stereokimyası ve topolojisi (zincir veya halka özellikleri) belirtilir. 16-18'e kadar olan satırlar ise daha önceki Ctab yapılandırmalarıyla uyumluluğu korurken gelecekteki Ctab özelliklerinin genişletilebilirliğini sağlar.

SMILES formatı ise bir molekülün bağlanabilirliği ve kiralitesine ilişkin bilgileri içeren genellikle QSAR uygulamalarında kullanılan doğrusal bir metin biçimidir. Örnek olarak Fludeoxyglucose ilaç molekülünün SMILES formatında gösterimi, [H]C(=O)[C@H]([18F])[C@@H](O)[C@H](O)[C@H](O)CO'şeklindedir.

3.3 Moleküllere Ait Özniteliklerin Hesaplanması

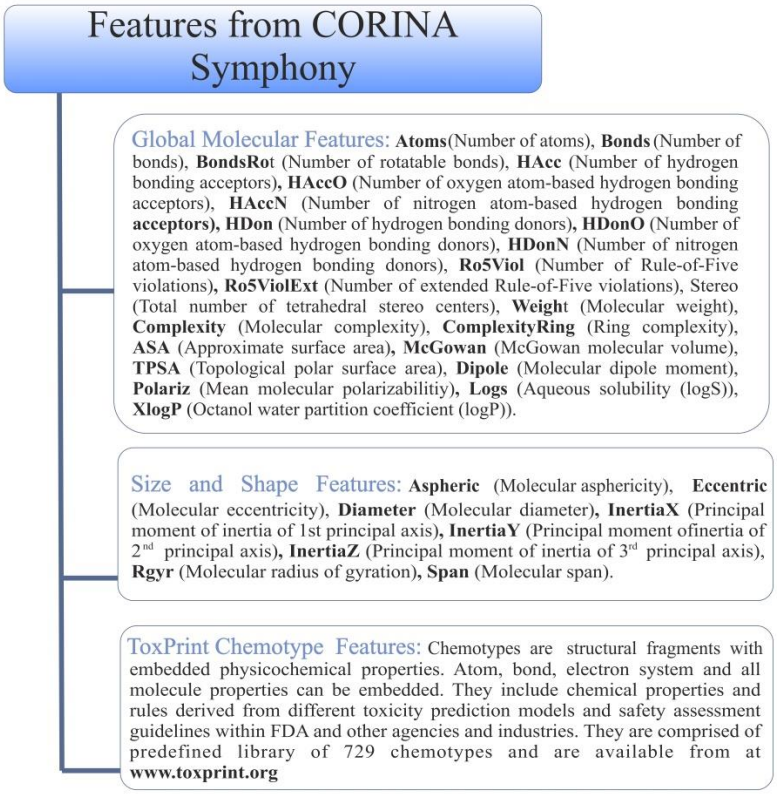
Moleküler tanımlayıcılar yani ilaç moleküllerine ait öznitelikler bir moleküle ait önemli yapısal özellikleri basit bir matematiksel gösterimle sunar. Çalışmada moleküler tanımlayıcıları hesaplamak amacı ile CORINA Symphony programı, Molecular Networks Inc. ilaç veri setlerine uygulanmıştır. Program SDF formatındaki ilaç verilerinin kimyasal bilgilerini alıp bunları her bir ilaç molekülü için, bazı standart deneylerin sonucunda elde dilen fiziko kimyasal özelliklere ve güçlü kemotip profilleme yaklaşımına dayanarak yararlı bir sayıya dönüştürür. Program aynı zamanda ToxPrint kemotip kütüphanesine erişim sağlar. Bu kemotip kütüphanesi toksisite uzayında kimyasal veri setlerinin profillemesi için 729 tane

önceden tanımlanmış kemotipler içermektedir. Kemotipler, elementler, bağ türleri, zincirler ve molekül halkaları, organik gruplar ve ligandlar, inorganik, organometalik vb. gibi değerli özellikleri içerir (Mitchell, 2014). Global molekül tanımlayıcıları molekülün büyük formüllerinden, iki boyutlu (2D yapı) ve üç boyutlu (3D) yapısından türetilir. Benzer şekilde, boyut ve şekil tanımlayıcıları bir molekülün 3D yapısından türetilir. Şekil (3.2) Fludeoxyglucose molekülünün özelliklerinin belirlenmesinde kullanılan 2D ve 3D yapılarını göstermektedir.



Şekil 3.2: Fludeoxyglucose molekülünün (A) 2D ve (B) 3D yapısını göstermektedir.

CORINA Symphony moleküler tanımlayıcıları altı kategoride hesaplar. Bunlar global moleküler (global molecular), şekil ve boyut (size and shape), 2D ve 3D otokorolasyon (2D and 3D autocorrelation), 3D özellik-ağırlaştırılmış RDF (3D property-weighted RDF) ve molekülün yüzey özelliklerinin otokorelasyonuna (autocorrelation of surface properties) ilişkin tanımlayıcılardır. Çalışmalarda ilaç molekülleri için hesaplanan 760 moleküler tanımlayıcı Şekil (3.3)'te verildi. ToxPrint kemotiplerine ait özellikler, FDA ve diğer federal ajanslar ve endüstrilerdeki çeşitli toksiklik tahmini modellerinden ve güvenlik değerlendirme kılavuzlarından türetilen bir dizi kimyasal özellik ve kurallardır. Bu özellikler www.toxprint.org'da (Yang ve diğ., 2015) ayrıntılı olarak verilmiştir. Kimyasal bileşiklerin içerdikleri kemotipler toksisitelerini tahmin etmede birer parmak izi gibidir. Bu nedenle bu özellikler kimyasalların risk ve güvenlik değerlendirmesi yapılırken onlara bir profil oluşturur.



Şekil 3.3: CORINA Symphony ile hesaplanan moleküler tanımlayıcılar.

Corina Symphony moleküler tanımlayıcıları hesaplamadan önce kimyasal bileşik setlerine kimyasal yapıların temizlenmesi ve standardizasyonu amacıyla bir dizi önceden tanımlanmış adım uygular. Bu adımlar, (i) kimyasal bir yapıya sahip olmayan kayıtları bağlantı tablosundan çıkartır, (ii) kimyasal kayıtlardan tuz veya çözücü moleküllerde bulunan karşıt iyonlar gibi küçük parçaları çıkartır (iii) kimyasal yapılarda formal yüklerin nötralize edilmesini sağlar (iv) yinelenen kimyasal yapıların tespiti ve çıkarılmasını sağlar ve son olarak (v) CORINA tarafından 3D yapılar üretilir. Sonrasında yukarıda anlatılan moleküler tanımlayıcılar hesaplanır.

3.4 Çalışmalarda Kullanılan İlaç Veri Bankaları

Onaylanmış ve geri çekilen ilaçların tümü, KEGG DRUG, PubChem ve DRUGBANK veri tabanlarından toplandı. KEGG DRUG Japonya, ABD ve Avrupa'da kimyasal yapılara ve kimyasal bileşenlere dayalı olarak birleştirilmiş, hedef, metabolize eden enzim ve diğer moleküler etkileşim ağı bilgileri ile ilişkili onaylanmış ilaçlar için kapsamlı bir ilaç bilgi kaynağıdır. Sadece reçeteli ilaçlar değil aynı zamanda OTC (over the counter, reçetesiz) ilaçları olmak üzere Japonya'daki

tüm pazarlanmış ilaçlar, KEGG DRUG'ta tam olarak temsil edilir. İlaçların etkinlik alanları ve terapötik kullanımının belirtilmesi amacıyla KEGG BRITE farklı ilaç sınıflandırma sistemlerini destekler. Çalışmalarda Anatomik Terapötik Kimyasal (ATC) sınıflamasına dahil ilaçlar kullanıldı. Sinir sistemi hastalıklarının tedavisinde kullanılan onaylanmış ilaçlar N Nervous System alt sınıfından toplandı. Bu sınıf NS ilaçlarını yedi gruba ayırmıştır. Aşağıda sinir sistemine ilişkin ilaç veri setlerinin ATC sınıflaması yer almaktadır, Çizelge (3.1).

Çizelge 3.1: Sinir sistemi ilaçlarına ilişkin ilaç veri setlerinin ATC sınıflaması.

KEGG Anatomical Therapeutic Chemical (ATC) Classification	
N	Nervous System
N01	Anesthetics
N02	Analgesics
N03	Anti-epileptics
N04	Anti-parkinson
N05	Psycholeptics
N06	Psychoanaleptics
N07	Other nervous system drugs

Çalışmalarda kullanılan diğer ilaç veri setleri sindirim sistemi ve metabolizma, kan ve kan yapıcı organlar, dolaşım sistemi, dermatolojik, üreme ve hormonel fonksiyonlar, sistematik kullanım için antienfektifler, kas-iskelet sistemi, antiparaziter ürünler, böcek ilaçları ve kovucular, solunum sistemi, duyu organları'na ilişkin hastalıkların tedavisinde kullanılan ilaçlarından oluşur. Bunun yanında yine çalışmalarda kullanılan tüm onaylanmış ve geri çekilen ilaç veri setleri 14.02.2016 tarihine kadar olan ve o tarihe kadar güncellenmiş ilaç veri bankalarından elde edilmiştir. Bir başka değişle ilaç veri bankaları yeni onaylanmış ve geri çekilen ilaçlar belirlendikçe sistem kendini yenilemektedir. İlaç moleküllerine ait .sdf ve .smiles dosyaları 14.02.2016 tarihinde ilaç veri tabanlarının yenilenen formundan elde edilmiştir.

PubChem küçük moleküllerin biyolojik aktiviteleri hakkında bilgi sağlar. PubChem, NCBI'nın Entrez bilgi alma sistemi içinde üç bağlantılı veri tabanı olarak düzenlenmiştir. Bunlar PubChem Substance, PubChem Compound ve PubChem BioAssay'dir. PubChem ayrıca kullanıcılara hızlı bir kimyasal benzerlik arama aracı da sağlar. DRUGBANK veritabanı, ilaç hedefi (dizi, yapı ve yolak) ile ayrıntılı ilaç

(kimyasal, farmakolojik ve farmasötikal) verilerini birleştiren benzersiz bir biyoenformatik ve kimyasalenformatik kaynaktır. DRUGBANK/Structures kısmında bulunan onaylanmış ve geri çekilen ilaçlar bu gruplara ilişkin ilaç tasarım problemlerinde bir model oluşturulmak üzere toplandı.

3.5 Modellerin Uygulama Sınırları

Bu çalışmada geliştirilen modeller, Lipinski'nin beş kuralına uyan ilaçlar için daha uygundur (Ro5). Çünkü modellerinin elde edilmesinde ilaç veri bankalarındaki onaylanmış ve geri çekilen ilaçlar kullanılmıştır. Ilaçların özellik dağılımlarının yüzde 85'i temel alınarak hesaplanan Ro5'e göre, moleküler kütle <500 dalton ve toplam azot-hidrojen ve oksijen-hidrojen bağ sayıları (hidrojen bağ donörleri) <5'tir. Ayrıca, toplam azot ve oksijen atomu sayısı (hidrojen bağı alıcıları) <10 ve, log P değeri 5'den fazla olmamalıdır (Lipinski ve diğ., 1997). İlaç molekülleri bu dört kriterin en az üçünü sağlamalıdır.

Etkili bir ilaç olmak için, bir maddenin hem suda hem de yağda çözülebilmesi gerekir. Ağızdan alınan ilaçlar bağırsak astarından geçmeli ve sulu kan içinde taşınmalı, daha sonra bir hücrenin iç kısmına ulaşmak için hücre zarına nüfuz etmelidir. Hücre zarı için model bileşik oktanol olup, log P_{ow} olarak bilinen oktanol/su bölme katsayısının logaritması çözünürlüğü hesaplamak için kullanılır. Suda çözünürlük, molekül içindeki hidrojen bağ donörlerinin sayısına karşı alkil yan zincirlerle hesaplanabilir. Düşük su çözünürlüğü, yavaş emilim ve etki anlamına gelir. Öte yandan çok fazla hidrojen bağı vericisi, düşük yağ çözünürlüğüne neden olduğundan, ilaç hücre duvarına nüfuz edemez. Molekül ağırlığı ne kadar düşükse, o kadar iyi olur. Toplam ilaçların% 80'inde moleküler ağırlık 450 dalton'un altındadır. Bir bileşik bu kuralları sağladığında hücre zarı geçirgenliğine sahip ve vücutta kolayca emilmesi daha olasıdır (Lipinski ve diğ., 1997; Lipinski ve diğ., 2001). Aday ilaçlarla ilgili genel olarak belirlenen bu kurallar tek başına farmakolojik aktiviteyi belirlemede yeterli değildir ancak yeni ilaç keşfi için molekül özelliklerini belirlemekte bir başlangıç noktası olarak kullanılabilir (Rester, 2008; Yusof ve diğ., 2013; Yusof ve diğ., 2014).

3.6 Moleküler Tanımlayıcılar İçin Boyutsal Küçültme

Özellik çıkarımı veya seçimi (feature extraction or feature selection) veri madenciliğinde model sınıflamasında ön işlem adımlarından biridir. Etkili bir boyut

azaltma tekniğidir ve gürültülü özellikleri kaldırmak için gerekli ön işleme yöntemidir (Krishnapuram ve diğ., 2004). Özellik seçimi algoritmalarının temel fikri, verilerin olası bileşim kombinasyonlarını araştırıp, hangi özelliklerin alt kümesinin tahmin için en uygun olduğunu bulmaktır. Seçim, özellik vektörlerinin özellik sayısını azaltarak, ilgisiz veya gereksiz olanları kaldırarak yapılır (Liu ve diğ., 2009). Özelliklerin altkümelerinin oluşturulması ve değerlendirilmesi sırasında artan özellik, sınıflandırma sorununun dezavantajlarını getirir.

Sınıflama modellerinin geliştirilmesinde kullanılan moleküler tanımlayıcıların hepsi onaylanmış ilaçları geri çekilen ilaçlardan ayırt etmek için uygun değildir. Bu nedenle sınıflandırıcıların tahmin performansını geliştirmek amacıyla gereksiz olan tanımlayıcılar ki-kare özellik seçme metodu (chi-square attribute selection method) ve tezde geliştirilen etkin öznitelik seçme stratejisi kullanılarak elendi. Seçilen moleküler tanımlayıcılar sınıflandırma için daha fazla bilgi sağlar. Buna ek olarak ki-kare özellik seçme metodunda arama metodu (search method) olarak ranker kullanılmıştır. Her bir veri seti için sınıflandırmada en etkin moleküler tanımlayıcılar belirlendi. Bu tanımlayıcılar ilaç adayı molekülleri sınıflandırma problemlerinde büyük önem taşımaktadır.

3.6.1 Ki-kare öznitelik seçme metodu

Ki-kare testi iki değişkenin birbiri arasındaki bağımlılık durumunu test eder. İlk olarak ki-kare istatistiğinin değerini sınıflara göre hesaplayarak bir özelliğin değerini belirler. Sonra saptanan önemlilik ve serbestlik derecesine göre ki-kare nicel değeri göz önüne alınarak veriler arasındaki birbiriyle uyuşmayan özellikler belirlenene kadar arka arkaya özellikler ayrıştırılır. Ki-kare değeri arttıkça özelliğin sınıf içerisindeki bağımlılığı artar, sıfır değerini aldığında ise o küme için bağımsız olduğunu belirtir. Eşitlik (3.1-3.2) ki-kare değerinin hesaplanmasında kullanılır.

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{k}\frac{\left(A_{ij}-E_{ij}\right)^2}{E_{ij}} \tag{3.1}$$

$$\mathrm{E_{ij}} = (\mathrm{R_i} * \mathrm{C_j})/\mathrm{N} \tag{3.2}$$

Eşitlik (3.1)'de k sınıf sayısını, i satır ve j sütun olmak üzere A_{ij} gözlenen frekansı, E_{ij} ise A_{ij}'nin beklenen frekansını belirtir. Eşitlik (3.2)'de ise R_i i'deki veri sayısını, C_j j'deki sınıfta gözlemlerin sayısını ve N sınıflardaki gözlemlerin

toplamını göstermektedir. Ranker, özellikleri bireysel değerlendirmelerine göre sıralar. Sinir sistemi ilaç veri setleri üzerinde ki-kare özellik seçimi metodu uygulandıktan sonra eğitim seti üzerinde özelliklerin sayısı azaldı. Bir özelliğin sınıf içerisindeki ki-kare değeri sıfır olduğunda bu özellik çıkarıldı ve tahmin için en uygun olan özellikler belirlendi. Sınıflandırma modelleri seçilen bu özelliklerle eğitildi. Çalışmalar ki-kare özellik seçimi metodunun sinir sistemi ilaç veri setleri üzerinde sınıflandırıcıların ayırma yeteneğini arttırdığını gösterdi.

3.6.2 Altküme seçimi metodu

Korelasyon tabanlı alt küme seçimi metodu (correlation-based feature subset selection, Cfs Subset Eval) Bölüm (6)'da farklı hastalık gruplarına ait ilaç veri setine uygulandı. Tezde Bölüm (6)'da geliştirilen dengesiz veri seti üzerinde etkin öznitelik seçme stratejisi ile belirlenen özniteliklerle elde edilen modelin performansı, altküme seçimi metodu sonrası belirlenen öznitelikler ile elde edilen modelin performansıyla karşılaştırıldı. Bir özellik alt kümesinin değeri belirlenirken her özellik için bireysel tahmin yeteneğine ve diğer özellikler arasındaki etkisine bakılır. Özelliklerin alt kümeleri belirlenirken, sınıfa bağımlılığı yüksek ancak düşük inter korelasyona sahip alt kümeler tercih edilir (Guyon and Elisseeff, 2003). Diğer bir değişle, bileşenler ve dış değişkenler arasındaki korelasyon ne kadar yüksek olursa, bileşik ile dış değişken arasındaki korelasyonu arttırır. Bileşenler arasındaki karşılıklı korelasyon ne kadar düşük olursa, bileşik ve dış değişken arasında korelasyonda o kadar yüksek olur. Bir başka önemli noktada, bileşikteki bileşen sayısı arttıkça bileşik ve dış değişken arasındaki korelasyon artar (Hall, 1999). Bir testte bileşenlerin herbiri ile dış değişken arasındaki korelasyon bilinirse ve bileşenlerin her bir çifti arasındaki karşılıklı korelasyon verilirse, toplanan bileşenlerden oluşan bileşik bir test ve dış değişken arasındaki korelasyon Eşitlik (3.3) ile verilir (Hall, 1999).

$$r_{zc} = \frac{k\overline{r_{zi}}}{\sqrt{k+k(k-1)\overline{r_{ii}}}} \tag{3.3}$$

Buna göre, r_{zc} toplanan bileşenler ve dış değişken arasındaki korelasyon, k bileşenlerin sayısı, $\overline{r_{zi}}$ bileşenler ve dış değişken arasındaki ortalama korelasyon ve $\overline{r_{ii}}$ bileşenler arasındaki karşılıklı korelasyon ortalamasıdır. Eşitlik (3.3) aslında tüm değişkenlerin bulunduğu standartlaştırılmış Pearson korelasyon katsayısıdır.

Arama metodu Bestfirst özellik alt küme uzayını geri izleme (back tracking) yeteneğiyle güçlendirilmiş aç gözlü tırmanış (greedy hill climbing) algoritmasıyla arar. Özellik seçimi metodunda doğrulama yöntemi olarak 10-kat çapraz doğrulama (10-fold cross validation) kullanılmıştır. Veri kümesi öncelikle 10 eşit parçaya bölünür. Parçalardan birisi test olarak seçilir ve geri kalan eğitim setini oluşturur. Buna göre sınıflandırma fonksiyonlarından elde edilen toplam performans k sayısına bölünerek tek bir sonuç elde edilir (Eşitlik (3.4)).

$$t_i \epsilon\, VK,\ Sınıflandırma\ performansı = \frac{\sum_{i=0}^{k} SF\ (t_{i,}\ VK - t_i)}{k} \tag{3.4}$$

Eşitlik (3.4)'te SF (test ve eğitim) sınıflandırma fonksiyonu, VK veri kümesi, k kat sayısı, t test kümelerinin herbirini göstermektedir. Tezde önerilen etkin öznitelik seçme stratejisi ile Çizelge (6.10) ve (6.11)'de karşılaştırma amaçlı kullandığımız diğer öznitelik seçme algoritmalarının performansı değerlendirilirken meta-sınıflandırma doğruluğu ve seçilen özniteliklerin sayısına bakılmıştır. Yine bu çizelgelerde kullanılan ve aynı veri seti üzerinde diğer öznitelik seçme algoritmaları ve meta-sınıflandırıcılarına ait ayrıntılı bilgilere Bouckaert ve diğ. (2015)'den ulaşılabilir. Seçilen özniteliklerin sayısı ise özellik seçimi sonuçlarının basitliğini ölçer. İki yöntem arasında sınıflandırma doğruluğu aynı olsa bile daha az özellik seçen yöntem tercih edilir. Özellik seçme öğrenme algoritmalarının performansını geliştirmeyi amaçlar (Wang ve diğ., 2013). Buda genellikle sınıflandırma doğruluğu ile ölçülür.

3.7 Sınıflandırma Metotları

Veri madenciliği bir veri kümesindeki saklı bilgileri çıkarır ve bu bilgileri anlaşılabilir bir yapıya dönüştürür (Wassermann ve diğ., 2015). Verilerin sınıflandırılması veri madenciliğinin çalışma alanlarından biridir. Çalışmalarımızda ilaçların onaylanmış ve geri çekilen kategorilerine sınıflandırılması için L SVM (linear SVM), MG SVM (medium gaussian SVM) ve CG SVM (coarse gaussian SVM) kullanıldı. Bunun yanında topluluk halinde kurulan (ensemble) güçlendirilmiş (boosted trees, BS T) ve torbalı karar ağacı (bagged trees, BG T) algoritmaları da sınıflandırma problemlerinde kullanılmıştır.

3.7.1 Destek vektör makinaları (SVMs)

Vapnik ve iş arkadaşları tarafından geliştirilen SVM'ler eğiticili (supervised) makine öğrenme algoritmalarıdır (Vapnik, 2000). SVM'ler bileşik sınıflandırma, aktivite tahmini, regresyon tabanlı özellik değer tahmini, vb. uygulamalarda kullanılırken öncelikle bileşik kütüphaneleri yüksek boyutlu bir özellik alanına yansıtılır. Burada moleküller tanımlayıcılar vektörler olarak temsil edilir ve bileşiklerin doğrusal olarak ayrılabilir hale gelmesi beklenir. Bu yansıtma bir kernel fonksiyonunun kullanılmasıyla sağlanır. Bu fonksiyon gruplarına örnek olarak doğrusal, polinom, sigmoid ve radyal tabanlı fonksiyon (radial basis functions, RBF) verilebilir. İlk üç fonksiyon global kernel sadece RBF lokal (local) kernel'dır. Yapılan çalışmalar RBF tabanlı SVM'nin, diğer üç kernel'dan daha iyi performans gösterdiğini bu nedenlede yaygın olarak kullanıldığını göstermiştir. (Camps-Valls and Bruzzone, 2005). Gaussian veya diğer polinomsal kernel fonksiyonları genellikle ligand tabanlı sanal tarama problemlerinde kullanılırlar (Hinselmann, 2011). SVM kernel'larının seçimi ve kernel parametrelerinin kurulumu büyük oranda ampirik ve deneysel analize bağımlıdır. SVM sınıflandırmasında ilk aşamada örnek veri vektörleri (bileşiklere ait moleküler tanımlayıcılar) kernel fonksiyonuyla çok yüksek boyutlu özellik alanına eşlenir. Bu alanın boyutu orijinal veri alanının boyutundan önemli derecede büyüktür. Doğrudan özellik fonksiyonunu sınıflandırma hiper düzlemini hesaplarken kullanmak pratik değildir. Bunun yerine, özellik fonksiyonlarıyla indüklenen doğrusal olmayan eşlem (nonlinear mapping) doğrusal olmayan özellikli kernel fonksiyonlarıyla hesaplanır. İkinci aşamada sınıflandırıcı veri sınıflarını ayıran en geniş marjınlı yüksek boyutlu özellik uzayında en geniş marjı olan bir hiper düzlem bulur. Yüksek boyutlu özellik alanında hiper düzlemi bulmak herzaman mümkün değildir. Marjın içindeki herbir vektör için ayırıcı marjın boyutu ile cezalar arasında ödünleşme başlar (Cortes and Vapnik, 1995). Kernel parametreleri ve hata ceza faktörü C seçilen kernel'ın sınıfından çok kullanıcı tarafından belirlenir. Bu parametreler ve C'nin seçimi SVM'nin performansındaki belirleyici faktörlerdir (Foody and Mathur, 2006).

SVM'nin temel teorisine göre en geniş marjine sahip ayırma hiper düzlemini bulurken, bir veri kümesindeki örneklerin hepsinin Eşitlik (3.5) ve Eşitlik (3.6)'yı sağlaması gerekir (Soman ve diğ., 2011).

$$f(x_i) = (w, x_i) + b \geq +1 \qquad y_i = +1 \tag{3.5}$$

$$f(x_i) = (w, x_i) + b \leq -1 \qquad y_i = -1 \tag{3.6}$$

Bunları tek bir eşitsizlikle belirtmek istersek Eşitlik (3.7)'yi elde ederiz.

$$\forall_i \qquad y_i\big((w, x_i) + b\big) - 1 \geq 0 \tag{3.7}$$

Bir hiper düzlem w.x+b=0 ise, w normali, $|b|/\|w\|$ orjinden dik uzaklığı verir (Schölkof ve Smola, 2002). Burada x yüksek boyutlu uzaya eşlenmiş örnek vektör, y x'in sınıf etiketi, w ve b SVM sınıflandırıcısının tahmin edeceği hiper düzlemin parametresini gösterir. Kanonik biçimde bir ayrıcı hiper düzlem aşağıdaki kısıtlamaları sağlamalıdır. Bunlar Eşitsizlik (3.8-3.11) ve Eşitlik (3.9-3.10) ile gösterilmiştir.

$$y^i\big[\big(w, x^i\big) + b\big] \geq |w|\tau, i = 1,2, \ldots., n \tag{3.8}$$

Bir eğitim örneğinin hiper düzleme olan uzaklığı Eşitlik (2.9) ile verilir (Gunn, 1998).

$$d(w, b, x) = \big|(w, x^i) + b\big|/\|w\| \tag{3.9}$$

Marjin minimal bir τ olarak ifade edilebilir. Probleme yeni kısıtlamalar gelmediği sürece marjin w'ya $|\, w\, |\tau = 1$ olacak şekilde bir kısıtlama uygular. $\|w\|$, w normal düzleminin normu yani ağırlık vektörüdür. Eğitim örneklerini ayıran en iyi hiper düzlem Eşitlik (3.10)'u minimize eden düzlemdir. (Cao ve diğ., 2012). Buradan SVM eğitimi aşağıdaki kısıtlamalarla bir fonksiyonun minimumunu bulma problemi haline gelir, Eşitlik (3.10) ve Eşitsizlik (3.11).

$$\eta(w) = 1/2\|w\|^2 \tag{3.10}$$

(kısıtlamalara bağlı) $\forall_i \qquad y^i\big[\big(w, x^i\big) + b\big] - 1 \geq 0$ (3.11)

Problem Lagrange çarpanlarını kullanarak ve fonksiyonun minimize edilmesiyle çözülür. Lagrange çarpanları problemi dual probleme dönüştürür ve kolaylıkla çözülmesine imkan verir (Gunn, 1998). Lagrange fonksiyonu Eşitlik (3.12) ile verilir.

$$L_P(w, b, \alpha) = 1/2\|w\|^2 - \sum_{i=1}^{n} \alpha_i \left\{y^i\big[\big(w, x^i\big) + b\big]\right\} + \sum_{i=1}^{n} \alpha_i \tag{3.12}$$

Burada α_i Lagrange çarpanlarıdır. L_P, w ağırlık vektörünü ve b sabitini küçükleyen α_i'nin büyük olmasını sağlayan bir foksiyondur. Lagrange fonksiyonunun w ve b'ye göre türevleri alınır ve elde edilen eşitlikler fonksiyona yerleştirilir. Problem artık dual Lagrange problemine ($L_D(\alpha)$) dönüşmüştür, Eşitlik (3.13-3.14).

$$Max\ L_D(\alpha) = L_D(w, \alpha, b) = \sum_{i=1}^{n} \alpha_i + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i\, \alpha_j\, y_i y_j\, (x_i\, x_j) \qquad (3.13)$$

$$\text{(kısıtlamalara bağlı)} \quad \sum_{i=1}^{n} \alpha_i\, y_i = 0 \quad ve\ \forall \alpha_i \qquad (3.14)$$

Eşitlik (2.13-2.14)'de $L_D(\alpha)$'yı maximum yapan α_i değerleri optimal hiper düzlemin belirlenmesi için elde edilir. Destek vektörlerinin α_i Lagrange çarpanları sıfırdan büyük değer alır. Optimal ayırma hiper düzlemi bunlarla belirlenir. α_i'yi çözerek optimal hiper düzlemin w ve b'si belirlenir, Eşitlik (3.15-3.16).

$$w^* = \sum_{i=1}^{n} y_i\ \alpha_i x_i \qquad (3.15)$$

$$b^* = -1/2\ (w^*, x_i) \qquad (3.16)$$

Buradan elde edilen hiper düzleme bağlı sınıflandırıcı Eşitlik (3.17)'de verildi.

$$f(x) = sign\left((w^*, x_i) + b^*\right) = sign\left(\sum_{i=1}^{n} y_i \alpha_i \left(x_i, x_j\right)\right) \qquad (3.17)$$

Bazı durumlarda, örneğin veri seti gürültülü, karmaşık ve çok boyutlu olduğunda iki sınıflı veri setini ayırmak için soft marjin yaklaşımı kullanılır. Belirli bir hata ile doğrusal ayrılma durumunda, bir örnek yanlış sınıflandırılırsa dahil olduğu karar sınırına olan uzaklığın ölçüsü olan aylak değişkeni ε_i(slack variable) eklenir (Cortes ve Vapnik, 1995). Bu durumda ayırma hiper düzlemini bulmak amacıyla veri kümesindeki örneklerin tümü Eşitlik (3.18-3.19)'u sağlamalıdır (Cortes and Vapnik, 1995).

$$f(x_i) = (\mathrm{w}, x_i) + \mathrm{b} \geq +1 - \varepsilon_i \qquad y_i = +1 \qquad (3.18)$$

$$f(x_i) = (\mathrm{w}, x_i) + \mathrm{b} \leq -1 + \varepsilon_i \qquad y_i = -1 \qquad (3.19)$$

Gerekli dönüşümler yapılarak problem Eşitlik (3.20-3.21)'deki kareli optimizasyon problemine dönüşür. Problemin çözümü için yine Lagrange fonksiyonu kullanılır. Bu doğru sınıflandırma olasılığını arttırır (Cortes ve Vapnik, 1995; Schölkopf ve Smola, 2002).

$$Min\ {}^{1}/_{2}\,\|w\|^2 + C\ \sum_{i=1}^{N} \varepsilon_i \qquad (3.20)$$

(kısıtlamalara bağlı) $\qquad y_i((\mathrm{w}, x_i)+\mathrm{b}) - 1 + \ \varepsilon_i \ \geq \ 0, \qquad \varepsilon_i \geq 0 \qquad (3.21)$

Burada C katsayısı ceza parametresini gösterir ve Lagrange çarpanının alabileceği üst sınır değeridir. $\alpha_i = C$ ise destek vektörleri ayırma hiper düzlemi üzerinde yer alır (Katagiri ve Abe, 2006). Sonuç olarak elde edilen hiper düzleme bağlı sınıflandırıcı Eşitlik (3.22)'de verildi (Burges, 1998).

$$f(x) = sign\ ((w^*, x_i) + b^*) = sign\ (\sum_{i=1}^{n} y_{(i)}\alpha_{(i)}\ (x_{(i)}, x_{(j)})) \qquad (3.22)$$

Çalışmalarda ilaç molekülleri eğitim ve test setlerini sınıflamak amacıyla doğrusal ve radyal tabanlı fonksiyonlar (Gaussian) kernel fonksiyonu olarak kullanıldı. İki sınıftan oluşan ilaç molekülleri doğrusal olarak ayrılabilir olduktan sonra özellik uzayında bir hiper düzlemle ayrılabilir. Aslında sonsuz sayıda bu tür hiper düzlem vardır. Ancak SVM bilinmeyen verilerle uğraşırken sınıflandırıcı hatasını minimize etmek için iki sınıf arasındaki marjini maksimize eden hiper düzlemi seçer. Bu hiper düzlem üzerinde yer alan veri noktaları destek vektörleri olarak adlandırılır. Birbirinden ayrışmayan sınıflar olduğunda yanlış sınıflandırılmış örneklerin sayısını en düşük seviyede tutmak amacıyla marjı maksimize edecek soft marjin yaklaşımı uygulanır.

3.7.2 Topluluk halinde kurulan karar ağaçları

Yapılan çalışmalar topluluk halinde kurulan karar ağacı algoritmalarının sınıflandırma doğruluğunu arttırdığını gösterdi. Çünkü karar ağacı modelleri, yüksek varyansa maruz kalan tahminlerde bulunur. Breiman tarafından 1996'da önerilen bir teknik olan torbalama (bagging) yöntemi, tahmine ilişkin varyansı azaltmaktadır (Breiman, 1996; Breiman, 1998). Freund ve Schapire tarafından 1995 yılında geliştirilen güçlendirme yöntemi, torbalama gibi komiteye dayalı bir yaklaşımdır. Torbalanmış ağaçlar Breiman'ın rastgele orman algoritmasını kullanmaktadır (Freund ve Schapire, 1996). Ayrıca, güçlendirilmiş ağaçlar 1996'da Freund ve Schapire tarafından geliştirilen AdaBoost algoritması ile ilişkilendirildi. Torbalama ve güçlendirme yöntemleri sınıflandırıcıların birleştirilmesinde kullanılan en yaygın metotlardan biridir. Sınıflandırıcılar ayrı ayrı eğitim setleri ile eğitilirler. Elde edilen tahminler doğrultusunda sınıflandırma gerçekleştirilir. Her bir sınıflandırıcı kendi

içinde değerlendirildiğinde, birleştirme ile elde edilen sınıflandırıcının doğruluk oranı daha yüksektir (Opitz ve Maclin, 1999). İki metot arasındaki en önemli fark güçlendirilmiş ağaç algoritmasının sürekli olarak birden çok sınıflandırıcı oluşturmasıdır. Böylelikle bir dizi sınıflandırıcı serisi oluşturur (Nanni ve Lumini, 2006). Torbalama metodunda sınıflandırıcıların her biri ayrı ayrı rastgele eğitim setleriyle eğitilirler. (Breiman, 1996; Efron ve Tibshirani, 1993). Bu yöntemde eğitim setinin belirlenmesinde önceki sınıflandırıcıların performansı dikkate alınmaz. Başlangiçta eğitim seti N boyutlu olsun, yeni eğitim setleride N boyutlu olur ve bu set başlangıçtaki veri setindeki bazı örnekleri içinde bulundurmayabilir yada bir çok örnek içinde tekrar edebilir. Sonuç olarak rastgele örnekleme yöntemiyle elde edilen eğitim verileri sınıflandırıcıların eğitimini gerçekleştirilir (Breiman, 1996).

3.7.3 Hiyerarşik çoklu-etiket sınıflaması

Hiyerarşik çoklu etiket sınıflaması (hierarchical multi-class classification, HMC), örneklerin aynı anda birden çok sınıfa ait olabileceği ve bu sınıfların bir hiyerarşide düzenlendiği bir sınıflandırma varyantıdır.

Sınıflandırma sınıflandırılmış örneklerden oluşan bir setten yeni örneklerin sınıfını öğrenme görevidir (Chen ve diğ., 2009). Hiyerarşik çok etiketli sınıflandırma normal sınıflandırmadan iki şekilde farklıdır. İlk olarak tek bir örnek aynı anda birden fazla sınıfa dahil olabilir. İkincisi sınıflar bir hiyerarşide organize edilir. Yani bir örnek bir sınıfa (class) aitken aynı anda o sınıfın tüm üst sınıflarınada (superclasses) aittir (Hiyerarşi kısıtlaması) (Freitas and Carvalho, 2007). HMC metin sınıflandırması (Rousu ve diğ., 2006), fonksiyonel genomik (Barutcuoğlu, 2006) ve nesne tanımlama alanlarında kullanılabilir. Burada fonksiyonel genomik genlerin fonksiyonlarını tahmin etme açısından önemli bir problemdir. Biyologlar genlerdeki olabilir fonksiyonların setine sahiptir. HMC algoritması genlerin sahip olabileceği işlevleri ve bu işlevleri bir hiyerarşide organize edebilir (Schietgat ve diğ., 2010). Tek bir gen birden fazla fonksiyona sahip olabilir. Bu nedenle, farklı genler arasındaki etkileşimleri anlamak amacıyla yorumlanabilir bir model elde etmek önemli bir problemdir. Buradan yola çıkarak ilaç adayı moleküller için bu yolla oluşturulan bir modelle moleküllerin onaylanmış, geri çekilen veya sinir sistemi ilacı olma durumunun yanında aynı zamanda sinir sistemi ilaçları ele alındığında ilacın hangi hastalık grubuna dahil olduğu önceden yorumlanabilir. Çalışmamızda HMC

algoritması ilaçların dahil olduğu grupları tahmin edip onları bir hiyerarşide organize etmiştir. HMC tüm sınıfları bir kerede öngören bir ağaç öğrenme algoritmasıdır. HMC'nin görevini şu şekilde tanımlayabiliriz (Barutcuoğlu ve diğ., 2006; Blockeel ve diğ., 2002; Vens ve diğ., 2008).

Verilenler:

- X bir örnek uzayı olsun
- Sınıf hiyerarşisi (C, $\leq_h$), C sınıfların bir seti ve $\leq_h$ partial order (hiyerarşi köklü bir ağaç olarak yapılandırıldı) üst sınıf ilişkilerini sunar. Her c_1 ve $c_2 \in C$ $için,\ c_1 \leq_h c_2$ ancak ve ancak c_1 c_2'nin bir super sınıfı ise geçerlidir.
- Örneklerin bir seti T olsun (x_i,S_i) , $x_i \in X$ ve $S_i \subseteq C$ buradan c $\in S_i \Rightarrow \forall c' \leq_h c : c' \in S_i$ ve
- Sınıflandırmanın kalitesi q ile belirtilsin burada tahmin yeteneği yüksek düşük karmaşıklık hedeflenir.

Bulunacak hiyerarşik kısıtlama:

- f bir fonksiyon f: X $\rightarrow 2^C$ burada 2^C, C'nin kuvvet seti, buradan f q'yu maksimize eder ve c$\in$ f(x) $\Rightarrow \forall c' \leq_h c : c' \in f(x)$ 'tir. f karar ağaçlarıyla ifade edilir.

Karar ağacı metotları tahmini kümelenme ağacı (predictive clustering tree, PCT) çerçevesinde belirlenmiştir. Buna göre ağaç bir kümelenme hiyerarşisi olarak karşımıza çıkar. En tepedeki düğüm bütün verileri içeren bir kümeye karşılık gelir. Bu küme yinelemeli olarak daha küçük kümelere bölünür (aşağı yönde). PCT'ler her bölünmede küme içi varyansı azami ölçüde azaltacak şekilde inşa edilmiştir ve burada yukarıdan-aşağı indüksiyon algoritmasıdır. PCT'ler standart karar ağacı algoritmalarına çok benzerler. Aralarındaki temel fark PCT'ler varyans ve prototip fonksiyonlarını parametre olarak görür ve bu parametreler eldeki öğrenme görevine dayanarak örneklendirilir. Bir regresyon ağacı oluşturmak için, varyans fonksiyonu, verilen örneklerin hedef değerlerinin varyansını döndürür ve prototip, hedef değerlerinin ortalamasıdır (Blockeel ve diğ., 1998; Struyf ve diğ., 2005). PCT yapısı Clus-sistemi içine uygulanmıştır (Struyf ve diğ., 2011). Clus bir karar ağacı ve kural öğrenme sistemidir. PCT'ler için yukarıdan aşağı indüksiyon algoritması Çizelge (3.2)'de gösterilmektedir (Schietgat ve diğ., 2010).

Çizelge 3.2: PCT'ler için yukarıdan-aşağı indüksiyon algoritması.

prosedure PCT (I) **returns** tree	**Procedure** BestTest (I)
1: $(t^*, \mathcal{P}^*)$ = BestTest (I)	1: $(t^*, h^* \mathcal{P}^*)$ = (*none*, 0, $\emptyset$)
2: **if** $t^* \neq none$	2: **for each** posible test t
3: **for each** $I_k \in \mathcal{P}^*$	3: $\mathcal{P}$ = partition induced by t on I
4: tree$_k$= PCT (I_k)	4: h = Var (I) -$\sum_{I_k \in \mathcal{P}} \frac{\lvert I_k \rvert}{\lvert I \rvert}$ Var (I_k)
5: **return** node (t^*, $\mathcal{U}_k\{tree_k\}$)	5: **if** ($h > h^*$) $\wedge$ $Acceptable$ $(t, \mathcal{P})$
6: **else**	6: $(t^*, h^* \mathcal{P}^*) = (t, h, \mathcal{P})$
7: **return** leaf (Prototype (I))	7: **return** (t^*, h^*)

Çizelge 3.2'de I eğitim örneklerini, t öznitelik-değer testi, $\mathcal{P}$ t tarafından I üzerinde ayrılmış bölüm ve h t nin sezgisel değeridir. Üst simge * ise mevcut en iyi testi ve onunla ilgili bölüm ve sezgisel metotları göstermektedir. En iyi testi seçmek için (BestTest) algoritma testleri varyans azalması ile puanlandırır. Varyans azalması maksimum olduğunda küme homojenliği en üst düzeye çıkar ve bu tahmin performansını arttırır. Kabul edilebilir herhangi bir test bulunamazsa yani varyansı önemli ölçüde azaltacak test yoksa algoritma bir yaprak (leaf) oluşturur ve onu verilen örneklerin temsili bir örneği veya prototipi ile etiketler.

PCT'ler HMC görevine uygulandığında varyans ve prototip parametreleri örneği aşağıdaki gibidir (Struyf ve diğ., 2011). Örnek bir etiket, Boolean bileşenleriyle bir vektör olarak gösterilir. Eğer vektörün i.ci bileşeni 1 ise örnek sınıf c_i'ye aittir aksi halde bileşenler 0 değerini alır. Örneklerin bir setinin varyansı Eşitlik (3.23) ile verilir.

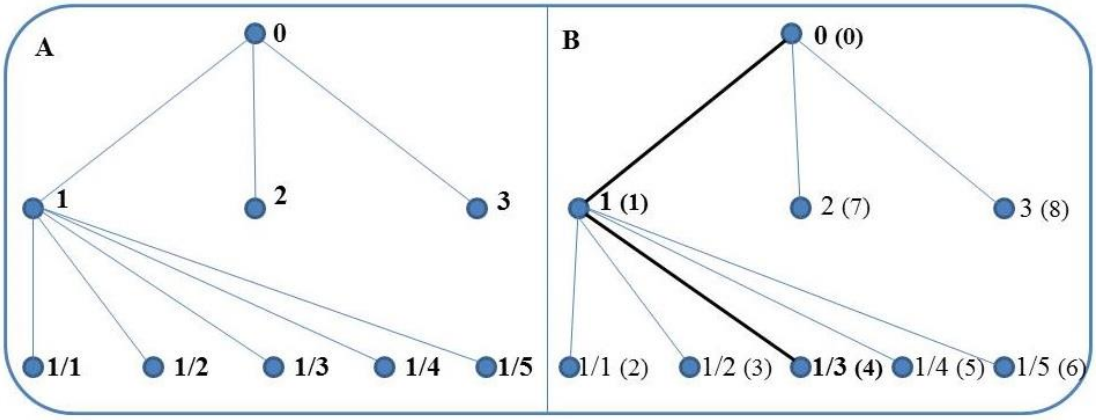
$$\text{Var (S)} = \frac{\sum i \, d(v_i, \bar{v})^2}{|S|} \tag{3.23}$$

Burada v_i her örnek vektör arasında ortalama kare mesafesi ve $\bar{v}$ setin ortalama vektörünü göstermektedir. HMC bağlamında hiyerarşinin yüksek seviyelerinde benzerlikleri dikkate almak alt düzeydekilerden çok daha önemlidir bu nedenle ağırlıklı öklid uzaklığı (weighted Euclidean distance) kullanılır (Struyf ve diğ., 2011). Clus-HMC sınıf vektörleri arasındaki öklid uzaklığına dayalı varyansı hesaplar. Bu uzaklık Eşitlik (3.24)'te verilmiştir.

$$d(v_1, v_2) = \sqrt{\sum i\, w(c_i).(v_{1,i} - v_{2,i})^2} \quad (3.24)$$

Eşitlik 3.24'te $v_{k,i}$, bir örnek x_k olmak üzere örneğin sınıf vektörünün (v_k) i.ci bileşenidir. W(c) sınıf ağırlıklarını temsil eder ve hiyerarşide sınıfın derinliğiyle birlikte azalır. ($w(c) = w_0^{depth(c)}$, with 0 <w_0<1) (Vens ve diğ., 2008). Bir sınıflandırma ağacı çoğunluk sınıfını bir yaprağa depolar ve bu sınıf ağaca varan örnekler için bir tahmin olur. Ancak bizim problemimizde bir örnek birden fazla sınıfa sahip olabileceğinden çoğunluk sınıfı kavramı geçerli olmaz. Bunun yerine, bu yaprakta depolanan örnek vektörlerinin ortalama $\bar{v}$ 'sidir. Yani prototip fonksiyonu $\bar{v}$ 'yi döndürür. Eğer $\bar{v_i}$ bazı eşik t_i değerinin üstünde olursa, örneğin ait olduğu sınıf c_i tahmin edilir. Tahminlerin hiyerarşi sınırlamasını yerine getirmesini sağlamak için $c_i \leq_h c_j$ olduğunda $t_i \leq t_j$ seçmek yeterlidir. Bu eşik değerlerin nasıl seçileceği sınıflandırmanın sonucunda tahmin doğruluğunu maksimum yapma, maksimum F1-skoruna ulaşma gibi hedefler için farklı seçilebilir.

Şekil (3.4)'te ilaçları hiyerarşik olarak gruplarına göre sınıflandırırken bir sınıf setinin vektör olarak gösterimi (v_k) örnek olarak verilmiştir. Yapılan çalışma 4. bölümde detaylı anlatıldı.



Şekil 3.4: İlaç grup hiyerarşisi. (A) Sınıf etiketlerini ve (B) bir sınıf seti örneğini göstermektedir, hiyerarşide kalın çizgiyle gösterilmiştir.

Sınıf etiketleri hiyerarşideki konumu yansıtır. Örneğin '1/3' '1'in bir alt sınıfıdır. Bir sınıf seti {1, 1/3} olsun. Şekil 2.4 (B)'de sınıfların bir setinin vektör olarak gösterimi $v_k = \begin{bmatrix} \overset{(1)}{\vec{1}} & \overset{(2)}{\vec{0}} & \overset{(3)}{\vec{0}} & \overset{(4)}{\vec{1}} & \overset{(5)}{\vec{0}} & \overset{(6)}{\vec{0}} & \overset{(7)}{\vec{0}} & \overset{(8)}{\vec{0}} \end{bmatrix}$ şeklindedir.

Uygulama alanlarına bağlı olarak sınıfların hiyerarşisi her bir sınıf en fazla bir tane parent'a sahipse hiyerarşi ağaç yapısındadır, eğer sınıf birden fazla parent'a sahipse hiyerarşi DAG (directed acyclic graphs) yapısındadır. İlaç moleküllerini sınıflandırma probleminde hiyerarşi ağaç yapısındadır.

Clus-HMC ile sınıflandırma yaparken diğer önemli bir konuda dengesiz sınıf dağılımlarıdır bu durumda Eşitsizlik (3.25) problemin çözümü için oldukça dengeli sınıf dağılımları oluşturur.

$$\frac{N_c}{N} + \frac{N_c}{N_{par(c)}} < 1 \tag{3.25}$$

Burada c sınıfları, N toplam örnek sayısını, *par (c)* c'nin parent sınıfını belirtmektedir. Buna göre N_c, c sınıfına ait örnek sayısını; $N_{par(c)}$, c'nin parent sınıfına ait örnek sayısını göstermektedir. İlaç molekülleri için hiyerarşik çoklu etiket sınıflaması yaparken Eşitsizlik (3.25) ile oldukça dengeli sınıf dağılımları oluşturduk.

Clus-HMC'nin tahmin edici performans ölçütleri olarak precision-recall eğrileri ve ROC analizleri kullanılır. Precision ve recall geleneksel olarak pozitif ve negatif sınıflara sahip bir ikili sınıflandırma görevi için tanımlanır. Precision pozitif tahminlerin doğru olduğu bölgedir ve recall pozitif olarak doğru tahmin edilen pozitif örneklerin oranıdır, Eşitlik (3.26-3.27).

$$\text{precision} = \frac{TP}{TP+FP} \tag{3.26}$$

$$\text{recall} = \frac{TP}{TP+FN} \tag{3.27}$$

Eşitlik (3.26-3.27)'de TP doğru tahmin edilen pozitif örneklerin sayısını, FP yanlış tahmin edilen pozitif örneklerin sayısını, FN negatif tahmin edilen pozitif örnekleri göstermektedir. Bir precision-recall eğrisi (PR eğrisi) bir modelin precision'nı recall'un bir fonksiyonu olarak çizer. Varsayalımki model yeni bir örneğin pozitif olma olasılığını tahmin etsin. Tahmin edilen sınıfı elde etmek için bu olasılığı bir eşik t ile eşleştirelim. Bu eşik PR alanındaki tek bir noktaya karşılık gelir ve eşiği değiştirerek bir PR eğrisi elde ederiz. T'yi 1.0'dan 0.0'a düşürürken, artan sayıda örnek pozitif olarak tahmin edilir, recall artar ancak precision artabilir veya azalabilir (normalde bir eğilim azalır). Bir PR eğrisi, modelin tahmin edici davranışını anlamaya yardımcı olsa da, modelleri karşılaştırmak için tek bir performans skoru

daha yararlıdır. Bu amaçla sıklıkla kullanılan bir skor, PR eğrisi ve recall ekseni arasındaki alan olup, PR eğrisinin altındaki alan AUPRC olarak adlandırılır. AUPRC 1.0'a ne kadar yakın olursa, model o kadar iyi olur (Schietgat ve diğ., 2010). PR eğrileri çoklu etiket sınıflandırma görevindeki her bir sınıf için sınıfa ait örnekleri pozitif olarak ve diğer örnekleri negatif olarak alarak oluşturulur.

Ortalama PR eğrisinin altındaki alan (Area Under the Average PR Curve) Clus-HMC'de genel bir performans skoru elde etmek için kullanılır. Başlangıçta belirli bir eşik değeri için bu PR alanında tek bir nokta ($\overline{Precision}$, $\overline{Recall}$) oluşturur, Eşitlik (3.28-3.29).

$$\overline{Precision} = \frac{\sum_i TP_i}{\sum_i TP_i + \sum_i FP_i} \tag{3.28}$$

$$\overline{Recall} = \frac{\sum_i TP_i}{\sum_i TP_i + \sum_i FN_i} \tag{3.29}$$

Eşitlik (3.28-3.29)'da i bütün sınıfları dolaşır. $\overline{Precision}$ doğru tahmin edilen etiketlerin oranına karşılık gelir ve $\overline{Recall}$ doğru tahmin edilen verilerin içindeki etiketlerin oranıdır. Yukarıda anlatılan eşik değeri çeşitlendirilerek ortalama bir PR eğrisi elde edilir. Bu eğri altındaki alan $AU(\overline{PRC})$ ile verilir.

PR eğrileri altındaki ortalama alan (Average Area Under the PR Curves) ise hesaplanan bireysel (her sınıf için) PR eğrileri altındaki alanların ağırlıklı (weighted) ortalamasını almakla elde edilir ve aşağıdaki şekilde hesaplanır, Eşitlik (3.30).

$$\overline{AUPRC_{W_1,\ldots,W_{|C|}}} = \sum_i w_i \,.\, AUPRC_i \tag{3.30}$$

Bu yaklaşımda tüm ağırlıklar genel olarak 1 / |C| belirlenir. C sınıfların kümesini ifade eder. Diğer yandan bir sınıfın ağırlığı frekansıyla belirlenebilir. $w_i = {}^{v_i}/_{\sum_j v_j}$ olmak üzere, v_i veri içinde c_i'nin frekansını belirtir. $\overline{AUPRC_{W_1,\ldots,W_{|C|}}}$ ifadesine alan olarak karşılık gelen PR eğrisi, recall ekseni üzerindeki her değer için sınıfsal precision değerlerin (ağırlıklı) noktasal ortalaması alınarak çizilir. Bu eğri üzerindeki her nokta, her sınıf için farklı bir eşiğe karşılık gelebilir (Schietgat ve diğ., 2010).

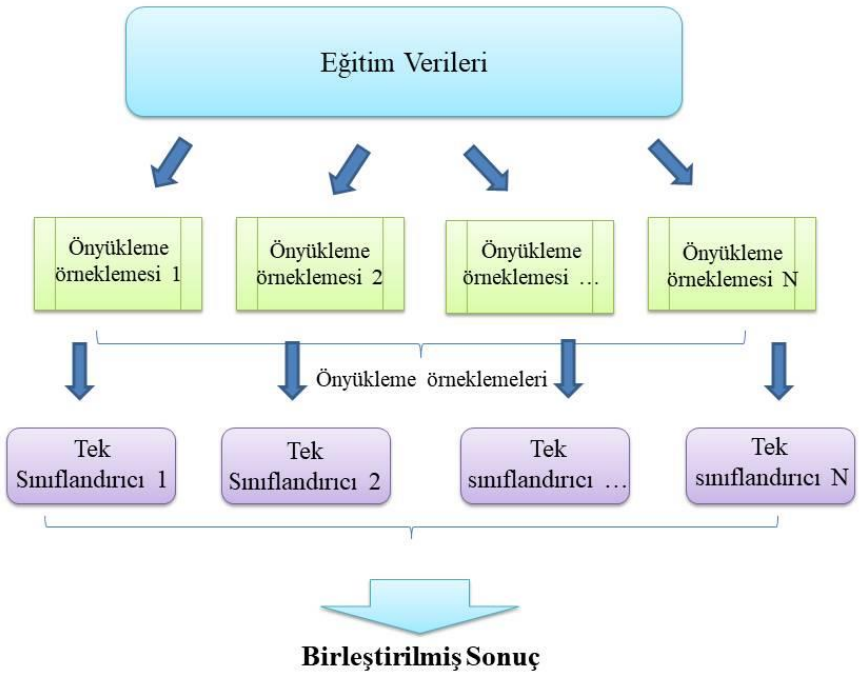
3.7.4 Dengesiz verileri tekrar örnekleme

Tezde Bölüm (6)'da topluluk sınıflandırması için önerilen modelin bir aşamasında farklı hastalık gruplarına ait (1200'den fazla ilaç) dengesiz ilaç veri setinde azınlık sınıfının örnek sayısını arttırmak için veri kümesine Synthetic Minority Over-sampling Technique (SMOTE) metodu uygulanmıştır (Galar ve diğ., 2011). Dengesiz veri setleri için sınıflandırma problemleri makine öğrenmede aktif alan araştırmasıdır (Kotsiantis ve diğ., 2006). Dengesiz verilerin tekrar örneklenmesi yöntemi kullanılan sınıflandırıcıdan bağımsız olduğu için dengesiz veri dağılımını işlemek için sıklıkla kullanılır. Çalışmanın odağı, sınıflandırma performansını arttırmak için hem veri seviyesi hem de sınıflayıcı topluluk yaklaşımının avantajlarını kazanmaktır. Dengesiz ilaç veri setimizde çoğunluk grubunda 1020 onaylanmış, azınlık grubunda ise 150 geri çekilen ilaç yer almaktadır. Metot ilaç veri kümesinde azınlık sınıfı ve çoğunluk sınıfı arasındaki dengesizlik oranının düşürülmesi aşamasında kullanılmıştır. Buna ek olarak under-sampling tekniğide veri setinin dengesizlik oranının azaltılması için kullanılmıştır. Amacımız over-sampling ve under sampling ile dengeli hale getirilen veri setlerinden geliştirilen modellerin sınıflandırma performanslarını karşılaştırmaktır. Under sampling metodunda algoritma çoğunluk sınıfının bazı örneklerini veri setini dengelemek amacıyla rastgele siler ancak silinen örnekler veri kümesi için gerekli örnekler olabilir. Buda metodun bir dezavantajıdır (Kotsiantis ve Pintelas, 2003).

3.7.5 Meta sınıflandırma

Tezde Bölüm(6)'da dengesiz veri setlerini sınıflandırıcı topluluk tasarımı için önerilen modelde son aşamada sınıflandırıcı topluluk oluşumunda meta-learner (meta-öğrenici) olarak bagging algoritması ve base-learner (temel-öğrenici) olarakta SVM+RBF Kernel kullanılmıştır. Bir başka değişle torba yapılacak (bagged) öğrenme planının adı SVM+RBF Kernel'dır. Bagging farklı eğitim verileri alt kümeleri kullanılarak oluşturulmuş bir sınıflandırıcı topluluktur. Bagging önyükleme (bootstrapping) ve ortalamaların (averaging) bir birleşimidir (Salunkhea ve Mali, 2016). Burada öngörücünün birden çok versiyonu oluşturulur ve birleştirilmiş tahmini üretmek için biraraya gelirler. Bireysel temel sınıflandırıcıları bağımsız olarak önyükleme örnekleri olarak bilinen farklı eğitim setleri üzerinde eğitilir. Önyükleme örnekleri bazı örneklerin rastgele bir şekilde eğitim setinin orijinal örnekleri değiştirilerek toplanmasıyla oluşturulur. Dolayısıyla, bagging'in kararsız

(unstable) algoritmalarla kullanılması tercih edilir. Burada eğitim setindeki küçük değişiklikler, o sistemin çıktısında büyük değişikliklerle sonuçlanır (Graczyk ve diğ., 2010). Çalışmalarımızda meta-sınıflandırma aşamasında Bagging ile SVM+RBF Kernel seçmemizin sebebi ilaç veri seti üzerinde geliştirilen modelde özellikle bağımsız test verilerini sınıflandırmada oldukça başarılı olmasıdır. Şekil (3.5) eğitim verilerine uygulanan bagging algoritmasının aşamalarını göstermektedir. Eğitim verilerinden önyükleme örneklemeleri elde edilir ve her bir önyükleme örneğine aynı meta-öğrenici (bagging) algoritması uygulanır. Sonuç olarak herbirinden gelen sonuçlar birleştirilir.



Şekil 3.5: Önyükleme birleştirme (Bootstrap aggregating, Bagging).

Çizelge (3.3)'te dengesiz ilaç veri seti üzerine uygulanan meta-öğrenici Bagging algoritması verilmiştir. Tezde onaylanmış ve geri çekilen ilaçlardan oluşan dengesiz bir veri setinin önerilen sınıflandırıcı topluluk tasarımı modeli ile sınıflandırılması sonuçları ve aynı veri seti üzerinde diğer öznitelik seçme algoritmaları ve farklı meta-sınıflandırıcıları kullanılarak elde edilen sonuçların eğitim seti ve bağımsız test seti için elde edilen sınıflandırma performans değerleri Bölüm (6.3.2)'de detaylı bir şekilde verilmiştir.

Çizelge 3.3: Bagging algoritması.

input E- Öğrenme seti, N- Önyükleme örneklemesi numarası, ÖA-Öğrenme Algoritması
output S*- Çoklu Sınıflandırıcı
for i=1 **to** N**do**
begin
E_i := E'den elde edilen önyükleme örneklemesi;
S_i := ÖA(E_i);
end;
S*(x) = $\text{argmax}_y \sum_{i=1}^{N}(S_i(x) = y)$

3.8 İlaç Molekülleri Üzerinde Sık Alt Çizge Madenciliği

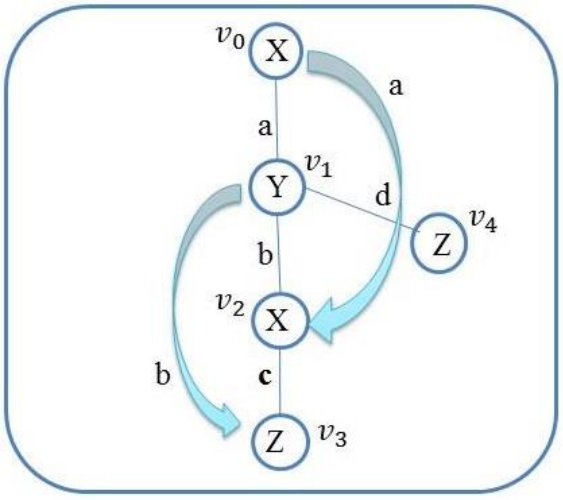
Moleküler veri tabanlarının bilgisayar destekli analizi, yeni ilaç adaylarının araştırılması için önemlidir. Yeni biyoaktif molekülleri veya kimyasal bileşikleri keşfetmek için, hem aktif hem de aktif olmayan moleküller içeren veritabanları, moleküler fragmanların havuzunda ayırt edici parçaları bulmak üzere aranır. Ardından moleküllerin yapısal özellikleri belirlenir ve moleküller etkin veya pasif olarak sınıflandırılır. Buradan yola çıkarak çalışmamızda sinir sistemi ilaç molekülleri (geri çekilen ve onaylanmış) üzerinde sık alt çizge madenciliği uygulayarak geri çekilen ilaç molekülleri yapısında sıklıkla tekrar eden, onaylanmış molekülerin yapısında ise nadir gözlenen ayırt edici fragmanlar belirlendi. Bu fragmanlar moleküllerin spesifik özellikleri hakkında önceden fikir verir ve ilaç adayı bir molekülün onaylanmış ve geri çekilen sınıfını belirlemede kullanılabilir. Çalışmada gSpan (çizge-tabanlı alt yapı örüntüsü madenciliği, Graph-Based Substructure Pattern Mining) algoritmasını içeren ParMol paketi (Paralel Moleküler Madencilik) kullanıldı.

3.8.1 gSpan

gSpan çizgeler arasında yeni bir sözlük sıralaması (lexicographic order) oluşturur ve her bir çizgeyi bir DFS (derinlik öncelikli arama) koduna kanonik etiket olarak eşler yani sözlük sıralamasına dayanarak derinlik öncelikli arama stratejisini benimser (Ramraj ve Prabhakar, 2015). Verilen bir çizge veri seti için D = $\{G_0, G_1, \ldots, G_n\}$ olsun. support(g) D içerisindeki çizgelerin sayısını ve g herhangi bir alt çizgeyi belirtir (Yan ve Han, 2002). İlaç moleküllerinin kimyasal yapısı çizgeye dönüştürülürken atomlar etiketlenmiş düğümlerle ve bağlar etiketlenmiş kenarlarla

temsil edilir. Problemlerde karmaşıklığı (complexity) düşürmek amacıyla sık bağlı alt çizgeler çalışıldı. İlaç molekülleri için sık alt çizge madenciliği problemi g herhangi bir moleküler fragman olmak üzere support (g) ≥ minSup (minimum destek eşik değeri) eşitsizliğini sağlamaktır. İlaç keşfinde moleküler veritabanlarının madenciliği benzer kimyasal özellikleri paylaşan moleküllerin içindeki fragmanları belirlemek amacıyla yapılır. Moleküllerin 2 boyutlu atom-bağ yapısı bu molekülleri içeren yönsüz etiketli çizge madenciliği için temel bir yapı oluşturur. Bu veritabanları en az belli sayıda molekülde gözlenen alt çizgeler için araştırılır. Alt çizgeler için bir çizge veritabanı (madencilik molekülleri) en az tüm çizgelerin belli bir yüzdesinde (support) veya sayısında (frequency) görülen alt çizgeler olacak şekilde azaltılabilir (Meinl ve diğ., 2006).

gSpan'de her çizge bir DFS koduna eşlenir, bu kodlar arasında yeni bir sözlük sıralaması inşa edilir ve bu sıralamaya dayalı bir arama ağacı oluşturulur. DFS sözlük sıralaması ve minimum DFS kod yeni bir kanonik etiketleme sistemi oluşturan iki tekniktir bu teknikler DFS arama'yı (DFS search) destekler. GSpan, aday oluşturma ve yanlış pozitif budama olmaksızın tüm sık alt çizgeleri keşfeder. Bir çizge (G) için verilen DFS ağacı (T) olsun. Kenar dizisi (e_i) $\prec_T$'ye dayandırılır. Örneğin $e_i \prec_T e_{i+1}$, burada i=0,…., $|E|-1$'dir. e_i DFS kod olarak adlandırılır ve kod (G, T) olarak belirtilir . Kenarlar (edges) 5 değişkenle belirtilir, $(i,\ j,\ l_i, l_{(i,j)}, l_j)$. Burada l_i v_i'nin, l_j v_j'nin etiketleri ve $l_{(i,j)}$ aralarındaki kenarın etiketidir (Yan ve Han, 2002) . Şekil (3.6)'da örnek bir çizge üzerinde (v_0, v_1), (0, 1, X, a, Y) ile temsil edilmektedir.



Şekil 3.6: Örnek bir çizge üzerinde (v_0, v_1), (0, 1, X, a, Y) ile temsil edilmektedir.

Z = $\{kod\ (G,T)/\ T\ \ G'nin\ DFS\ \ ağacı\}$ olsun, Z bütün bağlı etiketlenmiş çizgeler için bütün DFS kodlarını içerir. DFS sözlük sıralaması, doğrusal bir sırada şu şekilde tanımlanır. Eğer $\alpha = kod\ (G_\alpha, T_\alpha) = (a_0, a_1, \dots, a_m)$ ve $\beta = kod\ (G_\beta, T_\beta) = (b_0, b_1, \dots, b_m)$, $\alpha, \beta\ \epsilon\ Z\ \ ise$ ve aşağıdaki 2 koşulu sağlıyorsa $\alpha \leq \beta$'dır.

(i) $k < t, a_t \prec_e b_t\ için\ \exists t,\ 0 \leq t \leq \min(m,n),\ a_k = b_k$

(ii) $0 \leq k \leq m\ ve\ n \geq m$ için $a_k = b_k$

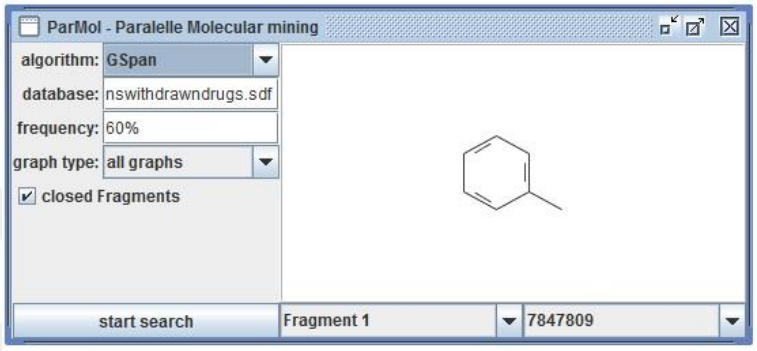
G bir çizge ve Z(G) = $\{kod\ (G,T)/\ T\ \ G'nin\ DFS\ \ ağacı\}$ (DFS sözlük sıralamasına dayalı) olsun. min (Z(G)) minimum DFS kod olarak adlandırılır. Bu aynı zamanda G'nin bir kanonik etiketidir. Bir DFS kod ağacında her bir düğüm bir DFS kod ile temsil edilir. Ebeveyn (parent) ve child (çocuk) düğümü arasındaki ilişki ebeveyn-çocuk ilişkisine uygundur ve kardeşler arasındaki ilişki DFS sözlük sıralamasıyla tutarlıdır. Kod ağacının derinlik öncelikli aranması sayesinde, sık alt çizgelerdeki tüm minimum DFS kodları keşfedilebilir. Bir başka değişle tüm sık alt çizgeler bu yolla keşfedilebilir (Yan ve Han, 2002).

gSpan çizgeleri depolamak için seyrek bitişiklik listesi (sparse adjacency list) kullanır. Düğümler (vertices) için bir etiket seti {A, B, C, …}ve kenarlar içinde {a, b, c, ..} olsun. Algoritma öncelikli olarak $A \overset{a}{\frown} A$ kenarlarını içeren sık alt çizgeleri keşfeder. Daha sonra $A \overset{a}{\frown} B$ kenarlarını içeren tüm sık alt çizgeler bulunur. Bu işlem tüm sık alt çizgeler bulunana kadar devam eder. Sık alt çizge madenciliği alt çizgeleri ve onların tüm sık torunlarını (descendants) büyütmek için özyinelemeli bir şekilde çağrılır. Alt çizge madenciliği support (g), minSup'tan küçük olduğunda durur. Bu tüm çizgelerin ve onların torunlarının üretildiğini ve önceden keşfedildiğini gösterir.

gSpan'de kalan fragman tekrarlamalarını elemek amacıyla arıtma üretimi iki şekilde kısıtlanmıştır. İlk olarak fragmanlar yalnızca derinlik öncelikli arama ağacının en sağdaki yol uzantısında (rightmost path extension) uzanan düğümlere genişletilebilirler. İkincisi fragman üretimi görünüm listelerindeki (appearance lists) oluşum tarafından yönlendirilir. Bu iki budama kuralı izomorfik fragman oluşumunu tam olarak engelleyemediğinden, gSpan her arıtma için kanonik (leksikografik olarak en küçük) DFS kodu hesaplar. Minimal olmayan DFS kod arınmaları budanabilir. Çünkü gSpan gömmeler (embeddings) yerine yalnızca her bir fragman için görünüm

listelerini depolar, açık alt çizge izomorfizm testi bu görünüm listelerindeki bütün çizgeler için yapılmak zorundadır (Meinl ve diğ., 2006).

Şekil (3.7)'de ParMol paketi içerisinde gSpan algoritması geri çekilen sinir sistemi ilaç moleküllerinden oluşan bir veri tabanına (N02, N03, N04, N05, N06 gruplarına dahil) frekans (frequency) değeri 60% olacak şekilde uygulanmış ve elde edilen çizge veri tabanında çizgelerin en az % 60'ında bulunan alt çizgelerden biri (toluene) gösterilmiştir.



Şekil 3.7: ParMol paketi kullanılarak ilaç molekülleri veri tabanında yaygın moleküler fragmanları belirlemede çalışan bir örnek.

Bir fragman gözönüne alınan çizge setinde aynı desteğe (support) sahip süper çizge yok ise "kapalı" (closed) olur. Bu bulunan fragmanların sayısını azaltır. Çalışmamızda ilaç moleküllerinden oluşan veri tabanları için bulunan alt çizgeler (moleküler fragmanlar) kapalıdır.

4. SİNİR SİSTEMİ İLAÇLARI ÜZERİNDE UYGULAMA

4.1 Giriş

Potansiyel olarak ilaç olmaya uygun olmayan molekülleri bileşik kütüphanelerinde belirlemek amacıyla bileşiklerin moleküler yapılarından elde edilen tanımlayıcılara dayalı bilgisayar destekli kimyasal bileşikleri sınıflandırmaya yönelik modeller geliştirilmiş ve ilaç keşfi sürecinde bileşiklere uygulanmıştır (Lavecchia, 2015). Hesaplamalı bir filtre yöntemi olan sanal tarama (VS) gelecek vaat eden bileşiklerin daha ileri testler için seçilmesini sağlar. Yapılan bu çalışmalar ilaç tasarımı sürecinde zamandan kazanmayı ve çabadan tasarruf etmeyi hedefler (Reddy ve diğ., 2007). Potansiyel olarak aktif bir ilaç olmak için ligandın bazı özelliklere sahip olması gerekir. İlaç etki göstereceği dokuya uygun ve oral olarak biyolojik elde edilebilirliğe sahip olmalıdır (Veber ve diğ., 2002). Toksisite düzeyi de en aza indirilmelidir (Xue ve diğ., 2004).

Sinir sistemi (NS) nöronların karmaşık yapılarına sahiptir. Özel hücreler adı verilen nöronlar, vücudun farklı bölümleri arasında sinyal iletir. NS'nin iki temel bileşeni vardır. Bunlar merkezi sinir sistemi (CNS) (Ghorbanzad'e ve Fatemi, 2012) ve periferik sinir sistemi (PNS) 'dir. CNS, beyin ve sinirlerden oluşur. PNS, duyusal nöron, ganglionlar ve sinirlerden oluşur. NS'nin hasarı epilepsi, alzheimer hastalığı, parkinson hastalığı, huntington hastalığı, çoklu skleroz (MS) (Pachner ve diğ., 2015), inme ve amiyotrofik lateral skleroz (ALS) gibi çeşitli sinir hastalıklarına neden olur. Kalsifikasyona uğramış parenkimal nöro-sistiserkoz, NS hasarı nedeniyle refrakter epilepsi ile sonuçlanır ve alzheimer hastalığı, insan nöron hasarıyla ilişkilidir (Leon ve diğ., 2015; Sullivan ve Young-Pearse, 2017). Ayrıca, nöron içerisinde α-sinüklein birikimi Parkinson hastalığına neden olur (Engelender ve Isacson, 2017). Ek olarak, literatürde Huntington hastalığına yol açan NS hasarı, çoklu skleroz (MS), inme ve amiyotrofik lateral skleroz (ALS) hakkında birkaç makale bulunmaktadır (Gendelman ve diğ., 2015; Kobal ve diğ., 2016; Lee ve diğ., 2017). Sinir sistemi ilaçları periferal etkilere sahiptir çünkü bu ilaçlar çoklu organları etkileyen beyine ulaşır. Dolayısıyla, sinir sistemi ilaçları sıklıkla marketlerden çekilir.

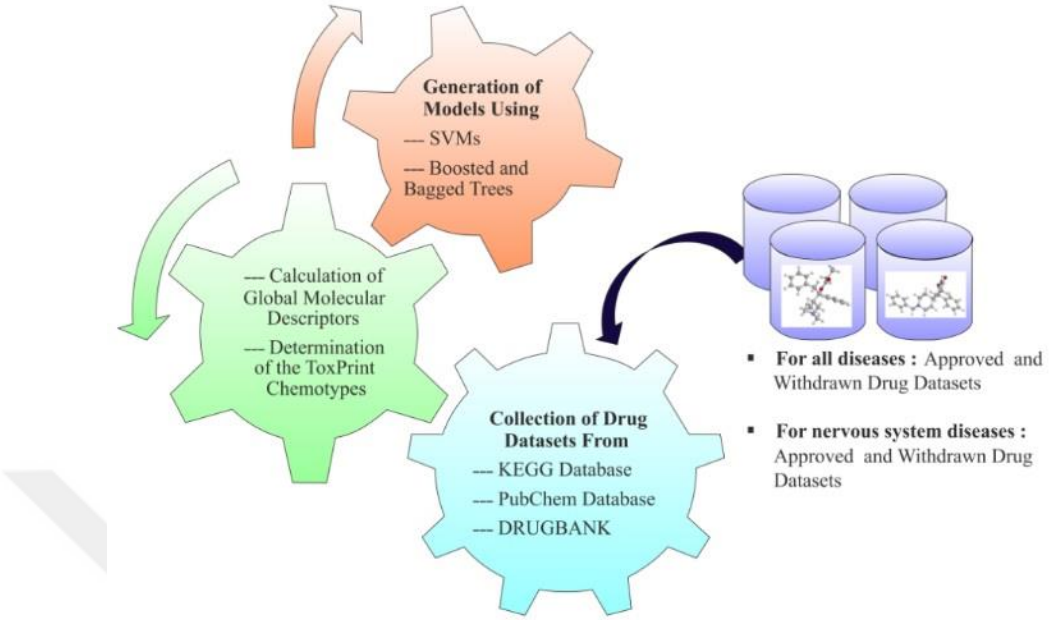
Çalışmada ToxPrint kemotipler kullanılarak onaylanmış ve geri çekilen sinir sistemi ilaçlarının makine öğrenmesi metotları kullanılarak sınıflandırılması yer almaktadır. Çalışmada ilaçları onaylanmış ve geri çekilen kategorilere sınıflandırmak amacıyla linear SVM (L SVM), medium gaussian SVM (MG SVM), coarse gaussian SVM (CG SVM), güçlendirilmiş (boosted trees) ve torbalanmış (bagged trees) karar ağaçları algoritmaları kullanıldı. 760 moleküler tanımlayıcı Molecular Networks Inc.'ten CORINA.Symphony programı kullanılarak veri kümelerindeki tüm moleküller için hesaplandı. Bunlar global moleküler, boyut ve şekil ve ToxPrint kemotip tanımlayıcılarıdır. ToxPrint kemotipleri, kimyasal özelliklerin ve kuralların bir setidir. Bu çalışmada kullanılan ilaç veri setleri, sinir sistemi hastalıkları için kullanılan onaylanmış ilaçlar (NSAD'ler) ve geri çekilen ilaçları (NSWD'ler) içermektedir. Buna ek olarak farklı hastalık gruplarına ait onaylanmış ilaçlar (AD) ve geri çekilen ilaçlardan (WD) oluşan bir set'te onaylanmış ve geri çekilen ilaçlar arasında ayrım yapmak için kullanıldı. Farklı hastalık gruplarına ait ilaç molekülleri Drugbank, KEGG DRUG ve PubChem veritabanından alındı.

İlaç sınıflandırma problemlerinde doğruluk oranının arttırılması ve sınıflandırmada daha etkin moleküler tanımlayıcıların belirlenmesi amacıyla ki-kare öznitelik seçme yöntemi kullanıldı (Kavitha ve diğ., 2012; Liu, 2004). Bu tanımlayıcılar ilaç adayı moleküllerini ön değerlendirmede büyük önem taşır. Buna ek olarak onaylanmış ve geri çekilen sinir sistemi ilaç molekülleri üzerinde ortak ve farklı (ayırt edici) moleküler fragmanları belirlemek amacıyla ilaç veri setlerine 60% destek (support) ile gSpan algoritması uygulandı. Bu moleküler fragmanlardan ayırt edici olanları kullanılarak ilaç adayı moleküllerin geri çekilmiş ve onaylanmış durumu hakkında fikir sahibi olabiliriz. Çalışmada MATLAB yazılım paketi (MATLAB & SIMULINK, R2015a) kullanıldı.

4.2 Materyaller Ve Yöntemler

Bu çalışma NS ilaç veri setleri üzerinde yürütüldü. İlaç veri setleri KEGG, PubChem ve DRUGBANK veritabanlarından toplandı. KEGG DRUG veritabanı onaylanmış sinir sistemi ilaçlarının Anatomik Terapötik Kimyasal Sınıflaması'na (ATC) sahiptir. Sınıflandırma problemlerinde SVM'lerin ve topluluk DT'lerin performansını değerlendirmek amacıyla farklı hastalık gruplarına ait ilaç veri setleride çalışıldı ve bu ilaçlar da DRUGBANK'tan toplandı. Ardından, SVM ve topluluk DT modellerini oluşturmak için CORINA Symphony programı kullanılarak DS_1'den (Dataset_1,

veri seti_1) DS_6'ya (Dataset_6, veri seti_6) kadar olan veri kümeleri için bir dizi moleküler tanımlayıcılar hesaplandı. Çalışmanın mimari yapısı Şekil 4.1'de verildi.



Şekil 4.1: Çalışmanın mimari yapısı.

4.2.1 Veri kümelerinin toplanması

400'den fazla onaylanmış ve geri çekilen ilaç molekülleri KEGG DRUG, PubChem ve DRUGBANK veritabanlarından toplandı. Bu ilaç veritabanlarında ilaç molekülleri SDF, SMILES vb. formatlarda bulunur. Bir SDF dosyası, molekülün adını, özelliklerini, atomların koordinatlarını, atomlar arasındaki bağ çeşitlerini, içinde bulunduğu ve depoladığı atomları ve bir molekül ile ilgili diğer bilgileri içerir. DS_1, DRUGBANK veritabanından alınan farklı hastalık gruplarına ait 220 ilaç molekülü içerir. Bunlardan 110 tanesi onaylanmış, geri kalan 110 geri çekilen ilaçlardır. Sindirim sistemi ve metabolizma, kardiyovasküler sistem, genito idrar sistemi, kas iskelet sistemi, solunum sistemi ve sinir sistemi vb. hastalıklarına ilişkin ilaçları içerir. Onaylanmış ilaçlar bir çok hastalığı tedavi etmek için onlarca yıldır insanlar tarafından kullanılmaktadır. Geri çekilen ilaçlar ise genellikle insan sağlığı üzerinde olumsuz ve beklenmedik yan etkileri nedeniyle geçmişten günümüze kadar olan süreçte marketlerden çekilen ilaçlardır. DS_1 sınıflandırma çalışmaları için 10-kat çapraz doğrulama metoduyla on farklı eğitim ve test setine bölündü. Buna göre her eğitim seti onaylanmış ve geri çekilen ilaçlar olmak üzere 198 ilaçtan, test setleri ise yine onaylanmış ve geri çekilen 22 ilaçtan oluşur. DS_2, N05 grubuna ait 15

onaylanmış ile 11 geri çekilen ilaç içerir. DS_3, N06 grubuna dahil 15 onaylanmış ve 12 geri çekilen ilaç içerir. İlaç veritabanlarında geri çekilen sinir sistemi ilaçlarının sınırlı sayıda olması nedeniyle (N01'den N07'ye kadar olan geri çekilen ilaçlar) DS_4, DS_5 ve DS_6 aynı geri çekilen sinir sistemi ilaçlarını içerir ancak bu veri setleri birbirinden tamamen farklı onaylanmış sinir sistemi ilaçlarına sahiptir. DS_2, DS_3, DS_4, DS_5 ve DS_6 sırasıyla 26, 27, 72, 72 ve 72 ilaç molekülünden oluşmaktadır. DS_1 ve diğer veri setlerinin içerdikleri ilaç moleküllerine ait bilgilerin tamamı DVD_Çizelge Ek.1'de verildi. Sinir sistemi ilaç veri setlerinin ATC sınıflaması Bölüm 2, Çizelge (2.1)'de verildi.

4.2.2 Moleküler tanımlayıcıların hesaplanması

Moleküler tanımlayıcılar bir molekülün önemli yapısal özelliklerini basit matematiksel bir gösterimle sunar. Veri matrisimiz 760 tanımlayıcıdan oluşur. Tanımlayıcılar CORINA Senfoni programı ile ilaç veri setleri için hesaplandı. Bu çalışmada 760 moleküler tanımlayıcı, 22'si global moleküler, 8'i boyut ve şekil, 729'u toxprint kemotip tanımlayıcılarından (DVD_Çizelge Ek.2) ve 1 kullanıcı özelliğinden oluşur. Global moleküler tanımlayıcılar bir molekülün kaba formülünden (gross formula), 2 boyutlu ve 3 boyutlu yapısından türetildi. Benzer şekilde, boyut ve şekil tanımlayıcıları da bir molekülün 3 boyutlu yapısından türetildi. CORINA Symphony programı tarafından hesaplanan 760 moleküler özellik Bölüm 2, Şekil (2.3)'te verildi.

4.2.3 Veri ön işleme ve özellik seçimi

Veri ön işlememizin amacı ham ilaç veri setlerinden yararlı veriler elde etmektir. CORINA programı tarafından önceden tanımlanmış bir takım basamaklar SDF formatındaki ilaç moleküllerine uygulanır. Öncelikle kimyasal bir yapıya sahip olmayan (2 boyutlu veya 3 boyutlu) ilaç molekülleri veri setinden çıkarılır. Sonrasında kimyasal kayıtlardaki küçük fragmanlar (tuz içerisindeki karşıt iyonlar, çözümleyiciler vb.) büyük fragmanları tutmak amacıyla kaldırılır. Ardından kimyasal yapılardaki formal yükler nötr hale getirilir. İlaç molekül setleri içerisinden eğer varsa aynı olanları seçilip kaldırılır. Son olarak, ilaç moleküllerinin üç boyutlu yapıları oluşturulur ve veri setindeki her ilaç molekülü için 760 moleküler özellik belirlenir. Hesaplanan moleküler tanımlayıcılar sınıflandırma modellerinin geliştirilmesinde kullanıldılar.

İlaç veri setindeki ilgisiz bilgiler nedeniyle, orijinal ham veriler direkt olarak tahmin işlemi için kullanılmadı. Ham veriler temizlendi, analiz edildi ve dönüştürüldü. Her satır verisi bir ilaç molekülüne ait bilgileri temsil eder ve sütunlarda 760 moleküler özellik listelenir. Son sütun ise ilacın geri çekilen / onaylanmış duruma karşılık gelen sınıf etiketini taşır.

Hesaplanan tanımlayıcıların tümü onaylanmış ve geri çekilen ilaçları ayırt edici nitelikte değildir. Gereksiz tanımlayıcıların ortadan kaldırılması sınıflandırıcıların öngörme performansını geliştirir. Bu nedenle ki-kare öznitelik seçme yöntemi moleküller için hesaplanan özelliklerin boyutunun azaltılması için veri setlerine uygulandı. Ki-kare bir tanımlayıcının değerini sınıfla ilgili olarak ki-kare istatistiğinin değerini hesaplayarak belirler. Ardından özellikler bu skora göre sıralanır. Mevcut çalışmada, bir tanımlayıcı için skor sıfır olduğunda, sınıflandırmada daha etkin olan özellikleri tutmak için bunlar veri setinden kaldırıldı. Altı özellik seti (feature sets, FSs) yukarıda anlatılan şekilde belirlendi. Skorları yüksek olan özelliklerden özellik setleri elde edildi. FS_1, FS_2, FS_3, FS_4, FS_5, FS_6 sırasıyla 44, 16, 14, 16, 18, 15 özellik içermektedir. Çizelge (4.1)'de deneylerde kullanılan özellik setleri sırasıyla verilmiştir. Sınıflandırma modelleri seçilen bu özelliklerle eğitildi. Ayrıca Ki-kare özellik seçimi metodu sinir sistemi ilaç veri setleri üzerinde sınıflandırıcıların ayırma yeteneğini arttırdı.

4.2.4 Veri madenciliği modellerinin geliştirilmesi

4.2.4.1 Sınıflandırma metotları

Bu çalışmada, ilaçların onaylanmış ve geri çekilen kategorilere sınıflandırılması için L SVM, MG SVM ve CG SVM kullanıldı. Bunun yanında BS T ve BG T metotlarıda sınıflandırma görevleri için kullanıldı. Onaylanmış ve geri çekilen ilaçlar için ayırt edici özellikleri belirlemek yeni bir ilacın keşfinden önce anlamlıdır. Pazarlamadan sonra ilaçların marketlerden geri çekilmesi önemli ilaç etkileşimleri bildirilen ölümler veya ciddi yan etkiler gibi çeşitli olaylarla ilişkilendirilebilir (Fliri ve diğ., 2005). Bu nedenle, mevcut ve gelecekteki ilaç keşfi için temel toksisite mekanizmalarını bulmak gerekir. Bu amaçla, onaylanmış ilaçların genel özellikleri ve piyasadan geri çekilen ilaçlar üzerinde yapılacak çalışmalar büyük önem taşımaktadır. Bu amaçla moleküllerin 760 tanımlayıcı özelliği kullanılarak sınıflandırma modelleri geliştirildi. Sınıflamanın amacı, veri setindeki her bir durum

Çizelge 4.1: Deneylerde kullanılan altı özellik seti.

FS_1 (44 features)
HaccO, HDonO, LogS, XlogP, bond:C(=O)N_carboxamide_(NH2), bond:C(=O)N_carboxamide_(NHR), bond:C(=O)N_carboxamide_generic, bond:C=O_acyl_hydrazide, bond:C=O_carbonyl_abunsaturated_aliphatic_(michael_acceptors), bond:CC(=O)C_ketone_alkene_cyclic_(C6), bond:CC(=O)C_ketone_alkene_cyclic_2-en 1-one, bond:CN_amine_pri-NH2_aromatic, bond:CN_amine_pri-NH2_generic, bond:COC_ether_aliphatic, bond:COH_alcohol_diol_(1_3-),bond:CX_halide_aromatic-X_halo_phenol_meta, bond:NC=O_aminocarbonyl_generic,bond:NN_hydrazine_acyclic_(connect_noZ),bond:PC_phosphorus_organo_generic,chain:alkaneCyclic_propyl_C3,chain:alkyne_ethyne_generic,group:aminoAcid_aminoAcid_generic,group:aminoAcid_asparagine,group:aminoAcid_leucine,group:carbohydrate_aldopentose,group:carbohydrate_ketohexose,group:carbohydrate_pentofuranose_2-deoxy, group:carbohydrate_pentofuranose, group:ligand_path_5_bidentate_aminopropanal, group:nucleobase_adenine,group:nucleobase_uracil,ring:hetero_[5]_N_pyrazole,ring:hetero_[5]_O_dioxolane_(1_3),ring:hetero_[5]_O_furan,ring:hetero_[5]_O_oxolane,ring:hetero_[5_6]_N_purine,ring:hetero_[6]_N_diazine_(1_3)_generic,ring:hetero_[6]_N_pyrimidine,ring:hetero_[6]_N_pyrimidine_2_4dione,ring:hetero_[6]_N_triazine_generic, ring:hetero_[6]_Z_1_2_4-, ring:hetero_[6]_Z_1_3-, ring:hetero_[6]_Z_generic, The number of total chemotypes.
FS_2 (16 features)
HDon,HDonN,Aspheric:Cor3D:ori1,bond:C=O_carbonyl_ab-unsaturated_generic,bond:C=O_carbonyl_generic, bond:CC(=O)C_ketone_aliphatic_acyclic,bond:CC(=O)C_ketone_aromatic_aliphatic,bond:CN_amine_alicyclic_generic, bond:CN_amine_aliphatic_generic,bond:CN_amine_ter-N_aliphatic,bond:CN_amine_ter N_aromatic, bond:CN_amine_ter-N_generic, chain:alkaneLinear_butyl_C4, ring:hetero_[6]_N_piperidine, ring:hetero_ [6]_ N_pyridine_generic, The number of total chemotypes.
FS_3 (14 features)
Atoms, Bonds, ASA, McGowan, Polariz, bond:C(=O)N_carboxamide_(NR2), bond:C=O_acyl_hydrazide, bond:NN_hydrazine_acyclic_(connect_noZ), bond:NN_hydrazine_alkyl_N(connect_Z=1), ring:hetero_[5]_N_pyrrole_generic, ring:hetero_[5]_Z_1-Z, ring:hetero_[5_6]_Z_generic, ring:hetero_[6_6]_Z_generic, The number of total chemotypes.
FS_4 (16 features)
bond:C=O_acyl_hydrazide,bond:CC(=O)C_ketone_aliphatic_acyclic,bond:CN_amine_aliphatic_generic, bond:CN_amine_ter-N_aliphatic,bond:CN_amine_ter-N_generic,bond:COC_ether_aliphatic, bond:COH_alcohol_aliphatic_generic,bond:COH_alcohol_generic,bond:COH_alcohol_sec-alkyl,bond:NN_hydrazine_acyclic_(connect_noZ),bond:NN_hydrazine_alkyl_N(connect_Z=1),chain:alkaneLinear_butyl_C4, group:ligand_path_5-7_bidentate, ring:hetero_[5]_O_oxolane, ring:hetero_[6]_Z_1_4-, The number of total chemotypes.
FS_5 (18 features)
bond:CN_amine_aliphatic_generic,bond:C=O_acyl_hydrazide,bond:CC(=O)C_ketone_aliphatic_acyclic, bond:CN_amine_ter-N_aliphatic,bond:CN_amine_ter-N_generic,bond:COH_alcohol_aliphatic_generic, bond:COH_alcohol_generic,bond:COH_alcohol_sec-alkyl,bond:NN_hydrazine_acyclic_(connect_noZ),bond:NN_hydrazine_alkyl_N(connect_Z=1), chain:alkeneCyclic_diene_cyclohexene, group:ligand_path_5-7_bidentate, ring:fused_[6_6]_tetralin, ring:hetero_[5]_O_oxolane, ring:hetero_[6]_N_piperazine, ring:hetero_[6]_Z_1_4-, ring:hetero_[6_6_6]_N_S_phenothiazine, The number of total chemotypes.
FS_6 (15 features)
bond:C=O_acyl_hydrazide,bond:CC(=O)C_ketone_aliphatic_acyclic,bond:CN_amine_aliphatic_generic, bond:CN_amine_ter-N_aliphatic,bond:CN_amine_ter-N_generic,bond:COH_alcohol_aliphatic_generic, bond:COH_alcohol_generic, bond:NN_hydrazine_acyclic_(connect_noZ), bond:NN_hydrazine_alkyl_N(connect_Z=1), group:ligand_path_5-7_bidentate,ring:fused_[6_6]_tetralin, ring:hetero_[6]_N_pyridine, ring:hetero_[6]_Z_1_4-,ring:hetero_[6_6_6]_N_S_phenothiazine, The number of total chemotypes.

için hedef sınıfın (geri çekilmiş/onaylanmış durumu) doğru bir şekilde tahmin edilmesidir. İlaç molekülleri için moleküler tanımlayıcılar kullanılarak ikili sınıflandırma problemleri çalışıldı. 10-kat çapraz doğrulama metodu ile veri seti on alt gruba ayrılır. Her seferinde, on alt kümeden biri test kümesi olarak kullanılır ve diğer alt kümeler ise bir eğitim seti oluşturmak üzere sisteme konur ve çapraz doğrulama işlemi on kez tekrarlanır. Daha sonra bu on sonucun ortalaması tek bir sonuç elde etmek için hesaplanır. Eğitim seti parametrelerini ayarlayarak farklı SVM ve topluluk DT modelleri üretmeye katılır ve test seti modellerin performansını değerlendirir.

Deneylerde kullanılan veri setleri (data set names), sınıf etiketleri (class labels), örnek sayısı (number of ınstance), veri seti boyutu (the data set size), uygulanan makine öğrenmesi algoritmaları (applied machine learning algorithms), özellik setleri (feature sets) olmak üzere, ilaç veri setleri için deneysel ayarlar ve uygulanan metotlar Çizelge (4.2)'de verildi.

Çizelge 4.2: İlaç veri setleri için deneysel ayarlar ve uygulanan makine öğrenme algoritmaları.

Data set Names	Class Labels	Number of Instances	The data sets size	Applied Machine Learning Algorithms	Feature Sets
DS_1	All Drugs	AD: 110 WD: 110	220	L SVM, MG SVM, CG SVM, BS T, BG T	FS_1
DS_2	N05	AD: 15 WD: 11	26	L SVM, MG SVM, CG SVM, BS T, BG T	FS_2
DS_3	N06	AD: 15 WD: 12	27	L SVM, MG SVM, CG SVM, BS T, BG T	FS_3
DS_4	N01 to N07	AD: 40 WD:32	72	L SVM, MG SVM, CG SVM, BS T, BG T	FS_4
DS_5	N01 to N07	AD: 40 WD:32	72	L SVM, MG SVM, CG SVM, BS T, BG T	FS_5
DS_6	N01 to N07	AD: 40 WD:32	72	L SVM, MG SVM, CG SVM, BS T, BG T	FS_6

L SVM, Linear Support Vector Machine; MG SVM, Medium Gaussian Support Vector Machine; CG SVM, Coarse Gaussian Support Vector Machine; BS T, Boosted Trees; BG T, Bagged Trees; AD, Approved Drug; WD, Withdrawn Drug.

SVM'nin avantajı, karar fonsiyonu için farklı kernel fonksiyonlarının (doğrusal, polinom, sigmoid ve radyal tabanlı vb.) belirtilebilmesidir. SVM kernel seçimi ve kernel parametrelerinin kurulumu büyük ölçüde ampirik ve deneysel analize bağımlıdır. Mevcut çalışmada hem eğitim hem de test kümelerini sınıflandırmak için kernel olarak doğrusal ve gaussian veya radyal tabanlı fonksiyonlar kullanılmıştır. L SVM sınıflar arasında basit bir doğrusal ayırma oluşturdu ve MG SVM için kernel ölçeği sqrt (P)'ye ayarlanarak sınıflar arasında ortalama ayırma oluşturdu. Ayrıca CG

SVM için kernel ölçeği sqrt (P)*4 olarak ayarlandı ve sınıflar arasında kaba ayrımlar yaptı. P tahmin edicilerin sayısıdır. Model üretiminde görev alan eğitim seti ilaçlarının sınıflandırılması için farklı kernel fonksiyonlarına sahip SVM modelleri üretildi, üretilen SVM modellerini doğrulamak için test ilaçları kullanıldı. Bu çalışmada, topluluk DT'lerinin sınıflandırmaların doğruluğunu geliştirdiği gözlendi. Topluluk metotlarında etkili ve yeterli modeller elde etmek amacıyla öğrenme oranı ve öğrenici sayısı sırasıyla 0.1 ve 200'e ayarlandı. Bu metotlarla ilaç molekülleri onaylanmış ve geri çekilmiş kategorilere sınıflandırılırken bütün ağaçlardan bireysel tahminler toplandı ve sınıflandırma için tek bir topluluk tahmini olarak birleştirildi (Breiman, 1996; Sutton, 2005). Özellik seçimi yöntemleri, SVM ve DT oluşumu ve topluluk teknikleri için MATLAB yazılım paketi kullanılmıştır.

4.2.4.2 Sinir sistemi ilaçları için sık alt çizge madenciliği

Çalışmada SDF formatında 32 geri çekilen (ilaç veri tabanındaki tüm geri çekilen ilaçlar) ve 145 onaylanmış sinir sistemi ilacını içeren veri setlerine gSpan algoritması uygulanarak her iki grupta göze çarpan, ayırt edici fragmanlar araştırıldı. gSpan algoritması destek (support) 60% ile veri setlerine uygulandı. Gereksiz ve fazla olan alt çizgeleri elemek amacıyla kapalı fragmanlar belirlendi.

4.2.4.3 Performans ölçümleri

Karışıklık matrisi (confusion matrix) bir sınıflandırma modelinin kalitesini değerlendirmek için kullanılan, doğru pozitif (TP), yanlış pozitif (FP), doğru negatif (TN) ve yanlış negatif (FN) sayısını bildiren iki sıra ve iki sütunlu bir tablodur. Ayrıntılı olarak, TP onaylanmış ilaçlar onaylanmış olarak doğru tanımlandı, FP onaylanmış ilaçlar geri çekilen olarak yanlış tanımlandı, TN geri çekilen ilaçlar geri çekilmiş olarak doğru tanımladı, FN geri çekilen ilaçlar onaylanmış olarak yanlış tanımlandı. Modellerin performansı, doğruluk oranı (accuracy rate, AR), eğri altındaki alan (area under the curve, AUC), pozitif öngörme değeri (positive predictive value, PPV), negatif öngörme değeri (negative predictive value, NPV), duyarlılık (sensitivity, SE), özgüllük (specificity, SP), F1-skoru (F1-score, F1-S) ve matthews korelasyon katsayısı (matthews correlation coefficient, MCC) aşağıdaki Eşitliklerle (4.1-4.7) ile hesaplanır.

$$\mathbf{AR} = (TP+TN) / (TP+TN+FP+FN) \quad (4.1)$$

$$\mathbf{PPV} = TP / (TP+FP) \quad (4.2)$$

$$\mathbf{NPV} = TN / (FN+TN) \quad (4.3)$$

$$\mathbf{SE} = TP / (TP+FN) \quad (4.4)$$

$$\mathbf{SP} = TN / (TN+FP) \quad (4.5)$$

$$\mathbf{F1\text{-}S} = 2TP / (2TP+FP+FN) \quad (4.6)$$

$$\mathbf{MCC} = (TP \times TN - FP \times FN) / \left(\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}\right) \quad (4.7)$$

MMC -1 ve +1 arasında değişen istatistiksel bir değerdir, burada +1 mükemmel bir tahmini, 0 ortalama rasgele bir tahmini ve -1 tersi bir tahmini gösterir ve bu değerler sınıflandırma modelinin kalitesinin bir ölçüsü olarak kullanılabilir.

4.3 Sonuçlar

4.3.1 Moleküler tanımlayıcıları sıralama

Hesaplanan tüm moleküler tanımlayıcılar arasından en önemli olanlarını seçme modellerin yorumlanmasına olanak sağlar. Ki-kare testi ki-kare istatistiğinin değerini sınıflara göre hesaplayarak her bir tanımlayıcının değerini belirler. Her veri seti için (DSs) en yüksek rank değerine sahip (ilk beş) olan moleküler tanımlayıcılar Şekil (4.2)'de gösterildi. Bu tanımlayıcılar sınıflandırma modelleri oluşturulurken oldukça etkindirler.

	Rankings	CSS	Rankings	CSS
	[DS_1]		[DS_2]	
1st	The number of total chemotypes	48.76	The number of total chemotypes	12.23
2nd	XlogP	24.73	Aspheric:Cor3D:ori1	10.09
3rd	Bond: C (=O) N_ carboxamide_(NHR)	21.07	Bond: CN_amine_alicyclic_generic	9.66
4th	LogS	17.82	Bond: C=O_carbonyl_generic	8.44
5th	Ring:hetero_ [6] _ Z_generic	14.42	Ring:hetero_ [6] _ N_pyridine_generic	7.72
	[DS_3]		[DS_4]	
1st	Bonds	14.21	The number of total chemotypes	15.40
2nd	The number of total chemotypes	14.21	Bond: CN_amine_aliphatic_generic	8.63
3rd	Atoms	13.23	Ring: hetero _ [6] _Z_1_4-	7.50
4th	ASA	11.81	Bond: CN_amine_ter-N_generic	7.30
5th	Polariz	9.64	Bond: CN_amine_ter-N_aliphatic	7.30
	[DS_5]		[DS_6]	
1st	The number of total chemotypes	19.60	The number of total chemotypes	22.05
2nd	Bond: CN_amine_aliphatic_generic	10.26	Bond: CN_amine_aliphatic_generic	10.26
3rd	Bond: COH_alcohol_generic	8.58	Bond: CN_amine_ter-N_generic	8.64
4th	Bond: COH_alcohol_aliphatic_generic	7.42	Bond: CN_amine_ter-N_aliphatic	8.64
5th	Bond: CN_amine_ter-N_aliphatic	7.30	Bond: NN_hydrazine_alkyl_N (connect_Z=1)	6.71

Şekil 4.2: Tüm DS'ler için rank değeri en yüksek ilk beş tanımlayıcı ve onların ki-kare istatistik değerleri (CSS).

Kemotiplerin molekül içerisindeki toplam sayısını belirten the number of total chemotypes, CSS değerlerine bakıldığında DS_1, DS_2, DS_4, DS_5 ve DS_6 için en etkin tanımlayıcı olarak belirlendi. Bonds olarak adlandırılan tanımlayıcı ise DS_3 için en önemli tanımlayıcıdır. Buna ek olarak ki-kare öznitelik seçme metodu XlogP, Aspheric:Cor3D:ori1 and the number of total chemotypes tanımlayıcılarını DS_1, DS_2 ve DS_3 için ikinci en etkin tanımlayıcı olarak belirledi. Bond:CN_amine_aliphatic_generic tanımlayıcısı ise DS_4, DS_5, ve DS_6 için ikinci en önemli tanımlayıcıdır. Hesaplamalara göre Şekil (3.2)'de gösterilen tanımlayıcılar veri setleri üzerinde sınıflandırma modelleri oluşturulurken etkin bir rol oynamaktadır.

Çizelge (4.3)'te sınıflama modelleri oluşturulurken en etkin moleküler tanımlayıcıların (Şekil (4.2)) ayrıntılı analizi verilmiştir. DS veri seti, SF seçilmiş özellikler, FS özellik seti'ni göstermektedir. Buna ek olarak GMF global moleküler, SSF boyut ve şekil, TCF ToxPrint kemotip ve UP kullanıcı özelliklerini belirtmektedir. Veri setlerinde ADs onaylanmış, WDs geri çekilen ilaçları göstermektedir. Bir moleküler tanımlayıcı hangi özellik grubunda ise +, aksi halde – ile temsil edilmektedir. Moleküler tanımlayıcı eğer bir kemotipse, mostly çoğunlukla hangi grupta onaylanmış/geri çekilen ilaçlarda bulunduğunu, only ise yalnızca onaylanmış/geri çekilen ilaçlarda bulunduğunu belirtir. Eğer moleküler tanımlayıcı, global moleküler, şekil ve boyut ve kullanıcı özelliklerinden biri ise sayısal bir değer alır. Bu durumda Range [a,b] for ADs ve Range [a, b] for WDs, veri setindeki bu tanımlayıcının aldığı maksimum ve minimum değerleri gösterir ve bir aralıkla temsil edilir. Amaç onaylanmış ve geri çekilen ilaç gruplarında bu tanımlayıcıların hangi değer aralığında bulunduğunu ortaya koymaktır. Bir başka önemli konuda ToxPrint kemotip özellikleri değerlendirilirken molekül içerisinde bu 729 kemotipten hangilerinin bulunduğu belirlenir. Buradan onaylanmış/geri çekilen ilaçlarda bulunan/bulunmayan kemotipler belirlenir. Bazı kemotipler sadece geri çekilen ilaçlarda bulunurken bazıları çoğunlukla onaylanmış ilaçlarda bulunur. Çizelge (3.3) aynı zamanda onaylanmış ve geri çekilen ilaç moleküllerinden oluşan veri setlerinin kemotip analizinide vermektedir. The number of total chemotypes moleküler tanımlayıcısı bir moleküldeki kemotiplerin toplam sayısını vermektedir. Bu özellik kullanıcı özelliği olarak tarafımızdan veri setine eklenmiştir ve özellikle sinir sistemi ilaçlarını onaylanmış ve geri çekilen sınıflara ayırırken etkin bir rol oynamaktadır.

Çizelge 4.3: Sınıflama modellerinde en etkin moleküler tanımlayıcıların ayrıntılı analizi.

DS_1/Model_1			
SF from FS_1	**GMF/SSF/TCF/UP (mostly/only located in ADs/WDs)**	**Range [a, b] for ADs**	**Range [a, b] for WDs**
The number of total chemotypes	- / - / - / + -	[10, 61]	[1, 29]
XlogP	+ / - / - / - -	[-4.1, 8.21]	[-2.72, 7.06]
bond:C(=O)N_carboxamide_(NHR)	- / - / + / - mostly located in ADs (38/110 ADs, 8/110 WDs)	-	-
LogS	+ / - / - / - -	[-8.22, 2.56]	[-7.42, 1.32]
ring:hetero_[6]_Z_generic	- / - / + / - mostly located in ADs (62/110 ADs, 35/110 WDs)	-	-
DS_2/Model_2			
SF from FS_2	**GMF/SSF/TCF/UP (mostly/only located in ADs/WDs)**	**Range [a, b] for ADs**	**Range [a, b] for WDs**
The number of total chemotypes	- / - / - / + -	[12, 36]	[10, 33]
Aspheric:Cor3D:ori1	- / + / - / - -	[0.06, 0.46]	[0.07, 0.29]
bond:CN_amine_alicyclic_generic	- / - / + / - mostly located in ADs (14/15 ADs, 4/11 WDs)	-	-
bond:C=O_carbonyl_generic	- / - / + / - mostly located in ADs (15/15 ADs, 6/11 WDs)	-	-
ring:hetero_[6]_N_pyridine_generic	- / - / + / - mostly located in ADs (11/15 ADs, 2/11 WDs)	-	-
DS_3/Model_3			
SF from FS_3	**GMF/SSF/TCF/UP (mostly/only located in ADs/WDs)**	**Range [a, b] for ADs**	**Range [a, b] for WDs**
Bonds	+ / - / - / - -	[42, 64]	[22, 68]
The number of total chemotypes	- / - / - / + -	[13, 30]	[7, 25]
Atoms	+ / - / - / - -	[39, 60]	[21, 65]
ASA	+ / - / - / - -	[350.14, 544.94]	[215.38, 616.9]
Polariz	+ / - / - / - -	[27.53, 51.41]	[18.3, 51.77]

Çizelge 4.3: (devam) Sınıflama modellerinde en etkin moleküler tanımlayıcıların ayrıntılı analizi.

DS_4/Model_4			
SF from FS_4	**GMF/SSF/TCF/UP (mostly/only located in ADs/WDs)**	**Range [a, b] for ADs**	**Range [a, b] for WDs**
The number of total chemotypes	- / - / - / + -	[12, 35]	[7, 28]
bond:CN_amine_aliphatic_generic	- / - / + / - mostly located in ADs (32/40 ADs, 16/32 WDs)	-	-
ring:hetero_[6]_Z_1_4-	- / - / + / - mostly located in ADs (15/40 ADs, 3/32 WDs)	-	-
bond:CN_amine_ter-N_generic	- / - / + / - mostly located in ADs (30/40 ADs, 14/32 WDs)	-	-
bond:CN_amine_ter-N_aliphatic	- / - / + / - mostly located in ADs (30/40 ADs, 14/32 WDs)	-	-
DS_5/Model_5			
SF from FS_5	**GMF/SSF/TCF/UP (mostly/only located in ADs/WDs)**	**Range [a, b] for ADs**	**Range [a, b] for WDs**
The number of total chemotypes	- / - / - / + -	[10, 35]	[7, 28]
bond:CN_amine_aliphatic_generic	- / - / + / - mostly located in ADs (34/40 ADs, 16/32 WDs)	-	-
bond:COH_alcohol_generic	- / - / + / - mostly located in ADs (16/40 ADs, 3/32 WDs)	-	-
bond:COH_alcohol_aliphatic_generic	- / - / + / - mostly located in ADs (13/40 ADs, 2/32 WDs)	-	-
bond:CN_amine_ter-N_aliphatic	- / - / + / - mostly located in ADs (30/40 ADs, 14/32 WDs)	-	-
DS_6/Model_6			
SF from FS_6	**GMF/SSF/TCF/UP (mostly/only located in ADs/WDs)**	**Range [a, b] for ADs**	**Range [a, b] for WDs**
The number of total chemotypes	- / - / - / + -	[10, 35]	[7, 28]
bond:CN_amine_aliphatic_generic	- / - / + / - mostly located in ADs (34/40 ADs, 16/32 WDs)	-	-
bond:CN_amine_ter-N_generic	- / - / + / - mostly located in ADs (31/40 ADs, 14/32 WDs)	-	-
bond:CN_amine_ter-N_aliphatic	- / - / + / - mostly located in ADs (31/40 ADs, 14/32 WDs)	-	-
bond:NN_hydrazine_alkyl_N(connect_Z=1)	- / - / + / - only located in WDs (0/40 ADs, 5/32 WDs)	-	-

DS_4, DS_5 ve DS_6 için Bond:CN_amine_aliphatic_generic en önemli tanımlayıcılardan biridir. Çizelge (4.3)'e göre onaylanmış ilaçlarda bulunma oranı geri çekilen ilaçlara göre daha yüksektir. DS_3 için kimyasal bağ sayısını tanımlayan bonds (<42) ise genellikle geri çekilen ilaçlar olarak kategorize edilir. Benzer şekilde DS_2 için Aspheric:Cor3D:ori1 değeri 0.29 ile 0.47 arasında onaylanmış ilaçları belirtir.

4.3.2 Sınıflandırma

Bu çalışmada sınıflandırma teknikleri L SVM, MG SVM, CG SVM, BS T ve BG T'dir. Geliştirilen modeller ilaçları onaylanmış ve geri çekilen kategorilere ayırmak için FS_1'dan FS_6'ya kadar olan özellik setlerini kullandı. Oluşturulan modeller on kat çapraz doğrulama yöntemi vasıtasıyla doğrulandı. K=10 değeri bize daha fazla örnekle çalışma imkanı sunar. Buda tahminlerimiz üzerinde daha doğru bir güven aralığı ve problemler üzerinde iyi bir denge elde etmemizi sağlar. Diğer yandan yüksek ve düşük K değerleri sırasıyla düşük bias, daha yüksek varyanslı tahmincileri ve yüksek bias, daha düşük varyanslı tahmincileri getirmektedir (Kohavi, 1995). Orijinal veri seti (eğitim seti) on alt gruba rastgele bölündü ve Çizelge (4.4)'teki sonuçlar, çeşitli eğitim ve test örnekleri ile on deneme boyunca çıkan sonuçların ortalaması alınarak elde edildi. DS_1-DS_6 modellerin doğruluk oranları (AR) Çizelge (4.4) 'te verildi.

Çizelge 4.4: Test setleri için doğruluk oranına dayalı modellerin performans karşılaştırması.

	L SVM	MG SVM	CG SVM	BS T	BG T
DS_1	0.76	0.78	0.77	0.73	0.73
DS_2	0.77	0.77	0.77	0.81	0.89
DS_3	0.89	0.88	0.74	0.85	0.82
DS_4	0.65	0.71	0.72	0.68	0.79
DS_5	0.71	0.65	0.71	0.72	0.74
DS_6	0.74	0.72	0.72	0.81	0.72

L SVM, Linear Support Vector Machine; MG SVM, Medium Gaussian Support Vector Machine; CG SVM, Coarse Gaussian Support Vector Machine; BS T, Boosted Trees; BG T, Bagged Trees.

AR sonuçlarına göre metotlar MG SVM, BG T, L SVM, BG T, BG T ve BS T, veri setleri DS_1, DS_2, DS_3, DS_4, DS_5 ve DS_6 için yüksek performans (sırasıyla %78, %89, %89, %79, %74 ve %81) göstermiştir. Bu sonuçlar, bu çalışmadaki sınıflandırma modellerinin onaylanmış ve geri çekilen ilaçların ayrılması için uygun olduğunu göstermektedir. Bu nedenle, ortaya çıkan modeller ilaç tasarım sürecinde basit filtreler olarak kullanılabilir. Veri setleri için (DSs) AR sonuçlarına göre en iyi elde edilen modeller ve onların AUC, PPV, NPV, SE, SP, F1-S ve MCC sonuçları Çizelge (3.5)'te verildi. Ayrıca sınıflandırma modelinin başarı indeksi olan ROC (Receiver Operating Characteristic) eğrilerinin altında kalan alan da Çizelge (4.5)'te verildi. Alan 1 ise mükemmel bir testi, 0.5 ise değersiz başarısız bir testi gösterir. Alanı hesaplamak için trapeziodların oluşturulmasına dayanan parametrik olmayan yöntem kullanıldı.

Çizelge 4.5: Test setleri için doğruluk oranı (AR), eğri altındaki alan (AUC), pozitif öngörme değeri (PPV), negatif öngörme değeri (NPV), duyarlılık (SE), özgüllük (SP), F1-skoru (F1-score) ve Matthews korelasyon katsayısına (MCC) dayalı sınıflandırıcı sonuçlarının performans karşılaştırması.

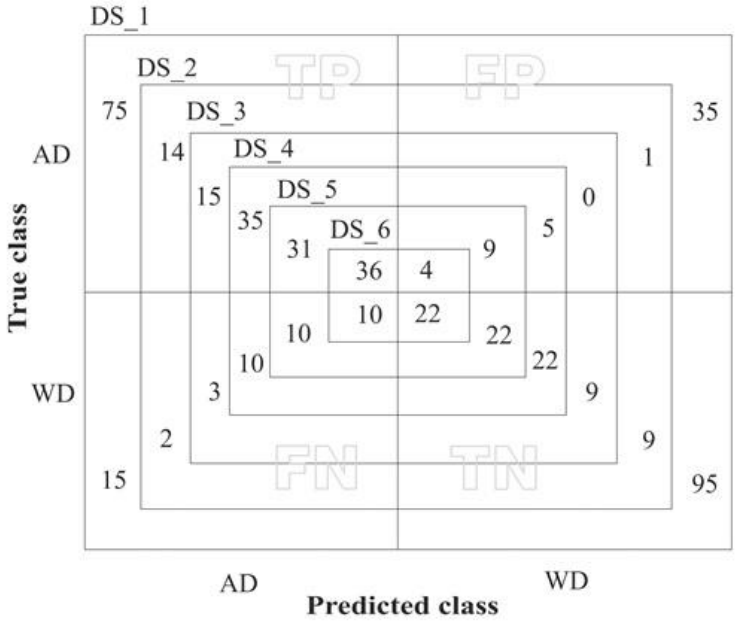
	DS_1 (MG SVM)	DS_2 (BG T)	DS_3 (L SVM)	DS_4 (BG T)	DS_5 (BG T)	DS_6 (BS T)
Accuracy rate	0.78	0.89	0.89	0.79	0.74	0.81
Area under curve	0.83	0.86	0.88	0.85	0.78	0.77
Positive predictive value	0.68	0.93	1.00	0.88	0.77	0.90
Negative predictive value	0.86	0.82	0.75	0.69	0.69	0.69
Sensitivity	0.83	0.88	0.83	0.78	0.76	0.78
Specificity	0.73	0.90	1.00	0.81	0.71	0.85
F1 score	0.75	0.90	0.91	0.82	0.77	0.84
Matthews correlation	0.55	0.76	0.79	0.58	0.46	0.61

L SVM, Linear Support Vector Machine; MG SVM, Medium Gaussian Support Vector Machine; CG SVM, Coarse Gaussian Support Vector Machine; BS T, Boosted Trees; BG T, Bagged Trees.

Çizelge (4.5)'e göre, AUC sonuçları veri setleri için 0.77 ile 0.88 arasında değer aldı. Elde edilen sonuçlara göre, L SVM metodu DS_3 için en yüksek performansı (0.88) gösterdi. MCC kriteri için, L SVM metodu DS_3 için en yüksek performansı gösterirken (0.79), BG T yöntemi DS_5 için en düşük performansı (0.46) gösterdi. L SVM metodu DS_3 için PPV sonuçlarına göre diğer dört metottan daha iyi performans gösterdi (L SVM = 100 ve diğer metotlar 0.68 ile 0.93 arasında değer aldı). NPV sonuçlarına göre, MG SVM metodu DS_1 için 0.86 iken, metotlar BG T,

BG T ve BS T, sırasıyla DS_4, DS_5 ve DS_6 için aynı NPV sonuçlarını (0.69) gösterdi. SE ve SP sonuçlarını göz önüne alındığında, BG T ve L SVM metotları sırasıyla DS_2 ve DS_3 için iyi sonuçlar (0. 88 ve 100) elde etti. F1-S için, L SVM metodu DS_3 için 0.91 değerini aldı.

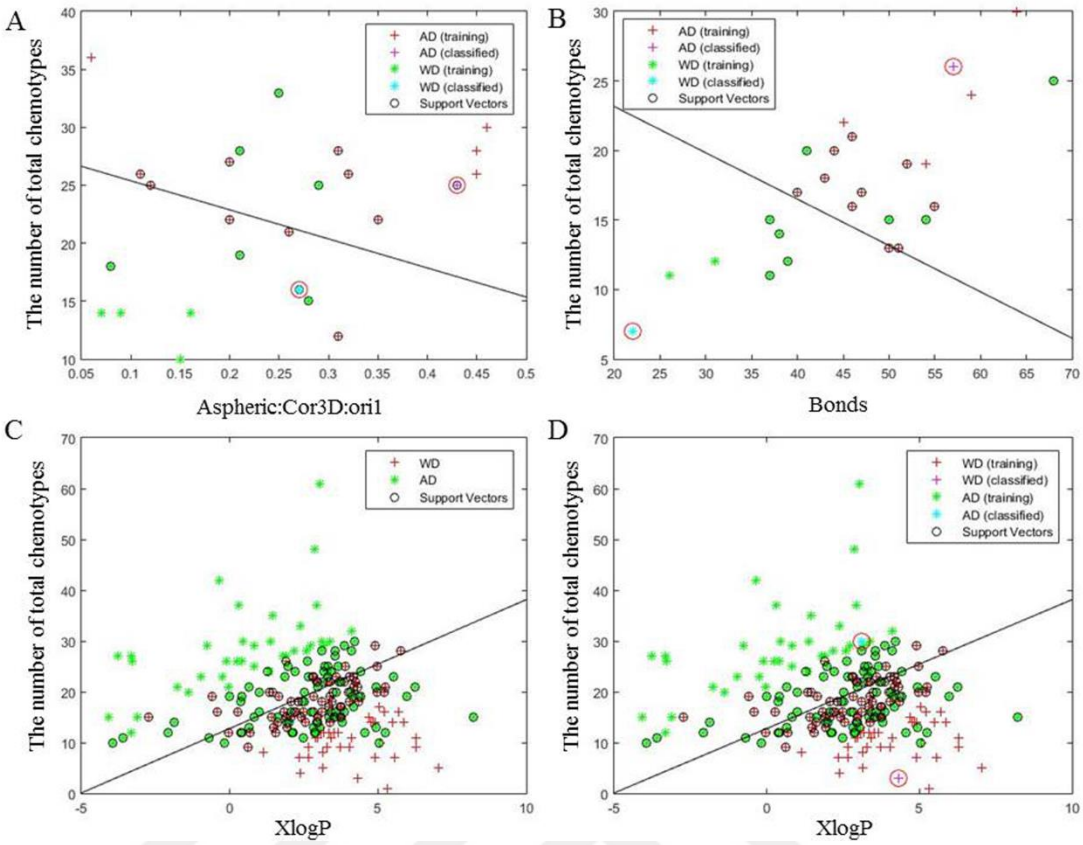
Veri setleri için geliştirilen SVM ve topluluk DT modellerinin her biri tutarlı AR sonuçları verdi. Şekil (4.3)'te AD onaylanmış, WD geri çekilen ilaçları temsil etmektedir. Buna göre DS_1 için 110 AD'den 75 tanesi (TP) doğru tahmin edildi ve 110 WD'den 95 tanesi (TN) MG SVM modeliyle doğru tahmin edildi. BG T metodu DS_2 için 23 ilacı (TP'ler ve TN'ler dahil) doğru olarak sınıflandırdı. Buna ek olarak, L SVM metodu DS_3 için 3 ilacı (FN) kaçırdı ve bunların hepsi WD'lerdi. 40 AD'nin 35 tanesi (TP) doğru tahmin edildi ve 32 WD'nin 22 tanesi (TN) DS_4 için BG T modeliyle doğru tahmin edildi. 40 AD'nin 31 tanesi (TP) doğru sınıflandırıldı ve 32 WD'nin 22'si (TN) DS_5 için BG T modeliyle doğru kategorize edildi. Ayrıca BS T metodu DS_6 için 58 ilacı doğru olarak sınıflandırdı (TP'ler ve TN'ler dahil) ve 14 ilacı kaçırdı (FN'ler ve FP'ler dahil). Bu açıkça kurulan SVM ve topluluk DT modellerinin test setlerinde AD'leri ve WD'leri sınıflandırabildiğini göstermektedir. Karmaşıklık matrisleri Şekil (4.3)'te verildi. TP, doğru pozitif; FP, yanlış pozitif; FN, yanlış negatif; TN, doğru negatif.



Şekil 4.3: Karmaşıklık matrislerinde DS_1 ile DS_6 arasındaki sınıflandırma sonuçlarının karşılaştırılması.

Çalışmada çeşitli hastalık gruplarına ait sinir sistemi (NS) ilaçlarını içeren veri kümelerinden bu molekülerin çok sayıda özelliği kullanılarak, NS ilaçlarının geri çekilen/onaylanmış durumunu tahmin etmek için yeni sınıflandırma modelleri geliştirildi. Geliştirilen modellerden biri, aday ilaç moleküllerini test etmeleri amacıyla araştırmacılar ve son kullanıcılar için verildi. Bu amaçla, araştırmacılar verilen prosedürü sırasıyla yerine getirmelidirler. İlk olarak, Molecular Networks Inc.'ten CORINA.Symphony programını kullanarak test veri setinde aday ilaç molekülleri için 18 ToxPrint Kemotip tanımlayıcıları (Çizelge (4.1)/FS_5) hesaplanmalıdır. Geliştirilen model DVD_ModelDosyası Ek.3'ten elde edilir. Ardından, MATLAB ModelDosyası.mat dosyası MATLAB yazılım paketi aracılığıyla çalışma alanına alınmalıdır. Yapı sınıflandırma nesnesini ve tahmin fonsiyonunu içerir. Son olarak, içe aktarılan sınıflayıcı yeni verilere ilişkin tahminler yapmak için aşağıdaki form kullanılabilir, yfit=predict(trainedClassifier,Thetestdata {:,trainedClassifier.PredictorNames}) Thetestdata burada (tablo) test dosyanızın adıdır. Tablo, eğitim verilerinizle aynı öngörücü isimleri içermelidir (Çizelge (4.1)/FS_5). Çıktı yfit her veri noktası için sınıf tahmini (AD veya WD) içerir.

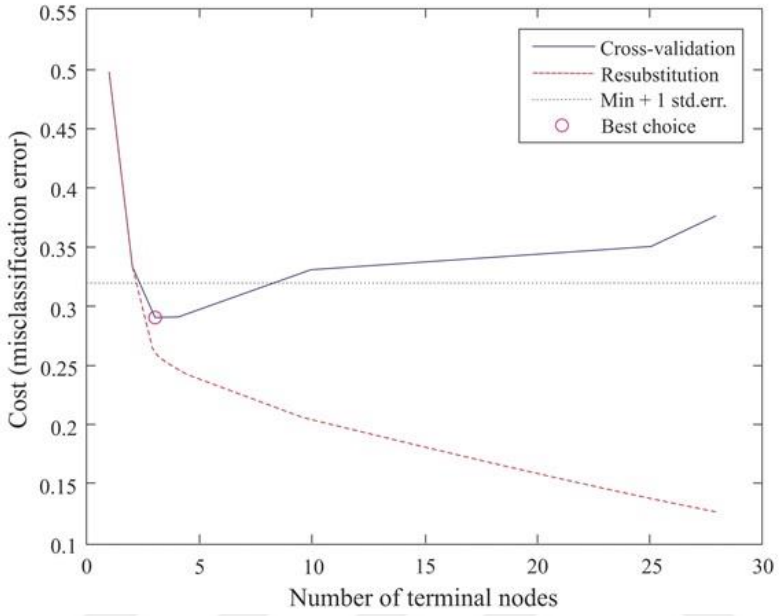
Eğitimli SVM sınıflandırıcısı sınıflandırma yaparken durumları hedef kategorilerine göre ayıran 1 boyutlu bir hiper düzlem (yani bir çizgi) bulur. SVM'nin amacı iki sınıf arasındaki marjı en yükseğe çıkaracak hiper düzlemi seçmek ve bilinmeyen verilerle baş ederken sınıflandırıcının hatasını azaltmaktır. Bu çalışmada, bir ilaç molekülünün sınıflandırılması için, DS_1, DS_2 ve DS_3 verilerini AD ve WD grupları üzerinde ayıran en iyi hiper düzlemler, veri setlerinde rank değerine göre ilk iki sırada yer alan tanımlayıcıların değerine göre tanımlandı ve Şekil (4.4)'te gösterildi. Her bir veri seti için en üstteki iki tanımlayıcı, ki-kare istatistik değerleri ile belirlendi, Şekil (4.2). Veri noktaları bir öngörücünün X ekseni üzerindeki değeri ve diğer öngörücünün Y ekseni üzerindeki değeri ile çizilmiştir. Bağımsız doğrulama verileri (her bir veri seti için 1 AD ve 1 WD) DS_1, DS_2 ve DS_3 için yeni üretilen SVM modellerinin öngörülebilirliği açısından test edildi. Bu iki yeni ilacın sınıflandırılması Şekil (4.4)'te gösterildi [(A) **x:**0.27, **y:**16 and **x':**0.43, **y':**25, (B) $\mathbf{x_1}$ **:**22, $\mathbf{y_1}$**:**7 and $\mathbf{x_1}$**':**57, $\mathbf{y_1}$**':**26, (D) $\mathbf{x_2}$**:**4.32, $\mathbf{y_2}$**:**3 and $\mathbf{x_2}$**':**3.11, $\mathbf{y_2}$**':**30]. SVM modelleri eğitim setlerinin dışındaki ilaçlar içinde önemli sınıflandırma yeteneklerine sahiptir.



Şekil 4.4: İki yeni ve mevcut ilaçların bir boyutlu hiper düzlem ile AD ve WD gruplarına sınıflandırılması (A) DS_2, (B) DS_3 ve (C-D) DS_1.

DS_1 farklı hastalık gruplarına ait onaylanmış ve geri çekilen ilaçlar içerdiğinden karar ağacı oluşturmak için ilaç moleküllerinin önemli yapısal özellikleri hakkında yeterli bilgiye sahiptir. Bu çalışmanın amacı, her ilaç adayı molekülü için XlogP (x) ve toplam kemotip sayısı (y) ile karar kurallarını bulmak ve sınıf atamalarını belirlemektir. Bir dizi kural her bir örneği 28 terminal düğümden birine sınıfladı. Aslında orijinal ağacın çeşitli alt kümeleri vardır. Daha basit bir ağaç elde etmek için yeniden yerleştirme hatası (resubstitution error, dtResubErr) ve çapraz doğrulama hatası (cross-validation error, dtCVErr) karar ağacı için hesaplandı. DtResubErr ve dtCVErr sonuçları sırasıyla 0.1273 ve 0.3182'dir. Bu basit ağaç yeni bir örneği sınıflandırırken karmaşık olanlardan daha iyi performans gösterdi. En küçük ağacı bulmak için kesme değeri (cutoff value) hesaplandı. Kesme değeri minimum maliyet (minimum cost) artı bir standart hataya (standard error) eşittir. Grafiğin en iyi seviyesi, bu sınırın altındaki en küçük ağaca karşılık gelir. Budanmış ağaç için tahmin edilen yanlış sınıflandırma hatası (0.2909) olarak hesaplandı. En iyi seviye (best level) = 0 budanmamış ağaca karşılık gelir bu yüzden bir indeks olarak kullanmadan önce 1 eklenmelidir. Şekil (4.5) en küçük çapraz doğrulama hatasını

göstermektedir. Bu hata AD ve WD grupları üzerindeki DS_1 verileri için budanmış ağacı elde etmede kullanılır. Aşağıdaki kural seti DS_1 için budama ağacından elde edilmiştir, (i) y < 23.5 ve x < 1.08 ise ilaç grubu = AD, (ii) y < 23.5 ve x >= 1.08 ise ilaç grubu = WD, (iii) Eğer y >= 23.5 ise ilaç grubu = AD. İlaçların sınıflandırılmasının amacı, budama ağacı ile ilaç grubunu belirlemektir.



Şekil 4.5: Orijinal ağacın çeşitli alt kümeleri için yeniden birleştirme hatası ve çapraz doğrulama hatasının hesaplanması ve AD ve WD grupları üzerindeki DS_1 verileri için en küçük çapraz doğrulama hatası ile budanmış ağaç için tahmin edilen yanlış sınıflandırma hatası.

4.3.2.1 Leave-one-out cross validation

K sayısının veri grubundaki örnek sayısına eşit olduğu K-kat çapraz doğrulamanın özel bir durumudur. Böylece, öğrenme algoritması her örnek için bir kez diğer tüm örnekleri bir eğitim seti olarak ve seçilen örneğide tek parçalık test seti olarak kullanır (Drehmer and Morris, 1981). Çalışmalarda kullandığımız DS_2 ve DS_3 veri kümelerinde sırasıyla 26 ve 27 ilaç bulunmaktadır. Veri kümelerinde örnek sayısının az olması sebebi ile sınıflandırma sonuçlarının güvenilirliği açısından leave-one-out cross validation yöntemi DS_2 ve DS_3 için denenmiş ve sonuçlar 10-kat çapraz doğrulama metodu ile uyumluluğu değerlendirilmiştir. Sınıflandırma teknikleri olarak yine aynı şekilde DS_2 ve DS_3 için BG T ve L SVM kullanıldı. Oluşturulan modeller leave-one-out cross validation yöntemi ile doğrulandı. K= 26 ve 27 değerleri deneylerimizde bize daha fazla örnekle çalışma imkanı sunar. Buda tahminlerimiz üzerinde daha doğru bir güven aralığı elde etmemizi sağlar. Bunun

yanında yüksek K değerleri düşük bias, daha yüksek varyanslı tahmincileri getirirki tahmin edici daha kesin sonuçlar vericektir. Çizelge (4.6)'da DS_2 ve DS_3 için leave-one-out cross validation yöntemi ile doğrulanan sınıflandırma modellerinin performans değerleri yer almaktadır.

Çizelge 4.6: Test setleri için doğruluk oranı (AR), eğri altındaki alan (AUC), pozitif öngörme değeri (PPV), negatif öngörme değeri (NPV), duyarlılık (SE), özgüllük (SP), F1-skoru (F1-score) ve Matthews korelasyon katsayısına (MCC) dayalı sınıflandırıcı performans sonuçları.

	DS_2 (BG T)	DS_3 (L SVM)
Accuracy rate	0.89	0.89
Area under curve	0.84	0.88
Positive predictive value	0.93	1.0
Negative predictive value	0.82	0.75
Sensitivity	0.88	0.83
Specificity	0.90	1.0
F1 score	0.90	0.91
Matthews correlation	0.76	0.79

L SVM, Linear Support Vector Machine; BG T, Bagged Trees.

Çizelge (4.5)'de DS_2 ve DS_3 için elde edilen sınıflandırma performans değerlerinin Çizelge (4.6) ile tamamen uyumlu olduğu gözlendi.

4.3.2.2 Sınıflandırma modelinin bir veri seti üzerinde doğrulanması

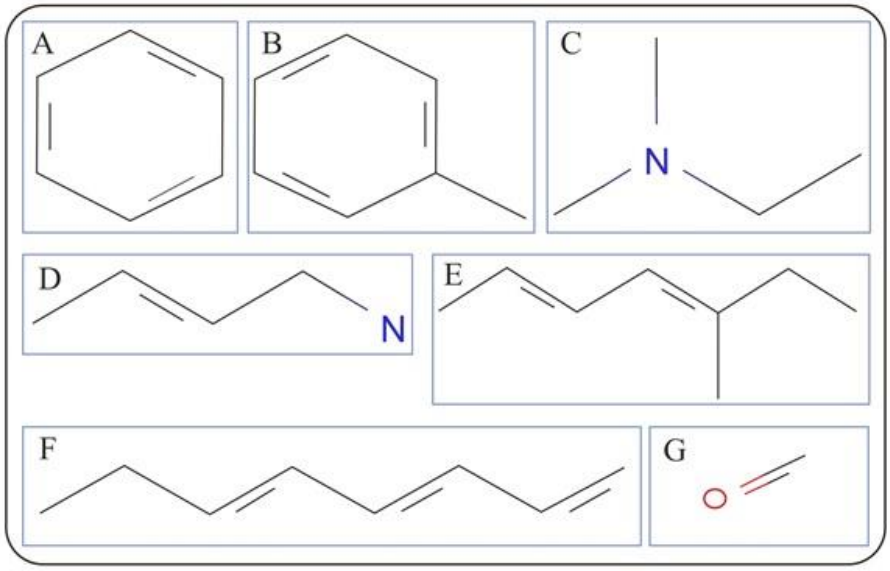
İlaç tasarımına standart bir yaklaşımda bileşik kütüphanelerinin hesaplamalı yöntemlerle taranmasıdır. Ligand tabanlı sanal taramada genel olarak spesifik biyolojik bir aktiviteye sahip olan az sayıda molekülün çok sayıda aktif olmayan bileşikler arasından ayırt edilmesi hedeflenir. Çalışmada kullanılan veri seti UCI machine learning repository'den (PubChem Bioassay veri setlerinden) AID362 veri setindeki aktif olmayan 20 bileşik ile oluşturulmuştur (ilaç benzeri moleküller ancak aktif bileşikler değil). Sınıflandırma modeli olarak DS_1 ile geliştirilen Model_1 kullanılmıştır. DS_1 farklı hastalık gruplarına ait 220 ilaç molekülü içermekteydi. Bölüm (4.2.1)'de modele ilişkin ayrıntılı bilgi yer almaktadır. Veri setinde aktif olmayan bileşikleri seçmemizin sebebi bu bileşiklerin ilaç tasarımında spesifik bir aktiviteye sahip olmadıklarının hesaplamalı olarak önceden belirlenmesidir. Bu

hesaplamalarla elde edilen sonuçlar yaklaşık 480000 ilaç benzeri molekülün sanal olarak taranmasına dayanmaktadır. DS_1 kullanılarak elde edilen model_1 MATLAB yazılım paketi ile geliştirilmiştir. Sonuç olarak 20 aktif olmayan bileşik model üzerinde test edildiğinde bunlardan 16 tanesi geri çekilen ilaçlar kategorisinde buna karşılık 4 tanesi onaylanmış ilaçlar kategorisinde yer almıştır. Önerilen model onaylanmış ve geri çekilen ilaçları sınıflandırma için gerekli öznitelikleri içermektedir yani model aktif ve aktif olmayan bileşikleri sınıflandırma problemine ait öznitelikleri içermemektedir ama mutlaka test setindeki aktif olmayan bileşeni bir gruba (onaylanmış, geri çekilen) atacaktır. Burada model aslında aktif olmayan bir bileşeni geri çekilen ilaç grubuna atarak aslında eldeki özniteliklerle kimyasal bileşiği baştan elimine ediyor. Bu nedenle elde edilen sonuç önerilen model için istenilen ve beklenen bir sonuçtur. Elde ettiğimiz bu sonuçla modelin farklı bir veri setine ilişkin sınıflandırma öngörüsünü de test etmiş olduk.

4.3.3 Alt çizge madenciliği

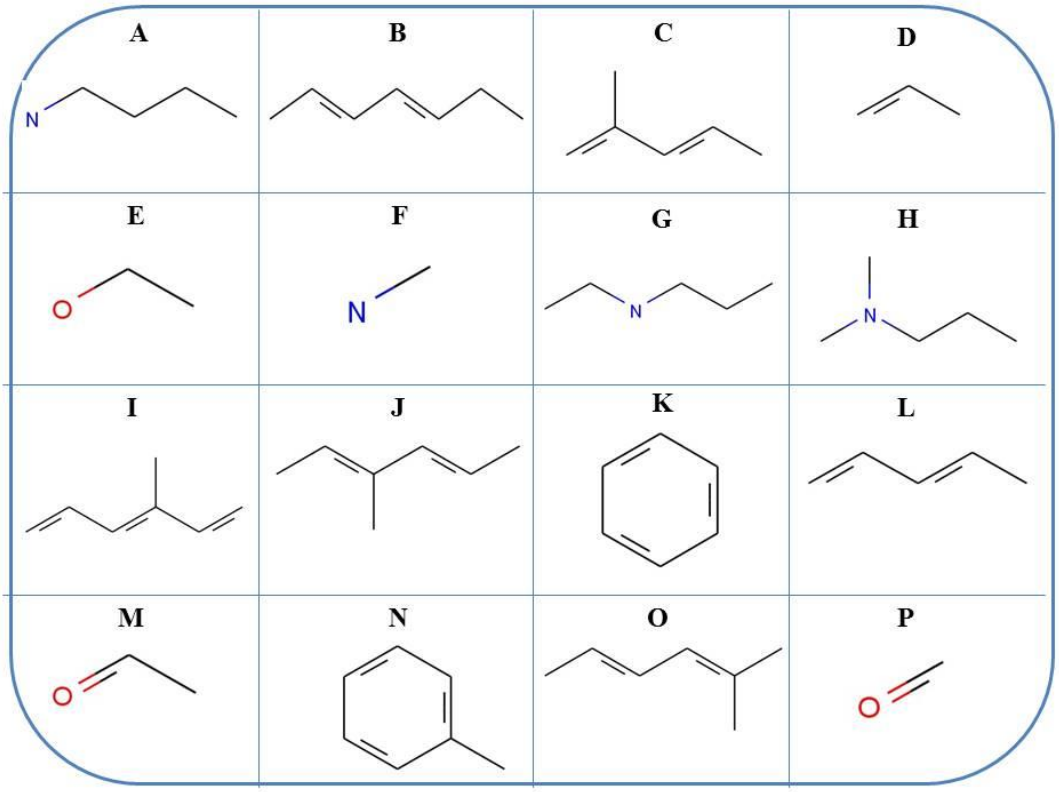
İlaçların onaylanmış veya geri çekilen durumunun belirlenmesi amacıyla ilaç veri setlerine sık alt çizge madenciliği uygulanıp yapılarında bulunan ayırt edici fragmanlar belirlenebilir. Geri çekilen sinir sistemi ilaç veri setine (32 geri çekilen ilaç) gSpan algoritması uygulanarak (minimum destek 60%) yapılarındaki göze çarpan fragmanlardan hangilerinin geri çekilmelerine sebep olabileceği araştırıldı. Bunları belirlemek amacıyla geri çekilen sinir sistemi ilaçlarının yanında onaylanmış sinir sistemi ilaçlarına da (145 onaylanmış ilaç) sık alt çizge madenciliği uygulandı. Her bir veri seti için sıklıkla tekrar eden fragmanlar belirlendi ve sadece onaylanmış/geri çekilen ilaç moleküllerinde bulunan/bulunmayan fragmanlar belirlendi. Bu fragmanlar ilaç tasarım çalışmalarına oldukça katkı sağlar. Çalışmada öncelikle 32 tane marketlerden geri çekilen ilaçtan yaklaşık 20 tanesinde bulunan kapalı fragmanlar belirlendi. Bu fragmanlar aşağıda açıklamalarıyla birlikte verilmiştir. Benzen, toluen, N, N-dimetiletilamin (DMEA), krotilamin, 5-metil-2,4-heptadien, oktatrien, karbonil grubudur. Bu fragmanlar geri çekilen sinir sistemi ilaçlarının kimyasal çizge gösterimlerinin çoğunda (en az % 60'ında) bulunmaktadır. Bu fragmanlar ilaçların marketlerden çekilmesine sebep olan yapılarda olabilir ve insan ilaç metabolizmasına zarar verebilir. Bu alt çizge fragmanlarından benzen, periferik kan lökositlerinde ve kemik iliğinde kromozomal anormalliklere yol açar ve DNA oluşumunu bozar. Toluen, bir fenil grubuna bağlı bir CH3 grubundan oluşur.

Toluen benzen grubundan daha az toksiktir. Toluen rekabetçi olmayan NMDA reseptör antagonisti ve GABAα reseptörü pozitif allosterik modülatör olarak işlev görür. N, N-dimetiletilamin (DMEA) sinir sistemi ilaçlarının yapısında bulunan diğer bir fragmandır. İnsanda amine bağımlı reseptör TAAR5'in agonistidir. Koriorilamin veya 2-buten-1-amin doymamış primer mono amindir ve DMEA gibi N grubu içerir. Buna ek olarak 5-metil-2,4-heptadien bir alken ve doymamış hidrokarbon'dur. İki karbon-karbon çift bağ içerir. 1,3,5 oktatrien üç çift bağa sahip dallanmamış bir sekiz karbonlu alkatriendir. Kahverengi alg türü tarafından üretilebilir. Ayrıca karbonil grubu bir oksijen atomuna çift bağlanmış bir karbon atomundan oluşur. Aldehid, keton, karboksilik asit, ester ve amid karbonil grubunu içerir. Altçizge fragmanları Şekil (4.6)'da gösterilmiştir. Sonuçlar geri çekilen sinir sistemi ilaç moleküllerinin kimyasal çizge gösterimlerinden elde edilmiştir.



Şekil 4.6: Geri çekilen sinir sistemi ilaç moleküllerinin kimyasal çizge gösterimlerinden elde edilen kapalı fragmanlar (A to G), A) Benzene B) Toluene C) N,N-Dimethylethylamine (DMEA) D) Crotylamine E) 5-Methyl-2,4-Heptadiene F) Octatriene G) Carbonyl group.

Çalışmanın diğer önemli kısmı onaylanmış 145 sinir sistemi ilacının en az 87'sinde ortak olarak bulunan fragmanları belirlemektir. Şekil (4.7)'de bu sık altçizge fragmanları verildi. Sonuçlar onaylanmış sinir sistemi ilaç moleküllerinin kimyasal çizge gösterimlerinden elde edilmiştir.



Şekil 4.7: Onaylanmış sinir sistemi ilaç moleküllerinin kimyasal çizge gösterimlerinden elde edilen kapalı fragmanlar (A to P), A) N-bütilamin B) 2,4, heptadien 3) 2-metil 1,3-pentadien D) Propilen E) Etanol F) Metilamin G) N-Etil-N-propilamin H) N,N dimetilpropilamin I) 3-Metil 1,3,5-hekzatrien J) 3-Metil 2,4-hekzadien K) Benzen L) 1,3-pentadien M) Asetaldehit N)Tolien O) 2- metil 2,4 hekzadien P) Karbonil Grubu.

Yukarıda belirlenen fragmanlardan n-bütilamin organic bir bileşiktir ve bütanın dört izomerik amininden bir tanesidir. Diğerleri sec-butilamin, tert-butilamin ve izobutilamindir. n-bütilamin diğer aminler gibi amonyak benzeri bir kokuya sahiptir. Hava ile temas haline geldiğinde sarı renk alır. n-bütilamin amin grubundan dolayı zayıf bir baz özelliği gösterir. Bu madde daha çok böcek öldürücü olan pestisitlerin yapısında bulunur. En önemlisi tiyokarbaazidlerdir. Bunlar karbondisüfit ile hidrazin reaksiyonu sonucu oluşurlar. Ayrıca, n-bütilamin farmakoloji ve kremlerde emilsiyon madde olarak kullanılabilir. Farmostetik katkıların % 50 sinden fazlası kiral amin birimine sahip en az bir adet kiral merkez içerir. Ayrıca 2020 yılına kadar da kullanılan tüm ilaçların % 95'i içinde bulunması beklenmektedir (Slabu ve diğ., 2017). n-bütilamin fazla dozda alındığında ölümcül olabilir. Sıçanlar ile yapılan çalışmalarda ölümcül dozu (LD_{50})= 366 mg/kg olarak bulunmustur. Ayrıca n-bütilamin alkolik içeceklerin, domates, buğday unu ve peynirin içinde bulunabilir.

Bu maddenin aşırı solunması kaşıntı, tahriş, başağrısı ve kusmaya neden olabilir. 2,4, heptadien iki adet çift bağ içeren alken grubu içeren bir bileşiktir. Çift bağlar karbon zincirlerinin doymamışlığını göstermektedir. Alkenler suda az çözünürler ve genellikle kokusuzdurlar. Propene kadar olanlar oda kosullarında gaz halinde bulunurlar. Beş karbondan onaltı karbona kadar sıvı ve daha fazla karbon atomlatı içerenler ise katıdırlar. Hayvanlarda ve insanlarda ateşli hastalıklarda kullanılan ilaçların içerisinde heptadien bulunur. Bunların TRPA1 ion kanalları üzerine etkileri araştırılmaktadır (Leamy ve diğ., 2011). Ayrıca heptadienler kurkuminoidlerin yapısına katılırlar. Kurkuminoidler çoğu sinir dokusu bozulması hastalıklarında sinyal iletiminde ve biyokimyasal reaksiyonların gerçekleşmesinde görev aldığı bilinmektedir. Ayrıca bunlar izositrat, NADH dehidrojenaz, sitokrom c oksidaz, kompleks 1 ve toplam ATP seviyelerini arttırarak yaşlanma ile meydana gelen mitokondri bozulmasına karşı sinir sistemi dokusunu korur (Dey ve diğ., 2017). 2-metil 1,3-pentadien bir adet metil grubu taşıyan alkenlerdir. Bunlarda alkenlerin genel özelliklerini taşırlar. 1,3-pentadienler hakkında ilaçlarla etkileşimleri ile ilgili fazla bilgi bulunmamaktadır. Fakat biz biliyoruz ki son birkaç yılda MAO olarak bildiğimiz metilaluminoksan üzerine sterospesifik etkilerinden dolayı 1,3-pentadienlerin fiziksel özellikleri sıklıkla çalışıldı. 1,3-pentadienlerden kristal yapılı polimerler üretildi (Bertini ve diğ., 2009). Propilen basit bir alkendir. Bir adet çift bağ içerir ve kısa zincire sahiptir. Propen olarakta adlandırılırlar. Propilen doğada fermantasyonun yan ürünü olarak açığa çıkarlar. Ayrıca propilen işlenmemiş fosil, petrol ve doğalgaz ile üretilebilir. Propilen petrol sanayide etilenden sonra ikinci en önemli başlangıç materyalidir. Propilen ayrıca imin yapısı ile birleşerek dentrimer olarak iş görebilir. Dentrimerler kemoterapik ajanları taşıyan ve kanser tedavisinde kullanılan polimerlerdir. Bu propilenimin içeren dentrimerler anyonik ilaçların aktif formlarının tümör hücrelerine etkili bir şekilde iletimini sağlarlar (Szulc ve diğ., 2016). Etanol diğer adıyla etil alkol en çok bilinen ve kullanılan alkol grubudur. Uçucu ve tutuşabilen özelliğe sahiptir. Petrokimya sanayi ve mayaların fermentasyonu sonucu son ürün olarak oluşabilirler. Uzun zamandan beri ilaç ve dezenfektan olarak kullanılmaktadır. Ayrıca sanayide yakıt olarakta kullanılmaktadır. Fazla etanolun hücre içine girmesi ile toksik etki yaptıgı bir çok çalışma ile kanıtlanmıştır. Son yapılan bir çalısmaya göre etanole maruz bırakılan sıçanların aort içerisindeki antioksidan kapasitesi ve protein sentezindeki etkileri gösterilmiştir (Ceron ve diğ., 2017). Metilamin bir organic moleküldür, amonyak molekülündeki

hidrojenin yerine bir metil grubu gelerek oluşmuştur. Renksizdir. Birincil amin grubundadır. Amonyak ile metanolun sentezi sonucu meydana gelir. Metilamin şeker hastalığında antidiabetic ilaç olarak dipeptil peptidaz IV (DPP IV) ün inhibitörü olarak iş görür (Namato ve diğ., 2014). Bu madde etkisini hiperglisemide doku zararını önleyerek etkilerini gösterdikleri bulunmuştur (Cioni ve diğ., 2006). Ayrıca sinir sisteminde noradrenalin tekrar alımının seçici inhibitörü olarak iş görerek depresyon tedavisinde kullanılmaktadır (Fish ve diğ., 2008). N-Etil-N-propilamin amin grubu bir bileşiktir. Renksiz ve uçucudur. Amin grubundan dolayı zayıf baz olarak kabul edilir. Propanol ve amonyum kloratın yüksek sıcaklık ve basınç altında demir kloratın katalizörlüğü eşliğinde oluşabilir. N,N dimetilpropilamin de amin grubu bir bileşiktir. Renksizdir ve bazik özelliğe sahiptir. Göz ve deri ile teması halinde tahrişe neden olur. 3-Metil-1,3,5-hekzatrien bir poli alkendir. Polialkenler görme pigmenti olan rodopsin içinde fotokimyasal rol oynarlar. Bunlar ayrıca membrane moleküllerinin sırasını belirlemek için florosens olarak görev alırlar. Ayrıca Vitamin E ve Trolox içerisinde antioksidan aktivitenin değerlendirilmesinde görev alabilirler (Labidi ve Djebaili, 2010). 3-Metil 2,4-hekzadien alken grubu bir bileşiktir ve alkenlerin özelliklerini taşır. İki adet çift bağ içerir ve doymamış bir bileşiktir. 1,3-pentadien bir alkendir ve diğer adı piperilendir. Uçucudur iki adet çift bağ içerir. İşlenmemiş yağdan etilen üretiminin yan ürünü olarak ortaya çıkar. Ayrıca plastik sanayide ham madde olarak kullanılır. Asetaldehit en önemli aldehitlerden biridir. Sanayide ve doğada yüksek oranlarda bulunabilir. Kahvede, ekmekte ve meyvede yüksek oranlarda bulunabilir. Bitkilerde fermantasyon sürecinde oluşturulabilir. Alkol dehidrojenazın etanolun oksidasyonu ile oluşturulabilir. Alkol tedavisi mekanizmasında görev alır. Kanserojendir. Yüksek miktarlarda alımları başağrısı, kusma ve sersemlik meydana getirebilir. Asetaldehitin asetik asite dönüşümü genetik olarak sağlayamayan insanların Alzheimer hastalığına daha yatkın olduğu ıspatlanmıştır. Ayrıca asetaldehitin karaciğer kanser hücrelerinde süperoksit dismutaz enzimini hedeflediği gösterilmiştir (Cornejo ve diğ., 2014). Diğer bir çalışmada nikotin verilen genç sıçanlarda asetaldehitin nikotinin etkisini arttırarak beyindeki sinyal iletimini etkilediği ve nikotinin etkisini arttırdığı bulunmuştur (Sershen ve diğ., 2009). 2- metil 2,4 hekzadien alken grubu bir bileşiktir. Iki adet çift bağ içermektedir. Alkenlerin bütün özelliklerini taşır. Metil grubu içerir.

Benzen, Toluen ve karbonil grup hem piyasadan geri çekilen hemde onaylanmış sinir sistemi ilaçlarında ortak olan fragmanlardır. Bunlar ilaç olmak için ilacın yapısında büyük olasılıkla bulunması gereken ve ilacın etki göstermesinde etkili olan ama ilacın piyasadan çekilip çekilmeyeceği hakkında öngörü sağlayacak gruplar değildirler. Diğer taraftan n-bütilamin, 2,4, heptadien, 2- metil 1,3-pentadien, Propilen, Etanol, Metilamin, N-Etil-N-propilamin, N,N dimetilpropilamin, 3-Metil-1,3,5-hekzatrien, 1,3-pentadien, Asetaldehit ve 2- metil 2,4 hekzadien piyasadan geri çekilen sinir sistemi ilaçlarının 20 tanesinin yapısında bulunmazken onaylanmış ilaçların yapılarında sıklıkla bulunan fragmanlardır. Sık alt çizge madenciliği ile belirlenen bu fragmanlar bir ilacın onaylanmış/geri çekilen durumu hakkında bilgi verir ve geri çekilen ilaçların yapılarında sıklıkla görülen fragmanlar aday ilaç molekülleri için sınıflandırma problemlerinde dikkatle ele alınmalıdır.

4.4 Tartışma

Burada ele alınan çalışmalar Bölüm 1, Literatür araştırması kısmında ayrıntılı olarak anlatıldı. Bu bölümde yapılan çalışma diğerleriyle karşılaştırıldı artıları ve varsa eksik yanları ele alınarak okuyucuya sunuldu.

Cao (2012) yaptığı çalışmada HDAC8 inhibitör ve inhibitör olmayanları 23 moleküler tanımlayıcı kullanarak ayırmaya çalıştı. Çalışmada molekül özellikleri olarak global moleküler, yüzey ve 2 boyutlu ve 3 boyutlu özellikleri veri setindeki tüm bileşikler için hesaplandı. HDon, HAcc, and NRotBond tanımlayıcıları sınıflandırmada en etkin faktörler olarak belirlendi ve model test setinde 0.75 doğruluk oranına ulaştı. Korkmaz (2014) aktif molekülleri aktif olmayanlardan ayırmak için farklı üç özellik seçimi metodu kullanıp SVM modellerini elde etti. Toplam 34 moleküler tanımlayıcı kullandı. Modellerin doğruluk oranları 0.76 ile 0.81 arasında değerler aldı. Zhang (2011) ilaçların nöbet yükümlülüğünü erken safhalarda belirlemek amacıyla bir model geliştirdi. Modelin doğruluk oranı 0.87'dir. Çalışmada nöbet yükümlülüğüne sahip bileşiklerin tahmini için, moleküler elektronik özellikleri, hidrojen bağlama özelliği, moleküler aromatik fonksiyonlar, lipofiliklik, moleküler polar yüzey alanı ve moleküler yapısal bilgi içeren 18 moleküler tanımlayıcı kullanıldı. Klekota ve Roth (2008) çalışmalarında çoklu bileşik kütüphaneleri için biyoaktiviteyi tanımladılar. Biyolojik aktivite ile ilgili alt yapıları

(fragmanlar) belirlemek için 4860 altyapı kümesinden alınan alt yapıları karar ağaçlarını kullanarak birbirinden ayırdılar.

Yaptığımız çalışma önceki çalışmalardan üstündür çünkü ilaç moleküllerin sınıflandırma problemlerinde kullanılmak üzere hesaplanan global moleküler, boyut ve şekil özelliklerinin yanında 729 ToxPrint kemotip özellikleride hesaplanmıştır ve elde edilen modellerin doğruluk oranı 0.89'a ulaşmıştır. Buda geliştirilen özellik setlerinin geri çekilen ilaçları onaylanmış olanlardan ayırmada başarılı olduğunu gösteriyor. Bir molekülün kimyasal yapısında bulunan ToxPrint kemotipleri belirlemek ilacın geri çekilen/onaylanmış durumu hakkında bize önceden bir bilgi verebilir. Buna ek olarak Klekota ve Roth'tan farklı olarak çalışmada sadece sinir sistemi ilaçları için belirlediğimiz ayırt edici fragmanlar mevcuttur ve veri setimiz sınırlı sayıda geri çekilen ilaç yapısı içermektedir.

Çalışmada kullanılan veri setleri DS_1, DS_2,…, DS_6 her biri farklı bir amaca hizmet etmektedir. Bunlardan DS_1, farklı hastalık gruplarına ait onaylanmış ve geri çekilen ilaçları içermektedir. Burada geliştirilen sınıflandırma modeli ilaç adayı bir molekülün gelecekte onaylanmış/geri çekilen durumu hakkında bize öngörüde bulunacaktır. DS_2 ve DS_3 ise sinir sistemi hastalıklarından sırasıyla psikoleptik ve psychoanaleptics'e ait onaylanmış ve geri çekilen ilaçları içermektedir. Burada amaç geliştirilen sınıflandırma modelinin hastalık bazında ilaç adayı bir molekülün onaylanmış/geri çekilen durumu hakkında tahmin yapabilmektir. İlaç veri bankasında geri çekilen ilaç sayısı oldukça sınırlıdır. Hastalık bazında ele aldığımzda ise bu sayı en fazla sinir sistemi hastalıklarının tedavisinde kullanılan ve daha sonra beklenmedik bir yan etki sonucu geri çekilen ilaçlardan oluşmaktadır. Sinir sistemi için toplamda bu sayı 32 dir. Diğer hastalıklara baktığımızda ise hastalık başına düşen geri çekilen sayısı 32'den çok daha azdır. Bu durumda ilaçları sınıflandırmak için geliştirdiğimiz modeller hastalık bazında ele aldığımızda en iyi sinir sistemi için elde edilir. Sinir sisteminden farklı bir hastalık için şu durumda bir model geliştirmek sağlıklı olmayacaktır.

DS_4, DS_5 ve DS_6 veri setleri N01, N02, …, N07 gruplarının hepsinden onaylanmış/geri çekilen ilaç içermektedir. Elimizde tüm sinir sistemi hastalıkları için 32 geri çekilen ilaç olduğundan bu gruplardan gelen onaylanmış ilaçlarla bunlardan dengeli bir veri seti elde edip yine genel olarak (özel bir sinir sistemi hastalığına ait olmayan, N05 gibi) sinir sistemi hastalıklarının tedavisinde kullanılan ilaçlardan bir

model oluşturulmaya çalışılmıştır. Bu model bize genel olarak sinir sistemi hastalıklarının tedavisinde kullanılmak istenen bir aday ilaç molekülünün (herhangi bir gruba ait olabilir) onaylanmış/geri çekilen durumu hakkında öngörüde bulunacaktır. Bu 3 modeli N01, N02, …, N07 gruplarının herhangi birinden olan aday ilaç molekülünü test etmede kullanabilirsiniz. DS_4, DS_5 ve DS_6 birbirinden tamamen farklı onaylanmış ancak aynı geri çekilen (32 ilaç) ilaçları içermektedir. Kullanıcı ilaç adayı molekülünü geliştirilen 3 modelden biri ile test edebilir. Bakılacak olursa her üç modelin NPV değeri aynıdır. Sonuç olarak deneylerdeki her bir veri seti farklı bir motivasyona hizmet etmektedir. Burada geliştirilen modeller ilaç aday moleküllerini sınıflandırma problemlerinde basit bir filtre olarak kullanılabilir.

5. İLAÇLAR ÜZERİNDE HİYERARŞİK ÇOKLU ETİKET SINIFLAMASI

5.1 Giriş

Farklı hastalık gruplarına ait 550'den fazla ilaç hiyerarşik çoklu etiket sınıflaması yöntemi kullanılarak dahil oldukları gruplar belirlenip bu gruplar bir hiyerarşide organize edildi. Çalışmamızda ilaç adayı moleküller için oluşturulan model moleküllerin onaylanmış, geri çekilen ve sinir sistemi ilacı olma durumu hakkında bilgi verirken aynı zamanda ilacın hangi sinir sistemi hastalık grubuna ait olduğunuda öngörür. HMC bir ağaç öğrenme algoritmasıdır ve tüm sınıfları bir kerede öngörür. Ayrıntılı bilgi bölüm (3.7.3)'te verildi. Burada bir ilaç aynı anda birden fazla sınıfa dahil olabilir. Yani bir ilaç sinir sistemi ilacı olup aynı anda N02 sınıfına da dahil olabilir. Uygulamaya bağlı olarak her sınıf bir tane parent'a sahip olduğundan elde edilen hiyerarşi ağaç yapısındadır.

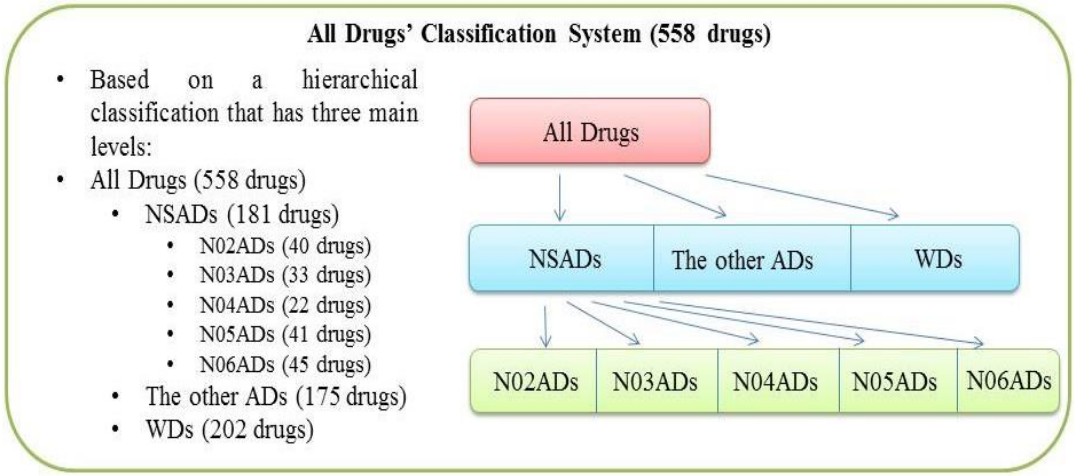
5.2 Materyaller Ve Yöntemler

Çalışmada farklı hastalık gruplarından ve sinir sistemi hastalıklarının tedavisinde kullanılan (N02, N03, N04, N05, N06) onaylanmış ve geri çekilen 558 ilaç Clus-HMC algoritması kullanılarak 3 temel seviyede sınıflandırıldı. İlaç molekülleri KEGG DRUG, PubChem ve DRUGBANK'tan toplandı. 558 ilaçtan oluşan veri seti için bir dizi moleküler tanımlayıcı CORINA Symphony programı ile hiyerarşik sınıflandırma modelleri oluşturmak üzere hesaplandı. Çalışmanın mimari yapısı Şekil (5.1)'de gösterildi. Şekil (5.1)'de onaylanmış sinir sistemi ilaçları NSADs, tüm ilaçları kapsayan en büyük sınıf All Drugs, geri çekilen ilaçlar WDs, farklı hastalık gruplarına ait onaylanmış ilaçlar the other ADs ile gösterilmiştir.

5.2.1 Veri kümelerinin toplanması

Çalışmada kullanılan onaylanmış sinir sistemi ilaçları KEGG DRUG veri tabanından toplandı. N01 ve N07 sınıflarında geri çekilen ilaç yoktur bu nedenle N02, N03, N04, N05, N06'daki ilaçlar çalışıldı. Çizelge (3.1)'de bu ilaçlara ilişkin veri setlerinin

ATC sınıflaması yer almaktadır. Geri çekilen ilaçlar ve diğer onaylanmış ilaçlar (bölüm (3.4)) KEGG DRUG, PubChem ve DRUGBANK'tan alındı.



Şekil 5.1: Farklı hastalık gruplarına ait ilaçların hiyerarşik çoklu etiket sınıflaması.

Çalışmada kullanılan (İlk seviye) HMC_DS veri seti 558 ilaçtan oluşur, Şekil (5.1). (İkinci seviye) 181'i onaylanmış sinir sistemi ilaçlarını, 175'i farklı hastalık gruplarına ait onaylanmış ilaçları ve 202'si geri çekilen ilaçları içerir. (Üçüncü seviye) N02, N03, N04, N05 ve N06 sınıfları sırasıyla 40, 33, 22, 41, 45 ilaçtan oluşur. Geri çekilen ilaçlar seti (202) farklı hastalık gruplarından tüm geri çekilen ilaçları içerir. HMC_DS sınıflandırma çalışmaları için 10-kat çapraz doğrulama metoduyla on farklı eğitim ve test setine bölündü. Her eğitim ve test seti onaylanmış ve geri çekilen ilaçlardan oluşmaktadır.

Buna ek olarak dışardan bağımsız bir test setiyle modelin performansını değerlendirmek amacıyla HMC_DS veri seti eğitim (446) ve test setlerine (112) bölündü. Eğitim ve test setleri hiyerarşide yer alan her sınıftan ilaç molekülüne ait veri içermektedir. Model 446 eğitim setiyle eğitildi ve sonrasında 112 ilaç molekülü üzerinde test edildi. Sonuçlar her iki doğrulama yöntemi için ayrı ayrı verildi.

5.2.2 Moleküler tanımlayıcıların hesaplanması

Tanımlayıcılar bir önceki uygulamada olduğu gibi CORINA Symphony programı ile ilaç veri seti için hesaplandı. Bu çalışmada da 760 moleküler tanımlayıcı, 22'si global moleküler, 8'i boyut ve şekil, 729'u toxprint kemotip tanımlayıcılarından (DVD_Çizelge Ek.2) ve bir kullanıcı özelliğinden oluşmaktadır.

5.2.3 Veri ön işleme ve özellik seçimi

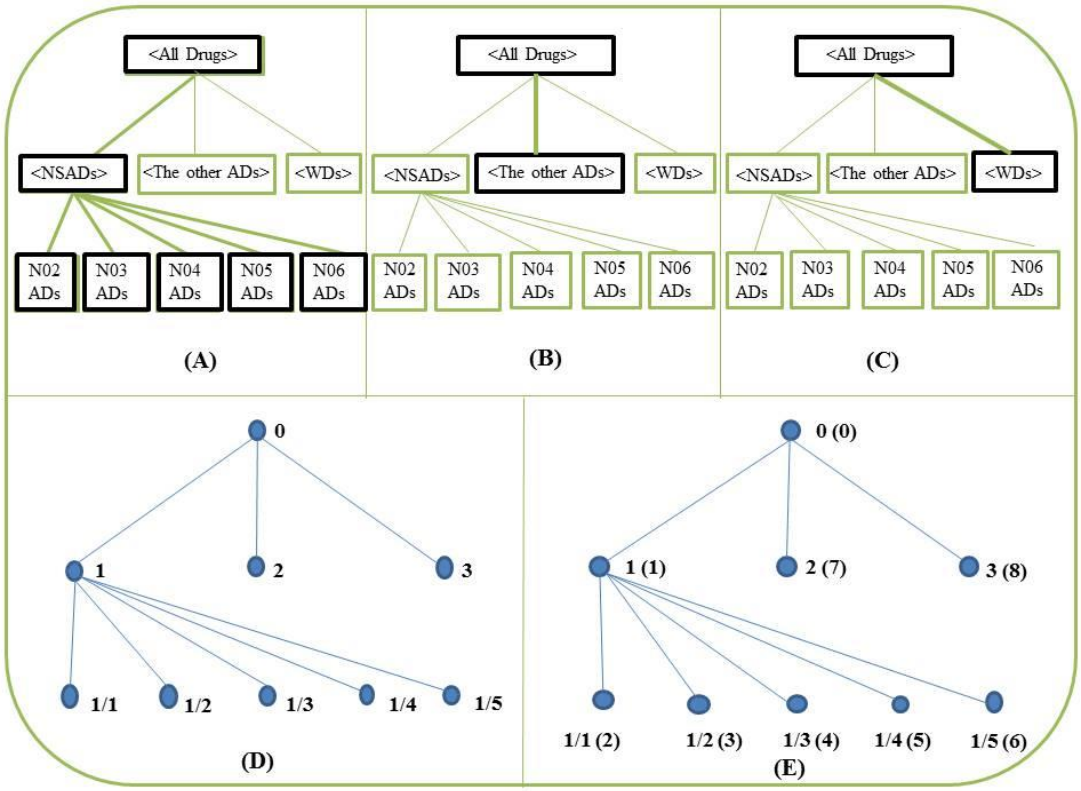
Veri ön işleme bölüm (4.2.3)'te anlatıldığı gibi burda da ilaç veri setine uygulanmıştır. Hesaplanan moleküler tanımlayıcıların hepsi ilaçları hiyerarşik olarak sınıflandırırken kullanılmıştır. Özellik seçimi yapılmamıştır. Her satır verisi bir ilaç molekülüne ait bilgileri temsil eder ve sütunlarda 760 moleküler özellik listelenir. Son sütun ise ilacın ait olduğu hiyerarşik sınıf etiketini taşır. Örnek olarak AllDrugs/NSADs/N03ADs hiyerarşik bir sınıf etiketini temsil eder. "/" sınıflar arasında ayırıcı olarak kullanılır. N03ADs'ın üst sınıfları NSADs ve AllDrugs'tır. Bu durumda N03ADs sınıfındaki bir ilaç aynı anda NSADs ve AllDrugs sınıflarına da aittir.

5.2.4 Hiyerarşik olarak organize edilen sınıflama modellerinin geliştirilmesi

Bu bölümde aday ilaç molekülleri için oluşturulan modeller moleküllerin onaylanmış, geri çekilen veya sinir sistemi ilacı kategorilerinden hangisine ait olduğu hakkında öngörüde bulunur. Deneylerde kullanılan veri seti HMC_DS 10-kat çapraz doğrulama metoduyla ve dışardan bağımsız bir test seti ile test edildi. Clus-HMC-Ens algoritması ile deneylerde kullanılan parametre ayarları detaylı bir şekilde anlatılmıştır. Hiyerarşik çoklu etiket sınıflandırmasını gerçekleştirmek için öncelikle HMC_DS için hmc_ds.arff ve hmc_ds.s iki girdi dosyası hazırlanır. Bunlardan hmc_ds.s dosyası parametre ayarlarını içermektedir. hmc_ds.arff ise eğitim setini içerir. Çıktı dosyaları hmc_ds.out ve hmc.xval'dır. Çalışmada doğrulama metodu olarak ayrı bir test sınıfı kullanıldığında hmc_ds_test_.arff dosyası hazırlanır ve çıktılar hmc_ds.out dosyasına yazılır. Eğer modeli test etmek için ayrı bir test sınıfı değil çapraz doğrulama metodu kullanıldıysa sonuçlar hmc_ds.xval dosyasına yazılır. İlaç tasarım problemlerine çözüm bulmak amacıyla çalışmada kullanılan hmc_ds.arff dosyası DVD_hmc_ds Ek.4'e ve hmc_ds.s dosyası ise Ek.1'e konulmuştur. Burada hmc_ds.arff eğitim seti (558 ilaç molekülü) 10-kat çapraz doğrulama metoduyla test edilmiştir.

hmc_ds.s dosyasında kullanılan en önemli parametrelerden biri F-testi'dir. Varyansı önemli ölçüde düşürecek bir test istatistiksel olarak F-testle ölçülür. Varyans en aza indiğinde küme homojenliği artar ve modelin tahmin performansı gelişir. FTest için altı olası değerden [0.001, 0.005, 0.01, 0.05, 0.1, 0.125] 0.125 ilaç moleküllerini sınıflandırmada oluşturulan modelin performansını arttırmıştır.

Hiyerarşik kısmında Type olarak geçen parametre Tree olarak belirlendi. Bu sınıf hiyerarşisinin bir ağaç olduğunu belirtir. Yapılan deneylerde w_0 sezgisel karar ağacında farklı sınıfların ağırlıklarını belirler. Çalışmamızda [0.25, 0.75, 1.0] değerlerinden 1.0 test verileri için hiyerarşide üçüncü düzeyde bulunan sinir sistemi ilaçları için tahmin performansını arttırmıştır. Buna ek olarak her bir karar ağacının yaprağındaki örneklerin sayısı için alt sınır [1.0, 2.0, 5.0] değerleri arasından 1.0 olarak belirlendi. k parametresi ise topluluk metodunda kullanılan ağaçların sayısını belirtir. [10, 50, 100] değerleri arasından k değeri için 100 alındı. Sınıflandırma performansı sadece bir parametreye güçlü bir şekilde bağlı değildir. Çalışmada ensemble metot olarak random forest kullanıldı. Şekil (5.2) ilaç moleküllerini hiyerarşik sınıflandırmada kullanılan karar ağacı ile sınıf etiketlerinin hiyerarşik yapısı gösterilmiştir.



Şekil 5.2: Karar ağacı ile sınıf etiketlerinin hiyerarşik yapısı. (A-C) sınıf seti örneklerini göstermektedir, hiyerarşide kalın çizgiyle belirtilmiştir. (D-E) sınıf etiketlerini göstermektedir.

Şekil (5.2)'ye göre hiyerarşik olarak organize edilmiş sınıflar ve etiketleri aşağıdaki gibi verilir;

(0) All Drugs
(1) All Drugs/NSADs
(2) All Drugs/NSADs/N02ADs
(3) All Drugs/NSADs/N03ADs
(4) All Drugs/NSADs/N04ADs
(5) All Drugs/NSADs/N05ADs
(6) All Drugs/NSADs/N06ADs
(7) All Drugs/TheotherADs
(8) All Drugs/WDs

Diğer bir önemli parametrede, Hiyerarşik'te sınıflandırma eşiğidir. Çalışmada eşik değerlerinin bir listesi [0.5, 0.75, 0.80, 0.90, 0.95] olarak belirlendi. Orijinal ağaç her bir yaprak içerisinde tahmin edilen olasılık vektörlerini içerir. Böylesi bir olasılık tahmini bir eşik t uyguluyarak bir etiket setine dönüştürülebilir. olasılık≥ t ile tahmin edilen tüm etiketler tahmin edilen set içerisindedir. Algoritma setteki her bir değer için bir çıktı verir. Ağaçta tahmin edilen etiket seti bu belirli eşik ile inşa edilir. Çıktı dosyamız bu eşik değerleriyle alakalı 5 ağaç içerir.

Hiyerarşik'te algoritma eğer m-estimate parametresini kullanılacak olursa m-estimate'i her yaprağın tahmin vektörüne uygular. Her bir yaprak ve her etiket için T= toplam eğitim örnekleri ve P= pozitif eğitim örneklerinin sayısı belirlenir. M-estimate, P/T'yi tahmin etmek yerine (P+p*T')/(T+T')'ı tahmin eder. Burada extradan (virtual) sanal örnekler görmüş gibi davranır. P pozitif bir değerdir ve T', p parametrelerdir. Burada T'= 1 ve p=tüm eğitim seti içerisindeki pozitif örneklerin oranıdır. Verilen bir etiket için yapraktaki tahminler (P+p)/(T+1) olarak yorumlanır. Biz çalışmamızda P/T tahmin değerini kullandık. 10 kat çapraz doğrulama metoduyla test edilen ilaç moleküllerinin hangi sınıfa (hiyerarşik yapıda organize edilmiş) ait olduğunu içeren tahmin dosyasında (Şekil 4.2) her bir yaprak ve her etiket için P/T değerlerine göre 0 ile 1 arasında bir tahmin yapılır. Bu değerlerden hangisi ilaç molekülü için yüksek ise ilaç o sınıfa aittir denir. Bazı durumlarda bu değer her iki sınıf için çok yakın olabilir o durumda ilacın hangi sınıfa daha yakın olduğu araştırılıp yorumlanabilir. Sonuçlar kısmında daha detaylı bilgi verildi. Bu kısımda anlatılan parametre ayarları 10-kat çapraz doğrulama metoduyla doğrulanan test seti

için verilmiştir. Bağımsız bir test setiyle model değerlendirilirken performansı en iyi elde ettiğimiz parametre değerleri arasında yukarıdakilerden farklı olarak sınıflandırma eşiğidir. Çalışmada eşik değerlerinin bir listesi [10.0, 20.0, 30.0, 40.0, 50.0, 60.0, 70.0, 80.0, 90.0, 100.0] olarak belirlendi. Her iki modelde de performans ölçümleri için FTest optimizasyon stratejisi yani (Pooled AUPRC, area under the average (or pooled) precision-recall curve) Precision-recall eğrisinin altındaki alan kullanıldı. Pooled AUPRC'de en iyi performans bütün parent'ların ağırlıklarının ortalama ağırlığı kullanıldığında elde edilir. Bunun yanında AU($\overline{PRC}$)'da daha sık fonksiyonların ağırlığı daha fazla olur. $\overline{AUPRC}$ her fonksiyonun önemini eşit olarak ele alır.

5.3 Sonuçlar

5.3.1 Çapraz doğrulama metodu kullanılarak test edilen modelin performansı

Çizelge (5.1) geliştirilen modelin HMC_DS için hiyerarşik hata oranını gösterir. Model eğitim verileri ile test edildiğinde elde edilen performans değerleri ortalama AUROC, AUPRC, AUPRC (weighted) ve Pooled AUPRC ile değerlendirildiğinde oldukça iyidir. Çizelge (5.1)'de yer alan Nodes (0), (1), …, (8) Şekil (4.2)'deki sınıf etiketlerini temsil etmektedir. Sonuçlar aşağıda her bir sınıf için (0), (1), …, (8)'e kadar verildi. En iyi performansı Average AUPRC (weighted) gösterdi. Test hatasına baktığımızda ise Average AUPRC her bir fonksiyonun önemini eşit olarak hasaba katar ve hesaplanırken bütün AUPRC skorlarının toplamı alınır ve toplam sınıf sayısına bölünür. Average AUPRC (weighted) ise daha sık olan fonksiyonlara daha fazla ağırlık verir ve performansı Average AUPRC'den daha iyidir. En iyi performansı Pooled AUPRC gösterdi, burada bütün parent'ların ağırlıklarının ortalama ağırlığı kullanıldı. Düğümlere bakıldığında ise (1).ci (7).ci ve (8).ci düğümlerde AUPRC'nin performansı diğer düğümlere göre daha yüksektir. Sinir sistemi ilaçları gözönüne alındığında ise (3.düzey) All Drugs/NSADs/N02ADs ve All Drugs/NSADs/N05ADs sınıflarının tahmin performansı diğer sinir sistemi ilaçlarına göre daha yüksektir. Kullanılan moleküler tanımlayıcılar ilaç moleküllerini hiyerarşik olarak sınıflandırmada oldukça başarılıdır. İlaç tasarım problemleri üzerinde çalışan araştırmacılar için çalışmada kullanılan 558 ilacın (her fold'taki test verilerinin toplamı) hiyerarşik çoklu etiket sınıflaması ile yapılan hiyerarşik organize edilmiş sınıf tahminleri DVD_hmc_ds_test_predictions Ek.5'te verildi (arff dosyası).

Bu dosyada 558 ilacın DRUG_ID numaraları, her bir sınıf etiketi için hesaplanan p değerleri ve ilacın sınıf tahmini yer almaktadır. p değerinden yola çıkıp ilacın onaylanmış ve geri çekilen sınıflarını tahmin etmenin yanında hangi sinir sistemi hastalık grubuna dahil olabileceğinide öngörebiliriz. Bir ilaç hedef hastalık dışında başka bir hastalık üzerinde de iyileştirici etki gösterebilir. p değerine göre yanlış sınıfa yerleştirilen ilaçların o sınıfla olan ilişkisi ayrıca incelenebilir.

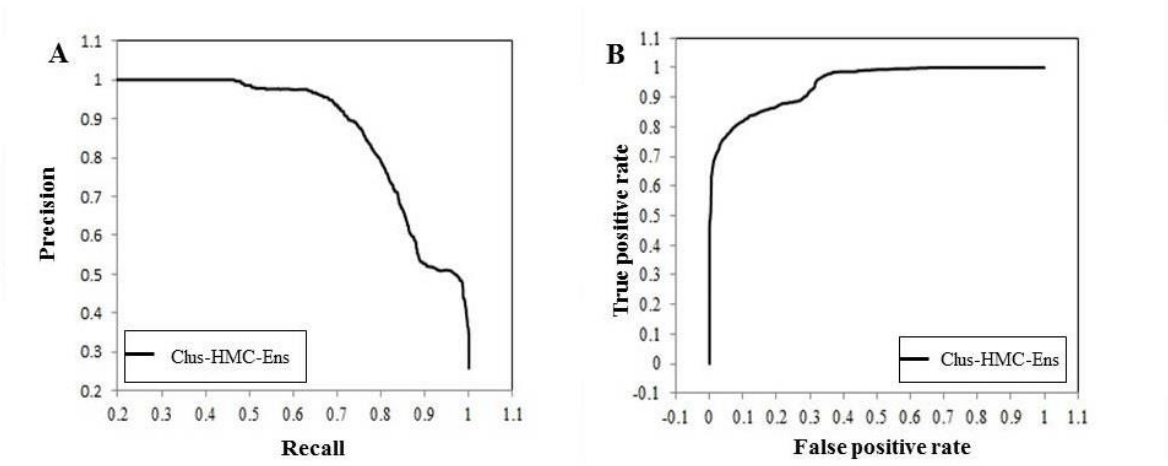
Çizelge 5.1: Geliştirilen modelin HMC_DS üzerindeki hiyerarşik hata ölçümleri.

Hiyerarşik hata ölçümleri: Eğitim hatası Örnek sayısı: 558			
Average AUROC: 0.90 Average AUPRC : 0.81 Average AUPRC : 0.91 (weighted) Pooled AUPRC : 0.89			
Nodes			
(0) : All Drugs,	AUROC: 0.50,	AUPRC:1,	Freq: 1
(1) : All Drugs/NSADs,	AUROC: 0.91,	AUPRC: 0.85,	Freq: 0.32
(2) : All Drugs/NSADs/N02ADs,	AUROC: 0.98,	AUPRC: 0.79,	Freq: 0.07
(3) : All Drugs/NSADs/N03ADs,	AUROC: 0.96,	AUPRC: 0.62,	Freq: 0.06
(4) : All Drugs/NSADs/N04ADs,	AUROC: 0.98,	AUPRC: 0.79,	Freq: 0.04
(5) : All Drugs/NSADs/N05ADs,	AUROC: 0.96,	AUPRC: 0.75,	Freq: 0.07
(6) : All Drugs/NSADs/N06ADs,	AUROC: 0.97,	AUPRC: 0.77,	Freq: 0.08
(7) : All Drugs/The other ADs,	AUROC: 0.95,	AUPRC: 0.88,	Freq: 0.31
(8) : All Drugs/WDs,	AUROC: 0.88,	AUPRC: 0.85,	Freq: 0.36

Hiyerarşik hata ölçümleri: Test hatası Örnek sayısı: 558			
Average AUROC: 0.76 Average AUPRC : 0.47 Average AUPRC : 0.75 (weighted) Pooled AUPRC : 0.85			
Nodes			
(0) : All Drugs,	AUROC: 0.50,	AUPRC:1,	Freq: 1
(1) : All Drugs/NSADs,	AUROC: 0.78,	AUPRC: 0.61,	Freq: 0.32
(2) : All Drugs/NSADs/N02ADs,	AUROC: 0.76,	AUPRC: 0.37,	Freq: 0.07
(3) : All Drugs/NSADs/N03ADs,	AUROC: 0.84,	AUPRC: 0.22,	Freq: 0.06
(4) : All Drugs/NSADs/N04ADs,	AUROC: 0.79,	AUPRC: 0.14,	Freq: 0.04
(5) : All Drugs/NSADs/N05ADs,	AUROC: 0.86,	AUPRC: 0.41,	Freq: 0.07
(6) : All Drugs/NSADs/N06ADs,	AUROC: 0.70,	AUPRC: 0.15,	Freq: 0.08
(7) : All Drugs/The other ADs,	AUROC: 0.88,	AUPRC: 0.75,	Freq: 0.31
(8) : All Drugs/WDs,	AUROC: 0.72,	AUPRC: 0.62,	Freq: 0.36

Average AUROC, average class-wise area under the ROC convex hull; Average AUPRC, average the area under the Precision-Recall Curve; Freq, frequency; NSADs, nervous system approved drugs; N(02,03,04,05,06)ADs, N(02,03,04,05,06)approved drugs; The other ADs, the other approved drugs; WDs, withdrawn drugs.

Burada modelin tahmin edici performans ölçütleri olarak precision-recall eğrisi (PR eğrisi) Şekil (5.3/A)'te verildi. Bölüm (3.7.3)'te ayrıntılı bir şekilde anlatıldı. PR eğrisinin altındaki alan AUPRC olarak adlandırılır. AUPRC 1.0'a ne kadar yakın olursa model o kadar iyi olur. PR eğrileri çoklu etiket sınıflandırma görevindeki her bir sınıf için sınıfa ait örnekleri pozitif olarak ve diğer örnekleri negatif olarak alarak oluşturulur. Bir başka sınıflandırma modelinin başarı indeksi ROC eğrisidir. ROC eğrisinin altında kalan alan AUROC ile belirtilir. Ortalama AUROC sınıfların ROC eğrilerinin altında kalan bütün alanların ortalamasıdır. Ancak ROC eğrisi model performansını tek başına değerlendirmeye uygun değildir. Çok sayıda sınıfın fazla örnek sayısına sahip olmadığı durumlarda (sınıf sık olmadığında), HMC veri setleri için PR tabanlı değerlendirme tipik HMC veri setlerinin özelliklerine daha uygundur. ROC eğrileri bazı durumlarda düşük yanlış pozitif oranı belirlediği için algoritmanın performansı hakkında iyimser bir bakış sunar. Bu olay veri setinde pozitif örneklerin sayısının negatiflerden az olduğu durumlarda pozitif örnekleri ayırt ederken yaşanır. Bizim veri setimizde de özellikle sinir sistemine ait ilaçların All Drugs/NSADs alt sınıflarında (3. düzey) az sayıda pozitif örnek çok sayıda negatif örnek vardır. Çünkü aynı şekilde ROC eğrisi çoklu etiket sınıflandırma görevindeki her bir sınıf için sınıfa ait örnekleri pozitif olarak ve diğer örnekleri negatif olarak alarak oluşturulur. Bu nedenle çalışmada geliştirilen modelin performansını PR tabanlı değerlendirmek daha doğru olur. Aşağıda modele ilişkin her iki eğride verilmiştir. Şekil (4.3/B)'de modele ilişkin ROC eğrisidir. Aynı şekilde bu alan 1.0'a ne kadar yakınsa model o kadar iyi olur.



Şekil 5.3: Çapraz doğrulama metodu kullanılarak test edilen modele ilişkin (A) PR eğrisi ve (B) ROC eğrisi.

5.3.2 Bağımsız bir test seti ile doğrulanan modelin performansı

Çizelge (5.2) 112 ilaç molekülünün (bağımsız test verisi) geliştirilen model üzerinde hiyerarşik bir şekilde sınıflandırma probleminde hata ölçümlerini gösterir. Model eğitim verileri ile test edildiğinde (446 ilaç molekülü) en iyi performansı ortalama AUPRC (weighted) elde etti. Aynı şekilde Çizelge (5.2)'de yer alan Nodes (0), (1), …,(8) hiyerarşide sınıf etiketlerini belirtmektedir. Test hatasına baktığımızda ise Pooled AUPRC en iyi performansı göstermiştir. Düğümler dikkate alındığında (1).ci (7).ci ve (8).ci düğümlerdeki tahmin performansı diğer düğümlere göre daha yüksektir. Geliştirilen modelde (1).ci düğümün tahmin performansına bakıldığında onaylanmış sinir sistemi ilaçlarının diğer hastalık gruplarına ait onaylanmış ve geri çekilen ilaçlardan ayrılabildiğini görüyoruz. Aynı şekilde model (8).ci düğümdeki geri çekilen ilaçları diğer ilaçlardan ayırt edebiliyor. Test seti için en kötü performanslar All Drugs/NSADs/N04ADs ile All Drugs/NSADs/N06ADs sınıfları için elde edildi. Sinir sistemi ilaçları arasında en iyi öngörü ise All Drugs/NSADs/N02ADs sınıfı için yapıldı. Test sınıfında kullanılan 112 ilaç ve onların hiyerarşik yapıdaki sınıf tahminleri konu üzerinde çalışanlar için Ek.2'de verildi. Bu dosyada 112 ilacın DRUG_ID numaraları, her bir sınıf etiketi için hesaplanan p değerleri ve ilacın sınıf tahmini yer almaktadır. Dosyada sınıfı şu an onaylanmış olan ilaçlardan geri çekilen olarak tahmin edilen ilaçlar ayrıca analiz edilebilirler.

Çizelge 5.2: Geliştirilen modelin test ve eğitim verileri üzerindeki hiyerarşik hata ölçümleri.

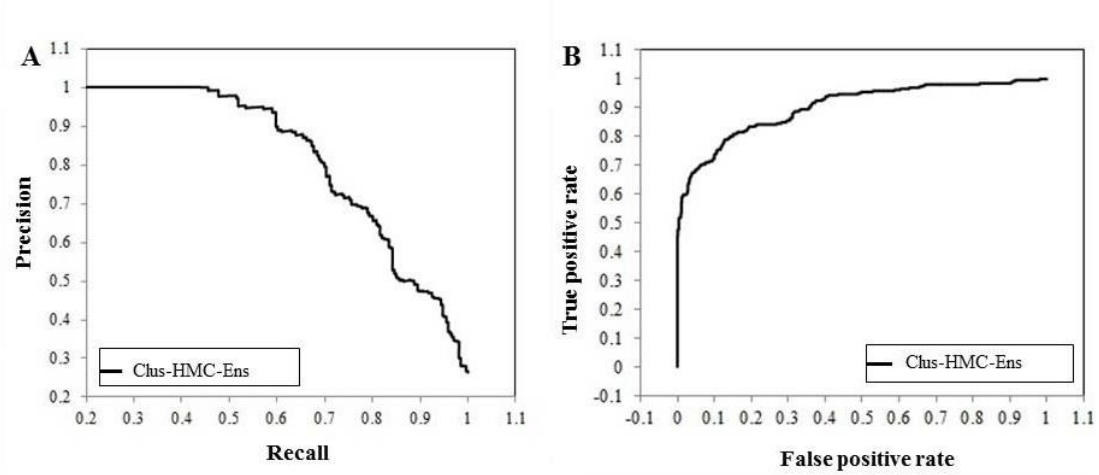
Hiyerarşik hata ölçümleri: Eğitim hatası Örnek sayısı: 446			
Average AUROC: 0.89 Average AUPRC : 0.81 Average AUPRC : 0.90 (weighted) Pooled AUPRC : 0.89			
Nodes			
(0)	: All Drugs,	AUROC: 0.50, AUPRC:1,	Freq:1
(1)	: All Drugs/NSADs,	AUROC: 0.91, AUPRC: 0.82,	Freq: 0.31
(2)	: All Drugs/NSADs/N02ADs,	AUROC: 0.97, AUPRC: 0.72,	Freq: 0.06
(3)	: All Drugs/NSADs/N03ADs,	AUROC: 0.96, AUPRC: 0.63,	Freq: 0.06
(4)	: All Drugs/NSADs/N04ADs,	AUROC: 0.98, AUPRC: 0.80,	Freq: 0.03
(5)	: All Drugs/NSADs/N05ADs,	AUROC: 0.97, AUPRC: 0.79,	Freq: 0.07
(6)	: All Drugs/NSADs/N06ADs,	AUROC: 0.96, AUPRC: 0.78,	Freq: 0.07
(7)	: All Drugs/The other ADs,	AUROC: 0.93, AUPRC: 0.86,	Freq: 0.30
(8)	: All Drugs/WDs,	AUROC: 0.86, AUPRC: 0.83,	Freq: 0.38

Çizelge 5.2: (devam) Geliştirilen modelin test ve eğitim verileri üzerindeki hiyerarşik hata ölçümleri.

Hiyerarşik hata ölçümleri: Test hatası Örnek sayısı: 112			
Average AUROC: 0.77			
Average AUPRC : 0.52			
Average AUPRC : 0.78 (weighted)			
Pooled AUPRC : 0.85			
Nodes			
(0)	: All Drugs,	AUROC: 0.50, AUPRC:1,	Freq:1
(1)	: All Drugs/NSADs,	AUROC: 0.77, AUPRC: 0.69,	Freq: 0.38
(2)	: All Drugs/NSADs/N02ADs,	AUROC: 0.81, AUPRC: 0.48,	Freq: 0.08
(3)	: All Drugs/NSADs/N03ADs,	AUROC: 0.90, AUPRC: 0.27,	Freq: 0.04
(4)	: All Drugs/NSADs/N04ADs,	AUROC: 0.59, AUPRC: 0.09,	Freq: 0.06
(5)	: All Drugs/NSADs/N05ADs,	AUROC: 0.87, AUPRC: 0.39,	Freq: 0.08
(6)	: All Drugs/NSADs/N06ADs,	AUROC: 0.76, AUPRC: 0.23,	Freq: 0.10
(7)	: All Drugs/The other ADs,	AUROC: 0.92, AUPRC: 0.86,	Freq: 0.33
(8)	: All Drugs/WDs,	AUROC: 0.80, AUPRC: 0.66,	Freq: 0.29

Average AUROC, average class-wise area under the ROC convex hull; Average AUPRC, average the area under the Precision-Recall Curve; Freq, frequency; NSADs, nervous system approved drugs; N(02,03,04,05,06)ADs, N(02,03,04,05,06)approved drugs; The other ADs, the other approved drugs; WDs, withdrawn drugs.

Geliştirilen modele ilişkin test verileri üzerinde tahmin edici performans ölçütleri PR eğrisi Şekil (5.4/A) ve ROC eğrisi Şekil (5.4/B) aşağıda verildi. Ortalama AUPRC'nin altında kalan alan test verileri için 0.52, eğitim verileri içinde 0.81 hesaplandı, Çizelge (5.2). Ortalama ROC eğrisinin altında kalan alan ise test verileri için 0.77 ve eğitim verileri için 0.89 hesaplandı, Çizelge (5.2).



Şekil 5.4: Bağımsız bir test seti ile doğrulanan modele ilişkin (A) PR eğrisi ve (B) ROC eğrisi.

Çizelge (5.1-5.2)'de test verileri için (1), (7) ve (8).ci düğümlere baktığımızda tahmin performansı diğer düğümlere göre daha yüksektir. Hiyerarşik yapıda 3.cü düzeyde sinir sistemi ilaçlarının All Drugs/NSADs/N04ADs ve All Drugs/NSADs /N06ADs alt sınıflarında ise tahmin performansı düşüktür. Modellerin gelişmesinde kullanılan moleküler tanımlayıcılar daha üst düzeydeki (1 ve 2.ci düzey) sınıfların öngörüsünde daha etkin bir rol oynadı ancak alt sınıflardaki sinir sistemi ilaçlarında (3.cü düzey) aynı başarıyı gösteremediler. Sonuç olarak aday ilaç molekülleri için geliştirilen modeller onaylanmış sinir sistemi ilaçları, diğer hastalık gruplarına ait onaylanmış ilaçlar ve geri çekilen ilaçları içeren sınıfları birbirinden ayırmada başarılıdır.

Geliştirilen model araştırmacıların aday ilaç moleküllerini test etmeleri için DVD_hmc_ModelDosyası Ek.6'de verilmiştir. Kullanıcılar öncelikle aday ilaç molekülleri için 760 moleküler tanımlayıcıyı CORINA Symphony programı ile hesapladıktan sonra (Bölüm(5.2.2)) Ek.6'de verilen model dosyası ve (arff dosya formatı) CLUS sistemini kullanıp kendi test verilerinin (ilaç adayı moleküller) sınıflarını öngörebilirler. Hazırlanacak test dosyası arff dosya formatında olup eğitim setiyle aynı moleküler tanımlayıcıları içermelidir.

5.4 İlaçların Farklı Hiyerarşik Yapılar Geliştirilerek Çoklu Etiket Sınıflaması

Farklı hastalık gruplarına ait ilaçların hiyerarşik çoklu etiket sınıflaması yapılırken Şekil (5.1)'den farklı hiyerarşik yapılar geliştirilebilir. Önerilen hiyerarşiye bağlı olarak elde edilen modelin performansı değişecektir. Örnek olarak ilaç moleküllerini sınıflandırırken aşağıdaki gibi bir hiyerarşik yapı kullanılırsa Çizelge (5.3), önerilen model üzerinde hiyerarşik bir şekilde sınıflandırma probleminde hata ölçümleri değişir.

Çizelge 5.3: İlaçların farklı bir hiyerarşik yapıda çoklu etiket sınıflaması_1

İlk seviye: (1) All Drugs (558)
İkinci seviye: (1_1) ADs (356), (1_2) WDs (202),
Üçüncü seviye : (1_1_1) NSADs (181), (1_1_2) The other ADs (175),
(1_2_1) NSWDs (32), (1_2_2) The other WDs (170),
Dördüncü seviye: (1_1_1_1) N02ADs (40), (1_1_1_2) N03ADs (33),
(1_1_1_3) N04ADs (22),
(1_1_1_4) N05ADs (41), (1_1_1_5) N06ADs (45),
(1_2_1_1) N02WDs (5), (1_2_1_2) N03WDs (1),
(1_2_1_3) N04WDs (3),
(1_2_1_4) N05WDs (11), (1_2_1_5) N06WDs (12)

Şekil (5.1)'e bakılacak olursak farklı hastalık gruplarına ait onaylanmış ve geri çekilen ilaçlar için önerilen hiyerarşide 3 seviye vardır. Son seviyede N02ADs, N03ADs, …, N06ADs grubuna ait sinir sistemi ilaçları yer almaktadır. Çizelge (5.3)'te ise dört seviye vardır ve son seviyede hem onaylanmış hem geri çekilen sinir sistemi ilaçları yer almaktadır. Önerilen bu modelde hiyerarşide seviye artarken geri çekilen sinir sistemi ilaçları (N02WDs, N03WDs, …, N06WDs) her biri ayrı bir düğüme karşılık gelecek şekilde yapıda yer alır. Şekil (5.1)'deki hiyerarşide sinir sistemine ait geri çekilen ilaçlar WDs grubu içerisinde yer almakta ve modelde sadece ilacın geri çekilenmi olduğu tahmin edilmektedir. Çizelge (5.4)'te geliştirilen modelin (Çizelge (5.3)'de verildi) HMC_DS üzerindeki hiyerarşik hata ölçümleri verilmiştir.

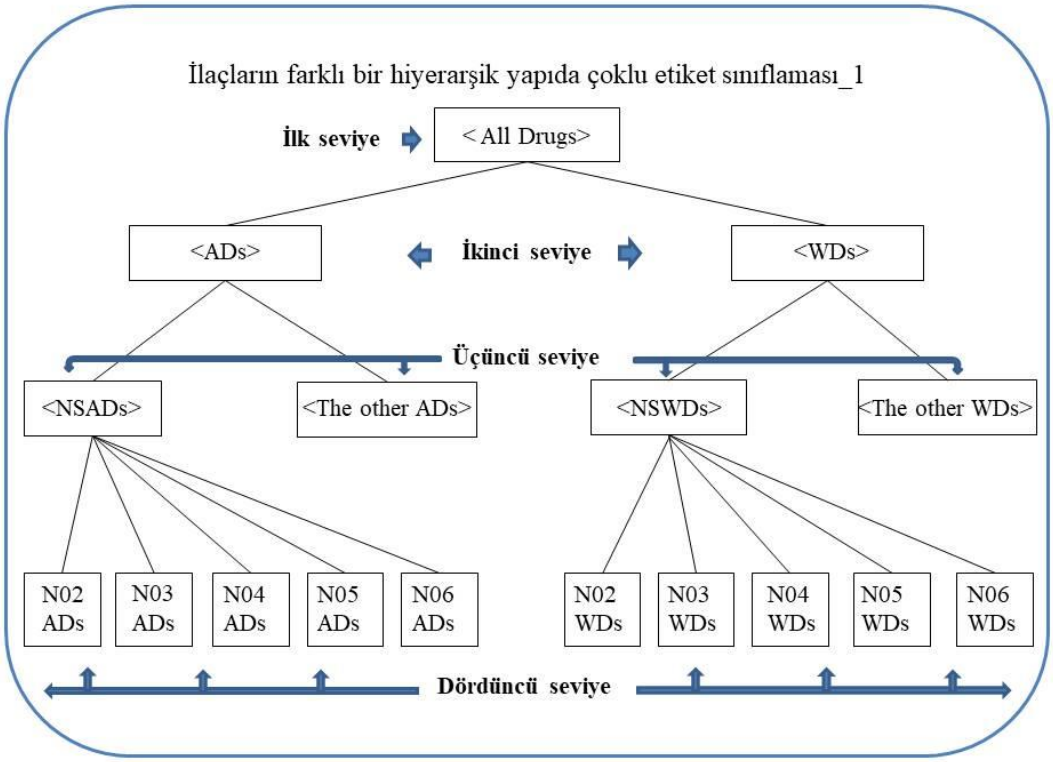
Çizelge 5.4: Geliştirilen modelin HMC_DS üzerindeki hiyerarşik hata ölçümleri.

Hiyerarşik hata ölçümleri: Eğitim hatası Örnek sayısı: 558
Average AUROC: 0.88 Average AUPRC : 0.56 Average AUPRC : 0.82 (weighted) Pooled AUPRC : 0.83
Hiyerarşik hata ölçümleri: Test hatası Örnek sayısı: 558
Average AUROC: 0.62 Average AUPRC : 0.31 Average AUPRC : 0.69 (weighted) Pooled AUPRC : 0.80

Average AUROC, average class-wise area under the ROC convex hull; Average AUPRC, average the area under the Precision-Recall Curve; Freq, frequency.

Çizelge (5.4)'deki Average AUPRC değerlerini Çizelge (5.1)'deki eğitim hatası ve test hatasındaki aynı değer ile karşılaştıracak olursak Şekil (5.1)'e ilişkin modelin HMC_DS üzerinde çok daha başarılı olduğu gözlenir. Average PRC değerinin Çizelge (5.3)'e ilişkin modelde daha düşük olmasının en büyük nedenlerinden biri 4.cü seviyede bulunan onaylanmış ve geri çekilen sinir sistemi gruplarına ait ilaçların sayısının oldukça az olmasıdır. Özellikle geri çekilen ilaçların sayısının çok az oluşu düğümlerde All Drugs/WDs/NSWDs/N02WDs, …, N06WDs (toplam 5 düğümde) hesaplanan Average AUPRC değerini oldukça düşürmektedir. Bu nedenle bu geri çekilen sinir sistemi ilaçlarının Şekil (5.1)'deki gibi WDs grubuna dahil edilmesi

modelin performansını arttırır. Şekil (5.5)'te Çizelge (5.3)'e ait ilaçların farklı bir hiyerarşik yapıda çoklu etiket sınıflaması_1 yer almaktadır.



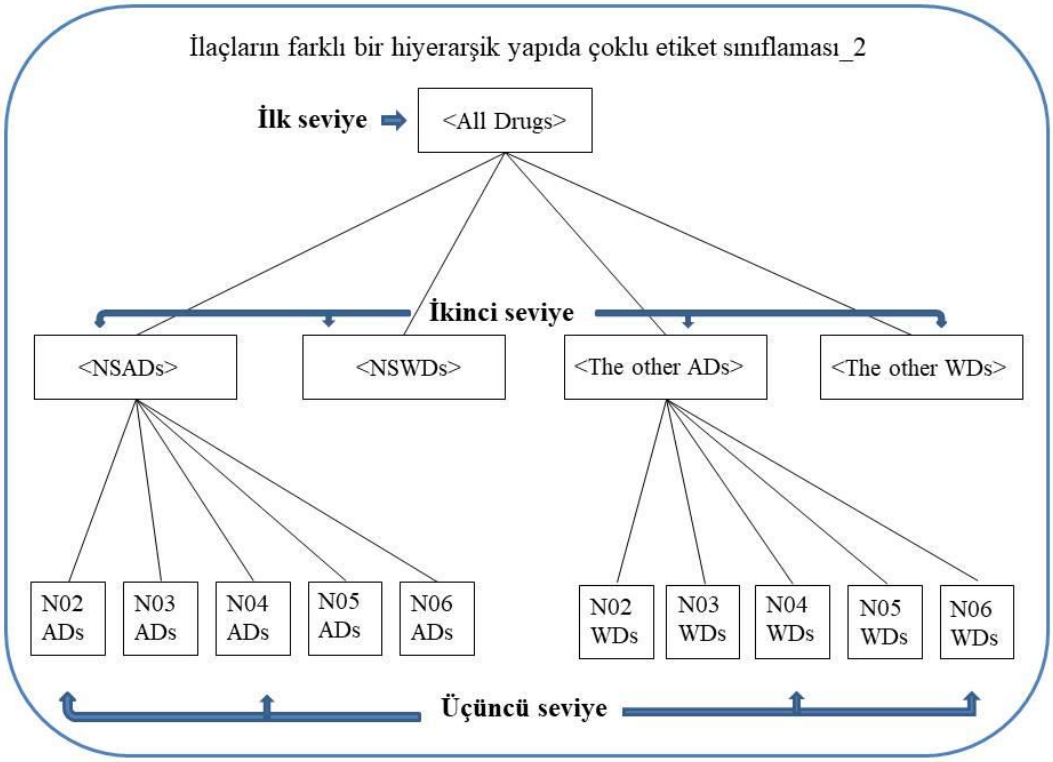
Şekil 5.5: İlaçların farklı hiyerarşik yapıda çoklu etiket sınıflaması_1.

Çizelge (5.5)'de HMC_DS'nin daha farklı bir hiyerarşik yapıda çoklu etiket sınıflaması yer almaktadır. Burada da yine Şekil (5.1)'deki gibi geliştirilen modelde 3 seviye vardır ancak ikinci seviyede sinir sistemine ait geri çekilen ilaçlar WDs'den ayrı bir düğümde yer almaktadır. Buna ek olarak 3.cü seviyede sinir sistemine ait geri çekilen ilaç grupları ayrı düğümlerde yer almaktadır.

Çizelge 5.5: İlaçların farklı bir hiyerarşik yapıda çoklu etiket sınıflaması_2

İlk seviye: (1) All Drugs (558)
İkinci seviye: (1_1) NSADs (181), (1_2) NSWDs (32),
(1_3) The other ADs (175), (1_4) The other WDs (170),
Üçüncü seviye : (1_1_1) N02ADs (40), (1_1_2) N03ADs (33),
(1_1_3) N04ADs (22), (1_1_4) N05ADs (41),
(1_1_5) N06ADs (45),
(1_2_1) N02WDs (5), (1_2_2) N03WDs (1),
(1_2_3) N04WDs (3), (1_2_4) N05WDs (11),
(1_2_5) N06WDs (12).

Şekil (5.6)'da Çizelge (5.5)'e ait ilaçların farklı bir hiyerarşik yapıda çoklu etiket sınıflaması_2 yer almaktadır.



Şekil 5.6: İlaçların farklı hiyerarşik yapıda çoklu etiket sınıflaması_2.

Çizelge (5.6)'da geliştirilen modelin (Çizelge (5.5)'de verildi) HMC_DS üzerindeki hiyerarşik hata ölçümleri verilmiştir. Geliştirilen modelde All Drugs/NSWDs /N02WDs, N03WDs, …, N06WDs düğümlerindeki (toplam 5 düğümde) hesaplanan Average AUPRC değerleri oldukça düşüktür. Bu nedenle geliştirilen modelde NSWDs düğümüne ait ilaçları WDs ilaçları içerisinde vermek Average AUPRC değerini arttırır. Çizelge (5.3) ve Çizelge (5.5)'de geliştirilen modeller veri seti HMC_DS üzerinde 10-kat çapraz doğrulama metoduyla test edildi. Sonuç olarak onaylanmış ve geri çekilen ilaçlar üzerinde hiyerarşik çoklu etiket sınıflaması gerçekleştirmek amacıyla farklı hiyerarşik yapılara sahip üç model geliştirdik. Bunlardan Şekil (5.1)'e ait olan model diğer modellere göre sınıflamada hiyerarşik hata ölçümleri gözönüne alındığında daha başarılı olduğu gözlenmiştir. Bunun en büyük nedenlerinden biri modelin 2.ci seviyede sinir sistemine ait geri çekilen ilaçların WDs'nin içerisinde yer almasıdır.

Çizelge 5.6: Geliştirilen modelin HMC_DS üzerindeki hiyerarşik hata ölçümleri.

Hiyerarşik hata ölçümleri: Eğitim hatası Örnek sayısı: 558
Average AUROC: 0.89 Average AUPRC : 0.56 Average AUPRC : 0.81 (weighted) Pooled AUPRC : 0.84
Hiyerarşik hata ölçümleri: Test hatası Örnek sayısı: 558
Average AUROC: 0.61 Average AUPRC : 0.27 Average AUPRC : 0.68 (weighted) Pooled AUPRC : 0.81

Average AUROC, average class-wise area under the ROC convex hull; Average AUPRC, average the area under the Precision-Recall Curve; Freq, frequency.

Önerilen bu model onaylanmış sinir sistemi ilaçlarını diğer onaylanmış ilaçlardan ayırırken aynı zamanda geri çekilen ilaçlarıda belirliyebilmektedir. Çalışmada ilaçları sınıflandırmak amacıyla daha bunlara benzer farklı hiyerarşide sınıflama modelleri geliştirilebilir. Bunların performansı düğümlerdeki örnek sayısı ve belirlenen seviyelere göre değişecektir. En önemlisi düğümlerdeki örnek sayısının her düğüm için yeterli sayıda olması ve düğümlerde dengesiz veri setlerinin olmamasıdır.

6. DENGESİZ İLAÇ SAYISI İÇİN BİR SINIFLANDIRMA YAKLAŞIMI

6.1 Giriş

Çalışmanın bu kısmında 4. Bölümden farklı olarak yalnızca spesifik bir hastalığa ait ilaçlar değil ilaç veri bankasında çok sayıda hastalığın tedavisinde kullanılan 1200'den fazla onaylanmış ve geri çekilen ilaç üzerinde çalışıldı. Bölüm (3.4)'te bu hastalıkların hangileri olduğuna geniş yer verildi. Burada, kullanılan moleküler tanımlayıcıların çeşitli hastalık gruplardan gelen ilaçları onaylanmış ve geri çekilen durumlarını tahmin etmede etkin olma durumları incelendi. Çalışmada ele aldığımız ilaç veri seti geri çekilen ilaçların sayısının onaylanmış ilaçların sayısına göre çok daha az olması nedeniyle ilaç veri kümesi oldukça dengesizdir. Dengesiz veri kümelerinin sınıflandırılması ve özniteliklerin seçilmesi makine öğrenme zorluklarından ikisidir. Sınıflandırmada etkin rol oynayan moleküler tanımlayıcılar tezde önerilen etkin öznitelik seçme stratejisi ile belirlendi. Amacımız dengesiz veri setleri için depolama gereksinimlerini sınırlamak ve algoritma hızını arttırmak için özellik alanının boyutsallığını azaltmaktır. Böylelikle gereksiz alakasız gürültülü verileri veri setimizden kaldırdık. Geliştirdiğimiz etkin öznitelik seçme stratejisi ile ortaya çıkan modelin doğruluğunu arttırdık. Amaç sınıflandırmada daha etkin bir rol oynayan moleküler tanımlayıcıları ilaç tasarım problemleri için belirlerken aynı zamanda ilaç aday moleküllerini onaylanmış ve geri çekilen olarak kategorize etmektir. Buradan yola çıkarak çalışmamızda dengesiz veri setleri için sınıflandırma problemlerine çözüm getirebilecek içinde etkin öznitelik seçme stratejisininde yer aldığı bir yaklaşım önerildi. Çalışmada deneysel tasarımın gerçekleştirilmesi için MATLAB yazılım paketi (MATLAB & SIMULINK, R2015a) ve Weka veri madenciliği uygulaması (weka.version 3.7.13, package manager) kullanıldı.

6.2 Materyaller Ve Yöntemler

1200'den fazla ilaç başta DRUGBANK olmak üzere KEGG ve PubChem veri tabanlarından toplandı. Çalışmada önerilen yaklaşım üç aşamada gerçekleştirilmektedir. Başlangıçta dengesiz ilaç veri seti için etkin öznitelikler

belirlenir bunun için tezde geliştirilen etkin öznitelik seçme stratejisi kullanıldı. Veri setinin dengeli hale getirilmesi amacıyla SMOTE (Synthetic Minority Over Sampling Technique) algoritması veri setine uygulandı. Sınıflandırma problemleri için Meta-sınıflandırıcı olarak Bagging algoritması ile temel sınıflandırıcı olarak SVM+RBF Kernel ilaç veri setine uygulandı. Sınıflandırma modellerini oluşturmak için CORINA Symphony programı kullanılarak tüm veri setleri için bir dizi moleküler tanımlayıcı hesaplandı. Bunlar 22'si global moleküler, 8'i boyut ve şekil, 729'u toxprint kemotip tanımlayıcılarından ve 1 kullanıcı özelliğini içermek üzere 760 tane özellikten oluşur (DVD_Çizelge_Ek.2).

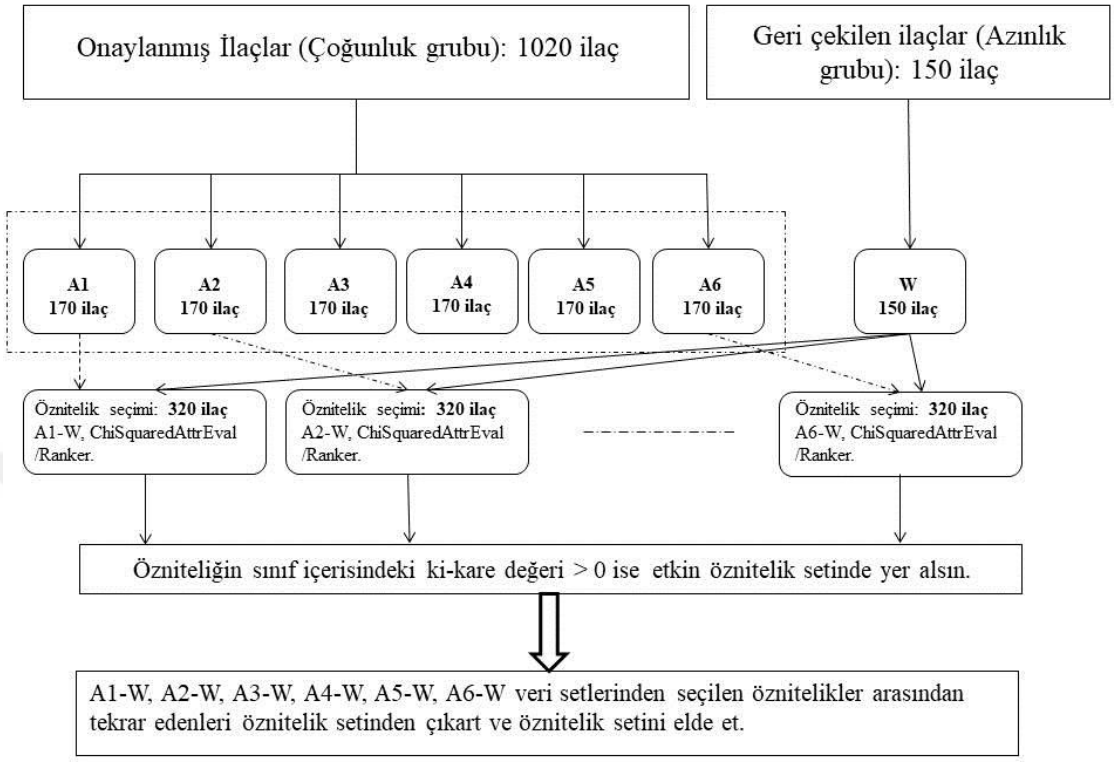
6.2.1 Veri kümelerinin toplanması

Sınıflandırma problemlerinde kullanmak üzere SDF formatında 1200'den fazla ilaç ilaç veri bankalarından toplandı. Toplamda 1050 onaylanmış ve 170 geri çekilen ilaç 1170'i eğitim setini ve 50'si bağımsız test setini oluşturmak için kullanıldığında, eğitim setinde 1020 onaylanmış ve 150 geri çekilen ilaç yer almaktadır. Eğitim setinde geri çekilen ilaçların sayısının onaylanmış olanlardan oldukça az olması nedeniyle ve dengesiz veriler üzerinde eğitilmiş model performansını arttırmak amacıyla geliştirilen yaklaşımın aşamalarından birinde veri setini dengelemek amacıyla SMOTE algoritması kullanıldı. Çalışmada geliştirilen stratejide etkin öznitelikler belirlenirken onaylanmış ilaçların tümü geri çekilen ilaç molekülleri sayısıyla dengeli olacak şekilde birbirinden bağımsız altı veri setine bölündü. Her bir veri seti 170 onaylanmış ve 150 geri çekilen ilaç molekülü olmak üzere 320 ilaç içermektedir. Geri çekilen ilaç molekülleri her veri setinde aynı ancak onaylanmış ilaç molekülleri her bir veri setinde birbirinden tamamen farklıdır. Bir onaylamış ilaç molekülü birden fazla veri setinde bulunmaz.

6.2.2 Veri ön işleme ve özellik seçimi

Ham ilaç veri setlerinden yararlı veriler elde etmek amacıyla CORINA programı tarafından önceden tanımlanan basamaklar SDF formatındaki ilaç moleküllerine bu çalışmada da uygulanmıştır. Ayrıntılara önceki bölümde yer verildi, Bölüm(4.2.3). Sınıflandırma çalışmalarında kullanılmak üzere hesaplanan moleküler tanımlayıcıların hepsi geri çekilen ve onaylanmış ilaçları ayırt edici nitelikte değildir. Burada ilaç molekülleri için etkin tanımlayıcıları belirlemek amacıyla etkin öznitelik

seçme stratejisi geliştirilmiştir. Şekil 6.1'de ilaç veri seti için sınıflandırmada etkin olan öznitelik setinin (FAW) elde edilmesi aşamaları gösterilmektedir.

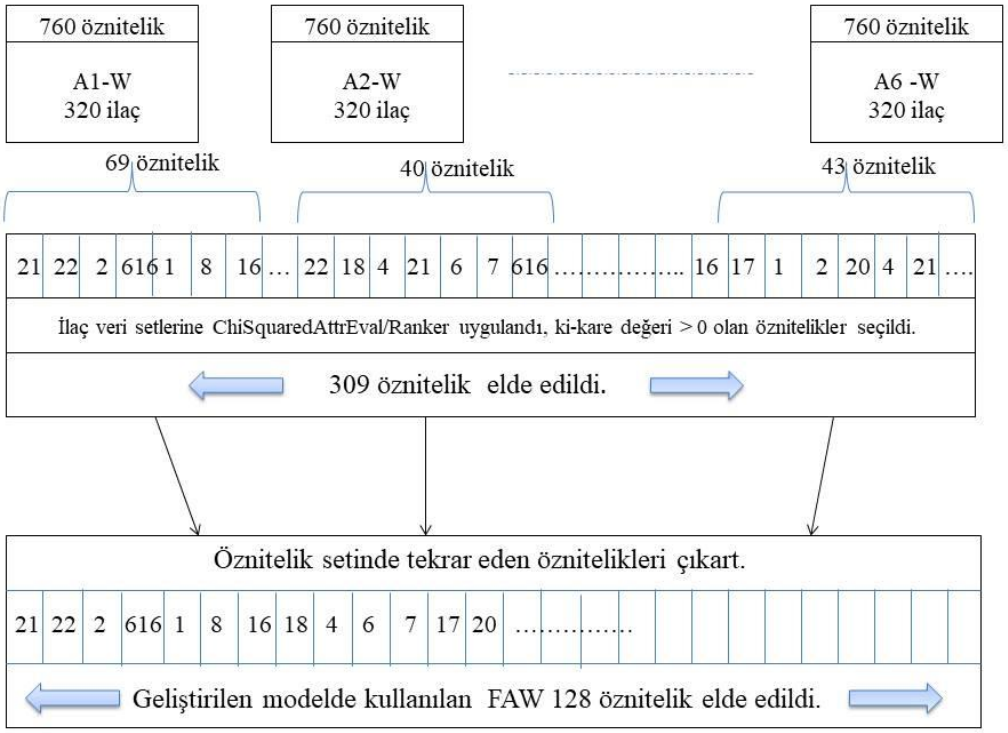


Şekil 6.1: Sınıflandırmada etkin olan öznitelik setinin (FAW) elde edilmesi aşamaları.

A1, A2…A6, onaylanmış ilaçlardan A'dan oluşan 6 grup. Bu grupların her biri (170 onaylanmış ilaç) geri çekilen (150 W) ilaç grubu ile birleşip A1-W, A2-W…A6-W ilaç veri setlerini oluşturmaktadır. Her bir set için etkin öznitelikler belirlenip bunlar son aşamada tek bir etkin öznitelik seti oluşturmak için birleştirilmektedir. İlaç veri setimiz başlangıçta toplam 1220 ilaç içermektedir. Bunlardan 1050'si onaylanmış, 170 tanesi ise geri çekilen ilaçlardan oluşmaktadır. Deneysel çalışmalara geçmeden önce bunlardan 50 tanesi dengesiz ilaç veri setinde sınıflandırıcı topluluk tasarımı için geliştirilen modelinin performansını değerlendirmek amacıyla bağımsız test seti olarak ayrıldı. Geriye kalan 1170 ilaç eğitim seti olarak kullanıldı. Eğitim setinde 1020 onaylanmış ve 150 geri çekilen ilaç yer almaktadır. Onaylanmış ilaçlar (1020 ilaç) A1'den başlayarak A6'ya kadar toplam 6 gruba ayrılmıştır. Her grupta 170 onaylanmış ilaç bulunmaktadır. Bu grupların her biri 150 geri çekilen ilaç içeren grup ile birleştirilir ve dengeli verilerden oluşan toplam 6 veri seti elde edilir. Her bir set toplamda 320 onaylanmış ve geri çekilen ilaçlardan oluşmaktadır. Burada

dengesiz ilaç veri seti için etkin öznitelikleri belirlemeden önce veri ön işleme yöntemlerinden olan alt örnekleme yöntemleri (undersampling), yüksek örnekleme (oversampling) ve hibrit yöntemler (her iki örnekleme yönteminin birleşiminden oluşan (hybrids methods) kullanılmamıştır. Sınıf dağılımını dengelemek amacıyla kullanılan bu metotlardan alt örnekleme yöntemi potansiyel olarak sınıflandırmada etkin olabilecek verileri atabilir, yüksek örnekleme yöntemi ise azınlık sınıf örneklerini rastgele çoğaltırken var olan örneklerin tam kopyalarını üretebilir. Bu nedenle geliştirilen öznitelik seçme metodunda sınıflandırmada etkin öznitelikleri belirlerken veri ön işleme yöntemleri kullanmanın yerine öncelikle çoğunluk grubu 6 parçaya ayrılıp her parça azınlık grubuyla birleştirildi. Bir sonraki adımda ise veri setlerine ki-kare öznitelik seçme yöntemi uygulandı. Burada özniteliğin sınıf içerisindeki ki-kare değeri > 0 ise öznitelik etkin öznitelik setinde yer alır (Şekil 6.1). A1-W, A2-W…A6-W ilaç veri setlerinden gelen etkin öznitelikler arasından tekrar eden öznitelikler öznitelik setinden çıkartılarak FAW (128) elde edildi. Şekil 6.2'de ilaç aday moleküllerinin onaylanmış/geri çekilen durumlarının karar verilmesi için geliştirilen modelde kullanılan FAW öznitelik seçimi stratejisi aşamaları ayrıntılı olarak verilmiştir. A1-W, A2-W…A6-W ilaç veri setleri başlangıçta 760 özniteliğe sahiptir ve bir öznitelik örneğin Atoms her veri setinde aynı index numarasıyla belirtilmiştir ve 760 öznitelik hepsi için birebir aynıdır. Veri setlerinden bu yöntemle toplam 309 öznitelik elde edildi. Bunlardan gruplardan gelen tekrar eden öznitelikler etkin öznitelik setinden çıkartılmıştır. Son durumda geliştirilen modelde kullanılmak üzere etkin öznitelik setinde FAW 128 öznitelik bulunmaktadır (Şekil 6.2). Şekilde 21, 22, 2, 616 ile devam eden sayılar sırasıyla özniteliklerin index numaralarıdır. Örnek olarak A1-W veri setine ki-kare öznitelik seçme metodu uygulandığında ki-kare değeri > 0 olan özniteliklerin sayısı 69'dur. Özniteliğin sınıf içerisindeki ki-kare değeri > 0 ise öznitelik FAW'da yer aldı.

Aşağıda ilaç veri setlerine Ki-kare öznitelik seçme metodu uygulanarak elde edilen sınıflandırmada etkin özniteliklerin sayıları, tüm setlerden gelen özniteliklerin toplam sayısı ve tekrar eden öznitelikler çıkarıldığında elde edilen etkin öznitelik (FAW) sayısı Çizelge 6.1'de ayrıntılı olarak verilmiştir.



Şekil 6.2: İlaç aday moleküllerinin onaylanmış/geri çekilen durumlarının karar verilmesi için geliştirilen modelde kullanılan FAW öznitelik seçimi stratejisi aşamaları.

Çizelge 6.1: A1-W, A2-W…A6-W ilaç veri setlerine Ki-kare öznitelik seçme metodu uygulanarak elde edilen sınıflandırmada etkin özniteliklerin sayıları, tüm setlerden gelen özniteliklerin toplam sayısı ve tekrar eden öznitelikler çıkarıldığında elde edilen etkin öznitelik (FAW) sayısı.

Veri Setleri	Veri setlerinden seçilen Etkin öznitelik sayısı	Elde edilen toplam Öznitelik sayısı	Kalan Öznitelik Sayısı
A1-W	69		
A2-W	40		
A3-W	36	309	128
A4-W	55		
A5-W	66		
A6-W	43		

Çizelge 6.1'e bakıldığında veri setlerinden gelen öznitelik sayısı toplamı başlangıçta 309'dur. Bunlardan aynı index numaralı olanlardan sette yalnızca 1 tane bırakıldığında 128 öznitelik kalmıştır. Yani çok sayıda öznitelik birden fazla veri setinde etkin öznitelik setinde yer almıştır. Bunlardan A1-W, A2-W…A6-W ilaç veri setlerinden en az üç veri setinde etkin öznitelik olarak seçilen 45 özniteliğin ayrıntılı analizi ayrıca sonuçlar kısmında verilmiştir. Çizelge 6.2'de öznitelik sıra no ile

belirtilen kolon özniteliğin aynı zamanda index numarasıdır. Çizelge veri setlerinden gelen etkin özniteliklerin seçilme stratejileri ile ilgili bilgi vermektedir. Herbir veri setinde 760 özniteliğin hepsi için ki-kare değeri hesaplanmış ve örnek olması amacıyla veri setinin adı, seçilen özniteliğin index numarası ve sınıf içerisindeki ki-kare değeri verilmiştir. Başlangıçta her veri seti için toplam 760 öznitelik kullanılmıştır. Her birinin öznitelik setinde bir sıra numarası yer almaktadır. Örneğin A1-W'de 21 numaralı öznitelik LogS' ye karşılık gelmektedir. Her veri setinde 21 numara aynı özniteliğe karşılık gelmektedir. LogS'nin sınıf içerisindeki ki-kare değeri 53.03'tür.

Çizelge 6.2: A1-W, A2-W…A6-W ilaç veri setlerine ki-kare öznitelik seçme metodu uygulandığında özniteliklerin sınıf içerisindeki ki-kare değerleri > 0 ise öznitelik etkin öznitelik setinde yer alır.

Öznitelik Sıra No	Öznitelik Adı		Veri Seti Adı	Öznitelik Sıra No	Ki-kare Değeri
1	Atoms		A1-W	21	53.03
2	Bonds			22	52.13
3	BondsRot			2	32.87
4	HAcc			…	…
5	HAccN		A2-W	22	30.77
6	HAccO			18	28.55
7	HDon			4	20.7
8	HDonN			…	…
9	HDonO		A3-W	27	36.34
10	Ro5Viol			29	34.29
…	…			24	32.97
…	…			…	…
…	…		A4-W	2	38.14
				16	37.38
				22	33.5
				…	…
			A5-W	27	66.67
				28	57.06
				30	53.77
				…	…
			A6-W	16	37.86
				17	37.5
				1	34.52
				…	…

Bunun yanında A1-W, A2-W…A6-W ilaç veri setlerine uygulanan değiştirilmiş ki-kare öznitelik seçme algoritması Çizelge 6.3'de verilmiştir.

Çizelge 6.3: Değiştirilmiş ki-kare algoritması ve etkin özniteliklerin belirlenmesi.

Değiştirilmiş Ki-kare algoritması

```
/* ki-dizisi: iki boyutlu dizi. İlk kolon veri seti (VS) içerisindeki özniteliklerin öznitelik
indeksini, ikinci kolon öznitelikler ve sınıf etiketleri (SE) için ki- kare değerini içermektedir
*/
ki-dizisi ← ∅
for i ← 1  to n do                          // Veri seti içerisindeki toplam öznitelik sayısı n'dir.
        ki- değeri ← ki-kare (VS [i], SE)   // Veri seti içerisindeki öznitelikler ve sınıf etiketleri
                                               arasındaki  ki-kare değerini hesaplar.
    if ki-değeri > 0 ise
    append (i, ki- değeri) to ki-dizisi
end for
sort ki-dizisi by ikinci kolon (ki-karedeğeri) azalan sırada
store ki-dizisindeki ilk kolon değerini to seçilen etkin öznitelikler
return seçilen etkin öznitelikler
```

Çalışmada sınıflandırıcı topluluk tasarımı için geliştirilen modeli eğitmek amacıyla kullanılan moleküler tanımlayıcılar sınıflandırma performansını arttırmada oldukça önemlidir. Yapılan deneylerde A1-W, A2-W…A6-W ilaç veri setlerine değiştirilmiş ki-kare (Çizelge 6.3) yerine başka öznitelik seçme algoritmalarıda uygulanmış ancak kullanılan yöntemin performansına bakarken sınıflandırma doğruluğu ve seçilen toplam özniteliklerin sayısı dikkate alındığında elde edilen modelin sınıflandırma doğruluğu uygulanan strateji doğrultusunda (etkin öznitelik seçme stratejisi) yüksek ve seçilen öznitelik sayısının daha az olduğu gözlemlenmiştir. Örnek olarak farklı hastalık gruplarına ait ilaçlar çalışılırken Cfs Subset Eval ve arama metodu olarak Bestfirst metodu ile seçilen moleküler tanımlayıcılarla elde edilen modelde seçilen öznitelik sayısı az, sınıflandırma doğruluğu daha düşüktür.

6.2.3 Sınıflandırıcı topluluk tasarımı için geliştirilen model

Önerilen model üç aşamada gerçekleştirilmektedir,

1. Dengesiz veri seti için etkin öznitelikler belirlenir.
2. Dengesiz veriler tekrar örneklenir. Veri seti dengeli hale getirildikten sonra veri setinde sadece örneklere ait 1 nolu aşamada belirlenen etkin özniteliklerle ilgili veriler yer alır.
3. Sınıflandırıcı topluluk oluşumu elde edilir.

Önerilen modelin aşamaları:

1. Dengesiz veri setleri için etkin özniteliklerin belirlenmesi,
   Girdi:

p, n ve 1 sırasıyla çoğunluk grubu, azınlık grubu ve öznitelik matrislerinde satır sayısını bir başka değişle veri setindeki örnek sayısını; d, tüm matrislerdeki sütun sayısını yani veri setindeki özniteliklerin toplam sayısını göstermektedir. Aşağıda verilen eşitliklerde i yine çoğunluk ve azınlık grubu matrislerindeki satır sayısını; j, çoğunluk ve azınlık grubu matrislerinde sütun sayısını temsil etmektedir. k ve l ise sırasıyla öznitelik matrisine ait satır ve sütun sayılarını belirtmektedir. Aşağıda ayrıca i, j, k, l'nin p, n, ve d cinsinden aldıkları değerler parantez içinde belirtilmiştir.

Çoğunluk grubu matrisi $G_ç$: $X_{pxd} = X_{i,j}$ (i = 1, 2, .., p ; j = 1, 2, .., d ve $X_{i,j} \in R$)

Azınlık grubu matrisi G_a: $X_{nxd} = X_{i,j}$ (i = 1, 2, .., n ; j = 1, 2, .., d ve $X_{i,j} \in R$)

Öznitelik matrisi O_f: $Y_{1xd} = Y_{k,l}$ (k= 1 ; l = 1, 2, .., d ve $Y_{k,l} \in N$) olmak üzere,

$G_ç = [X_{ij}]_{pxd}$, $G_a = [X_{ij}]_{nxd}$, $O_f = [Y_{1l}]_{1xd}$'dir.

Öznitelik matrisi O_f özniteliklerin sırasıyla indeks numaralarını içermektedir.

Buna göre;

- İlk olarak k = $\#G_ç$ / $\#G_a$ olacak şekilde belirlenir. k, çoğunluk grubu matrisindeki toplam örnek sayısının azınlık grubu matrisindeki toplam örnek sayısına bölünmesi sonucu elde edilir."k" belirlenirken bölüm sonucu alt sınır seçilecek şekilde tamsayıya yuvarlanır. Örnek olarak bölüm sonucu 6.8 çıkmış ise k = 6 tamsayısı alınır.
- Eğer k değeri k ≥ 2 ise;

  Çoğunluk grubu $G_ç$ matrisi k eşit parçaya bölünür ve k eşit parçanın her biri azınlık grubunu olan G_a matrisi ile birleştirilir. Sonuç olarak k adet veri matrisimiz olur. Elde edilen k adet matris etkin özniteliklerin belirlenmesi amacıyla öznitelik matrisi ile birleştirilir. Bu matrislerin herbiri aynı sayıda özniteliğe sahiptir (d adet). p ve n değerleri sırasıyla $G_ç$ ve G_a içerisindeki örneklerin sayısıdır.

  Eğer k değeri k = 1 ise;

  Eğitim setine direk olarak etkin öznitelik seçme stratejisi uygulanır.
- Özniteliklerideiçeren k adet matrise bir başka değişle veri setlerine etkin öznitelik seçme stratejisi (ki-kare) uygulanarak her biri için etkin öznitelikler belirlenir.
- Son olarak k adet matristen (veri setinden) gelen etkin öznitelikler birleştirilerek etkin öznitelik seti elde edilir. Elde edilen etkin öznitelik setinde tekrar eden özniteliklerden yalnızca birtanesi kullanılır.

2. Başlangıçta $G_ç$ ve G_a grubu matrisleri "d" adet öznitelik içermekteydi. 1 nolu aşamadan sonra belirlenen etkin öznitelik setine göre $e \leq d'dir$. Burada "e" seçilen etkin özniteliklerin sayısıdır. Dengesiz veri seti başta O_f, $G_ç$ ve G_a matrislerinin birleşiminden oluşmaktadır. Burada ç $> a$'dır. ç çoğunluk grubu matrisindeki toplam örnek sayısı ve a azınlık grubu matrisindeki toplam örnek sayısıdır. Dengesiz veri seti tekrar örneklenip dengeli hale getirildikten sonra veri setinde sadece örneklere ait 1 nolu aşamada belirlenen etkin özniteliklerle ilgili veriler yer alır.

Dengesiz verilerin tekrar örneklenmesi ve etkin özniteliklerin veri setinden seçilmesi,

Girdi:
Çoğunluk grubu matrisi $G_ç$: $X_{pxd} = X_{i,j}$ (i = 1, 2, .., p ; j = 1, 2, .., d ve $X_{i,j} \in R$)
Azınlık grubu matrisi G_a: $X_{nxd} = X_{i,j}$ (i = 1, 2, .., n ; j = 1, 2, .., d ve $X_{i,j} \in R$)
Öznitelik matrisi O_d: $Y_{1xd} = Y_{k,l}$ (k= 1 ; l = 1, 2, .., d ve $Y_{k,l} \in N$) olmak üzere,
$G_ç = [X_{ij}]_{pxd}$, $G_a = [X_{ij}]_{nxd}$, $O_d = [Y_{1l}]_{1xd}$'dir. Buradan girdi matrisi (veri seti),
$OG_{dça} = [X_{ij}]_{(1+p+n)xd}$ şeklinde gösterilebilir. Burada (p+ n) girdi matrisindeki satır sayısını yani çoğunluk ve azınlık grubu matrislerindeki toplam örnek sayısını, 1 ise girdi matrisinin ilk satırı olan öznitelik matrisine ait satır sayısını gösterir ve d matrislerdeki toplam öznitelik sayısını göstermektedir. Etkin öznitelik matrisi O_d özniteliklerin sırasıyla indeks numaralarını içermektedir.

Buna göre;

- İlk olarak azınlık sınıfının örnek sayısını arttırmak ve veri kümesinin dengesizlik oranını düşürmek amacıyla SMOTE (Over-sampling using Synthetic Minority Over-sampling Technique) tekniği eğitim setine uygulanmaştır.

Bu durumda veri seti matrisi,

$OG_{dça} = [X_{ij}]_{(1+p+n)xd}$ iken $\frac{p}{n_{son}} < 2$ olacak şekilde n, azınlık gurubu örnek sayısı veri kümesini dengelemek amacıyla arttırılır. Son durumda $n_{son} > n$ 'dir. Buna göre veri seti matrisi ilk satır etkin öznitelik indeks numaralarını içerecek şekilde son durumda,

$OG_{dça} = [X_{ij}]_{(1+p+n_{son})xd}$ şeklinde gösterilir.

- Sonuç olarak etkin öznitelikleri içeren ve dengeli veri setine sahip $OG_e = [X_{ij}]_{(1+p+n_{son})xe} = X_{i,j}$ (i = 1, 2, .., $(1 + p + n_{son})$; j = 1, 2, .., e ve $X_{i,j} \in R$) veri matrisi elde edilir. Burada e, 1 nolu aşamada belirlenen etkin özniteliklerin sayısını belirtir. Son durumda çoğunluk grubu matrisindeki toplam örnek sayısının azınlık grubu matrisindeki toplam örnek sayısına bölümü 2'den küçük olmalıdır ($\frac{p}{n_{son}} < 2$).

3. Sınıflandırıcı topluluk oluşumu,
   Girdi:

   Etkin özniteliklerden oluşan dengeli veri setine sahip eğitim seti matrisi
   OG_e: $X_{(1+p+n_{son})xe} = X_{i,j}$ (i = 1, 2, .., $(1 + p + n_{son})$; j = 1, 2, .., e ve $X_{i,j} \in R$)
   Buna göre;

   - Meta-sınıflandırıcı (Bagging algoritması temel sınıflandırıcı olarak SVM+RBF Kernel) OG_e eğitim setine uygulanır. Geliştirilen bu model ile ikili sınıflandırma problemlerinde eğitim seti 10-kat çapraz doğrulama metodu ile test edilebilirken, bağımsız test setleri için de sınıf belirlemede etkili bir modeldir.

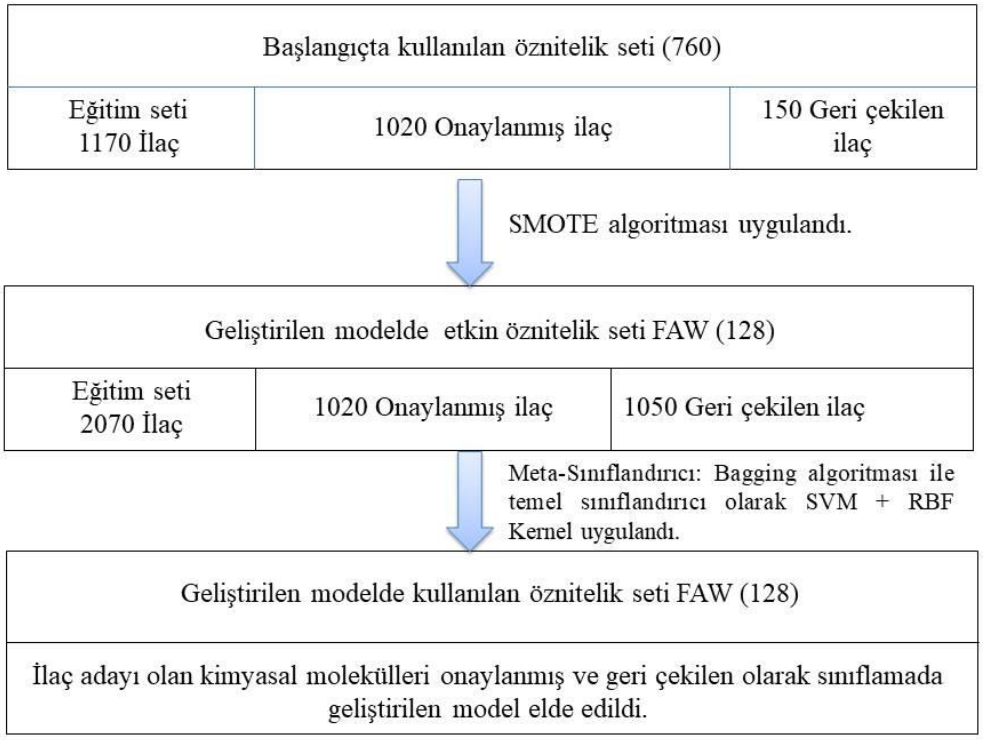
Çalışmada sınıflandırıcı topluluk tasarımı için önerilen modelde dengesiz veri setine ait çoğunluk, azınlık grubu ve özniteliklerin sayısal analizinin modelin her aşamasında yapılabilmesi amacıyla herbiri birer matris olarak ifade edilmiştir böylelikle her aşama sonrasında çoğunluk, azınlık grubu ve öznitelik matrisinin boyutsal olarak değişimi kolayca takip edilebilmektedir. Geliştirilen bu modelle veri setindeki her bir durum için hedef sınıfın (onaylanmış/geri çekilen durumu) doğru bir şekilde tahmin edilmesi hedeflenir. Deneylerde kullanılan ilaç veri setinin ve bağımsız test setinin özellikleri Çizelge 6.4'te verilmiştir. Veri setindeki toplam örnek sayısı, # Örnekler; Çoğunluk sınıfındaki örnek sayısı, # Çoğunluk; Azınlık sınıfındaki örnek sayısı, # Azınlık; Dengesizlik oranı, # DO ile gösterilmektedir. Tek yıldızla işaretlenen veri seti (eğitim seti) SMOTE algoritması uygulandıktan sonra çoğunluk ve azınlık grubu sayısını, çift yıldızla işaretlenen veri seti (eğitim seti) SpreadSubsample algoritması uygulandıktan sonra çoğunluk ve azınlık grubu sayısını temsil etmektedir. Çizelge 6.4'e göre başlangıçta dengesiz ilaç veri seti 1170 ilaç içermektedir. Veri seti için etkin öznitelikler belirlendikten sonra veri setine SMOTE algoritması uygulanarak

veri setinin dengelenmesi sağlanmıştır. Bu aşamadan sonra eğitim setimiz 2070 ilaç ve 760 öznitelik içermektedir. Veri seti dengelendikten sonra 1 nolu aşamada belirlenen etkin öznitelikler (FAW/128) veri setinden seçilerek içerisinde sadece etkin öznitelikleri içeren dengeli veri seti elde edilmiştir bu durumda veri setinde 2070 ilaç ve 128 öznitelik yer almaktadır. Elde edilen bu sayıca dengeli eğitim seti AWD1, meta-sınıflandırıcı olarak Bagging algoritması ile temel sınıflandırıcı olarak SVM+RBF Kernel metot kullanılarak 10-kat çapraz doğrulama metodu ve bağımsız test seti de AWD3 kullanılarak test edilmiş ve her iki durum içinde sınıflandırma performansları sonuç kısmında verilmiştir. Buna ek olarak yine dengesiz ilaç veri seti üzerinde (1170 ilaç) farklı bir model geliştirmek amacıyla veri setine SpreadSubsample algoritması uygulanmış ve AWD2 eğitim seti elde edilmiştir. AWD2 veri seti bu durumda 300 ilaç ve 760 öznitelik içermektedir. Sonrasında sadece 1 nolu aşamada seçilen etkin öznitelikler (FAW/128) veri setinde kalacak şekilde AWD2 eğitim seti elde edilir bu durumda AWD2, 300 ilaç ve 128 öznitelik içermektedir. Sonuç olarak dengesiz ilaç veri seti için belirlenen etkin öznitelikler hem AWD1 hemde AWD2 ile elde edilen modellerde kullanılmış ve modellerin performansları hem eğitim hem test seti üzerinde karşılaştırılmıştır. Önerilen modelin dengesiz veri setlerinde ikili sınıflandırma problemleri çalışılırken, eğitim seti 10-kat çapraz doğrulama metodu ile test edildiğinde ve bağımsız test setleri için de sınıf belirlemede etkili bir model olduğu gözlemlenmiştir. Çalışmada ayrıca onaylanmış ve geri çekilen ilaçlardan oluşan dengesiz bir veri setinin önerilen sınıflandırıcı topluluk tasarımı modeli ile sınıflandırılması sonuçları ve aynı veri seti üzerinde diğer öznitelik seçme algoritmaları ve meta-sınıflandırıcıları kullanılarak elde edilen sonuçların eğitim seti ve bağımsız test seti için doğruluk oranı (AR), eğri altındaki alan (AUC), pozitif öngörme değeri (PPV), negatif öngörme değeri (NPV), duyarlılık (SE), özgüllük (SP) değerleride hesaplanarak karşılaştırılmış ve sonuçlar kısmında verilmiştir. Şekil 6.3'te ilaç adayı kimyasal molekülleri onaylanmış ve geri çekilen sınıflarına ayırmada kullanılacak olan modelin geliştirilme aşamaları gösterilmektedir.

Çizelge 6.4: Deneylerde kullanılan ilaç veri setlerinin ve bağımsız test setinin özellikleri.

Veri seti (Eğitim seti)	**# Örnekler**	**# Çoğunluk (onaylanmış ilaçlar)**	**# Azınlık (geriçekilen ilaçlar)**	**# DO**
İlaç veri seti	1170	1020	150	6.8
*** Veri seti (Eğitim seti)**	**# Örnekler**	**#Onaylanmış ilaçlar**	**# Geri çekilen ilaçlar**	**# DO**
AWD1	2070	1020	1050	Dengeli
****Veri seti (Eğitim seti)**	**# Örnekler**	**#Onaylanmış ilaçlar**	**# Geri çekilen ilaçlar**	**# DO**
AWD2	300	150	150	Dengeli
Veri seti (Test seti)	**# Örnekler**	**#Onaylanmış ilaçlar**	**# Geri çekilen ilaçlar**	
AWD3	50	30	20	

Veri setindeki toplam örnek sayısı, # Örnekler; Çoğunluk sınıfındaki örnek sayısı, # Çoğunluk; Azınlık sınıfındaki örnek sayısı, # Azınlık; Dengesizlik oranı, # DO. *Veri setine (eğitim seti) SMOTE algoritması uygulandıktan sonra çoğunluk ve azınlık grubu sayısı.**Veri setine (eğitim seti) SpreadSubsample algoritması uygulandıktan sonra çoğunluk ve azınlık grubu sayısı.



Şekil 6.3: İlaç adayı kimyasal molekülleri onaylanmış ve geri çekilen sınıflarına ayırmada kullanılacak olan modelin geliştirilme aşamaları.

SMOTE algoritması 1020 onaylanmış ve 150 geri çekilen eğitim setine uygulandı. Meta-sınıflandırıcı ise 2070 ilaç ve 128 etkin öznielik içeren (FAW) eğitim setine uygulandı.

Çalışmada geliştirilen etkin öznitelik seçme stratejisi ilaç veri setinden farklı olarak PubChem biyoassay veri setinden biri olan AID 1284'de uygulanmıştır. Veri seti UCI makine öğrenme ambarında (machine learning repository) bulunmaktadır (Schierz, 2009). PubChem biyoassay veri setlerinin özelliği dengesiz veri setleri içermesidir. Veri setleri ilaç benzeri küçük moleküllere (bileşikler) ilişkin öznitelikler içermektedir ve bu öznitelikler (ilaç benzeri özellikler) kullanılarak bir bileşiğin sınıflandırma sonrasında aktif veya aktif olmadığına karar verilir. Tezde geliştirilen etkin öznitelik seçme stratejisi veri setine uygulanmış sonrasında aynı veri seti üzerinde diğer öznitelik seçme algoritmaları ve meta-sınıflandırıcıları kullanılarak elde edilen sonuçların eğitim seti ve bağımsız test seti için doğruluk oranı (AR), eğri altındaki alan (AUC), pozitif öngörme değeri (PPV), negatif öngörme değeri (NPV), duyarlılık (SE), özgüllük (SP) değerleride hesaplanarak karşılaştırılmış ve sonuçlar kısmında verilmiştir. Çizelge 6.5'te AID 1284 veri setine ilişkin özellikler yer almaktadır. Başlangıçta veri seti 286 veri içermektedir. Tek yıldızlı veri seti AID 1284_E, dengesiz veri setine (AID 1284) SMOTE algoritması uygulandıktan sonra elde edilen dengeli eğitim setini temsil etmektedir. AID 1284_T ise bağımsız test setini göstermektedir. Yapılan deneylere ilişkin veriler sonuçlar kısmında yer almaktadır. Çalışma MATLAB yazılım paketi (MATLAB & SIMULINK, R2015a) ve Weka veri madenciliği uygulaması ile gerçekleştirildi (weka.version 3.7.13, package manager).

Veri seti aktif ve aktif olmayan bileşikler (ilaç benzeri küçük moleküller) ile bunlara ait özellikleri içermektedir. Aşağıda bu bileşiklere ait veri setleri ve bağımsız test setine ait özellikler yer almaktadır.

Çizelge 6.5: PubChem biyolojik analizler (biyo-deney) veri setinin (AID1284) özellikleri. Veri seti UCI Machine Learning Repository'den tezde önerilen öznitelik seçme stratejisinin veri seti üzerindeki performansının diğer yöntemlerle karşılaştırılması amacıyla alındı.

Veri seti	**# Örnekler**	**# Çoğunluk (aktif olmayan bileşikler)**	**# Azınlık (aktif olan bileşikler)**	**# DO**
AID 1284	286	240	46	5.2
*** Veri seti**	**# Örnekler**	**#Aktif olmayan bileşikler**	**# Aktif bileşikler**	**# DO**
AID 1284_E	470	240	230	Dengeli
Veri seti (Test seti)	**# Örnekler**	**#Aktif olmayan bileşikler**	**# Aktif bileşikler**	
AID 1284_T	72	61	11	

Veri setindeki toplam örnek sayısı, # Örnekler; Çoğunluk sınıfındaki örnek sayısı, # Çoğunluk; Azınlık sınıfındaki örnek sayısı, # Azınlık; Dengesizlik oranı, # DO. *Veri setine SMOTE algoritması uygulandıktan sonra çoğunluk ve azınlık grubu sayısı.

6.3 Sonuçlar

6.3.1 Sınıflandırmada etkin olan moleküler tanımlayıcılar

İlaç molekülleri için hesaplanan moleküler tanımlayıcılar arasından sınıflandırmada en etkin olanlarını belirleme ilaç tasarım problemlerinde önemli bir rol oynar. Bu nedenle Bölüm (6.2.2)'de geliştirilen etkin öznitelik seçme metodu çeşitli hastalık gruplarına ait dengesiz ilaç veri setine uygulandı. Etkin öznitelik seçme stratejisinin dengesiz ilaç veri setine uygulanma aşamasında oluşturulan A1-W, A2-W…A6-W ilaç veri setlerinden en az üç veri setinde etkin öznitelik olarak seçilen 45 öznitelik Çizelge (6.6)'da verilmiştir. Etkin öznitelikler belirlenirken A1-W, A2-W…A6-W ilaç veri setlerine değiştirilmiş ki-kare öznitelik seçme algoritması uygulanmış ve bir özniteliğin sınıf içerisindeki ki-kare değeri > 0 ise etkin öznitelik setinde yer almıştır. Bu tanımlayıcılara ait veri setinde aldıkları alt ve üst sınır değerler aday ilaç moleküllerinin özellikleri için sınır koşullarını belirlemede kullanılabilir. Aşağıdaki çizelgede etkin öznitelik setinde yer alan 128 öznitelikten 45'i yer almaktadır.

Çizelge (6.7)'de Çizelge (6.6)'da yer alan moleküler tanımlayıcıların (SFs) ayrıntılı analizi verildi. Burada bir ilaç molekülünün içerdiği toplam kemotip sayısını veren the number of total chemotypes tanımlayıcısı belirlenen etkin öznitelik setinde yer almamıştır. Çizelge (6.7)'de başına '*' konarak belirtilmiştir ve dengesiz ilaç veri

seti içerisinde onaylanmış ve geri çekilen ilaçlar için aldığı alt ve üst değerler de yer almaktadır.

Çizelge 6.6: İlaç veri seti için sınıflandırma modellerinin geliştirilmesinde en etkin olan moleküler tanımlayıcılar (öznitelikler).

SFs (45) from FAW (128)
Atoms, Bonds, HAcc, HaccN, HAccO, HDon, Ro5Viol, Ro5ViolExt, Weight, ASA, McGowan, TPSA, Polariz, LogS, XlogP, Diameter:Cor3D:ori1, InertiaY:Cor3D:ori1, InertiaZ:Cor3D:ori1,Rgyr:Cor3D:ori1,Span:Cor3D:ori1,bond:C(=O)N_carboxamide_(NH2), bond:C(=O)N_carboxamide _(NHR), bond:C=O_acyl_hydrazide, bond:C=O_carbonyl_ab-unsaturated_generic, bond:CC(=O)C_ ketone_aromatic_aliphatic, bond:CN_amine_pri-NH2_generic,bond:COH_alcohol_diol_(1_3-), bond:NN_hydrazine_acyclic_(connect_noZ), bond:NN_hydrazine_alkyl_HH2,bond:NN_hydrazine_alkyl_N(connect_Z=1),bond:P=O_phosphorus_oxo,bond:PC_phosphorus_organo_generic,568_group:carbohydrate_aldohexose, group:carbohydrate_aldopentose,group:carbohydrate_hexopyranose_fructose,group:carbohydrate_hexopyranose_glucose,group:carbohydrate_ketohexose,group:carbohydrate_pentopyranose, ring:aromatic_benzene, ring:aromatic_phenyl, ring:hetero_[5] _N_S_thiazole, ring:hetero_[5]_O_oxolane, ring:hetero_[6]_N_diazine_(1_3-)_generic, ring:hetero_[6]_N_pyrimidine, ring:hetero_[6]_Z_1_3-
SFs, Seçilen özellikler; FAW, Öznitelik seti.

Özniteliklerin solunda yer alan sayılar index numaralarını göstermektedir. Çizelge (6.7)'de verilen etkin öznitelikler dengesiz ilaç veri setine (1170 ilaç analiz edildi) aittir. Buna ek olarak sınıflandırıcı topluluk tasarımı için önerilen modeli test etmek amacıyla kullanılan bağımsız test seti AWD3 aynı öznitelikler kullanılarak analiz edilmiştir. Çizelge (6.8)'de AWD3'e ait onaylanmış ve geri çekilen ilaçların (50 ilaç) aynı öznitelikler doğrultusunda ayrıntılı analizi yer almaktadır.

Çizelge 6.7: Seçilen etkin moleküler tanımlayıcıların dengesiz ilaç veri seti üzerinde (1170 ilaç) ayrıntılı analizi.

SFs from AWD	GMF/SSF/TCF/UP (mostly/only located in ADs/WDs)			Range [a, b] for ADs	Range [a, b] for WDs
*The number of total chemotypes	- / - / - / +	-		[1, 61]	[5, 41]
1_Atoms	+ / - / - / -	-		[2, 227]	[26, 152]
2_Bonds	+ / - / - / -	-		[1, 236]	[25, 159]
4_HAcc	+ / - / - / -	-		[0, 52]	[1, 24]
5_HAccN	+ / - / - / -	-		[0, 19]	[0, 8]
6_HAccO	+ / - / - / -	-		[0, 49]	[0, 24]
7_HDon	+ / - / - / -	-		[0, 26]	[0, 11]
10_Ro5Viol	+ / - / - / -	-		[0, 4]	[0, 3]
11_Ro5ViolExt	+ / - / - / -	-		[0, 5]	[0, 4]
13_Weight	+ / - / - / -	-		[16.03, 1793.1]	[144.26, 1085.15]
16_ASA	+ / - / - / -	-		[47.78, 2193.3]	[252.4, 1372.33]
17_McGowan	+ / - / - / -	-		[14.58, 1245.86]	[136.7, 767.44]
18_TPSA	+ / - / - / -	-		[0, 805.48]	[0, 358.2]
20_Polariz	+ / - / - / -	-		[1.02, 180.45]	[18.08, 104.32]
21_LogS	+ / - / - / -	-		[-16.12, 3.61]	[-7.42, 0.18]
22_XlogP	+ / - / - / -	-		[-18.17, 11.06]	[-2.55, 7.06]
24_Diameter:Cor3D:ori1	- / + / - / -	-		[0.97, 41.21]	[6.76, 38.96]
27_InertiaY:Cor3D:ori1	- / + / - / -	-		[0.89, 185536]	[717.42, 69450.1]
28_InertiaZ:Cor3D:ori1	- / + / - / -	-		[0.89, 206747]	[790.53, 70850.6]
29_Rgyr:Cor3D:ori1	- / + / - / -	-		[0.23, 11.63]	[2.31, 10.77]
30_Span:Cor3D:ori1	- / + / - / -	-		[0.52, 22.63]	[3.79, 20.77]
64_bond:C(=O)N_carboxamide_(NH2)	- / - / + / -	only located in ADs	(38/1020 ADs, 0/150 WDs)	-	-
65_bond:C(=O)N_carboxamide_(NHR)	- / - / + / -	mostly located in ADs	(250/1020 ADs, 17/150 WDs)	-	-

Çizelge 6.7: (devam) Seçilen etkin moleküler tanımlayıcıların dengesiz ilaç veri seti üzerinde (1170 ilaç) ayrıntılı analizi.

SFs from AWD	GMF/SSF/TCF/UP (mostly/only located in ADs/WDs)	Range [a, b] for ADs	Range [a, b] for WDs
92_bond:C=O_acyl_hydrazide	- / - / + / - mostly located in WDs (5/1020 ADs, 6/150 WDs)	-	-
99_bond:C=O_carbonyl_ab-unsaturated_generic	- / - / + / - mostly located in WDs (57/1020 ADs, 19/150 WDs)	-	-
116_bond:CC(=O)C_ketone_aromatic_aliphatic	- / - / + / - mostly located in WDs (51/1020 ADs, 19/150 WDs)	-	-
132_bond:CN_amine_pri-NH2_generic	- / - / + / - mostly located in ADs (157/1020 ADs, 10/150 WDs)	-	-
158_bond:COH_alcohol_diol_(1_3-)	- / - / + / - only located in ADs (66/1020 ADs, 0/150 WDs)	-	-
248_bond:NN_hydrazine_acyclic_(connect_noZ)	- / - / + / - mostly located in WDs (7/1020 ADs, 10/150 WDs)	-	-
253_bond:NN_hydrazine_alkyl_HH2	- / - / + / - only located in WDs (0/1020 ADs, 3/150 WDs)	-	-
254_bond:NN_hydrazine_alkyl_N(connect_Z=1)	- / - / + / - mostly located in WDs (9/1020 ADs, 10/150 WDs)	-	-
282_bond:P=O_phosphorus_oxo	- / - / + / - only located in ADs (35/1020 ADs, 0/150 WDs)	-	-
284_bond:PC_phosphorus_organo_generic	- / - / + / - only located in ADs (15/1020 ADs, 0/150 WDs)	-	-
568_group:carbohydrate_aldohexose	- / - / + / - only located in ADs (29/1020 ADs, 0/150 WDs)	-	-
569_group:carbohydrate_aldopentose	- / - / + / - only located in ADs (48/1020 ADs, 0/150 WDs)	-	-
573_group:carbohydrate_hexopyranose_fructose	- / - / + / - only located in ADs (28/1020 ADs, 0/150 WDs)	-	-
575_group:carbohydrate_hexopyranose_glucose	- / - / + / - only located in ADs (23/1020 ADs, 0/150 WDs)	-	-
578_group:carbohydrate_ketohexose	- / - / + / - only located in ADs (48/1020 ADs, 0/150 WDs)	-	-
582_group:carbohydrate_pentopyranose	- / - / + / - only located in ADs (27/1020 ADs, 0/150 WDs)	-	-
616_ring:aromatic_benzene	- / - / + / - mostly located in WDs (670/1020 ADs, 124/150 WDs)	-	-
618_ring:aromatic_phenyl	- / - / + / - mostly located in WDs (157/1020 ADs, 46/150 WDs)	-	-
653_ring:hetero_[5]_N_S_thiazole	- / - / + / - only located in ADs (27/1020 ADs, 0/150 WDs)	-	-
657_ring:hetero_[5]_O_oxolane	- / - / + / - mostly located in ADs (70/1020 ADs, 2/150 WDs)	-	-
680_ring:hetero_[6]_N_diazine_(1_3-)_generic	- / - / + / - mostly located in ADs (117/1020 ADs, 4/150 WDs)	-	-
687_ring:hetero_[6]_N_pyrimidine	- / - / + / - mostly located in ADs (59/1020 ADs, 1/150 WDs)	-	-
706_ring:hetero_[6]_Z_1_3-	- / - / + / - mostly located in ADs (149/1020 ADs, 7/150 WDs)	-	-

Çizelge 6.8: AWD3 test setindeki ilaçların (50 ilaç) etkin moleküler tanımlayıcılar kullanılarak ayrıntılı analizi.

SFs from AWD	GMF/SSF/TCF/UP (the number of chemotypes in ADs/WDs)		Range [a, b] for ADs	Range [a, b] for WDs
*The number of total chemotypes	- / - / - / +	-	[1, 16]	[5, 15]
1_Atoms	+ / - / - / -	-	[8, 21]	[13, 26]
2_Bonds	+ / - / - / -	-	[7, 22]	[12, 28]
4_HAcc	+ / - / - / -	-	[1, 8]	[0, 5]
5_HAccN	+ / - / - / -	-	[0, 5]	[0, 2]
6_HAccO	+ / - / - / -	-	[0, 4]	[0, 4]
7_HDon	+ / - / - / -	-	[0, 5]	[0, 5]
10_Ro5Viol	+ / - / - / -	-	[0, 1]	[0, 1]
11_Ro5ViolExt	+ / - / - / -	-	[0, 1]	[0, 1]
13_Weight	+ / - / - / -	-	[60.05, 214.05]	[89.09, 361.39]
16_ASA	+ / - / - / -	-	[101.13, 275.88]	[147.35, 352.17]
17_McGowan	+ / - / - / -	-	[46.48, 138.87]	[70.55, 209.47]
18_TPSA	+ / - / - / -	-	[9.23, 126.44]	[9.23, 89.34]
20_Polariz	+ / - / - / -	-	[5.17, 18.13]	[8.21, 31.78]
21_LogS	+ / - / - / -	-	[-2.10, 2.56]	[-6.19, 1.32]
22_XlogP	+ / - / - / -	-	[-4.44, 2.1]	[-2.72, 5.5]
24_Diameter:Cor3D:ori1	- / + / - / -	-	[4.11, 9.79]	[5.63, 12.28]
27_InertiaY:Cor3D:ori1	- / + / - / -	-	[56.55, 2323.96]	[248.63, 6328.95]
28_InertiaZ:Cor3D:ori1	- / + / - / -	-	[95.39, 2562.25]	[293.54, 6979.49]
29_Rgyr:Cor3D:ori1	- / + / - / -	-	[1.27, 3.48]	[1.83, 4.40]
30_Span:Cor3D:ori1	- / + / - / -	-	[2.13, 5.24]	[2.99, 7.45]
64_bond:C(=O)N_carboxamide_(NH2)	- / - / + / -	(2/30 ADs, 0/20 WDs)	-	-
65_bond:C(=O)N_carboxamide_(NHR)	- / - / + / -	(0/30 ADs, 1/20 WDs)	-	-

Çizelge 6.8: (devam) AWD3 test setindeki ilaçların (50 ilaç) etkin moleküler tanımlayıcılar kullanılarak ayrıntılı analizi.

SFs from AWD	GMF/SSF/TCF/UP (the number of chemotypes in ADs/WDs)		Range [a, b] for ADs	Range [a, b] for WDs
92_bond:C=O_acyl_hydrazide	- / - / + / -	(2/30 ADs, 0/20 WDs)	-	-
99_bond:C=O_carbonyl_ab-unsaturated_generic	- / - / + / -	(0/30 ADs, 1/20 WDs)	-	-
116_bond:CC(=O)C_ketone_aromatic_aliphatic	- / - / + / -	(0/30 ADs, 1/20 WDs)	-	-
132_bond:CN_amine_pri-NH2_generic	- / - / + / -	(18/30 ADs, 3/20 WDs)	-	-
158_bond:COH_alcohol_diol_(1_3-)	- / - / + / -	(0/30 ADs, 0/20 WDs)	-	-
248_bond:NN_hydrazine_acyclic_(connect_noZ)	- / - / + / -	(1/30 ADs, 2/20 WDs)	-	-
253_bond:NN_hydrazine_alkyl_HH2	- / - / + / -	(0/30 ADs, 2/20 WDs)	-	-
254_bond:NN_hydrazine_alkyl_N(connect_Z=1)	- / - / + / -	(0/30 ADs, 2/20 WDs)	-	-
282_bond:P=O_phosphorus_oxo	- / - / + / -	(0/30 ADs, 0/20 WDs)	-	-
284_bond:PC_phosphorus_organo_generic	- / - / + / -	(0/30 ADs, 0/20 WDs)	-	-
568_group:carbohydrate_aldohexose	- / - / + / -	(0/30 ADs, 0/20 WDs)	-	-
569_group:carbohydrate_aldopentose	- / - / + / -	(0/30 ADs, 0/20 WDs)	-	-
573_group:carbohydrate_hexopyranose_fructose	- / - / + / -	(0/30 ADs, 0/20 WDs)	-	-
575_group:carbohydrate_hexopyranose_glucose	- / - / + / -	(0/30 ADs, 0/20 WDs)	-	-
578_group:carbohydrate_ketohexose	- / - / + / -	(0/30 ADs, 0/20 WDs)	-	-
582_group:carbohydrate_pentopyranose	- / - / + / -	(0/30 ADs, 0/20 WDs)	-	-
616_ring:aromatic_benzene	- / - / + / -	(3/30 ADs, 14/20 WDs)	-	-
618_ring:aromatic_phenyl	- / - / + / -	(0/30 ADs, 5/20 WDs)	-	-
653_ring:hetero_[5]_N_S_thiazole	- / - / + / -	(0/30 ADs, 0/20 WDs)	-	-
657_ring:hetero_[5]_O_oxolane	- / - / + / -	(1/30 ADs, 0/20 WDs)	-	-
680_ring:hetero_[6]_N_diazine_(1_3-)_generic	- / - / + / -	(2/30 ADs, 0/20 WDs)	-	-
687_ring:hetero_[6]_N_pyrimidine	- / - / + / -	(1/30 ADs, 0/20 WDs)	-	-
706_ring:hetero_[6]_Z_1_3-	- / - / + / -	(2/30 ADs, 1/20 WDs)	-	-

Sinir sistemi ilaçlarından oluşan veri setlerinde the number of total chemotypes modellerin geliştirilmesinde önemli bir tanımlayıcıdır. Çizelge (6.7)'e bakıldığında onaylanmış ilaç molekülleri için moleküllerdeki toplam kemotip sayısı 1 ve 61 arasında değerler alırken, geri çekilen ilaçlar için bu değerler 5 ile 41 arasında değişmektedir bir başka değişle onaylanmış ilaçlarda bu sayı daha yüksektir.

SFs ile temsil edilen 45 etkin özniteliktir. Çizelgedeki diğer başlık etiketleri Bölüm (4.3.1) ile aynıdır ve bunlara ait ayrıntılı açıklamada aynı bölümde yer almaktadır. Çizelge (6.7)'ye göre bond:P=O_phosphorus_oxo, bond:PC_phosphorus_organo _generic,group:carbohydrate_aldohexose,group:carbohydrate_aldopentose,group:carbohydrate_hexopyranose_fructose,group:carbohydrate_hexopyranose_glucose,group:carbohydrate_ketohexose,group:carbohydrate_pentopyranose,ring:hetero_[5]_N_S_thiazole, bond:C(=O)N_carboxamide_(NH2), bond:COH_alcohol_diol_(1_3-) kemotipleri yalnızca onaylanmış ilaçların kimyasal yapısında gözlendi. Buna karşılık bond:NN_hydrazine_alkyl_HH2 kemotipi ise yalnız geri çekilen ilaçların kimyasal yapısında bulundu. Geri çekilen ilaçların yapısında ring:aromatic_benzene ve ring:aromatic_phenyl kemotipleri onaylanmış ilaçlara göre daha fazla gözlendi. ring:aromatic_benzene kemotipi ise hem onaylanmış hem geri çekilen ilaçların yapısında çok sayıda gözlendi.

Yukarıda 1170 ilaçtan oluşan dengesiz veri seti üzerinde etkin öznitelik seçme stratejisi uygulanarak seçilen özniteliklerden yola çıkarak onaylanmış/geri çekilen ilaçların kimyasal yapısında bulunan/bulunmayan ToxPrint kemotiplerini belirlemek bu bileşiklerin sınıflandırılması problemlerinde aday ilaç moleküllerinin geri çekilen/onaylanmış durumu hakkında bize bilgi verir. Ayrıca Çizelge (6.7) çoğunlukla onaylanmış/geri çekilen ilaçların yapısında bulunan kemotipleride gözlemlememizde yardımcı olur. Buna ek olarak ilaç veri seti göz önüne alındığında geri çekilen ve onaylanmış ilaç molekülleri için belirlenen TPSA değer aralıklarına bakılacak olursa bir ilaç molekülü için TPSA > 358.2 ise onaylanmış olarak kategorize edilir. Yine veri setinde geri çekilen ve onaylanmış HAcc değer aralıklarına bakıldığında ise bir ilaç molekülü HAcc > 24 ise onaylanmış olarak sınıflandırılır. HaccN > 8 ve HAccO > 24 olan ilaç molekülleri de onaylanmış olarak kategorize edilir.

6.3.2 Meta sınıflandırma

Çalışmada sınıflandırıcı topluluk tasarımı için önerilen modelde meta-sınıflandırıcı olarak bagging algoritması temel sınıflandırıcı olarak SVM+RBF Kernel ilaçları onaylanmış ve geri çekilen kategorilerine ayırmada kullanıldı. Dengesiz veri setleri için tezde geliştirilen etkin öznitelik seçme stratejisi ile seçilen öznitelikler (FAW/128) geliştirilen modelde kullanıldı. Egitim setleri 10-kat çapraz doğrulama metoduyla doğrulanırken, eğitim setleriyle geliştirilen modelbağımsız test setiyle de doğrulandı. Bunun yanında geliştirilen etkin öznitelik seçme stratejisi dengesiz ilaç veri seti dışında PubChem biyoassay veri setinden biri olan AID 1284 üzerinde de uygulandı ve geliştirilen modelAID 1284 veri seti için de doğrulandı.

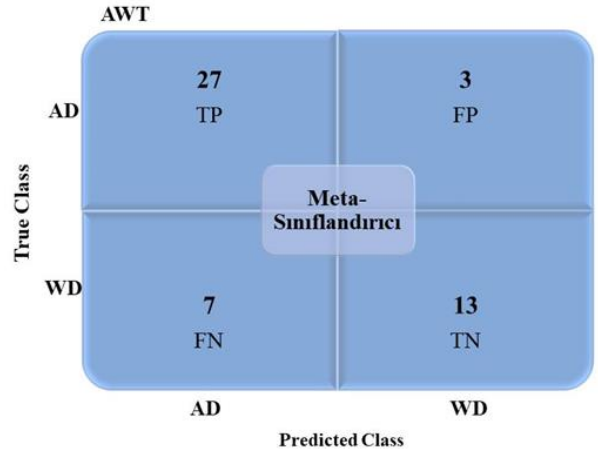
Çizelge (6.9)'da AWD1 veri seti üzerinde ilaçları onaylanmış ve geri çekilen olarak sınıflandırmada eğitim seti (AWD1) 10-kat çapraz doğrulama metoduyla ve test seti (AWD3) ile doğrulandı ve sınıflandırma modellerinin başarı indeksleri olan AUC, PPV, NPV, SE, SP, F1-S ve MCC sonuçları eğitim ve test seti için verildi.

Çizelge 6.9: Eğitim seti ve bağımsız test seti için doğruluk oranı (AR), eğri altındaki alan (AUC), pozitif öngörme değeri (PPV), negatif öngörme değeri (NPV), duyarlılık (SE), özgüllük (SP), F1-skoru (F1-score) ve Matthews korelasyon katsayısına (MCC) dayalı meta-sınıflandırıcı (bagging algoritması temel sınıflandırıcı olarak SVM+RBF Kernel) performansı.

[Meta-Sınıflandırıcı]	Eğitim Seti, AWD1 [2070 drugs]	Test Seti, AWD3 [50 drugs]
Doğruluk oranı	0.74	0.80
Eğri altındaki alan	0.78	0.79
Pozitif Öngörme Değeri	0.63	0.90
Negatif Öngörme Değeri	0.85	0.65
Duyarlılık	0.80	0.79
Özgüllük	0.70	0.81
F1-skor	0.70	0.84
Matthews korelasyon	0.50	0.58

Çizelge (6.9)'a göre AR sonuçları eğitim seti için 0.74 ve test seti için 0.80 değerini aldı. 10-kat çapraz doğrulama metoduyla doğrulanan eğitim setinde PPV değeri 0.63 iken bağımsız test seti için 0.90 değerini aldı. NPV sonuçlarına bakacak olursak eğitim seti için 0.85 ve test seti için 0.65'dir. Bu sonuçlar çalışmada elde edilen modelin ilaçları geri çekilen ve onaylanmış sınıflarına ayırmada başarılı olduğunu gösterir ve ilaç aday moleküllerini onaylanmış ve geri çekilen olarak sınıflandırmada basit bir filtre olarak kullanılabilir.

Şekil (6.4)'te AWD3 test seti için karmaşıklık matrisi verilmiştir. Buna göre AD onaylanmış, WD ise geri çekilen ilaçları temsil etmektedir. Verilen karmaşıklık matrislerinde yatay eksenler tahmin edilen sınıfları, düşey eksenler ise doğru sınıfı göstermektedir. AWD3 (50 ilaç) için 30 AD'den 27 tanesi (TP) ve 20 WD'den 13 tanesi (TN) modelle doğru olarak tahmin edildi. Yine AD'lerden 3 ilaç ve WD'lerden 7 ilaç (FP ve FN) yanlış sınıflandırıldı bunların çoğunluğu FN'lerden oluşmaktadır.

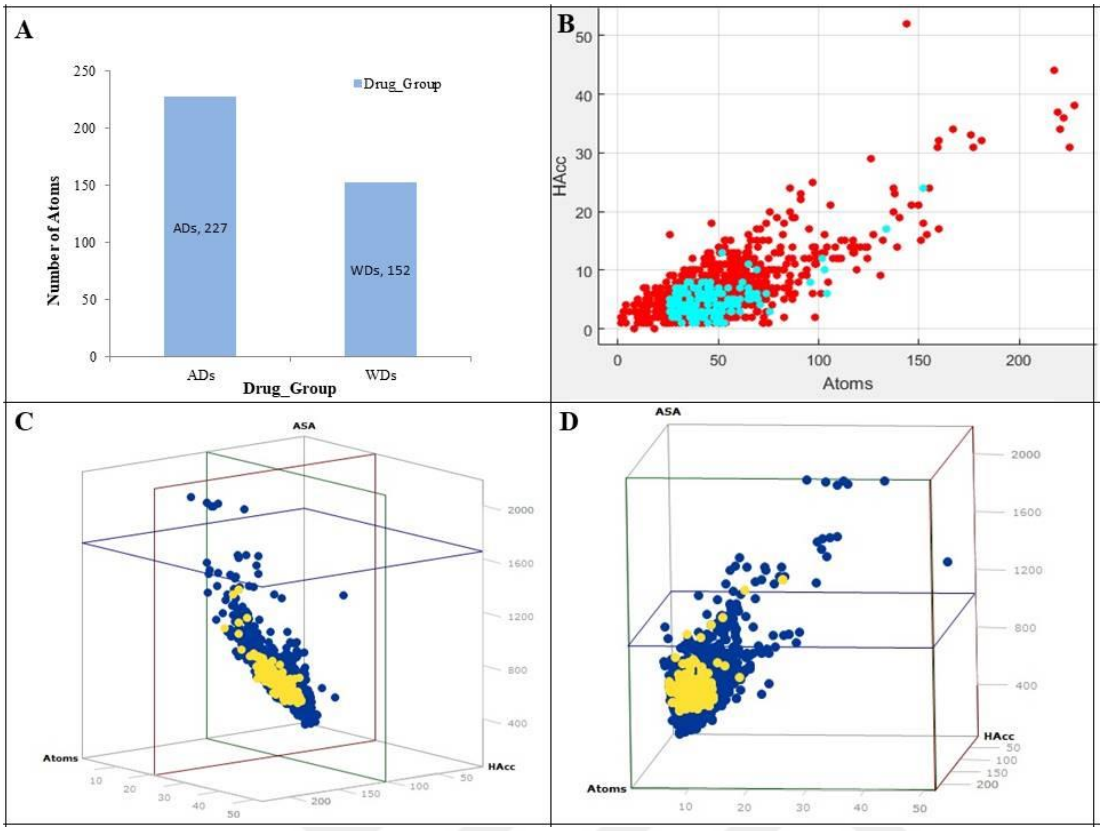


Şekil 6.4: Karmaşıklık matrisinde AWT sınıflandırma sonuçları. TP, doğru pozitif; FP, yanlış pozitif; FN, yanlış negatif; TN, doğru negatif.

Farklı hastalık gruplarına ait 1200'den fazla onaylanmış ve geri çekilen ilaç kullanılarak sınıflandırıcı topluluk tasarımı için önerilen modeladay ilaç moleküllerinin geri çekilen/onaylanmış durumunu önceden belirlemek amacıyla çalışmalarda kullanılabilir. Bu nedenle dengesiz ilaç veri seti için önerilen modelDVD_HybridModelDosyası Ek.7 adı altında araştırmacılara verildi. HybridModel.csv dosyası içerisinde ilaçları sınıflandırmada etkin rol oynayan öznitelikleri hazır olarak içermektedir. Bu öznitelikler tezde önerilen etkin öznitelik seçme stratejisi ile belirlenmiştir (FAW/128). Sonraki aşamada dengesiz ilaç veri setine SMOTE algoritması ile veri örneklemesi yapıldı ve dosyadaki 2070 ilaç örneği (1020 onaylanmış 1050 geri çekilen) eğitim setini oluşturmak üzere elde edildi. Araştırmacı öncelikle test dosyasında yer alan aday ilaç moleküllerinin her biri için CORINA Symphony programı kullanarak HybridModel.csv dosyasındaki ilk kolonda bulunan ToxPrint kemotip, global moleküler ve boyut ve şekil özelliklerini hesaplamalıdır. Sonrasında geliştirilen HybridModel.csv dosyası MATLAB yazılım

paketi veya Weka veri madenciliği uygulaması ile çalışma alanına alınmalıdır. Burada önemli bir nokta içe aktarılan sınıflayıcı yeni verilere ilişkin tahminler yapacağından ilaç aday molekülleri için hazırlanan test dosyasıda eğitim verilerinizle aynı öngörücü (öznitelik) isimlerini içermelidir. Son olarak meta-sınıflandırıcı bagging algoritması temel sınıflandırıcı olarak SVM+RBF Kernel eğitim seti (HybridModel.csv) kullanılarak araştırmacının test setindeki ilaçların onaylanmış/geri çekilen sınıf tahmini elde edilir. Araştırmacı eğer farklı bir dengesiz veri seti için bir eğitim ve test dosyası hazırlayacaksa Bölüm (6.2.3)'deki aşamaları takip etmelidir.

Çalışmada ayrıca farklı hastalık grupları için kullanılan 1020 onaylanmış ve 150 geri çekilen ilaçtan oluşan veri seti için sınıflandırmada etkin olan öznitelik setinin (FAW) elde edilmesi aşamasında 6 grup (A1-W, …, A6-W) elde edilmişti. Bu 6 grubun en az 5 inde etkin öznitelik olarak belirlenen moleküler tanımlayıcılar arasında en yüksek rank değerine sahip (ki-kare istatistik değerlerine göre) ilk üç tanımlayıcı için Şekil (6.5)'te sırasıyla 1D (Atoms), 2D (Atoms, HAcc) ve 3D (Atoms, HAcc, ASA) dağılım grafikleri elde edilmiştir. Buna göre A'da Atoms moleküler tanımlayıcısının onaylanmış ve geri çekilen ilaç veri setleri ele alındığında ilaç moleküllerinin aldığı maximum değerler verilmiştir. B'de yine onaylanmış ve geri çekilen ilaç veri setleri için Atoms'a karşılık HAcc dağılım grafiği verilmiştir. Kırmızı noktalar onaylanmış, mavi noktalar ise geri çekilen ilaçları temsil etmektedir. C-D'de Atoms, HAcc ve ASA değerlerinin onaylanmış ve geri çekilen ilaç veri setleri için 3D dağılımı yer almaktadır. C'de cut plane YZ (yeşil), cut plane XZ (kırmızı) ve cut plane XY (mavi) kullanılarak ilaçların 3D dağılımı belirginleştirilmiştir. D ise C'nin z-ekseni etrafında (ASA) döndürülmesi (rotate z) ve cut planeler kullanılarak elde edilen dikdörtgenler prizması içerisinde Atoms, HAcc ve ASA değerlerinin ilaç molekülleri için 3D dağılımı gösterilmiştir. C-D'de lacivert ile gösterilen noktalar onaylanmış, sarı ile gösterilen noktalar ise geri çekilen ilaçları göstermektedir.



Şekil 6.5: Farklı hastalık grupları için kullanılan 1020 onaylanmış ve 150 geri çekilen ilaçlara ait öznitelik değerleri kullanılarak elde edilen (A) 1D, ilaç grubuna göre ilaç moleküllerinin maximum atom sayısını, (B) 2D, onaylanmış ve geri çekilen ilaç moleküllerine ait Atoms'a karşılık HAcc grafiğini, kırmızı noktalar onaylanmış ve mavi noktalar geri çekilen ilaç moleküllerini temsil etmektedir, (C) 3D, onaylanmış ve geri çekilen ilaç moleküllerine ait Atoms, HAcc ve ASA değerlerinin dağılımını, (D) 3D, C'nin z ekseni etrafında döndürülmesiyle elde edilmiştir. C-D'de lacivert noktalar onaylanmış ve sarı noktalar geri çekilen ilaç moleküllerine ait değerleri göstermektedir.

Çizelge (6.10)'da onaylanmış ve geri çekilen ilaçlardan oluşan dengesiz bir veri setinin önerilen sınıflandırıcı topluluk tasarımı modeli ile sınıflandırılması sonuçları ve aynı veri seti üzerinde diğer öznitelik seçme algoritmaları ve meta-sınıflandırıcıları kullanılarak elde edilen sonuçların eğitim seti ve bağımsız test seti için doğruluk oranı (AR), eğri altındaki alan (AUC), pozitif öngörme değeri (PPV), negatif öngörme değeri (NPV), duyarlılık (SE), özgüllük (SP) değerleri hesaplanarak karşılaştırılması verilmiştir. Deney No.1 ile gösterilen tezde geliştirilen etkin öznitelik seçme stratejisi kullanılarak elde edilen sınıflandırma performans değerlerini göstermektedir. Toplamda aynı ilaç veri seti ile 13 deney yapılmıştır. Bu deney sonuçlarına göre, no.2'de yer alan CfsSubsetEval+BestFirst öznitelik seçme alg. ile seçilen öznitelik sayısı eğitim seti üzerinde 59'dur. Dengesiz veri seti için başlangıçta 760 öznitelik tanımlanmıştır. Eğitim seti ve test seti üzerinde no.1'in AR

sonuçları no.2 ile karşılaştırılacak olursak test seti AWD3 üzerinde daha yüksek doğruluk oranın elde edildiği gözlemlendi. Eğitim seti AWD1 üzerinde ise birbirine yaklaşık sonuçlar elde edildi. Özellikle NPV oranına bakacak olursak geri çekilen ilaçların sınıflarını belirlemede no.1'in hem eğitim hem test setinde daha başarılı olduğu açıktır. Buda yaptığımız çalışmada istenilen bir sonuçtur çünkü modellerin genel olarak geri çekilen ilaçları tahmin etme başarısı onaylanmış ilaçlara göre daha düşüktür. No.3'te veri seti dengelendikten sonra bir öznitelik seçme algoritması kullanılmadan 760 özniteliğin hepsi kullanılarak sınıflama sonuçları elde edildi. No.3'ün AWD1 üzerinde NPV değeri (0.91) no.1'den daha yüksektir ancak no.1'in de AWD3 üzerinde NPV değeri (0.65) daha yüksektir. No.5 ve 6 deneylerinde tezde geliştirilen etkin öznitelik seçme stratejisi dengeli veri seti üzerinde uygulandı ancak meta sınıflandırıcı olarak no.1 den farklı algoritmalar kullanıldı. Her iki modelde de AWD3 veri seti üzerinde NPV değeri oldukça düşüktür buda modellerde kullanılan meta-sınıflandırıcılardan kaynaklanmaktadır. Buda bize gösteriyorki ilaç veri seti üzerinde ilaçların onaylanmış/geri çekilen tahmini yapılırken meta-sınıflandırıcı olarak Bagging alg. ile SVM+RBF Kernel algoritması daha başarılıdır. Bu nedenle geliştirilen modelde meta-sınıflandırıcı olarak kullanıldı. No. 7, 11, 12 ve 13'e bakıldığında seçilen öznitelik sayısı no.1'dekine göre oldukça fazladır oysaki geliştirdiğimiz modellerde öznitelik sayısının az buna karşılık modelin sınıflandırma doğruluk oranının yüksek olmasını bekleriz. No.7, 11, 12 ve 13'ün AWD1 veri seti üzerinde aldığı NPV değeri no.1'den daha yüksektir. No. 4, 8, 9 ve 10 deneylerinde diğerlerinden farklı olarak veri setinin dengelenmesi için SpreadSubsample alg. kullanıldı. Bu deneylerin AWD2 veri seti üzerindeki sınıflandırma performansıda Çizelge (6.10)'da verildi. Buna ek olarak No.1'in AWD3 üzerinde aldığı AR değeri no.4'ten daha yüksektir. Tablodan çıkarılacak başka bir sonuçta veri setinin dengelenmesi için kullanılan algoritmalardan SMOTE ile geliştirilen sınıflandırma modelleri SpreadSubsample alg. ile olanlardan AR değerlerine bakıldığında daha başarılı olduğu gözlendi. Onaylanmış ve geri çekilen ilaçlardan oluşan dengesiz bir veri setinin önerilen sınıflandırıcı topluluk tasarımı modeli ile sınıflandırılması sonuçları ve aynı veri seti üzerinde diğer öznitelik seçme algoritmaları ve meta-sınıflandırıcıları kullanılarak elde edilen sonuçların eğitim seti ve bağımsız test seti için doğruluk oranı (AR), eğri altındaki alan (AUC), pozitif öngörme değeri (PPV), negatif öngörme değeri (NPV), duyarlılık (SE), özgüllük (SP) değerleri hesaplanarak

karşılaştırılması. Eğitim seti (ES), test seti (TS), toplam öznitelik sayısı (ÖS), seçilen özniteliklerin sayısı (SÖS) ile gösterilmektedir.

Çizelge 6.10: Onaylanmış ve geri çekilen ilaçlardan oluşan dengesiz bir veri setinin önerilen sınıflandırıcı topluluk tasarımı modeli ile sınıflandırılması sonuçları.

Deney No.	Kullanılan öznitelik seçme algoritması	Veri setinin dengelenmesi için kullanılan algoritma	Kullanılan Meta-Sınıflandırıcı
1	Tezde önerilen etkin öznitelik seçme stratejisi	SMOTE	Bagging alg. ile SVM+RBF Kernel
2	CfsSubsetEval+BestFirst	SMOTE	Bagging alg. ile SVM+RBF Kernel
3	Uygulanmadı	SMOTE	Bagging alg. ile SVM+RBF Kernel
4	Uygulanmadı	SpreadSubsample	Bagging alg. ile SVM+RBF Kernel
5	Tezde önerilen etkin öznitelik seçme stratejisi	SMOTE	Bagging alg. ile RandomForest
6	Tezde önerilen etkin öznitelik seçme stratejisi	SMOTE	AdaBoostM1 ile SVM+ RBF Kernel
7	ChiSquaredAttributeEval+Ranker	SMOTE	Bagging alg. ile SVM+RBF Kernel
8	ChiSquaredAttributeEval+Ranker	SpreadSubsample	Bagging alg. ile SVM+RBF Kernel
9	WrapperSubsetEval+GeneticSearch	SpreadSubsample	Bagging alg. ile SVM+RBF Kernel
10	WrapperSubsetEval+GeneticSearch	SpreadSubsample	AdaBoostM1 ile RandomForest
11	WrapperSubsetEval+GeneticSearch	SMOTE	Bagging alg. ile SVM+RBF Kernel
12	WrapperSubsetEval+GeneticSearch	SMOTE	Bagging alg. ile RandomForest
13	WrapperSubsetEval+GeneticSearch	SMOTE	AdaBoostM1 ile SVM+ RBF Kernel

Deney No.	ES/TS	ÖS	SÖS	AR	AUC	PPV	NPV	SE	SP
1	ES: AWD1	760	128	0.74	0.78	0.63	0.85	0.80	0.70
1	TS: AWD3	760	128	0.80	0.79	0.90	0.65	0.79	0.81
2	ES: AWD1	760	59	0.75	0.80	0.70	0.80	0.77	0.73
2	TS: AWD3	760	59	0.76	0.77	0.97	0.45	0.73	0.90
3	ES: AWD1	760	760	0.87	0.91	0.82	0.91	0.90	0.84
3	TS: AWD3	760	760	0.68	0.73	0.97	0.25	0.83	0.65
4	ES: AWD2	760	760	0.67	0.69	0.57	0.76	0.71	0.64
4	TS: AWD3	760	760	0.76	0.85	0.87	0.60	0.76	0.75
5	ES: AWD1	760	128	0.92	0.97	0.94	0.90	0.91	0.94
5	TS: AWD3	760	128	0.60	0.75	1.0	0.0	0.60	0.0
6	ES: AWD1	760	128	0.77	0.86	0.70	0.84	0.81	0.74
6	TS: AWD3	760	128	0.68	0.82	1.0	0.20	0.65	1.0
7	ES: AWD1	760	310	0.86	0.91	0.83	0.90	0.89	0.84
8	ES: AWD2	760	71	0.65	0.73	0.41	0.89	0.79	0.60
9	ES: AWD2	760	348	0.68	0.68	0.57	0.79	0.73	0.65
10	ES: AWD2	760	348	0.69	0.71	0.62	0.76	0.72	0.67
11	ES: AWD1	760	427	0.84	0.89	0.78	0.90	0.88	0.81
12	ES: AWD1	760	427	0.92	0.97	0.96	0.89	0.90	0.96
13	ES: AWD1	760	427	0.86	0.94	0.81	0.91	0.90	0.83

AWD1, 1020 onaylanmış+1050 geri çekilen = 2070 ilaç içerir (eğitim seti); AWD2, 150 onaylanmış+150 geri çekilen = 300 ilaç içerir (eğitim seti); AWD3, 30 onaylanmış+20 geri çekilen = 50 ilaç içerir (bağımsız test verisi). ES'ler (eğitim setleri) 10-kat çapraz doğrulama yöntemi kullanılarak test edilmiştir. TS'ler (test setleri) kendi deney numarasındaki eğitim setleri ile geliştirilen modeller kullanılarak test edilmiştir. Örneğin 1 numaralı deney no'daki AWD3 test seti yine 1 nolu deney no'daki AWD1 eğitim seti model olarak kullanılarak test edilmiştir.

Çizelge (6.11)'de PubChem biyolojik analizler (biyo-deney) aktif (active) ve aktif olmayan (inactive) bileşiklerden (compounds) oluşan dengesiz bir veri setinin (AID1284) önerilen sınıflandırıcı topluluk tasarımı modeli ile sınıflandırılması sonuçları ve aynı veri seti üzerinde diğer öznitelik seçme algoritmaları ve meta-

sınıflandırıcıları kullanılarak elde edilen sonuçların eğitim seti ve bağımsız test seti için doğruluk oranı (AR), eğri altındaki alan (AUC), pozitif öngörme değeri (PPV), negatif öngörme değeri (NPV), duyarlılık (SE), özgüllük (SP) değerleri hesaplanarak karşılaştırılması verilmiştir. Çizelgede AID1284_E, 240 aktif olmayan+230 aktif olan= 470 bileşikten oluşan eğitim setini ve AID1284_T, 61 aktif olmayan+11 aktif olan= 72 bileşikten oluşan bağımsız test setini temsil etmektedir. ES'ler (eğitim setleri) 10-kat çapraz doğrulama yöntemi kullanılarak test edilmiştir. TS'ler (test setleri) kendi deney numarasındaki eğitim setleri ile geliştirilen modeller kullanılarak test edilmiştir. Örneğin 1 numaralı deney no'daki AID1284_T yine 1 nolu deney no'daki AID1284_E model olarak kullanılarak test edilmiştir. No.1'de AID1284_E veri setine etkin öznitelik seçme stratejisi uygulandığında seçilen öznitelik sayısı 81'dir. Başlangıçta veri setinde 915 öznitelik yer almaktadır. No. 5'te ise seçilen öznitelik sayısı 51'dir. No. 1 ve no. 5 AID1284_T veri setine uygulandığında no.1'den elde edilen AR değeri no. 5'ten daha yüksektir. NPV değerine bakıldığında no.1, PPV değerine bakıldığında ise no.5 daha yüksektir. No.9 ve no.13, AID1284_E veri setine uygulandığında elde edilen AR değerleri no.1'den yüksektir. Ancak NPV değerine bakacak olursak no.1 her iki deneyde elde edilen NPV değerinden yüksektir. Buna ek olarak no. 9 ve no.13'te seçilen öznitelik sayısı no.1'den oldukça yüksektir. No. 4 ve no. 8'in AID1284_T veri seti üzerinde AR değerine bakacak olursal no.4'e ait AR değeri daha yüksektir. Tezde önerilen etkin öznitelik stratejisi kullanılarak geliştirilen no.1'den no. 4'e kadar olan deneylere bakacak olursak no.4 en yüksek AUC değerine sahiptir. Bu deneyler içerisinde yine en yüksek PPV değerine sahip no.4'tür bu deneyde meta sınıflandırıcı olarak AdaBoostM1 ile SVM+ RBF Kernel kullanılması PPV değerinin artmasına neden olmuştur. No. 1, 5, 9 ve 13 AID1284_E veri setine uygulandığında AR sonuçlarına göre no.1 diğer 3 deneydeki AR değerine göre düşüktür ancak burada no.9 ve no.13'te öznitelik algoritmalarıyla seçilen özniteliklerin sayısı (sırasıyla 247 ve 515) no.1'den (81) oldukça yüksektir. Genel olarak bakıldığında tezde önerilen öznitelik seçme stratejisi hem dengesiz ilaç veri seti üzerinde hemde AID1284 veri seti üzerinde sınıflandırma modelleri oluşturulurken oldukça başarılıdır. Çalışmada buna ek olarak TP'ler sınıflandırmada doğru tahmin edilen geri çekilen ilaçları (WDs) ve TN'ler ise sınıflandırmada doğru tahmin edilen onaylanmış ilaçları (ADs) göstermek üzere sınıflandırıcı topluluk tasarımı için ayrı bir model daha önerilmiştir.

Çizelge (6.11)'de PubChem biyolojik analizler (biyo-deney) aktif (active) ve aktif olmayan (inactive) bileşiklerden (compounds) oluşan dengesiz bir veri setinin (AID1284) önerilen sınıflandırıcı topluluk tasarımı modeli ile sınıflandırılması sonuçları ve aynı veri seti üzerinde diğer öznitelik seçme algoritmaları ve meta-

Çizelge 6.11: PubChem biyolojik analizler (biyo-deney) aktif (active) ve aktif olmayan (inactive) bileşiklerden (compounds) oluşan dengesiz bir veri setinin (AID1284) önerilen sınıflandırıcı topluluk tasarımı modeli ile sınıflandırılması sonuçları.

Deney No.	Kullanılan öznitelik seçme algoritması	Veri setinin dengelenmesi için kullanılan algoritma	Kullanılan Meta-Sınıflandırıcı
1	Tezde önerilen etkin öznitelik seçme stratejisi	SMOTE	Bagging alg. ile SVM+RBF Kernel
2	Tezde önerilen etkin öznitelik seçme stratejisi	SMOTE	Bagging alg. ile RandomForest
3	Tezde önerilen etkin öznitelik seçme stratejisi	SMOTE	AdaBoostM1 ile RandomForest
4	Tezde önerilen etkin öznitelik seçme stratejisi	SMOTE	AdaBoostM1 ile SVM+ RBF Kernel
5	CfsSubsetEval+BestFirst	SMOTE	Bagging alg. ile SVM+RBF Kernel
6	CfsSubsetEval+BestFirst	SMOTE	Bagging alg. ile RandomForest
7	CfsSubsetEval+BestFirst	SMOTE	AdaBoostM1 ile RandomForest
8	CfsSubsetEval+BestFirst	SMOTE	AdaBoostM1 ile SVM+ RBF Kernel
9	ChiSquaredAttributeEval+Ranker	SMOTE	Bagging alg. ile SVM+RBF Kernel
10	ChiSquaredAttributeEval+Ranker	SMOTE	Bagging alg. ile RandomForest
11	ChiSquaredAttributeEval+Ranker	SMOTE	AdaBoostM1 ile RandomForest
12	ChiSquaredAttributeEval+Ranker	SMOTE	AdaBoostM1 ile SVM+ RBF Kernel
13	WrapperSubsetEval+GeneticSearch	SMOTE	Bagging alg. ile SVM+RBF Kernel
14	WrapperSubsetEval+GeneticSearch	SMOTE	Bagging alg. ile RandomForest
15	WrapperSubsetEval+GeneticSearch	SMOTE	AdaBoostM1 ile RandomForest
16	WrapperSubsetEval+GeneticSearch	SMOTE	AdaBoostM1 ile SVM+ RBF Kernel

Deney No.	ES/TS	ÖS	SÖS	AR	AUC	PPV	NPV	SE	SP
1	ES: AID1284_E	915	81	0.71	0.74	0.52	0.88	0.81	0.66
1	TS: AID1284_T	915	81	0.78	0.66	0.27	0.87	0.27	0.87
2	ES: AID1284_E	915	81	0.86	0.93	0.83	0.89	0.88	0.85
2	TS: AID1284_T	915	81	0.78	0.53	0.27	0.87	0.27	0.87
3	ES: AID1284_E	915	81	0.86	0.92	0.84	0.87	0.86	0.85
3	TS: AID1284_T	915	81	0.80	0.60	0.27	0.90	0.33	0.87
4	ES: AID1284_E	915	81	0.75	0.83	0.68	0.81	0.77	0.73
4	TS: AID1284_T	915	81	0.72	0.71	0.45	0.77	0.26	0.89
5	ES: AID1284_E	915	51	0.74	0.81	0.70	0.77	0.74	0.73
5	TS: AID1284_T	915	51	0.70	0.80	0.63	0.72	0.29	0.91
6	ES: AID1284_E	915	51	0.89	0.95	0.83	0.96	0.95	0.85
6	TS: AID1284_T	915	51	0.86	0.57	0.27	0.96	0.60	0.88
7	ES: AID1284_E	915	51	0.90	0.95	0.83	0.96	0.95	0.86
7	TS: AID1284_T	915	51	0.88	0.60	0.36	0.98	0.80	0.90
8	ES: AID1284_E	915	51	0.76	0.85	0.75	0.77	0.76	0.76
8	TS: AID1284_T	915	51	0.62	0.64	0.45	0.66	0.20	0.87
9	ES: AID1284_E	915	247	0.83	0.90	0.80	0.85	0.84	0.82
10	ES: AID1284_E	915	247	0.90	0.96	0.85	0.95	0.95	0.87
11	ES: AID1284_E	915	247	0.90	0.96	0.85	0.96	0.95	0.87
12	ES: AID1284_E	915	247	0.88	0.95	0.88	0.88	0.88	0.88
13	ES: AID1284_E	915	515	0.81	0.88	0.79	0.85	0.83	0.80
14	ES: AID1284_E	915	515	0.91	0.96	0.85	0.96	0.95	0.87
15	ES: AID1284_E	915	515	0.91	0.96	0.85	0.96	0.95	0.87
16	ES: AID1284_E	915	515	0.87	0.92	0.90	0.85	0.85	0.90

sınıflandırıcıları kullanılarak elde edilen sonuçların eğitim seti ve bağımsız test seti için doğruluk oranı (AR), eğri altındaki alan (AUC), pozitif öngörme değeri (PPV), negatif öngörme değeri (NPV), duyarlılık (SE), özgüllük (SP) değerleri hesaplanarak karşılaştırılması. Eğitim seti (ES), test seti (TS), toplam öznitelik sayısı (ÖS), seçilen özniteliklerin sayısı (SÖS) ile gösterilmektedir.

Önerilen modelde yine meta-sınıflandırıcı olarak bagging algoritması temel sınıflandırıcı olarak SVM+RBF Kernel ilaçları onaylanmış ve geri çekilen kategorilerine ayırmada kullanılmıştır. Dengesiz veri setleri için tezde geliştirilen etkin öznitelik seçme stratejisi ile seçilen öznitelikler (FAW/128) geliştirilen modelde kullanılmıştır. Egitim setleri 10-kat çapraz doğrulama metoduyla doğrulanırken, eğitim setleriyle geliştirilen model bağımsız test setiyle de doğrulanmıştır. Çizelge (6.12)'de TP'ler sınıflandırmada doğru tahmin edilen geri çekilen ilaçları ve TN'ler ise sınıflandırmada doğru tahmin edilen onaylanmış ilaçları göstermek üzere AWD1 veri seti üzerinde ilaçları onaylanmış ve geri çekilen olarak sınıflandırmada eğitim seti (AWD1) 10-kat çapraz doğrulama metoduyla ve test seti (AWD3) ile doğrulandı ve sınıflandırma modellerinin başarı indeksleri olan AUC, PPV, NPV, SE, SP, F1-S ve MCC sonuçları eğitim ve test seti için verildi.

Çizelge 6.12: TP'ler sınıflandırmada doğru tahmin edilen WDs ve TN'ler ise sınıflandırmada doğru tahmin edilen ADs göstermek üzere, eğitim seti ve bağımsız test seti için meta-sınıflandırıcı performansı.

[Meta-Sınıflandırıcı]	Eğitim Seti, AWD [2070 drugs]	Test Seti, AWT [50 drugs]
Doğruluk oranı	0.74	0.78
Eğri altındaki alan	0.78	0.79
Pozitif Öngörme Değeri	0.87	0.60
Negatif Öngörme Değeri	0.60	0.90
Duyarlılık	0.70	0.80
Özgüllük	0.82	0.77
F1-skor	0.77	0.69
Matthews korelasyon	0.49	0.54

Çizelge (6.12)'de TP'ler sınıflandırmada doğru tahmin edilen WDs ve TN'ler ise sınıflandırmada doğru tahmin edilen ADs göstermek üzere, eğitim seti ve bağımsız test seti için doğruluk oranı (AR), eğri altındaki alan (AUC), pozitif öngörme değeri (PPV), negatif öngörme değeri (NPV), duyarlılık (SE), özgüllük (SP), F1-skoru (F1-score) ve Matthews korelasyon katsayısına (MCC) dayalı meta-sınıflandırıcı

performansını göstermektedir. PPV geri çekilen, NPV ise onaylanmış ilaçlara ait performans değerlerini göstermektedir. Çizelge (6.12)'ye göre AR sonuçları eğitim seti için 0.74 ve test seti için 0.78 değerini aldı. 10-kat çapraz doğrulama metoduyla doğrulanan eğitim setinde PPV değeri 0.87 iken bağımsız test seti için 0.60 değerini aldı. NPV sonuçlarına bakacak olursak eğitim seti için 0.60 ve test seti için 0.90'dir. Çizelge (6.9) ile sonuçları karşılaştıracak olursak eğitim setinde NPV değeri 0.85 iken (sınıflandırmada doğru tahmin edilen geri çekilen ilaçlara ait performans değeri) Çizelge (6.12)'de bu sonuç PPV değeri 0.87'dir. Test setlerinde ise Çizelge (6.9)'da geri çekilen ilaçlar için NPV 0.65 iken, Çizelge (6.12)'de PPV 0.60'tır. Sonuçlara göre önerilen her iki modelde ilaç moleküllerini onaylanmış ve geri çekilen olarak sınıflandırmada başarılıdır.

7. SONUÇ VE ÖNERİLER

Çalışmalarımızda genel olarak geri çekilen ve onaylanmış ilaç moleküllerini makine öğrenmesi metotlarını kullanıp kategorize ederken aynı zamanda geri çekilen ilaçları onaylanmış olanlardan ayırmak amacıyla bir dizi kurallar belirlenmeye çalışıldı.

Öncelikle çok sayıda geri çekilen ve onaylanmış sinir sistemi ilaçları ve farklı hastalık gruplarından ilaçlar için 760 moleküler tanımlayıcı hesaplandı. Her bir ilaç molekülünün ToxPrint kemotip analizi sınıflandırma çalışmalarında kullanılmak üzere yapıldı. Sınıflandırma problemlerinde SVM ve topluluk metotları ilaç veri setleri üzerine uygulandı. Potansiyel bileşiklerin belirlenmesi amacıyla onaylanmış/geri çekilen sinir sistemi ilaç veri setleri üzerine gSpan algoritması uygulayıp her iki kategori için ayırt edici fragmanlar belirlendi. Çalışmada göze çarpan sonuçlara bakacak olursak bir moleküldeki toplam kemotiplerin sayısını belirten the number of total chemotypes, bond CN_amine_aliphatic_generic, XlogP, aspheric: Cor3D:ori1ve Bonds tanımlayıcıları sinir sistemi ilaçlarını onaylanmış/geri çekilen kategorilerine ayırmada oldukça etkindir. Bu tanımlayıcıların aldıkları değerler kimyasal bileşikleri sınıflandırırken bir model oluşturmada önem taşır. İlaç moleküllerinin kemotip analizleri yapılırken sadece geri çekilen/onaylanmış ilaç moleküllerinde bulunan/bulunmayan kemotiplerden bond:NN_hydrazine_alkyl_N (connect_Z=1) ilaç veri setinde yalnızca geri çekilen ilaçların kimyasal yapısında bulundu. İlaçları onaylanmış ve geri çekilen olarak sınıflandırırken test setleri için doğruluk oranı 0.74 ile 0.89 arasında değerler aldı. İlaç veri setine alt çizge madenciği uygulayarak geri çekilen sinir sistemi ilaçlarının minimum %60'ında bulunan fragmanlar benzene, toluene, n,n-dimethylethylamine (DMEA), crotylamine, 5-methyl-2,4-heptadiene, octatriene, carbonyl group olarak belirlendi. Çalışma spesifik bir hastalığa ait ilaçlardan oluşan veri setlerinde geri çekilen ilaçları onaylanmış olanlardan ayırmada yapılan ilk çalışmadır. Çalışma diğer hastalık gruplarına da genişletilebilir ve farklı hastalık grupları için modeller oluşturulabilir.

İlaçlar üzerine yapılan bir diğer uygulamada toplamda 558 ilaç hiyerarşik çoklu etiket sınıflaması ile Clus-HMC-Ens algoritması kullanılıp 3 temel seviyede

sınıflandırıldı. HMC bir ağaç öğrenme algoritmasıdır. Geliştirilen model için seçilen parametreler ilaç veri setleri için geliştirilen modelin tahmin performansını arttırdı. Bu değerlerden FTest, w_0, k, sınıflandırma eşiği, m-estimate parametre değerleri özellikle hiyerarşide üçüncü düzeyde bulunan sinir sistemi ilaçları için tahmin performansını önemli ölçüde etkiledi. Modelin performans ölçümleri için FTest optimizasyon stratejisi yani precision-recall eğrisinin altındaki alan kullanıldı. Geliştirilen modeller aday ilaç moleküllerini test etmeleri amacıyla araştırmacılara için verildi. Modeller ilaçları farklı hastalık gruplarına ait onaylanmış ve geri çekilen olarak kategorize etmesinin yanında onaylanmış sinir sistemi ilacı olma durumu hakkında da öngörüde bulunur. Gelecekte yapılacak çalışmalarda hiyerarşik olarak organize edilen sınıf sayısı (3. düzeyde sinir sistemine ek olarak dolaşım sistemi, solunum sistemi vb. hastalıklarına ait ilaç veri setleri) arttırılabilir. Bunun sonucunda ilaçların onaylanmış ve geri çekilen sınıflarını tahmin etmenin yanında hangi hastalık grubuna dahil olabileceğinide öngörebiliriz. Bir ilaç hedef hastalık dışında başka bir hastalık üzerinde de iyileştirici etki gösterebilir.

Yapılan son uygulamada 1200'den fazla onaylanmış/geri çekilen ilaç üzerinde çalışıldı. Çalışılan ilaç veri seti dengesiz olması nedeni ile sınıflandırmada etkin öznitelikleri belirlemek amacıyla bir öznitelik seçme stratejisi geliştirildi. Seçilen bu moleküler tanımlayıcılardan ToxPrint kemotiplere ait olanlar onaylanmış ve geri çekilen ilaçlar üzerinde bir dizi kurallar belirlememizi sağladı. Böylece sadece onaylanmış/geri çekilen ilaçlarda bulunan/bulunmayan kemotipler analiz edildi. Bu analiz sonuçlarına göre, bond:P=O_phosphorus_oxo, bond:PC_phosphorus_organo _generic,group:carbohydrate_aldohexose,group:carbohydrate_aldopentose,group:carbohydrate_hexopyranose_fructose,group:carbohydrate_hexopyranose_glucose,group:carbohydrate_ketohexose,group:carbohydrate_pentopyranose,ring:hetero_[5]_N_S_thiazole, bond:C(=O)N_carboxamide_(NH2), bond:COH_alcohol_diol_(1_3-) kemotipleri yalnızca onaylanmış ilaçların kimyasal yapısında gözlendirken, bond:NN_hydrazine_alkyl_HH2 kemotipi ise yalnız geri çekilen ilaçların kimyasal yapısında yer aldığı gözlendi. Dengesiz ilaç veri seti üzerinde sınıflandırıcı topluluk tasarımı için geliştirilen model aday ilaç moleküllerinin onaylanmış/geri çekilen durumları hakkında araştırmacılara erken dönemde öngörüde bulunur. Çalışmada elde edilen modeller ilaç aday moleküllerini elemek için ilaç tasarım evrelerinin erken dönemlerinde basit bir filtre olarak kullanılabilirler.

Çalışmanın bu kısmında dengesiz ilaç veri seti üzerinde belirlenen etkin özniteliklerin onaylanmış ve geri çekilen ilaçlar için hesaplanan değer aralıklarına bakacak olursak bir molekülün ilaç benzerliği konusu ele alındığında benzerlik analizleri için bir dizi kural elde edilmiştir. Geri çekilen ve onaylanmış ilaç molekülleri için belirlenen TPSA değer aralıklarına bakılacak olursa bir ilaç molekülü için TPSA > 358.2 ise onaylanmış olarak kategorize edilir. Yine veri setinde geri çekilen ve onaylanmış HAcc değer aralıklarına bakıldığında ise bir ilaç molekülü HAcc > 24 ise onaylanmış olarak sınıflandırılır. HaccN > 8 ve HAccO > 24 olan ilaç molekülleri de onaylanmış olarak kategorize edilir. Bu sonuçlar DRUGBANK'taki 1170 ilaç üzerinden elde edilmiştir. İlaç güvenliği açısından konuya bakılacak olursa aday ilaç moleküllerinin toxprint özellikleri onların ters ilaç etkileşimleri hakkında bize bilgi verir. Bu nedenle bir bileşik için hesaplanan 760 özniteliğin aldığı değerler onun fizikokimyasal yapısı hakkında bilgi verir. Yine çalışmada sadece geri çekilen ilaç yapısında bulunan bir toxprint kemotip olan bond:NN_hydrazine_alkyl_HH2 ilaç adayı molekülün yapısında olması istenmeyen bir toxprint kemotiptir. Çünkü bu yapılar insanlar üzerinde beklenmedik ters ilaç reaksiyonlarına neden olabilirler. İlaç tasarım evrelerinin erken evrelerinde bu elde ettiğimiz kurallar aday ilaç molekülleri için hem sayısal veri analizlerini hemde kimyasal yapılarının analizlerini gerçekleştirmek için bir rol modeldirler. Bu tezde sinir sistemi ilaçlarının da yer aldığı sınıflandırma modelleri içinde geçerlidir. Tezde ilaç veri setleri için geliştirilen modellerin hepsi bu konuda çalışan araştırmacılara katkı sağlamak amacıyla verildi.

Çalışmaya genel olarak bakacak olursak piyasaya çıkmış ilaç veri tabanlarından elde edilen onaylanmış ve geri çekilen ilaçlara ait hesaplamalı öznitelikler kullanılarak bunları onaylanmış/geri çekilen olarak kategorize edecek sınıflandırma modelleri geliştirilmiştir. Tez boyunca geliştirilen sınıflandırma modellerindeki hatalar onaylanmış/geri çekilen ilaç molekülleri için hesaplanan özniteliklerin birbirine çok yakın değerler alması durumunda artmaktadır yani model gerçekte geri çekilen bir ilaç molekülünü onaylanmış sınıfına atabilir veya geri çekilen bir ilaç molekülünü onaylanmış sınıfına atabilir. Bu durum onaylanmış bir ilaç molekülünün kimyasal yapısında bulunan fragmanların geri çekilen ilaç molekülünün kimyasal yapısında bulunan ortak fragmanlara (veya tam tersi) sahip olmasının yanı sıra Bölüm (3.3)'te anlatılan özniteliklerin her iki ilaç molekülü içinde birbirine çok yakın olmasından

kaynaklanmaktadır. Bu özniteliklerin neler olduğu tez içerisinde ayrıntılı olarak anlatılmaktadır.

Sınıflandırma modelleri MATLAB yazılım paketi (MATLAB & SIMULINK, R2015a), CLUS sistemi ve Weka veri madenciliği uygulaması (weka.version 3.7.13, package manager) ile gerçekleştirildi.

KAYNAKLAR

Albert, B., Bray, D., Johnson, A., Lewis, J., Raff, M., (1998). Essential Cell Biology, Garland, NewYork.Bhahin, G.A.,The effect of pulsatingflow on forcedconvectiveheat transfer, *M.Sc. thesis*, University of Western Ontario, Ontario, (1998).

Amasyalı, M.F., (2008). New machine learning algorithms and applications to drug design, Doctoral thesis, Yıldız Technical University, Istanbul.

Bajorath, J., (2001). Selected concepts and investigations in compound classification, molecular descriptor analysis, and virtual screening, J. Chem. Inf. Comput, 41, 233–245.

Baloğlu, U.B., (2006). DNA Sıralarındaki tekrarlı örüntülerin ve potansiyel motiflerin veri madenciliği yöntemiyle çıkarılması, Yüksek lisans tezi, Elazığ.

Barutcuoglu Z., Schapire, R.E., Troyanskaya, O.G., (2006). Hierarchical multi-label prediction of gene function, Bioinformatics, 22, 830-836.

Bertini, F., Canetti, M., Ricci, G., (2009). Thermal behavior of syndiotactic E-1,2-poly(3-methyl-1,3-pentadiene), European Polymer Journal, 45, 923–931.

Blockeel, H., Bruynooghe, M., Dzeroski, S., Ramon, J., Struyf, J., (2002). Hierarchical multi-classification. In: Proceedings of the First SIGKDD Workshop on Multi Relational Data Mining (MRDM-2002), 21-35.

Bouckaert, R. R., Frank, E., Hall, M., Kirkby, R., Reutemann, P., (2015). WEKA Manual for Version 3-7-13, University of Waikato, Hamilton, New Zealand.

Breiman, L., (1996a). Bagging predictors, Machine Learning., 24, 123–140.

Breiman, L., (1998). Arcing classifiers (with discussion), Ann Statist., 26, 801-849.

Burbidge, R., Trotter, M., Buxton, B., Holden, S., (2001). Drug design by machine learning: support vector machines for pharmaceutical data analysis, Comput. Chem., 26, 5-14.

Burges, C. J. C., (1998). A tutorial on support vector machines for pattern recognition, data mining and knowledge discovery, Kluwer Academic Publishers, 2, 121-167.

Camps-Valls, G., Bruzzone, L., (2005). Kernel-based methods for hyperspectral image classification, IEEE Trans. Geosci. Remote Sens, 43, 1351–1362.

Cao, G.P., Thangapandian, S., John, S.,. Lee, K.W, (2012). Classification of HDAC8 inhibitors and non-inhibitors using support vector machines, IBC. , 4, 1-7.

Ceron, C.S., doVale, G.T., Simplicio, J.A., Passaglia, P., Ricci, S.T., Tirapelli, C.R., (2017). Data on the effects of losartan on protein expression,vascular reactivity and antioxidant capacity in the aorta of ethanol-treated rats, Data in Brief, 11, 111–116.

Chawla, N.V., Bowyer, K.W, Hall, L.O., Kegelmeyer, W.P., (2002). SMOTE: Synthetic Minority Over-sampling Technique, Journal of Artificial Intelligence Research, 16, 321–357.

Chen, B., Harrison, R. F., Papadatos, G., Willett, P., Wood, D. J., Lewell, X. D., (2007). Evaluation of machine-learning methods for ligand-based virtual screening, J. Comput. Aided Mol. Des., 21, 53–62.

Chen, Y.E., Hu, H.W., Tang, K., (2009). Constructing a decision tree from data with hierarchical class labels, Expert Systems with Applications, 36, 4838–4847.

Cioni, L., De Siena, G., Ghelardini, C., Sernissi, O., Alfarano, C., Pirisino, R., Raimondi, L., (2006). Activity and expression of semicarbazide-sensitive benzylamine oxidase in a rodent model of diabetes: Interactive effects with methylamine and alpha-aminoguanidine, European Journal of Pharmacology, 529, 179–187.

Clark, D. E. and Pickett, S. D., (2000). Computational methods for the prediction of drug-likeness, Drug Discovery Today, 5, 49–58.

Corbo-Dorca, R., Amat, L., Besalu, E., Girone's, X., Robert, D., (2000). Quantum mechanical origin of QSAR: theory and applications, Journal of Molecular Structure, 504, 181-228.

Cornejo, D.C., Carrera, M.G., Domínguez, M.P., Perez, M.D., Nuño, N., (2014). Acetaldehyde targets superoxide dismutase 2 in liver cancer cells inducing transient enzyme impairment and a rapid transcriptional recovery, Food and Chemical Toxicology, 69, 102–108.

Cortes, C., Vapnik, V., (1995). Support-vector networks, Mach. Learn., 20, 273-297.

de Cerqueira Lima, P., Golbraikh, A., Oloff, S., Xiao, Y., Tropsha, A., (2006). Combinatorial QSAR modeling of P-glycoprotein substrates, J. Chem. Inf. Model, 46, 1245–1254.

Dey, A., Bhattacharya, R., Mukherjee, A., Pandey, D.K., (2017). Natural products against Alzheimer's disease: Pharmaco-therapeutics and biotechnological interventions, Biotechnology Advances, 35, 178–216.

Diniz, E.M.L.P., Poiani, J.G.C., Taft, C.A., da Silva, C.H.T.P., Structure-based drug design, Molecular Dynamics and ADME/Tox to investigate protein kinase anti- cancer agents, 12, 1-10.

Drehmer, D. E. and Morris, G. W., (1981). Cross-validation with small samples: An algorithm for computing Gollob's estimator. Educational and Psychological Measurement, 41, 195-200.

Drews, J., (2000). Drug discovery: a historical perspective, Science, 287, 1960–1964.

Efron, B., Tibshirani, R.J., (1993). An Introduction to the Bootstrap, Chapman & Hall, New York.

Ekins, S., Shimada, J., Chang, C., (2006). Application of data mining approaches to drug delivery, Adv. Drug Deliver. Rev., 58, 1409-1430.

Elayaraja, E., Thangavel, K., Ramya, B., Chitralegha, M., (2012). Extraction of Motif Patterns from Protein Sequence Using Rough –K-Means Algorithm, Procedia Engineering, 30, 814-820.

Embrechts, M.J., Ozdemir, M., Lockwood, L., Breneman, C., Bennett, K., Devogelaere, D., Rijckaert, M., Feature selection methods based on genetic algorithms for in silico drug design, Evolutionary Computation in Bioinformatics, 1st Edition, (Sf. 317-339) USA, Elsevier (2003).

Engelender S. and Isacson, O., (2017). The Threshold Theory for Parkinson's Disease, Trends Neurosci., 40, 4-14.

Evens, R.P., (2007). Drug and biological development, USA, Springer.

Fish, P.V., Ryckmans, T., Stobie, A., Wakenhut, F., (2008). [4-(Phenoxy)pyridin-3-yl]methylamines: A new class of selective noradrenaline reuptake inhibitors, Bioorganic & Medicinal Chemistry Letters, 18, 1795–1798.

Fliri, A.F., Loging, W. T., Thadeio, P.F., Volkmann, R.A., (2005). Analysis of drug-induced effect patterns to link structure and side effects of medicines, Nat. Chem. Biol., 1, 389-397.

Fogel, G.B., (2008). Computational intelligence approaches for pattern discovery in biological systems, Brief Bioinform, 9, 307–316.

Foody, G.M., Mathur, A., (2006). The use of small training sets containing mixed pixels for accurate hard image classification: training on mixed spectral responses for classification by a SVM. Remote Sens. Environ. 103, 179–189.

Fourches, D., Muratov, E., Tropsha, A., (2010). Trust, but verify: On the importance of chemical structure curation in cheminformatics and QSAR modeling research, J. Chem. Inf. Model., 50, 1189–1204.

Freire, E., (2005) Thermodynamics Guide to Affinity Optimization of Drug Candidates, Protein Reviews vol 3, ed. J.E. Ladbury, New York: Kluwer/Plenum.

Freitas, A.A., de Carvalho, ACPLF, (2007). Research and Trends in Data Mining Technologies and Applications, Idea Group, chap A Tutorial on Hierarchical Classification with Applications in Bioinformatics, 175-208.

Freund Y., Schapire R. E., (1996). Experiments with a new boosting algorithm, In: Saitta, L. (Ed.), Proceedings of the Thirteenth International Conference on Machine Learning (ICML96), Morgan Kaufmann, 148-156.

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., Herrera, F., (2011). A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches, IEEE transactions on systems, man, and cybernetics-Part C: applications and reviews, 1094.

Garcia-Serna, R., Vidal, D., Remez, N., Mestres, J., (2015). Large-Scale Predictive Drug Safety: From Structural Alerts to Biological Mechanisms, Chem. Res. Toxicol., 28, 1875–1887.

Gendelman, H. E., Anantharam, V., Bronich, T., Ghaisas, S., Jin, H., Kanthasamy, A. G., (2015). Nanoneuromedicines for degenerative, inflammatory, and infectious nervous system diseases, Nanomed-Nanotechnol., 11, 751–767.

Geppert, H., Vogt, M., Bajorath, J., (2010). Current trends in ligand-based virtual screening: molecular representations, data mining methods, new application areas, and performance evaluation, J. Chem. Inf. Model, 50, 205–216.

Ghorbanzad'e, M., Fatemi, M. H., (2012). Classification of central nervous system agents by least squares support vector machine, Chemometr. Intell. Lab. Syst., 110, 102-107.

Gillet, V.J., Willett, P., Bradshaw, J., (2003). Similarity searching using reduced graphs, J. Chem. Inf. Comput. Sci, 43, 338–345.

Gleeson, M.P., Water, N. J., Paine, S. W., Davis, A. M., (2006). In silico human and rat Vss quantitative structure–activity relationship models, J. Med. Chem., 49, 1953-1963.

Graczyk, M., Lasota, T., Trawiński, B., Trawiński, K., (2010). Comparison of Bagging, Boosting and Stacking Ensembles Applied to Real Estate Appraisal, Intelligent Information and Database Systems, Springer Berlin Heiderberg, 340-350.

Gunn, S.R., (1998). Support Vector Machines for Classification, Regression, University of Southhampton, England.

Guyon, I., Elisseeff, A., (2003). An Introduction to Variable and Feature Selection, Journal of Machine Learning Research, 3, 1157-1182.

Hall, M.A., (1999). Correlation- Based Feature Selection for Machine Learning, the Degree of Doctor of Philosophy, the University of Waikato, NewZealand.

Hand, D., (2001). Principles of Data Mining. MIT Press.

Hawkins, P.C.D., Skillman, A.G., Nicholls, A., (2007). Comparison of shape-matching and docking as virtual screening tools, J. Med. Chem, 50, 74–82.

Hendlich, M., Bergner, A., Günther, J., Klebe, G., (2003). Relibase: design and development of a database for comprehensive analysis of protein-ligand interactions, Journal of Molecular Biology, 326, 607-620.

Hinselmann, G., Rosenbaum, L., Jahn, A., Fechner, N., (2011). Large-scale learning of structure–activity relationships using a linear support

vector machine and problem-specific metrics, J. Chem. Inf. Model, 51, 203–213.

Hou, T., Wang, J., Li, Y., (2007). ADME evaluation in drug discovery. 8. The prediction of human intestinal absorption by a support vector machine, J. Chem. Inf. Model., 47, 2408–2415.

Huang, L. C., Wu, X., Chen, J. Y., (2011). Predicting adverse side effects of drugs, BMC Genomics, 12, 5.

Jayaprakash, A., Arjunan, V., Jose, S.P., Mohan, S., (2011). Vibrational and electronic investigations, thermodynamic parameters, HOMO and LUMO analysis on crotonaldehyde by ab initio and DFT methods, Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 83, 411-419.

Jeffrey, G.A., (1997). An Introduction to Hydrogen Bonding, Oxford University Press, USA.

Jónsdóttir, S. O., Jørgensen, F. S., Brunak S., (2005). Prediction methods and databases within chemoinformatics: emphasis on drugs and drug candidates, Bioinformatics, 21, 2145–2160.

Jorissen, R.N. and Gilson, M.K., (2005). Virtual screening of molecular databases using a support vector machine, J. Chem. Inf. Model, 45, 549–561.

Katagiri, S., Abe, S., (2006). Incremental training of support vector machines using hyperspheres, Pattern Recognition Letters, 27, 1495-1507.

Kauffman, G.W. and Jurs, P. C., (2001). QSAR and k-nearest neighbor classification analysis of selective cyclooxygenase-2 inhibitors using topologically-based numerical descriptors, J. Chem. Inf. Comp. Sci., 41, 1553–1560.

Kavitha, A. S., Kavitha, R., Gripsy, J. V., (2012). Empirical evaluation of feature selection technique in educational data mining., 2, 1103- 1112.

Klekota, J., Roth, F. P., (2008). Chemical substructures that enrich for biological activity, Bioinformatics., 24, 2518- 2525.

Kobal, J., Cankar, K., Pretnar, J., Zaletel, M., Kobal, L., (2016). Functional impairment of precerebral arteries in Huntington disease, J. Neurol. Sci., 372, 363-368.

Kohavi, R., (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection, IJCAI.

Konovalov, D.A., Coomans, D., Deconinck, E., (2007). Benchmarking of QSAR models for blood–brain barrier permeation, J. Chem. Inf. Comp. Sci., 47, 1648–1656.

Korkmaz, S., Zararsiz, G., Goksuluk, D., (2014). Drug/nondrug classification using Support VectorMachines with various feature selection strategies, Comput. Methods Programs Biomed., 117, 51-60.

Kotsiantis, S., Kanellopoulos, D., Pintelas, P., (2006). Handling imbalanced datasets: A review, GESTS International Transactions on Computer Science and Engineering, 30.

Kotsiantis, S., Pintelas, P., (2003). Mixture of Expert Agents for Handling Imbalanced Data Sets, Annals of Mathematics, Computing & TeleInformatics, 1, 46-55.

Krishnapuram, B., Hartemink, A.J., Carin, L., Figueiredo, M.A.T., (2004). A Bayesian Approach to Joint Feature Selection and Classifier Design, IEEE Transactions on Pattern Analysis and Machine Intelligence, 26, 9.

Labidia, N.S., Djebaili, A., (2010). Enhancement of molecular polarizabilities by the push–pull mechanism: A DFT study of substituted hexatriene, Materials Science and Engineering B, 169, 28–32.

Lamanna, C., Bellini, M., Padova, A., Westerberg, G., Maccari, L., (2008). Straightforward recursive partitioning model for discarding insoluble compounds in the drug discovery process, J. Med. Chem, 51, 2891–2897.

Lavecchia A., (2015). Machine-learning approaches in drug discovery: methods and applications, Drug Discovery Today, 20, 318-331.

Lavecchia, A. and Di Giovanni, C., (2013). Virtual screening strategies in drug discovery: a critical review, Curr. Med. Chem, 20, 2839–2860.

Leamy, A.W., Shuklaa, P., McAlexandera, M.A., Carra, M.J., Ghattaa, S., (2011). Curcumin ((E,E)-1,7-bis(4-hydroxy-3-methoxyphenyl)-1,6-heptadiene-3,5-dione) activates and desensitizes the nociceptor ion channel TRPA1, Neuroscience Letters, 503, 157– 162.

Lee, J. M., Tan, V., Lovejoy, D., Braidy, N., Rowe, D. B., (2017). Involvement of quinolinic acid in the neuropathogenesis of amyotrophic lateral sclerosis, Neuropharmacology., 112, 346-364.

Leon, A., Saito, E. K., Mehta, B., McMurtray, A. M., (2015). Calcified parenchymal central nervous system cysticercosis and clinical outcomes in epilepsy, Epilepsy & Behav., 43, 77–80.

Liao, S., Chu, P., Hsiao, P., (2012). Data mining techniques and applications – a decade review from 2000 to 2011, Expert Syst. Appl., 39, 11303-11311.

Lipinski, C., Lombardo, F., Dominy, B., Feeney, P., (2001). Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings, Adv. Drug. Deliv. Rev, 46, 3-26.

Lipinski, C.A., (2000). Drug-Like Properties and the Causes of Poor Solubility and Poor Permeability, J. of Pharm. and Tox Methods, 44, 235-249.

Lipinski, C.A., Lombardo, F., Dominy, B.W., Feeney, P.J., (1997). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings, Adv. Drug Deliv. Rev., 23, 3-25.

Lipinski, C. A., (2004). Lead- and drug-like compounds: the rule-of-five revolution, Drug Discovery Today: Technologies, 1, 337–341.

Liu, H., Sun, J., Liu, L., Zhang, H., (2009). Feature selection with dynamic mutual information, Pattern Recognition, 42, 1330–1339.

Liu, Y., (2004). A Comparative study on feature selection methods for drug discovery, J. Chem. Inform. Comput. Sci., 44, 1823-1828.

Longdon, W.B., Barrett, S.J., (2004). Genetic Programming in Data Mining for Drug Discovery, Evolutionary Computing in Data Mining, 211-235.

Lowe, R., Lowe, R., Mussa, H.Y., Nigsch, F., Glen, R.C., (2012). Predicting the mechanism of phospholipidosis, J. Cheminformatics 4, 2.

Mamitsuka, H., (2003). Empirical Evaluation of Ensemble Feature Subset Selection Methods for Learning from a High-Dimensional Database in Drug Design, Proceedings of the Third IEEE International Symposium on Bioinformatics and Bioengineering (BIBE 2003) , 253-257.

McNaughton, R., Huet, G., Shakir, S., (2014). An investigation into drug products withdrawn from the EU market between 2002 and 2011 for safety reasons and the evidence used to support the decision making, BMJ Open., 4, 4221-4226.

Meinl, T., Wörlein, M., Urzova, O., Fisher, I., Philippsen, M., (2006). The parmol package for frequent subgraph mining, ECEASST 1.

Mente, S.R., Lombardo, F., (2005). A recursive-partitioning model for blood–brain barrier permeation, J. Comput. Aided Mol. Des, 19, 465–481.

Mitchell, J.B.O., (2014). Machine learning methods in chemoinformatics, WIREs Comput. Mol. Sci, 4, 468–481.

Namoto, K., Sirockin, F., Ostermann, N., Gessier, F., Flohr, S., (2014). Discovery of C-(1-aryl-cyclohexyl)-methylamines as selective, orally available inhibitors of dipeptidyl peptidase IV, Bioorganic & Medicinal Chemistry Letters, 24, 731–736.

Nanni, L., Lumini, A., (2006). An experimental comparison of ensemble of classifiers for biometric data, Neurocomputing, 69, 1670-1673.

Onakpoya, I.J., Heneghan, C.J., Aronson, J.K., (2015). Delays in the post-marketing withdrawal of drugs to which deaths have been attributed: a systematic investigation and analysis, BMC Med., 13:26.

Opitz D., Maclin R., (1999). Popular ensemble methods: An empirical study, Journal of Artificial Intelligence Research, 11, 169–198.

Özkan, Y., (2008). Veri madenciliği yöntemleri, İstanbul, Papatya yayıncılık eğitim.

Pachner, A. R., Li, L., Gilli, F., (2015). Chemokine biomarkers in central nervous system tissue and cerebrospinal fluid in the Theiler's virus model mirror those in multiple sclerosis, Cytokine., 76, 577–580.

Patel J. and Chaudhari, C., (2005). Introduction to the artificial neural networks and their applications in QSAR studies, ALTEX., 22, 271.

Pauwels, E., Stoven, V., Yamanishi, Y., (2011). Predicting drug side-effect profiles: a chemical fragment-based approach, BMC Bioinformatics, 12, 1-13.

Ramraj, T., Prabhakar, R., (2015). Frequent subgraph mining algorithms- A survey, Procedia Comput. Sci., 47, 197- 204.

Reddy, S., Pati, S. P., Potukuchi, P. K., Sastry, G. N., (2007). Virtual screening in drug discovery: a computatinal perspective, Curr. Protein Pept. Sci., 8, 329-351.

Rousu, J., Saunders, C., Szedmak, S., Shawe-Taylor, J., (2006). Kernel-based learning of hierarchical multilabel classification models, Journal of Machine Learning Research, 7,1601-1626.

Sakiyama, Y., Yuki, H., Moriya, T., Hattori, K., Suzuki, M., (2008). Predicting human liver microsomal stability with machine learning techniques, J. Mol. Graph. Model, 26, 907–915.

Salunkhe, U. R., Mali, S. N., (2016). Classifier Ensemble Design for Imbalanced Data Classification: A Hybrid Approach, International Conference on Computational Modeling and Security (CMS 2016), 85, 725 – 732.

Schietgat, L., Vens, C., Struyf, J., Blockeel, H., Kocev, D., Džeroski, S., (2010). Predicting gene function using hierarchical multi-label decision tree ensembles, BMC Bioinformatics, 11:2.

Schnur, D., Bena, B.R., Good, A., Tebben, A., (2004). Approaches to target class combinatorial library design. Chemoinformatics, Methods Mol. Biol, 275, 355–378.

Schölkopf, B., Smola, A.J., (2002). Learning with Kernels, the MIT Press, England.

Seiffert, C., Khoshgoftaar, T., Hulse, J.V., Napolitano, (2008). RUSBoost: Improving classification performance when training data is skewed, ICPR, 19th International Conference on Pattern Recognition, IEEE.

Sershen, H., Shearman, E., Fallon, S., Chakraborty, G., Smiley, J., Lajtha, A., (2009). The effects of acetaldehyde on nicotine-induced transmitter levels in young and adult brain areas, Brain Research Bulletin, 79, 458–462.

Sharif, M. A., Tsakovoska, I., Pajeva, I., Alov, P., Fioravanzo, E., Bassan, A., Kovarich, S., (2015). The application of molecular modelling in the safety assessment of chemicals: A case study on ligand-dependent PPAR-γ dysregulation, Toxicology, 16, 30009-9.

Siramshetty, V. B., Nickel, J., Omieczynski, C., Gohlke, B. O., Drwa, M. N., Robert, P., (2016). WITHDRAWN—a resource for withdrawn and discontinued drugs, Nucleic Acids Research, 44, D1080–D1086.

Slabua, I., Galmana, J.L., Iglesiasb, C., Weisea, N.J., Lloydc, R.C., (2017). n-Butylamine as an alternative amine donor for the stereoselectivebiocatalytic transamination of ketones, Catalysis Today, xxx, xxx–xxx.

Sneader, W., (2005). Drug discovery a history, John Wiley & Sons Ltd., England.

Stelle, D., Barioni, M. C., Scott, L. P., (2011). Using data mining to identify structural rules in proteins, Applied Mathematics and Computation, 218, 1997–2004.

Struyf, J., Džeroski, S., Blockeel, H., Clare, A., (2005). Hierarchical multi-classification with predictive clustering trees in functional genomics. Progress in Artificial Intelligence: 12th Portugese Conference on Artificial Intelligence Lecture Notes in Computer Science, Springer, 3808, 272-283.

Struyf, J., Zenko, B., Blockeel, H., Vens, C., Dzeroski, S., (2011). Clus: User's Manuel.

Sullivan, S. E., Young-Pearse, T. L., (2017). Induced pluripotent stem cells as a discovery tool for Alzheimer's disease, Brain Res. 1656, 98-106.

Sutton, C. D., (2005). Classification and Regression Trees, Bagging and Boosting, Handbook of statist., 24, 303-329.

Szulc, A., Pulaski, L., Appelhans, D., Voit, B., Maculewicz, B.K., (2016). Sugar-modified poly (propylene imine) dendrimers as drug delivery agents for cytarabine to overcome drug resistance, International Journal of Pharmaceutics, 513, 572–583.

Vapnik, V.N., (2000). The Nature of Statistical Learning Theory, Springer.

Veber, D.F., Johnson, S.R., Cheng, H., Smith, B.R., Ward, K.W., Kopple, K.D., (2002). Molecular Properties that Influence the Oral Bioavailability of Drug Candidates, J. Med. Chem., 45, 2615-2623.

Vens, C., Struyf, J., Schietgat, L., Džeroski S., Blockeel, H., (2008). Decision trees for hierarchical multi-label classification, Machine Learning, 73, 185-214.

Vogel, H.G., Maas, J., Hock, F.J., Mayer, D.,(2013). Drug Discovery and Evaluation: Safety and Pharmacokinetic Assays, Heidelberg, Second Edition, Springer.

von Korff, M., Sander, T., (2006). Toxicity-indicating structural patterns, J. Chem. Inf. Model, 46, 536–544.

Votano, J.R., Parham, M., Hall, L.H., Kier, L.B., Oloff, S., (2004). Three new consensus QSAR models for the prediction of Ames genotoxicity, Mutagenesis., 19, 365–377.

Wang, G., Song, Q., Sun, H., Zhang, X., (2013). A Feature Subset Selection Algorithm Automatic Recommendation Method, Journal of Artificial Intelligence Research, 47, 1-34.

Wang, Y., Xing, J., Xu, Y., Zhou, N., Peng, J., Xiong, Z., (2015). In silico ADME/T modelling for rational drug design, Q. Rev. Biophys., 2, 1-28.

Warmuth, M.K., Liao, J., Rätsch, G., Mathieson, M., Putta, S., Lemmen, C., (2003). Active learning with support vector machines in the drug discovery process, J. Chem. Inf. Comput. Sci, 43, 667–673.

Weaver, D.C., (2004). Applying data mining techniques to library design, lead generation and lead optimization, Curr. Opin. Chem. Biol, 8, 264–270.

Willett, P., (2005). Searching techniques for databases of two- and three-dimensional chemical structures, J. Med. Chem, 48, 4183–4199.

Witten, I.H., Frank, E., (2005). Data Mining: Practical MachineLearning Tools and Techniques, 2nd ed., Morgan KaufmannPublishers, San Francisco, CA.

Xue, Y., Li, Z. R., Yap, C. W., Sun, L.Z., Chen, X., Chen, Y.Z., (2004). Effect of molecular descriptor feature selectionin support vector machine classification of pharmacokinetic and toxicological properties of chemical agents, J. Chem. Inf.Comput. Sci., 44, 1630–1638.

Yan, X., Han, J., (2002). gspan: Graph-based substructure pattern mining. Technical Report UIUCDCS-R-2002-2296, Department of Computer Science, University of Illinois at Urbana Champaign.

Yang, C., Tarkhov, A., Marusczyk, J., (2015). New Publicly Available Chemical Query Language, CSRML, To Support Chemotype Representations for Application to Data Mining and Modeling, J. Chem. Inf. Model, 55, 510-528.

Yusof, I., Segall, M.D., (2013). Considering the impact drug-like properties have on the chance of success, Drug Discovery Today., 18, 659-666.

Yusof, I., Shah, F., Hashimoto, T., Segall, M.D., Greene, N., (2014). Finding the rules for successful drug optimization, Drug Discovery Today., 19, 680-687.

Zernov, V.V., Balakin, K.V., Ivaschenko, A.A., Savchuk, N.P., (2003). Drug discovery using support vector machines. The case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions, J. Chem. Inf. Comput. Sci, 43, 2048–2056.

Zhang, B., Shi, Z.L., Liu, B., Yan, X.B., Feng, J., Tao, H.M., (2010). Enhanced anticancer effect of gemcitabine by genistein in osteosarcoma: the role of Akt and nuclear factor-kappaB, Anticancer Drugs, 21, 288-296.

Zhang, H., Li, W., Xie, Y., Wang, W.J., Li, L.L., Yang, S.Y., (2011). Rapid and accurate assessment of seizure liability of drugs by using an optimal support vector machine method, Toxicol. In vitro., 25, 1848-1854.

Zhang, MQ., Wilkinson, B., (2007). Drug discovery beyond the rule-of-five, Curr. Opin. Biotechnol., 18, 478-488.

Zheng, M., Hu, X., Yan, A., Bajorath, J., (2013). Computational Methods for Drug Design and Discovery : focus on China, Trends Pharmacol. Sci., 34, 549-559.

EKLER

EK 1: Bir hmc_ds.s dosyasının içeriği ve ilaç veri seti için optimum parametre ayarları.

EK 2: 112 test verisinin tahmin edilen hiyerarşik sınıfları ve her sınıfa ait p değerleri.

EK 1

```
[General]
Verbose = 1
Compatibility = Latest
RandomSeed = 0
ResourceInfoLoaded = No

[Data]
File = hierarchical.arff
TestSet = None
PruneSet = None
PruneSetMax = Infinity
XVal = 10
RemoveMissingTarget = No
NormalizeData = None

[Attributes]
Target = 761
Clustering = 761
Descriptive = 2-760
Key = 1
Disable = None
Weights = Normalize
ClusteringWeights = 1.0
ReduceMemoryNominalAttrs = Yes

[Constraints]
Syntactic = None
MaxSize = Infinity
MaxError = 0.0
MaxDepth = Infinity


[Nominal]
MEstimate = 1.0

[Model]
MinimalWeight = 1.0
MinimalNumberExamples = 0
MinimalKnownWeight = 0.0
ParamTuneNumberFolds = 10
ClassWeights = 0.0
NominalSubsetTests = Yes

[Tree]
Heuristic = VarianceReduction
PruningMethod = C4.5
FTest = [0.001,0.005,0.01,0.05,0.1,0.125]
BinarySplit = Yes
ConvertToRules = Leaves
AlternativeSplits = No
Optimize = {}
MSENominal = No
SplitSampling = None
InductionOrder = DepthFirst
```

EK 1 (devam)

```
[Hierarchical]
Type = Tree
Distance = WeightedEuclidean
WType = ExpAvgParentWeight
WParam = 1.0
HSeparator = /
EmptySetIndicator = n
OptimizeErrorMeasure = PooledAUPRC
DefinitionFile = None
NoRootPredictions = No
PruneInSig = 0.0
Bonferroni = No
SingleLabel = No
CalculateErrors = Yes
ClassificationThreshold = [0.5,0.75,0.80,0.90,0.95]
RecallValues = None
EvalClasses = None
MEstimate = No

[Output]
ShowModels = {Default, Pruned, Others}
TrainErrors = Yes
ValidErrors = Yes
TestErrors = Yes
AllFoldModels = Yes
AllFoldErrors = Yes
AllFoldDatasets = No
UnknownFrequency = No
BranchFrequency = No
ShowInfo = {Count}
PrintModelAndExamples = No
WriteErrorFile = No
WritePredictions = {Test}
ModelIDFiles = No
WriteCurves = Yes
OutputPythonModel = No
OutputDatabaseQueries = No

[Ensemble]
Iterations = 100
EnsembleMethod = RForest
VotingType = Majority
SelectRandomSubspaces = 0
PrintAllModels = No
PrintAllModelFiles = No
Optimize = No
OOBestimate = No
FeatureRanking = No
WriteEnsemblePredictions = No
EnsembleRandomDepth = No
BagSelection = -1
BagSize = 0
```

EK 2

```
@RELATION '"htest"-predictions'

@ATTRIBUTE DRUGID
key
@ATTRIBUTE class-a
string
@ATTRIBUTE class-a-DGs
{1,0}
@ATTRIBUTE class-a-DGs/NSADs
{1,0}
@ATTRIBUTE class-a-DGs/NSADs/N02ADs
{1,0}
@ATTRIBUTE class-a-DGs/NSADs/N03ADs
{1,0}
@ATTRIBUTE class-a-DGs/NSADs/N04ADs
{1,0}
@ATTRIBUTE class-a-DGs/NSADs/N05ADs
{1,0}
@ATTRIBUTE class-a-DGs/NSADs/N06ADs
{1,0}
@ATTRIBUTE class-a-DGs/TheotherADs
{1,0}
@ATTRIBUTE class-a-DGs/WDs
{1,0}
@ATTRIBUTE Original-p-DGs
numeric
@ATTRIBUTE Original-p-DGs/NSADs
numeric
@ATTRIBUTE Original-p-DGs/NSADs/N02ADs
numeric
@ATTRIBUTE Original-p-DGs/NSADs/N03ADs
numeric
@ATTRIBUTE Original-p-DGs/NSADs/N04ADs
numeric
@ATTRIBUTE Original-p-DGs/NSADs/N05ADs
numeric
@ATTRIBUTE Original-p-DGs/NSADs/N06ADs
numeric
@ATTRIBUTE Original-p-DGs/TheotherADs
numeric
@ATTRIBUTE Original-p-DGs/WDs
numeric
@ATTRIBUTE Original-models
string


@DATA
DB00580,DGs/WDs,1,0,0,0,0,0,0,0,1,1.0000000000000007,0.3399243797967
333,0.0758581163027852,0.08340187931591492,0.03213691603720566,0.069
05658228218117,0.07947088585864634,0.21374555257606626,0.44633006762
72004,DGs/WDs.

DB08944,DGs/WDs,1,0,0,0,0,0,0,0,1,1.0000000000000007,0.3937701639217
965,0.09789559421713423,0.0763961398374852,0.05599435000074982,0.066
17791297650204,0.09730616688992545,0.16195560506106163,0.44427423101
71417,DGs/WDs.
```

EK 2 (devam)

DB09023,DGs/WDs,1,0,0,0,0,0,0,0,1,1.0000000000000007,0.3628134954356
186,0.08811850084790658,0.051731989639843254,0.06117005612021777,0.0
7277109429602419,0.08902185453162692,0.22511965620725619,0.412066848
35712504,DGs/WDs.

D00709,DGs/NSADs/N03ADs,1,1,0,1,0,0,0,0,0,1.0000000000000007,0.41080
4196409679,0.08458163054216779,0.12365518821147124,0.043395673410497
604,0.06980853990017863,0.08936316434536389,0.21190198419235245,0.37
729381939796824,DGs/NSADs/N03ADs.

D01044,DGs/WDs,1,0,0,0,0,0,0,0,1,1.0000000000000007,0.33584422399125
674,0.06625587559815739,0.035713801875510875,0.037550414980047256,0.
13504261927844025,0.061281512259101036,0.2763745297844296,0.38778124
62243137,DGs/WDs.

D01267,DGs/NSADs/N02ADs,1,1,1,0,0,0,0,0,0,1.0000000000000007,0.23524
846043556918,0.057986186658254765,0.04069307120192814,0.021265255667
82292,0.0549290713541372,0.0603748755534262,0.34985863679801055,0.41
489290276642005,DGs/WDs.

D03165,DGs/NSADs/N04ADs,1,1,0,0,1,0,0,0,0,1.0000000000000007,0.27034
459897249685,0.09991829129082788,0.021080435866009262,0.015552519402
970368,0.09021044252038575,0.04358290989230363,0.48019871181722007,0
.24945668921028316,DGs/TheotherADs.

DB00584,DGs/TheotherADs,1,0,0,0,0,0,0,1,0,1.0000000000000007,0.26336
01832772487,0.062431642274049896,0.04287149646946293,0.0216503608352
47885,0.06917411755136761,0.06723256614712042,0.38240906618240333,0.
35423075054034775,DGs/TheotherADs.

D07310,DGs/NSADs/N05ADs,1,1,0,0,0,1,0,0,0,1.0000000000000007,0.31724
079190652354,0.07443168973861145,0.045257037871218195,0.045981415072
41732,0.0833533966943596,0.06821725252991703,0.29286322921411706,0.3
898959788793592,DGs/WDs.

D00338,DGs/WDs,1,0,0,0,0,0,0,0,1,1.0000000000000007,0.38978619901978
284,0.06520092304417212,0.10756945192413318,0.03518658938096667,0.09
905624847342752,0.08277298619708336,0.17730487425496508,0.4329089267
25252,DGs/WDs.

DB01244,DGs/WDs,1,0,0,0,0,0,0,0,1,1.0000000000000007,0.3065372456881
726,0.07484080731969335,0.03204590174830451,0.033107987827275476,0.1
0020873510836042,0.06633381368453892,0.3412600815640277,0.3522026727
4779954,DGs/WDs.

DB00642,DGs/TheotherADs,1,0,0,0,0,0,0,1,0,1.0000000000000007,0.21856
038560276897,0.05066205784298987,0.03640237376783694,0.0150196369534
68569,0.05996936204705297,0.056506954991420597,0.4659868395129977,0.
31545277488423334,DGs/TheotherADs.

DB00671,DGs/TheotherADs,1,0,0,0,0,0,0,1,0,1.0000000000000007,0.16940
400380407675,0.036954999405995836,0.033931577787987895,0.01161691205
9772175,0.045214209506907056,0.041686305043413745,0.5505389340371429
,0.2800570621587804,DGs/TheotherADs.

DB01060,DGs/TheotherADs,1,0,0,0,0,0,0,1,0,1.0000000000000007,0.22203
777181395234,0.04862824837099828,0.04548139022767077,0.0215027478506
18478,0.04648432502054433,0.05994106034412044,0.4146829367249496,0.3
6327929146109794,DGs/TheotherADs.

EK 2 (devam)

DB00801,DGs/WDs,1,0,0,0,0,0,0,0,1,1.0000000000000007,0.3897861990197
8284,0.06520092304417212,0.10756945192413318,0.03518658938096667,0.0
9905624847342752,0.08277298619708336,0.17730487425496508,0.432908926
725252,DGs/WDs.

D03215,DGs/NSADs/N06ADs,1,1,0,0,0,0,1,0,0,1.0000000000000007,0.37710
325112400556,0.08386988175951154,0.10457656324049604,0.0338593860682
8071,0.05771712526881006,0.09708029478690743,0.191184629903309,0.431
7121189726853,DGs/WDs.

D02096,DGs/NSADs/N03ADs,1,1,0,1,0,0,0,0,0,1.0000000000000007,0.30093
75908583687,0.06025487664693262,0.07573690740113466,0.02254943985315
5797,0.06496756498420789,0.07742880197293779,0.32150831495970505,0.3
775540941819259,DGs/WDs.

D00498,DGs/NSADs/N02ADs,1,1,1,0,0,0,0,0,0,1.0000000000000007,0.44518
79714269012,0.13893046823260657,0.05565728616643118,0.09029259891178
21,0.059506742289861005,0.10080087582622038,0.17568341161875994,0.37
912861695433897,DGs/NSADs/N02ADs.

D02578,DGs/NSADs/N06ADs,1,1,0,0,0,0,1,0,0,1.0000000000000007,0.35238
316114869256,0.08879448019782567,0.08766520904539496,0.0388014544235
7095,0.05971746884074283,0.07740454864115831,0.1966353295563852,0.45
09815092949222,DGs/WDs.

DB01160,DGs/TheotherADs,1,0,0,0,0,0,0,1,0,1.0000000000000007,0.19366
253635628622,0.044524261658634365,0.04192908160401961,0.016817451755
02571,0.05109516688463791,0.03929657445396866,0.5035034482132902,0.3
0283401543042354,DGs/TheotherADs.

D05592,DGs/NSADs/N06ADs,1,1,0,0,0,0,1,0,0,1.0000000000000007,0.35528
309835047744,0.08617481672537416,0.07456223178710224,0.0382408154710
53034,0.06828931833675861,0.08801591603018953,0.22932619912790725,0.
41539070252161514,DGs/WDs.

DB01251,DGs/TheotherADs,1,0,0,0,0,0,0,1,0,1.0000000000000007,0.19383
002227759813,0.052283996151212514,0.02684845628750169,0.012038089690
592028,0.06622524265249302,0.036434237495798875,0.5147018905924066,0
.2914680871299953,DGs/TheotherADs.

D02575,DGs/NSADs/N06ADs,1,1,0,0,0,0,1,0,0,1.0000000000000007,0.26144
173907112306,0.04145571551193354,0.041740382367144696,0.021213601957
23416,0.09552851350021742,0.061503525734593265,0.38153237997472655,0
.3570258809541502,DGs/TheotherADs.

D00228,DGs/NSADs/N06ADs,1,1,0,0,0,0,1,0,0,1.0000000000000007,0.36241
836244064085,0.0758025862522148,0.08104552037670602,0.03618861101111
143,0.0773502485319578,0.09203139626865091,0.17943668943226693,0.458
14494812709217,DGs/WDs.

DB00323,DGs/WDs,1,0,0,0,0,0,0,0,1,1.0000000000000007,0.3020472032326
9285,0.08318971739152589,0.061062786800355144,0.026748117341718388,0
.053352117710316775,0.0776944639887767,0.2614432516172928,0.43650954
5150014,DGs/WDs.

EK 2 (devam)

D00987,DGs/NSADs/N04ADs,1,1,0,0,1,0,0,0,0,1.0000000000000007,0.32533534197025615,0.07433519680901134,0.04849405581550574,0.029993530723745384,0.10836129744794273,0.06415126117405101,0.3372991616767046,0.33736549635303925,DGs/WDs.

D04226,DGs/NSADs/N02ADs,1,1,1,0,0,0,0,0,0,1.0000000000000007,0.3053125337668589,0.06434183784568981,0.05398818491178492,0.025440723746021392,0.07606400676924578,0.0854777804941171,0.3057747789791388,0.38891268725400213,DGs/WDs.

DB08969,DGs/WDs,1,0,0,0,0,0,0,0,1,1.0000000000000007,0.3540013485386568,0.07980769676524839,0.0896114629765259,0.0376541641315469,0.05657704234878726,0.09035098231654826,0.25042459383414073,0.39557405762720244,DGs/WDs.

DB00581,DGs/TheotherADs,1,0,0,0,0,0,0,1,0,1.0000000000000007,0.20352722481063976,0.053013167290687016,0.047042292143684016,0.01856202872101727,0.03667111176576804,0.04823862488948347,0.5062645588943366,0.29020821629502364,DGs/TheotherADs.

DB01088,DGs/TheotherADs,1,0,0,0,0,0,0,1,0,1.0000000000000007,0.22049616604024122,0.0508360817394513,0.04293273096770143,0.021860590092860954,0.05598043410867826,0.04888632913154934,0.44929938310181505,0.33020044508579438,DGs/TheotherADs.

DB04743,DGs/WDs,1,0,0,0,0,0,0,0,1,1.0000000000000007,0.3172259262127293,0.07540279571942048,0.07823972932998309,0.03555449509862799,0.05566370086347688,0.07236520520122085,0.27714891715593276,0.4056251566313378,DGs/WDs.

D07335,DGs/WDs,1,0,0,0,0,0,0,0,1,1.0000000000000007,0.31470001474734144,0.0556154765689357,0.058099956990759795,0.03185857616977733,0.08266664522575639 9,0.08645955276030463,0.27550515501843725,0.40979483023422114,DGs/WDs.

D02625,DGs/NSADs/N05ADs,1,1,0,0,0,1,0,0,0,1.0000000000000007,0.39941177479614404,0.07050064286401711,0.044695842428829284,0.04045043521231774,0.17458233422102823,0.06918252006995174,0.24249675024251913,0.3580914749613369,DGs/NSADs/N05ADs.

D04747,DGs/NSADs/N06ADs,1,1,0,0,0,0,1,0,0,1.0000000000000007,0.33011288203871314,0.07095035797846917,0.0720452400418496 5,0.0397276689702 9813 4,0.060725827178209235,0.08666378786988706,0.27536028097430415,0.3945268369869826,DGs/WDs.

DB01117,DGs/TheotherADs,1,0,0,0,0,0,0,1,0,1.0000000000000007,0.2514399840228903,0.0558831190396229,0.04832234598962025,0.023016585230721535,0.06875019042559254,0.05546774333733308,0.334111524457257,0.414484915198526,DGs/WDs.

DB00848,DGs/WDs,1,0,0,0,0,0,0,0,1,1.0000000000000007,0.3615162564747042,0.07385616096883996,0.10304845757424949,0.03700145769666638,0.05801013234242889,0.08960004789251957,0.1799633577734136,0.45852038575188203,DGs/WDs.

EK 2 (devam)

D02068,DGs/WDs,1,0,0,0,0,0,0,0,1,1.0000000000000007,0.34970708423558
91,0.09135404939120581,0.08694075460159806,0.03568255205518925,0.052
1664474026838,0.08356328078491208,0.19468739590994652,0.455605519854
4643,DGs/WDs.

DB00734,DGs/TheotherADs,1,0,0,0,0,0,0,1,0,1.0000000000000007,0.37762
02546920979,0.05750696499591117,0.042557129345066565,0.0254388404481
50073,0.171236574404438,0.08088074549853208,0.2555384894181252,0.366
8412558897769,DGs/NSADs/N05ADs.

D07132,DGs/NSADs/N02ADs,1,1,1,0,0,0,0,0,0,1.0000000000000007,0.31872
422944075746,0.11383824029834119,0.03066540816571874,0.0437845807035
0748,0.0627571679138212,0.06767883235936901,0.3362527981070829,0.345
0229724521595,DGs/WDs.

D05200,DGs/WDs,1,0,0,0,0,0,0,0,1,1.0000000000000007,0.37702028483165
37,0.0978668690780777,0.07068119148086952,0.05153229616598243,0.0627
7589246024937,0.0941640356464747,0.16201435550623838,0.4609653596621
078,DGs/WDs.

DB00597,DGs/TheotherADs,1,0,0,0,0,0,0,1,0,1.0000000000000007,0.23406
82545565998,0.06507827285527276,0.030337609408433233,0.0251305677573
61454,0.0671915129166088,0.046330291618923575,0.4746437402891596,0.2
9128800515424047,DGs/TheotherADs.

DB01145,DGs/TheotherADs,1,0,0,0,0,0,0,1,0,1.0000000000000007,0.24464
65172747905,0.05496209299961027,0.04120246307985484,0.01813078866271
101,0.06403778915976435,0.06631338337285002,0.40135164820553954,0.35
400183451966966,DGs/TheotherADs.

D07302,DGs/NSADs/N04ADs,1,1,0,0,1,0,0,0,0,1.0000000000000007,0.31692
48217164979,0.06816589702077244,0.04456322196662893,0.04152554769599
96,0.09060843539928977,0.0720617196338071,0.27687324054473517,0.4062
01937738767,DGs/WDs.

D04924,DGs/NSADs/N02ADs,1,1,1,0,0,0,0,0,0,1.0000000000000007,0.41950
022148470123,0.11824368853998506,0.07818964911849956,0.0660961882551
0022,0.0630140030959503,0.09395669247516637,0.1641914228949918,0.416
3083556203068,DGs/NSADs/N02ADs.

D00786,DGs/WDs,1,0,0,0,0,0,0,0,1,1.0000000000000007,0.30204720323269
285,0.08318971739152589,0.061062786800355144,0.026748117341718388,0.
053352117710316775,0.0776944639887767,0.2614432516172928,0.436509545
150014,DGs/WDs.

D05623,DGs/NSADs/N06ADs,1,1,0,0,0,0,1,0,0,1.0000000000000007,0.38414
563661900153,0.09194336919385634,0.07037067671785777,0.0594034052744
3601,0.06490307072444396,0.09752511470840759,0.1655655071037696,0.45
028885627722876,DGs/WDs.

D02609,DGs/NSADs/N05ADs,1,1,0,0,0,1,0,0,0,1.0000000000000007,0.40215
06010686958,0.06544934261694216,0.04236209560048197,0.04973149527286
604,0.16470730064833577,0.07990036693007008,0.21328013191421108,0.38
456926701709276,DGs/NSADs/N05ADs.

EK 2 (devam)

D07299,DGs/NSADs/N03ADs,1,1,0,1,0,0,0,0,0,1.0000000000000007,0.34761
67670176354,0.07486711633125256,0.09185094915080137,0.04238960796292
945,0.056054569028823525,0.08245452454382861,0.23518304812475907,0.4
172001848576054,DGs/WDs.

DB00372,DGs/WDs,1,0,0,0,0,0,0,0,1,1.0000000000000007,0.3766751705525
717,0.06447535787116897,0.03680021134093415,0.04111135212067547,0.16
49459805547068,0.06934226866508646,0.2757501974652108,0.347574631982
2173,DGs/NSADs/N05ADs.

D02574,DGs/NSADs/N06ADs,1,1,0,0,0,0,1,0,0,1.0000000000000007,0.37909
73206067947,0.08933876144869621,0.08429949820687305,0.04977020776511
461,0.060695775645114754,0.09499307754099627,0.17610494050907813,0.4
44797738884127,DGs/WDs.

DB00790,DGs/TheotherADs,1,0,0,0,0,0,0,1,0,1.0000000000000007,0.26357
79148721642,0.05105907555027411,0.04320840350879526,0.02026985007350
973,0.061884766973790695,0.08715581876579429,0.39194709917185844,0.3
4447498595597736,DGs/TheotherADs.

D02623,DGs/NSADs/N05ADs,1,1,0,0,0,1,0,0,0,1.0000000000000007,0.37361
154505512484,0.06820897619735046,0.04330110558672402,0.0402662246860
0195,0.15327113956772825,0.06856409901732015,0.2516698914288182,0.37
4718563516057,DGs/WDs.

DB01238,DGs/TheotherADs,1,0,0,0,0,0,0,1,0,1.0000000000000007,0.39564
601628196544,0.05253820120009315,0.04026362827183691,0.0244644964377
34826,0.2077102921490558,0.07066939822324497,0.2918485585722532,0.31
250542514578106,DGs/NSADs/N05ADs.

DB00709,DGs/TheotherADs,1,0,0,0,0,0,0,1,0,1.0000000000000007,0.33340
1902172877,0.07557220286299483,0.09937679101438972,0.028837604863880
56,0.05772635785222428,0.07188894557938769,0.3105002945576586,0.3560
9780326946433,DGs/WDs.

D10509,DGs/NSADs/N06ADs,1,1,0,0,0,0,1,0,0,1.0000000000000007,0.38776
24605308574,0.08011159320362254,0.11470709880938745,0.03579679670107
622,0.06011726908450574,0.09702970273226566,0.1952013044601114,0.417
03623500903086,DGs/WDs.

D02680,DGs/NSADs/N05ADs,1,1,0,0,0,1,0,0,0,1.0000000000000007,0.33619
277076149795,0.04665726335664178,0.02481520348057759,0.0259894669823
29613,0.18302103878281314,0.05570979815913587,0.35017671502057884,0.
31363051421792315,DGs/TheotherADs.

DB01112,DGs/TheotherADs,1,0,0,0,0,0,0,1,0,1.0000000000000007,0.19168
55460853707,0.040747903386684194,0.0337679782481194,0.01261465226825
29,0.044965751738299205,0.05958926044401495,0.5208843453005267,0.287
4301086141028,DGs/TheotherADs.

DB08994,DGs/WDs,1,0,0,0,0,0,0,0,1,1.0000000000000007,0.3641251135609
7715,0.09120775453388646,0.054187653566180445,0.06847364632791751,0.
06549648253495981,0.08475957659803302,0.22748561254903085,0.40838927
38899918,DGs/WDs.

EK 2 (devam)

DB08905,DGs/WDs,1,0,0,0,0,0,0,0,1,1.0000000000000007,0.2991430790215
116,0.09685548342785244,0.06146059142138474,0.029051443418119056,0.0
4869387566256572,0.06308168509158966,0.23411000591004172,0.466746915
0684468,DGs/WDs.

DB00728,DGs/TheotherADs,1,0,0,0,0,0,0,1,0,1.0000000000000007,0.20858
499967293886,0.05774051781895473,0.02453969418736349,0.0209665700226
4096,0.05417111949712219,0.051167098146857444,0.4652820094627906,0.3
261329908642705,DGs/TheotherADs.

D07285,DGs/NSADs/N02ADs,1,1,1,0,0,0,0,0,0,1.0000000000000007,0.30126
363762893804,0.08788581215434485,0.03019219246150666,0.0296639877302
01295,0.0828459017147006,0.07067574356818479,0.3986759665070417,0.30
00603958640201,DGs/TheotherADs.

DB00753,DGs/TheotherADs,1,0,0,0,0,0,0,1,0,1.0000000000000007,0.37386
35254846819,0.0827597566273761,0.09730803988197172,0.040001591738587
1,0.061556681045137215,0.09223745619160972,0.24345906979623724,0.382
6774047190809,DGs/WDs.

D04882,DGs/NSADs/N05ADs,1,1,0,0,0,1,0,0,0,1.0000000000000007,0.43127
11360059752,0.08512460900821744,0.09282609999504313,0.06561345073993
022,0.0886466529054508,0.09906032335733383,0.14785078819352476,0.420
8780758004997,DGs/NSADs/N06ADs.

D01811,DGs/NSADs/N02ADs,1,1,1,0,0,0,0,0,0,1.0000000000000007,0.40159
657245143837,0.12339375791160608,0.10466426872617872,0.03092425571750
1313,0.06443482316913259,0.07817946692701976,0.22310032383045816,0.
3753031037181034,DGs/NSADs/N02ADs.

DB01141,DGs/TheotherADs,1,0,0,0,0,0,0,1,0,1.0000000000000007,0.17622
531339414352,0.04316553215086204,0.023507826766494683,0.010051537939
924743,0.05972456858432059,0.039775847952541465,0.5669199877927846,0
.2568546988130718,DGs/TheotherADs.

D07348,DGs/NSADs/N06ADs,1,1,0,0,0,0,1,0,0,1.0000000000000007,0.34897
27383585105,0.07964066624580333,0.09115182459892666,0.03083165280865
5685,0.05295897252740155,0.0943896221777234,0.2014024627898794,0.449
6247988516099,DGs/WDs.

D00059,DGs/NSADs/N04ADs,1,1,0,0,1,0,0,0,0,1.0000000000000007,0.33369
07175660102,0.07928182112075227,0.0857378306740387,0.028472468752885
83,0.05181908786301607,0.08837950915531735,0.30607419429881816,0.360
23508813517136,DGs/WDs.

DB06789,DGs/WDs,1,0,0,0,0,0,0,0,1,1.0000000000000007,0.21406152931628
927,0.043744927808053495,0.03302173373024354,0.015883021816481095,0
.08021287531837622,0.041198970643134936,0.4800705527938217,0.3058679
178898891,DGs/TheotherADs.

D03089,DGs/NSADs/N05ADs,1,1,0,0,0,1,0,0,0,1.0000000000000007,0.43058
84466199023,0.13030348726152435,0.09620355340735268,0.05072239731683
357,0.0592260466978561,0.09413296193633577,0.1860622556993626,0.3833
492976807349,DGs/NSADs/N02ADs.

EK 2 (devam)

D07304,DGs/NSADs/N04ADs,1,1,0,0,1,0,0,0,0,1.0000000000000007,0.35655
31310364909,0.08727046415588638,0.08884677011326898,0.03088839511598
0082,0.05209184362944839,0.09745565802190718,0.26927727521222167,0.3
741695937512873,DGs/WDs.

D00681,DGs/NSADs/N02ADs,1,1,1,0,0,0,0,0,0,1.0000000000000007,0.35266
253400749537,0.08937544793010592,0.04718824599532103,0.0422950320149
7059,0.09837682653642336,0.07542698153067458,0.2827733429004841,0.36
45641230920204,DGs/WDs.

DB00688,DGs/TheotherADs,1,0,0,0,0,0,0,1,0,1.0000000000000007,0.27577
008264208436,0.05895618037257964,0.03907715727551826,0.0268629437670
0278,0.08271426540252837,0.06815953582445528,0.4095874046833834,0.31
464251267453214,DGs/TheotherADs.

D00556,DGs/WDs,1,0,0,0,0,0,0,0,1,1.0000000000000007,0.34920260754900
02,0.10552372536090773,0.06262753627975334,0.038898771975190975,0.06
399692604154264,0.0781556478916056,0.166316938752401,0.4844804536985
988,DGs/WDs.

D02608,DGs/NSADs/N05ADs,1,1,0,0,0,1,0,0,0,1.0000000000000007,0.40156
982424008913,0.07268703120031665,0.04649721926548092,0.0529536365189
2405,0.14368275041617753,0.08574918683919017,0.21162711040665702,0.3
868030653532536,DGs/NSADs/N05ADs.

DB01546,DGs/WDs,1,0,0,0,0,0,0,0,1,1.0000000000000007,0.3760409834246
0063,0.0808214010834255,0.08317441464948228,0.039934142766151114,0.0
6099873505091111,0.11111228987463076,0.2001071745291619,0.4238518420
4623727,DGs/WDs.

DB00588,DGs/TheotherADs,1,0,0,0,0,0,0,1,0,1.0000000000000007,0.13089
30518814395,0.036503941883129894,0.023125503191138968,0.011365024451
670336,0.034921527567283235,0.02497705478821704,0.5729683653773925,0
.296138582741168,DGs/TheotherADs.

D06147,DGs/NSADs/N02ADs,1,1,1,0,0,0,0,0,0,1.0000000000000007,0.43935
887589465306,0.12604562389815038,0.06677129641745053,0.0721338572152
635,0.06743117019373554,0.10697692817005348,0.14892089317430196,0.41
17202309310448,DGs/NSADs/N02ADs.

DB00532,DGs/WDs,1,0,0,0,0,0,0,0,1,1.0000000000000007,0.4200986556369
616,0.07438194013110254,0.1488160956285493,0.03724800206007722,0.073
45323121944027,0.0861993865977924,0.17521878957992026,0.404682554783
11793,DGs/NSADs/N03ADs.

DB04813,DGs/WDs,1,0,0,0,0,0,0,0,1,1.0000000000000007,0.3044475231229
128,0.0753440177781157,0.07145495379644781,0.03446796243683974,0.052
227938617174736,0.0709526504943348,0.2185084668909354,0.477044009986
15176,DGs/WDs.

DB09185,DGs/WDs,1,0,0,0,0,0,0,0,1,1.0000000000000007,0.3610018413828
5696,0.07977497989961645,0.0808475544659474,0.03880472408379426,0.05
442654170896746,0.10714804122453149,0.20240249020674908,0.4365956684
1039393,DGs/WDs.

EK 2 (devam)

DB00615,DGs/TheotherADs,1,0,0,0,0,0,0,1,0,1.0000000000000007,0.20237
697998104964,0.0679302108731796,0.02013281977790428,0.01239873411881
0721,0.05865650320922199,0.043258712001933014,0.5288054524043136,0.2
6881756761463693,DGs/TheotherADs.

DB01245,DGs/TheotherADs,1,0,0,0,0,0,0,1,0,1.0000000000000007,0.31577
336071123163,0.054882150053177564,0.052401322060965684,0.02881387252
023643,0.08578249481459949,0.09389352126225231,0.2892450618892829,0.
3949815773994856,DGs/WDs.

DB00614,DGs/TheotherADs,1,0,0,0,0,0,0,1,0,1.0000000000000007,0.29259
66766272949,0.06272760459635714,0.08540129545771107,0.02321496413941
3867,0.052470458795909655,0.06878235363790314,0.35009942270898314,0.
3573039006637219,DGs/WDs.

DB01193,DGs/TheotherADs,1,0,0,0,0,0,0,1,0,1.0000000000000007,0.26495
96522446206,0.060927931547871024,0.04319976346664711,0.02474872189739
5687,0.07372509261699967,0.062358142715707145,0.40804568262581065,0
.32699466512956865,DGs/TheotherADs.

DB01388,DGs/WDs,1,0,0,0,0,0,0,0,1,1.0000000000000007,0.2576919544268
141,0.06241480264791561,0.031053724300881023,0.024708650581746292,0.
08598639029014117,0.053528386606130025,0.416029434310537,0.326278611
2626488,DGs/TheotherADs.

DB01079,DGs/WDs,1,0,0,0,0,0,0,0,1,1.0000000000000007,0.2467353062063
8115,0.05443780796395976,0.048022339239211134,0.02219526969192996,0.
0532922309573508,0.06878765835392947,0.38388148148997875,0.369383212
30364013,DGs/TheotherADs.

DB09004,DGs/WDs,1,0,0,0,0,0,0,0,1,1.0000000000000007,0.3784378241325
744,0.10930276393993947,0.057592594188108685,0.05826376427778617,0.0
6372578201606263,0.08955291971067746,0.2037360822962125,0.4178260935
7121303,DGs/WDs.

DB00646,DGs/TheotherADs,1,0,0,0,0,0,0,1,0,1.0000000000000007,0.14617
7284344115,0.043830471925634044,0.025244826436591202,0.0091132323254
59946,0.03570258015935128,0.032286173497078546,0.6178836845303438,0.
2359390311255414,DGs/TheotherADs.

D02671,DGs/NSADs/N04ADs,1,1,0,0,1,0,0,0,0,1.0000000000000007,0.39725
777827490044,0.07976924988927289,0.05012595729784935,0.0572969539275
5007,0.13124827924064653,0.07881733791958184,0.18662357872608626,0.4
161186429990131,DGs/WDs.

DB01172,DGs/TheotherADs,1,0,0,0,0,0,0,1,0,1.0000000000000007,0.15617
610027702178,0.0333572497217145,0.027834993090369776,0.0129949552111
64281,0.04010482243727744,0.041884079816495834,0.5942895796515164,0.
24953432007146192,DGs/TheotherADs.

DB00431,DGs/WDs,1,0,0,0,0,0,0,0,1,1.0000000000000007,0.3842472466634
24,0.08283103212733854,0.1137852427986724,0.039463850302544956,0.064
17595433809684,0.08399116709677135,0.19385658843566825,0.42189616490
09075,DGs/WDs.

EK 2 (devam)

DB00684,DGs/TheotherADs,1,0,0,0,0,0,0,1,0,1.0000000000000007,0.15037
46145386743,0.03354084733009609,0.02646850708916264,0.01225997877754
2913,0.03832981083606789,0.03977547050580482,0.6103415981861546,0.23
928378727517108,DGs/TheotherADs.

D00394,DGs/NSADs/N06ADs,1,1,0,0,0,0,1,0,0,1.0000000000000007,0.43568
58177573898,0.08789965347434671,0.07488551895586165,0.07993506866903
63,0.08951708826138612,0.1034484883967593,0.16618810986659638,0.3981
260723760136,DGs/NSADs/N06ADs.

DB01101,DGs/TheotherADs,1,0,0,0,0,0,0,1,0,1.0000000000000007,0.25962
104850621065,0.047532568319822256,0.05026065315843046,0.017302028846
419754,0.06455670854483998,0.07996908963669819,0.4474872638438948,0.
2928916876498944,DGs/TheotherADs.

DB08989,DGs/WDs,1,0,0,0,0,0,0,0,1,1.0000000000000007,0.3996004520023
659,0.11644704617599162,0.09947710894533325,0.03621988585499972,0.06
520301226523673,0.0822533987608048,0.18784956763995694,0.41254998035
7677,DGs/WDs.

D01190,DGs/NSADs/N03ADs,1,1,0,1,0,0,0,0,0,1.0000000000000007,0.39892
74846791635,0.08190492179898315,0.12997002858058965,0.04053846007354
964,0.07373697380534543,0.07277710042069568,0.19654244413147254,0.40
45300711893638,DGs/WDs.

DB01118,DGs/TheotherADs,1,0,0,0,0,0,0,1,0,1.0000000000000007,0.28794
169722661,0.05530470884816864,0.029847120339729586,0.025743116251905
594,0.10929461107085618,0.06775214071595008,0.3432432891525249,0.368
8150136208649,DGs/WDs.

DB00701,DGs/TheotherADs,1,0,0,0,0,0,0,1,0,1.0000000000000007,0.18237
763200245602,0.04721112781983154,0.02578284373495691,0.0103182602384
40933,0.046120240866533456,0.052945159342693136,0.5391935444949374,0
.27842882350260684,DGs/TheotherADs.

DB00631,DGs/TheotherADs,1,0,0,0,0,0,0,1,0,1.0000000000000007,0.24870
469363884679,0.051266057636983664,0.04998220664878899,0.017992993560
901088,0.045706596997876735,0.0837568387942963,0.44373658612077227,0
.30755872024038106,DGs/TheotherADs.

D07306,DGs/NSADs/N04ADs,1,1,0,0,1,0,0,0,0,1.0000000000000007,0.39726
96414920533,0.0994319483774853,0.057294212501425426,0.07207139292911
686,0.06911180441621527,0.09936028326781067,0.18912851199474856,0.41
360184651319803,DGs/WDs.

DB00729,DGs/WDs,1,0,0,0,0,0,0,0,1,1.0000000000000007,0.3635556720876
783,0.07685721703335685,0.08730602144855826,0.04472414326904948,0.05
878064267799156,0.09588764765872229,0.1918461295406036,0.44459819837
17178,DGs/WDs.

DB04832,DGs/WDs,1,0,0,0,0,0,0,0,1,1.0000000000000007,0.3521987821456
936,0.07890097935891764,0.060089874505905916,0.04889107423460938,0.0
7302983748931222,0.09128701655694854,0.17174383212312289,0.476057385
73118334,DGs/WDs.

EK 2 (devam)

DB01211,DGs/TheotherADs,1,0,0,0,0,0,0,1,0,1.0000000000000007,0.18825
24698770836,0.05080657619982978,0.021846600089823174,0.0314593858085
3735,0.050977217782711984,0.03316268999618124,0.5049272317008353,0.3
068202984220809,DGs/TheotherADs.

D02536,DGs/NSADs/N06ADs,1,1,0,0,0,0,1,0,0,1.0000000000000007,0.25911
221888667174,0.06727503343546601,0.04935343790158225,0.0235313071008
5734,0.05314029911101652,0.06581214133774956,0.3758675561724061,0.36
5020224940922,DGs/TheotherADs.

DB04898,DGs/WDs,1,0,0,0,0,0,0,0,1,1.0000000000000007,0.2150070967695
3286,0.05850023712684523,0.028178050260463196,0.013973371645584048,0
.063484798790649,0.05087063894599134,0.5226552425497504,0.2623376606
8071684,DGs/TheotherADs.

D07313,DGs/NSADs/N05ADs,1,1,0,0,0,1,0,0,0,1.0000000000000007,0.42882
410761032697,0.0857627435773065,0.06341012940921822,0.07929228752987
938,0.10557759133220422,0.09478135576171896,0.16818396903761457,0.40
29919233520582,DGs/NSADs/N05ADs.

DB01259,DGs/TheotherADs,1,0,0,0,0,0,0,1,0,1.0000000000000007,0.22719
426590927502,0.0373411431459561,0.0246947517832895,0.012834390135380
491,0.07935407639317835,0.07296990445147056,0.47140448706240223,0.30
14012470283228,DGs/TheotherADs.

D00538,DGs/NSADs/N03ADs,1,1,0,1,0,0,0,0,0,1.0000000000000007,0.35363
508698993984,0.08137901390401842,0.09585748703877134,0.02846549089923
822,0.06810462685183426,0.07982846829607773,0.24605459104346153,0.4
003103219665985,DGs/WDs.

DB01218,DGs/TheotherADs,1,0,0,0,0,0,0,1,0,1.0000000000000007,0.30226
056421730796,0.06588108336326863,0.026456990624716556,0.039350896245
08045,0.1219617447082529,0.048609849275989436,0.37837214734032215,0.
31936728844236983,DGs/TheotherADs.

DB01220,DGs/TheotherADs,1,0,0,0,0,0,0,1,0,1.0000000000000007,0.16996
679685143967,0.05334460292879058,0.021156127281114653,0.008639909044
568739,0.04415063549931373,0.04267552209765197,0.5820202094899813,0.
24801299365857907,DGs/TheotherADs.

DB04824,DGs/WDs,1,0,0,0,0,0,0,0,1,1.0000000000000007,0.3460914048680
8154,0.11904365351900777,0.0667877825521533,0.03442094603272652,0.05
3157541570116776,0.07268148119407726,0.24712961822833493,0.406778976
90358344,DGs/WDs.

DB01200,DGs/TheotherADs,1,0,0,0,0,0,0,1,0,1.0000000000000007,0.28111
2157500925,0.100252739451363,0.021080435866009262,0.0155525194029703
68,0.09841780037991753,0.045808662400664835,0.4714880094760828,0.247
39983302299218,DGs/TheotherADs.

ÖZGEÇMİŞ

Ad-Soyad : **Aytun Onay**

Uyruğu : **Türkiye (T.C.)**

Doğum Tarihi ve Yeri : **Ağustos 9, 1979, Landshud, Almanya**

E-posta : **koaytun@gmail.com, aonay@etu.edu.tr**

ÖĞRENİM DURUMU:

- **Lisans :** 2003, Ege Üniversitesi, Fen Fakültesi, Astronomi ve Uzay Bilimleri Bölümü ve Fen Fakültesi, Matematik Bölümü Çift Anadal
- **Yüksek Lisans :** 2008, Orta Doğu Teknik Üniversitesi, Fen Bilimleri Enstitüsü, Fizik Bölümü
- **Doktora :** 2017, TOBB Ekonomi ve Teknoloji Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Bölümü

MESLEKİ DENEYİM :

Yıl	Yer	Görev
2006-2008	**Orta Doğu Teknik Üniversitesi**	**Proje Asistanlığı**
2010-2016	**TOBB Ekonomi ve Teknoloji Üniversitesi**	**Burslu Doktora Öğrencisi**

BURSLAR VE ÖDÜLLER :

- 2006-2008, TÜBİTAK 1001-Proje Bursu, TBAG-107T142, Nanosistemlerin Hidrojen Depolama Kapasitesi: Moleküler Dinamik Simülasyonları.
- 2006-2008, ODTÜ BAP-Proje Bursu, 2006-07-02-00-01, Farklı Yapılardaki Boron Nitrid Katkılı Karbon Nanotüplerin Hidrojen Depolama Kapazitesinin DFT Hesaplama ile İncelenmesi.
- 2010-2016, TOBB Ekonomi ve Teknoloji Üniversitesi, Tam-Burslu Doktora Programı.

- 2003, Ege Üniversitesi, Fen Fakültesi, Astronomi ve Uzay Bilimleri Bölüm 3.cülüğü.
- 2000-2003, İzmir Ticaret Odası, Üniversite Öğrencileri İçin Üstün Başarı Bursu.

- 2006, İstanbul'da Nükleer Kolektif Dinamikler Yaz Okulu III Seyahat Bursu.
- 2011, İtalya 36.FEBS Kongresi Seyahat Bursu.

- 2012, San Diego, CA, ABD'de Biyoyakıt ve Biyoürünlerle ilgili 2. Uluslararası Konferans için Seyahat Bursu.

- 2012, İspanya'daki 37. FEBS Kongresi Seyahat Bursu.
- 2013, St. Petersburg'daki 38. FEBS Kongresi için Seyahat Bursu.

YABANCI DİL:

ÜDS İngilizce Sınavı : 87.5/100

TEZDEN TÜRETİLEN YAYINLAR VE SUNUMLAR :

- **Onay, A.**, Onay, M., Abul, O., 2016. Determination of subgraph fragments and categorization of drugs via data mining approaches, 1st International Mediterranean Science and Engineering Congress (IMSEC 2016), October 26-28, Paper ID 680, Pages 2338-2347, Adana, Turkey.
- **Onay, A.**, Onay, M., Abul, O., 2017. Classification of Nervous System Withdrawn and Approved Drugs with ToxPrint Features via Machine Learning Strategies, Computer Methods and Programs in Biomedicine, 142, 9-19.

DİĞER YAYINLAR VE SUNUMLAR:

- **Koyuncular, O. A.,** Erkoç, Ş., 2009. Enhancement of H_2 Storage in Carbon Nanotubes via Doping with a Boron Nitride Ring, Journal of Computational and Theoretical Nanoscience, 6, 933-941.

- Molecular Dynamics Workshop-I, 2010. TOBB University of Economics and Technology, September 13-17, Ankara, Turkey.

- Summer Scholl III on Nuclear Collective Dynamics, 2006. Feza Gürsey Instıtute, l, June 16, İstanbul,Turkey.

- **Koyuncular, O.A.**, Erkoç, Ş., 2010. Increasing of H_2 Storage in Carbon Nanotubes via Doping with a Boron Nitride Ring by Molecular Mechanics Methods, Turkish Journal of Biochemistry, volume 35, Special Issue, 2010. 22th National Biochemistry Congress, October 27-30, Eskişehir, Turkey.

- **Koyuncular, O.A.**, Erkoç, Ş., Investigation for Hydrogen Storage Capacity of Different Structures of Carbon Nanotubes Substitutionally Doped with Boron

Nitride (CBN Nanotubes) with DFT Calculation, 2011. Turkish Journal of Biochemistry, volume 36, Special Issue. 23th National Biochemistry Congress, 29 November-2 December, Adana, Turkey.

- **Koyuncular, O. A.,** Erkoç, Ş., 2011. Increasing of H_2 Storage in Carbon Nanotubes CNT(7,0), CNT(4,4), CNT(4,2) via Doping with a Boron Nitride Ring by Molecular Dynamics Simulation Studies, The FEBS Journal, Volume 278, Supplement 1, 36th FEBS Congress Biochemistry for Tomorrow's Medicine, June 25-30, Torino, Italy.
- **Onay, A.,** Abul, O.,2012. Computational Investigation on Flavonoids in Microalgae: A DFT Study, The 2nd International Conference on Algal Biomass, Biofuels and Bioproducts, June 10-13, San Diego, California, USA.

- **Onay, A.,** Abul, O., 2012. Theoretical investigations on o-substitue derivatives of ısoflavans: SAR and QSAR Study. The FEBS Journal, Volume 279, Supplement 1, 37th FEBS Congress from Single Molecules to Systems Bıology, September 4-9, Seville, Spain.
- **Onay, A.,** Abul, O., 2013. The computational investigation of new inhibitors for aldose reductase, The FEBS Journal, Volume 280, Supplement 1, 38th FEBS Congress Mechanisms in Biology, July 6-11, St. Petersburg, Russia.