T.C.
NECMETTİN ERBAKAN
ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

VERİ MADENCİLİĞİ KULLANARAK AChE ENZİMİYLE ETKİLEŞEN MOLEKÜLLERİN BAĞLANMA EĞİLİMİNİN TAHMİNİ

Merve YENEN

YÜKSEK LİSANS TEZİ
Bilgisayar Mühendisliği Anabilim Dalı

Eylül-2024
KONYA

TEZ KABUL VE ONAYI

Merve YENEN tarafından hazırlanan "Veri Madenciliği Kullanarak AChE Enzimiyle Etkileşen Moleküllerin Bağlanma Eğiliminin Tahmini" adlı tez çalışması 24/09/2024 tarihinde aşağıdaki jüri tarafından oy birliği ile Necmettin Erbakan Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı'nda YÜKSEK LİSANS TEZİ olarak kabul edilmiştir.

Jüri Üyeleri	**İmza**
Başkan Prof. Dr. Gülay TEZEL	……..
Danışman Dr. Öğr. Üyesi Özlem ERDAŞ ÇİÇEK	……..
Üye Dr. Öğr. Üyesi Ayşe Merve ACILAR	……..

Fen Bilimleri Enstitüsü Yönetim Kurulu'nun …./…/20.. gün ve …….. sayılı kararıyla onaylanmıştır.

Prof. Dr. Havvanur UÇBEYİAY
FBE Müdürü

TEZ BİLDİRİMİ

Bu tezdeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edildiğini ve tez yazım kurallarına uygun olarak hazırlanan bu çalışmada bana ait olmayan her türlü ifade ve bilginin kaynağına eksiksiz atıf yapıldığını bildiririm.

DECLARATION PAGE

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Merve YENEN

Tarih:

ÖZET

YÜKSEK LİSANS TEZİ

VERİ MADENCİLİĞİ KULLANARAK AChE ENZİMİYLE ETKİLEŞEN MOLEKÜLLERİN BAĞLANMA EĞİLİMİNİN TAHMİNİ

Merve YENEN

Necmettin Erbakan Üniversitesi Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Danışman: Dr. Öğr. Üyesi Özlem ERDAŞ ÇİÇEK
2024, 83 Sayfa

Jüri
Prof. Dr. Gülay TEZEL
Dr. Öğr. Üyesi Ayşe Merve ACILAR
Dr. Öğr. Üyesi Özlem ERDAŞ ÇİÇEK

İlaç keşfi için kullanılan veri setlerinin analizi, verilerin kalitesini ve uygunluğunu değerlendirmek için gereklidir. Veri setlerindeki eğilimlerin, dağılımların ve ilişkilerin analizi, doğru sonuçlar elde etmek için önemlidir. Ayrıca, potansiyel ilaç adaylarını belirlemek için kullanılan moleküler özelliklerin ve etkileşimlerin anlaşılmasına yardımcı olmaktadır. Bu çalışmada, AChE enzimini inhibe eden bileşiklerin IC50 değerlerini tahmin etmek amacıyla veri madenciliği ve makine öğrenme algoritmaları kullanılmıştır. Biyolojik aktivite verileri, kanonik SMILES dizileri ve Lipinski'nin 5 kuralına dayalı özelliklerle genişletilmiştir. Çalışmada, moleküllerin biyolojik aktivitelerini ve ilaç benzerliklerini tahmin etmek için moleküler parmak izi hesaplamaları yapılmıştır. Farklı uzunluklardaki bit vektörleriyle oluşturulan veri setleri üzerinde Rassal Orman, XGBOOST, Ridge, SVR ve PLS regresyon algoritmaları ile tahminlemeler gerçekleştirilmiştir. Performans değerlendirmesi için k-katlı çapraz geçerleme kullanılmıştır. Elde edilen sonuçlara göre söz konusu algoritmalardan bazıları seçilerek toplu öğrenim yönteminde tahminleyici olarak kullanılmıştır. Sonuç olarak XGBOOST, PLS ve Ridge tahminleyicilerinin kullanıldığı toplu öğrenim yöntemiyle 0.75 korelasyon ve 0.63 ortalama kare hata değeri ile literatürdeki örneklerine kıyasla daha iyi bir sonuç elde edilmiştir. Ayrıca, standart sapması 0.3'ten küçük olan bit sütunlarının elenmesi ile veri seti küçültülmüş ve modellerin çalışma hızları arttırılmıştır. Bu çalışmada, AChE enzimiyle etkileşen bileşiklerin etkinliğinin belirlenmesinde makine öğrenimi algoritmalarının performansını karşılaştırmalı analiz ederek en iyi IC50 değeri tahmin sonucuna ulaşmak hedeflenmiştir. En iyi IC50 değerini bulmak, bir ilacın hedef enzimi ne kadar etkili durdurduğunu gösterir ve bu sayede, hastalıkların tedavisinde kullanılabilecek en güçlü ilaç adaylarını seçmeye yardımcı olabilir.

Anahtar Kelimeler: Bağlanma afinitesi, ilaç keşfi, lipinski'nin 5 kuralı, makine öğrenmesi, moleküler parmak izi, rassal orman regresyonu, xgboost regresyonu, destek vektör regresyonu, ridge regresyonu, pls regresyon, topluluk öğrenimi, k-katlı çapraz geçerleme

ABSTRACT

MS THESIS

PREDICTION OF BINDING AFFINITY OF MOLECULES INTERACTING WITH AChE ENZYME USING DATA MINING

Merve YENEN

THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCE OF NECMETTİN ERBAKAN UNIVERSITY
THE DEGREE OF MASTER OF SCIENCE
IN COMPUTER ENGINEERING

Advisor: Asst. Prof. Dr. Özlem ERDAŞ ÇİÇEK

2024,83 Pages

Jury
Prof. Dr. Gulay TEZEL
Asst. Prof. Dr. Ayse Merve ACILAR
Asst. Prof. Dr. Ozlem ERDAS CICEK

Analysis of datasets used for drug discovery is essential to assess the quality and relevance of the data. Analysis of trends, distributions, and relationships in data sets is important to obtain accurate results. It also helps understand molecular properties and interactions used to identify potential drug candidates. In this study, data mining and machine learning algorithms were used to estimate the IC50 values of compounds that inhibit the AChE enzyme. Biological activity data are augmented with features based on canonical SMILES sequences and Lipinski's rule of 5. In the study, molecular fingerprint calculations were performed to estimate the biological activities and drug similarities of molecules. Predictions were performed using Random Forest, XGBOOST, Ridge, SVR and PLS regression algorithms on datasets created with bit vectors of different lengths. K-fold cross-validation was used for performance evaluation. According to the results obtained, some of the algorithms were selected and used as predictors in the ensemble learning method. As a result, a better result was obtained compared to the examples in the literature with 0.75 correlation and 0.63 mean square error value with the ensemble learning method using XGBOOST, PLS and Ridge estimators. In addition, the data set was reduced and the working speed of the models was increased by eliminating the bit columns with standard deviation less than 0.3. This study aims to reach the best IC50 value prediction result by comparatively analyzing the performance of machine learning algorithms in determining the effectiveness of compounds that interact with the AChE enzyme. Finding the best IC50 value indicates how effectively a drug inhibits a target enzyme and can help select the most potent drug candidates to treat diseases.

Keywords: Affinity binding, drug discovery, Lipinski's Rule of Five, machine learning, molecular fingerprint, Random Forest Regression, SVR, XGBoost regression, Ridge Regression, PLS regression, ensemble learning, k-fold cross validation

ÖNSÖZ

Bu tez çalışmasının gerçekleştirilmesinde bana rehberlik eden değerli danışmanım Dr. Öğr. Üyesi Özlem ERDAŞ ÇİÇEK'e teşekkürlerimi sunarım. Ayrıca, sürecin her aşamasında manevi destekleriyle yanımda olan aileme ve eşime minnettarım.

Bu çalışma, AChE enzimi inhibisyonu üzerine yapılan tahmin uygulamalarının etkinliğini artırmak amacıyla veri madenciliği yöntemlerini kullanmıştır. Bu tezin, ilaç keşfi alanında çalışan araştırmacılara faydalı olmasını temenni ederim.

Merve YENEN
KONYA-2024

İÇİNDEKİLER

SİMGELER VE KISALTMALAR

AChE: Asetilkolinesteraz
ChEMBL: Chemical Biology Database
IC50: Half-maximal inhibitory concentration
PLS: Partial Least Squares
SVR: Support Vector Regression
SVM: Support Vector Machine
XGBOOST: Extreme Gradient Boosting
MSE: Mean Squared Error
DNN: Deep Neural Networks
RMSE: Root Mean Squared Error

ŞEKİLLER LİSTESİ

ÇİZELGELER LİSTESİ

DENKLEMLER LİSTESİ

1. GİRİŞ

1.1. Motivasyon

İlaç, besin veya temel diyet bileşeni dışında, bilinen yapıya sahip, canlı bir organizmaya uygulandığında biyolojik etki meydana getiren kimyasal madde olarak tanımlanabilir. Hastalıkları tedavi etmek veya önlemek, semptomları hafifletmek veya iyileştirmek için kullanılabilir. İlaçlar doğal (bitkisel ilaçlar gibi), sentetik (laboratuvarlarda üretilen) veya biyolojik (aşılar veya insülin gibi canlı organizmalardan türetilen) olabilir (Baron vd., 2023).

İlaç tasarımı bağlamında, bir ilaç genellikle bir hastalıkla ilgili bir durumu düzeltmek veya iyileştirmek için belirli bir biyolojik hedefle etkileşime girmek üzere yaratılan veya seçilen bir moleküldür. İlaç tasarımındaki bir hedef, genellikle bir hastalık sürecinde önemli bir rol oynayan belirli bir protein veya enzimdir. Bir ilacın amacı, bu hedefle etkileşime girerek aktivitesini değiştirmektir (Schmidt, 2022).

Proteinler ve enzimler, vücuttaki birçok hayati süreci düzenledikleri için yaygın hedeflerdir. Bir protein düzgün çalışmıyorsa (aşırı aktif veya inaktifse), sorunlara neden olabilir. Hücre yüzeyindeki reseptörler de bir diğer yaygın hedeftir. Bu reseptörler hücrenin dışından sinyaller alır ve hücrenin içinde tepkileri tetikler. Arızalı veya yanlış düzenlenmişlerse, hastalıklar meydana gelebilir. Bilim insanları, bir ilacı hedefle etkileşime girecek şekilde tasarlayarak, normal işlevi geri kazandırmak için aktivitesini engellemeyi, etkinleştirmeyi veya düzenlemeyi amaçlar (Schmidt, 2022).

Bilim insanları son yıllarda genomik, proteomik ve tıp alanlarında yeni teknikler geliştirerek hastalıkların anlaşılmasında ilerleme kaydetti (Erdaş, 2013). Makine öğrenimi gibi yenilikçi yöntemlerle ilaç keşfi süreçleri daha hızlı ve maliyet etkin hale getirilmiştir. Bu alanda çalışarak, bilimsel ilerlemeye katkıda bulunabilir ve insan hayatını iyileştiren çözümler üretilebilmektedir. Moleküler düzeydeki bilgi toplanıp harmanlandığı sürece tıbbi tedavide güvenli ve etkili ilaçların bulunması umut verici olacaktır (DiMasi vd., 2003). Geleneksel ilaç keşfi süreci Şekil 1.1.'de görüldüğü gibi 5 adımı içerir (Singh vd., 2023):

1. Hastalıklara yol açan mekanizmaları anlamaya çalışmak ve olası hedefler (örneğin proteinler) önermek için temel araştırmaların yapıldığı ön keşif aşaması;

2. Bilim insanlarının araştırılan hastalığa müdahale eden veya onu iyileştiren veya en azından semptomları hafifleten molekülleri (diğer adıyla ligand) veya diğer terapötik stratejileri aradığı ilaç keşif aşaması;

3. İlaç adaylarının etki biçimini açıklığa kavuşturmaya odaklanan, olası toksisiteyi araştıran, çeşitli hücre içi (in vivo) veya hayvansal (in vitro) deneyler ile etkinliği doğrulayan ve formülasyonu değerlendirmeye başlayan klinik öncesi geliştirme aşaması;

4. İlaç adayının etkilerini insanlarda araştıran klinik deney aşaması;

5. İlacın onaylandığı veya onaylanmadığı inceleme, onay ve piyasaya sürülme sonrası izleme aşaması.

Şekil 1.1. Geleneksel ilaç keşfi süreci

İlaç keşfinin 5 adımının başarıyla tamamlanması 12-15 yıllık çalışma ve yaklaşık 2.8 milyar dolar bütçe gerektirmektedir. Geleneksel ilaç keşfinde ilk aşamalar binlerce aday ilaç molekülü üzerinde protein-ligand bağlanma deneylerinin laboratuvarda yapılmasını içermektedir (Singh vd., 2023). Buna rağmen, ilaç keşfinde birçok başarısızlık vardır. Üzerinde çalışılan binlerce bileşik arasından yalnızca bir tanesi onay alabilmektedir. Protein-ligand etkileşimlerini anlamak, güvenli ve etkili yeni ilaçlar tasarlamak ve ilaç keşfine ve geliştirilmesine yardımcı olmak için önemlidir. Hesaplamalı yöntemler, özellikle moleküler yerleştirme, protein-ligand etkileşimlerini araştırmak için faydalıdır. Ancak etkileşimin gücünü tahmin etmek için kullanılan yerleştirme

programlarının puanlama fonksiyonları her zaman güvenilir değildir. Son 20 yılda ilaç tasarımında akıllı yöntemler popüler hale gelmiştir (Erdaş, 2013).

Makine öğrenmesi, ilaç keşfi sürecinde önemli bir rol oynamakta ve bu süreci hızlandırmakta, verimliliği artırmakta ve maliyetleri düşürmekte yardımcı olmaktadır. Makine öğrenmesi, büyük veri analitiği ve örüntü tanıma yetenekleri sayesinde ilaç keşfinde önemli bir araç haline gelmiştir. İlaç araştırmalarında kullanılan yüksek hacimli taramalar ve geniş veri setleri, makine öğrenmesi algoritmalarının etkin şekilde kullanılabilmesini sağlar. Bu algoritmalar, moleküler yapılar, bileşik aktiviteleri, ilaç etkileşimleri ve yan etkiler gibi birçok faktörü analiz ederek ilaç keşfi sürecinde değerli bilgiler sunabilir.

Bileşiklerin sanal olarak taranması ve özelliklerinin tahmin edilmesi, potansiyel ilaç adaylarının belirlenmesinde akıllı algoritmalar kullanmanın yardımı olur. Ayrıca, makine öğrenmesi, mevcut veri tabanlarında veya literatürde bulunan bilgilere dayanarak yeni ilaç kombinasyonlarını veya etkin bileşikleri önerme yeteneğine sahiptir. Öğrenen modellerin kullanımı, ilaç keşfinde zaman ve maliyet tasarrufu sağlarken, aynı zamanda doğruluk oranını da artırır. Bu teknoloji, yüksek boyutlu ve karmaşık verileri analiz ederek, verilerdeki desenleri ve ilişkileri tespit edebilir ve potansiyel ilaç adayları hakkında değerli bilgiler sağlayabilir. Örneğin Tian vd. (2010) çalışmasında proteinlerin ısı kararlılığı veya termal stabilitesi üzerindeki tek ve çok bölgeli mutasyonların etkilerini tahmin etmek ve proteinlerde stabilizasyon sağlayan ve istikrarsızlaştıran mutasyonlar arasında ayrım yapmak için makine öğrenme algoritmalarından yararlanmıştır. Sonuç olarak, ilaç adaylarının belirlenmesi, sanal tarama ve tasarım süreçleri, ilaç kombinasyonları önerisi gibi alanlarda makine öğrenmesi yöntemleri etkili bir şekilde kullanılarak, ilaç keşfinin başarısını ve etkinliğini arttırmaktadır. Hedeflenen bu tez çalışması ile Şekil 1.1.'de görülen ilaç keşfinin ilk iki adımının veri madenciliği ve makine öğrenmesi yöntemleri kullanılarak hızlandırılmasını amaçlamaktadır.

1.2. Temel Konseptler

1.2.1. Protein

Protein, hücre içinde veya dışında çeşitli biyolojik işlevlere sahip olan büyük ve karmaşık bir moleküldür. Ligand ise, bir proteinin belirli bir bölgesine bağlanabilen ve spesifik bir etki meydana getiren bir moleküldür. Ligandlar küçük moleküller veya diğer

proteinler olabilirler. Protein-ligand molekül etkileşimlerini anlamak için protein, bağlayıcı molekül ve etkileşim terminolojisinin tanımlanması gerekir (Erdaş, 2013). Protein-ligand etkileşimi, bir ligandın belirli bir proteinin aktif bölgesine bağlanmasıyla gerçekleşir. Bu etkileşimler, hücresel sinyal iletimi, enzimatik reaksiyonlar, ilaç etkileşimleri ve diğer biyolojik süreçlerde önemli roller oynar (Chaffey, 2003).

Proteinler, canlı organizmaların temel yapı taşlarından biridir ve birçok biyolojik süreçte önemli roller oynarlar. Tekli oksijen aracılı protein oksidasyonu, amino asitlerde, peptitlerde ve proteinlerde değişikliklere neden olabilir ve biyolojik sistemlerde yan etkilere ve karanlık reaksiyonlara potansiyel olarak yol açabilir (Davies, 2003). Tek protein terimi, moleküler biyoloji ve biyokimyada kullanılan bir terimdir ve tek bir protein molekülünü ifade eder. Bu çalışmada kullanılan tek protein asetilkolinesteraz, sinir hücreleri arasında iletişim sağlayan özel bir alanda asetilkolini hidrolize eder. Bu mekanizma, sinir sisteminin düzgün çalışmasını sağlar ve sinirsel iletişimin kontrol altında tutulmasına yardımcı olur (Yağmuroğlu & Emir Diltemiz, 2020). Asetilkolinesteraz, sinir sisteminde önemli bir enzim olarak görev yapar. Bu enzim, asetilkolin adı verilen bir sinir sinyali aktarıcısının parçalanmasında rol oynar. Asetilkolin, sinir hücreleri arasında iletişimi sağlayan bir kimyasal habercidir. Asetilkolinesterazın hidroliz reaksiyonu sayesinde asetilkolin seviyeleri düşer ve sinir iletimi sonlanır. Hidroliz reaksiyonu sonucunda asetik asit (CH_3COOH) ve kolin ($(CH_3)_3N^+CH_2CH_2OH$) molekülleri açığa çıkar. Bir adet asetilkolin bileşiği, bir adet metil grubu (CH_3), bir adet karboksilat grubu (COO), bir adet etilen grubu (CH_2CH_2) ve bir adet trimethylammonium grubu ($N^+(CH_3)_3$) içermektedir. Asetilkolin bileşiğinin formülü $CH_3COOCH_2CH_2N^+(CH_3)_3$ şeklinde ifade edilmektedir.

Asetilkolinesteraz proteinin aşırı aktivitesi, asetilkolin bileşiğinin hızla parçalanmasına ve sinir iletiminin hızlı bir şekilde sonlanmasına yol açar. Bu durum, kasların kontrolsüz kasılmalarına, kramplara ve hatta solunum yetmezliğine neden olabilir. Diğer yandan, asetilkolinesteraz proteinin yetersiz aktivitesi, asetilkolin molekülünün yeterince parçalanamamasına ve sinir iletiminin aşırı uyarılmasına neden olur. Hesaplamalı ilaç keşfi süreciyle enzim baskılayıcıları geliştirilerek bu sorunların tedavisi mümkün olabilir.

Tek protein her bir protein molekülü, amino asit adı verilen küçük moleküllerin zincirlerini oluşturan bir polipeptit zinciridir. Proteinler, bu amino asit zincirlerinin belirli

bir düzenlenmesiyle oluşur ve genellikle bir veya daha fazla fonksiyonel bölgeye sahiptirler. Bu fonksiyonel bölgeler, proteinin belirli bir işlevini yerine getirmesini sağlar. Bir araştırmacı belirli bir hastalıkla ilişkilendirilen bir proteinin yapısını çözmek ve bu proteinin işlevini anlamak için tek protein çalışmaları yapabilir. Hücreler içinde tek protein izleme, daha önce düşünülenden daha geniş bir difüzyon sabitleri dağılımını ortaya çıkararak hücresel makromoleküllerin dinamiklerini ortaya koyar (Goulian & Simon, 2000). Bir proteinin yapısı, amino asit dizilimi ve üç boyutlu katlanması ile belirlenir. Bu yapı, proteinin fonksiyonunu belirler ve hücresel süreçlere katılımını sağlar. Tek protein çalışmaları, bu yapıyı ve işlevi anlamak için çeşitli yöntemler kullanır ve bu şekilde hücresel düzeydeki biyolojik süreçlerin anlaşılmasına katkıda bulunur.

1.2.2. Protein Ligand

Protein-ligand etkileşimi, biyokimyada ve ilaç tasarımında önemli bir konudur. Bu etkileşim, bir protein molekülü ile bir ligand molekülü arasındaki kimyasal bağlanmayı ifade eder. Ligand kimyası ile proteinleri ilişkilendirmek, ilaç hedefleri arasında beklenmedik ilişkileri ve bağlantıları ortaya çıkararak yeni tedaviler için potansiyel ilaç hedeflerini açığa çıkarır (Keiser vd., 2007). Protein-ligand etkileşimlerini anlamak, biyolojiyi anlamak ve ilaç keşfi ve geliştirmeye yardımcı olmak için hayati öneme sahiptir (Du vd., 2016). Ligand bağlanması, proteinlerde küçük yan zincir yeniden düzenlemelere yol açar, büyük, polar amino asitlerin aromatik olanlardan bağlanma ceplerinde daha esnek olduğunu gösterir (Najmanovich vd., 2000). Protein-ligand etkileşimi belirli bir proteinin yapısını ve işlevini anlamak, ilaç tasarımında yeni bileşiklerin geliştirilmesi veya mevcut ilaçların optimize edilmesi amaçları ile araştırılır. İlaç tasarımında, bir ligandın bir proteinin aktif bölgesine uygun şekilde bağlanması, istenilen biyolojik etkiyi meydana getirmek için önemlidir.

1.2.3. Moleküler Ağırlık

Moleküler ağırlık, bir molekülün içindeki tüm atomların nükleer kütlelerinin toplamıdır. Bu değer, bir molekülün kimyasal ve fiziksel özellikleri üzerinde önemli bir etkiye sahiptir ve bir bileşiğin yapısını ve davranışını anlamak için önemlidir. İlaç keşfi çalışmalarında moleküler temsiller, ilaç kombinasyon duyarlılığını ve ilaç sinerji skorlarını tahmin etmeye yardımcı olur, ancak modelin yorumlanabilirliği ve sağlamlığı gibi niteliksel hususlar da önemlidir (Zagidullin vd., 2021). Moleküler ağırlık, bir bileşiğin formülündeki her bir atomun atomik kütlesini alarak hesaplanır ve bu değerlerin

toplamı ile elde edilir. Moleküler ağırlık, biyoyararlanımı etkilediği ve ilaç güvenliğini etkileyebileceği için ilaç keşfi çalışmalarında önemlidir (Tsantili-Kakoulidou & Demopoulos, 2021). Moleküler ağırlık, bir bileşiğin fiziksel özelliklerini etkiler. Örneğin, moleküler ağırlığı artan bileşikler genellikle daha yoğun olur ve daha yüksek kaynama ve erime noktalarına sahip olabilirler. Ayrıca, moleküler ağırlık, bir bileşiğin çözünürlüğünü, buharlaşma hızını ve polarite gibi kimyasal özelliklerini de etkileyebilir. Bir bileşiğin tanımlanması, saflığı belirlenmesi ve laboratuvar koşullarında doğru dozların hesaplanması gibi birçok alanda önemlidir. Bir bileşiğin kimyasal formülünün belirlenmesi için de kullanılabilir, bileşiğin kimyasal ve fiziksel özelliklerini anlamak ve karakterize etmek için önemli bir parametredir. Bu nedenle, bir bileşiğin moleküler ağırlığının doğru bir şekilde belirlenmesi, birçok bilimsel ve endüstriyel uygulamada rol oynar. Moleküler dinamik yöntemler, ilaç-hedef tanıma ve bağlanma ile ilişkili termodinamik ve kinetiği daha doğru bir şekilde tahmin etmeyi sağlar, bu da ilaç etkinliğini artırır (De Vivo vd., 2016).

1.2.4. Hidrojen Bağ Sayısı

İlaçlarda hidrojen bağ sayısı bir ilacın hidrojen bağlama donörlerinin sayısını ifade eden bir parametredir. Hidrojen bağları, hidrojen atomunun bir elektronegatif atom ile oluşturduğu zayıf bir bağdır. Hidrojen bağları sayısı, bir moleküldeki hidrojen atomlarının sayısını belirtir, bu hidrojen atomlarının potansiyel olarak hidrojen bağları oluşturabilecekleri gruplara bağlıdır. Özellikle, bir hidrojen atomu, bir elektronegatif atomla doğrudan bağlı olduğunda, hidrojen bağları oluşturma potansiyeline sahiptir. Hidrojen bağları sayısı, bir molekülün su içinde çözünme yeteneği, biyoyararlanımı ve biyolojik etkileri gibi birçok farmakolojik özelliği üzerinde etkili olabilir. Bir ilacın hidrojen bağı sayısı değeri, ilacın farmakokinetik özellikleri üzerinde önemli bir etkiye sahiptir. Özellikle, bir ilacın hidrojen bağlama donörlerinin sayısı, ilacın vücutta absorbe edilme, dağılma, metabolize edilme ve atılma gibi süreçlerdeki davranışını etkileyebilir. Hidrojen bağ sayısı değeri, bir ilacın diğer moleküllerle etkileşimini etkileyebilir. Örneğin, bir ilacın hidrojen bağlama donörlerinin sayısı, ilacın hedef proteinlerle veya diğer moleküllerle nasıl etkileşim kuracağını ve bu etkileşimlerin gücünü belirleyebilir. Bu nedenle, hidrojen bağ sayısı değeri, ilaçların tasarımı ve optimizasyonu sürecinde dikkate alınması gereken önemli bir moleküler özelliktir. Hidrojen bağ sayısı parametresi, ilaç tasarımı sürecinde ilacın farmakokinetik ve farmakodinamik özelliklerini anlamak ve optimize etmek için kullanılır.

1.2.5. Hidrojen Bağı Alıcısı

Hidrojen bağı alıcısı, bir ilacın hidrojen bağlama kabul edici gruplarının sayısını ifade eden bir parametredir. Hidrojen bağları, hidrojen atomunun bir elektronegatif atomla oluşturduğu zayıf bağlardır. Hidrojen bağı alıcısı, bir moleküldeki hidrojen bağlama kabul edici gruplarının sayısını belirtir. Bu gruplar, hidrojen atomuyla birlikte, bir hidrojen bağının oluşabileceği elektronegatif atomlardır. Hidrojen bağ alıcı değeri, bir ilacın hidrojen bağlama kabul edici gruplarının sayısını ifade eder ve ilacın kimyasal yapısının bir özelliğidir. Bir ilacın hidrojen bağ alıcı değeri, ilacın su içinde çözünme yeteneği, vücutta absorbe edilme oranı, hedef proteinlerle etkileşim kabiliyeti ve biyolojik aktivitesi gibi farmakolojik özelliklerini etkileyebilir. İlaç tasarımı aşamasında, hidrojen bağ alıcı değeri, ilaç adaylarının seçiminde ve moleküler modifikasyonların planlanmasında dikkate alınır. Steiner ve Koellner (2001) çalışmalarında, düzenli ikincil yapı elemanlarını stabilize ederek heliks uçları, iplik uçları ve düzenli dönüşlerde rol oynayan proteinlerdeki hidrojen bağlarının, her 10.8 aromatik kalıntıda bir oluştuğunu belirtmiştir. Han vd. (2015) çalışmasında hidrojen bağı alıcıların, organo-enamin katalizinde kritik bir rol oynadığını ve enamonyum ile hidrojen bağı etkileşimleri yoluyla reaksiyon hızlarını etkilediğini gözlemlemiştir.

1.2.6. LogP Katsayısı

LogP, bir bileşiğin yağda ve suda çözünürlüğü arasındaki dengenin ölçüsüdür ve bu, bir ilacın biyoyararlanımını ve hedefe ulaşma yeteneğini belirlemede rol oynar. İlaç tasarımı sürecinde, LogP'nin optimizasyonu, ilacın hücre membranlarını geçme yeteneğini ve biyolojik etkinliğini etkileyebilir. Poulin ve Theil (2000) çalışmalarında, ilaçların doku-plazma bölünme katsayılarını tahmin etmek için iki mekanik denklem geliştirerek yeni ilaç adaylarının erken farmakokinetik taranmasını mümkün kılmıştır. Genellikle, bir ilacın LogP değeri arttıkça, ilacın yağda çözünürlüğü artar ve bu da ilacın hücre membranlarını daha kolay geçmesine ve hücre içine ulaşmasına olanak tanır. Ancak, çok yüksek bir LogP değeri, ilacın su içinde çözünürlüğünün azalmasına ve dolayısıyla biyoyararlanımının düşmesine neden olabilir. İdeal olarak, bir ilacın LogP değeri, hedeflenen biyolojik aktiviteyi etkilemeden, hücre membranlarını geçme yeteneğini artıracak kadar yüksek ve aynı zamanda istenmeyen toksik etkilere veya düşük su çözünürlüğüne neden olmayacak kadar düşük olmalıdır. Martel vd. (2013) çalışmasında erken ilaç keşfinde kimyasal bileşiklerin oktanol/su dağılım katsayılarını

tahmin etmek için yeni yaklaşımlar geliştirmek ve karşılaştırmak amacıyla logP değerinden oluşan büyük ve çeşitli bir veri seti ile çalışmıştır. İki katmanlı lipit membranlardaki ilaçların bölünme katsayısı, ilaç tasarımı için uygun bir yöntem olan membran kırılma indeksindeki değişiklik kullanılarak belirlenebilir (Ramsden, 1993). Bu süreçte, çeşitli bilgisayar destekli tasarım araçları ve hesaplama yöntemleri kullanılabilir. Bu araçlar, LogP'nin hesaplanmasını, ilaç adaylarının tasarımını ve optimizasyonunu hızlandırabilir ve geliştirebilir. LogP'nin optimize edilmesi, ilaçların biyoyararlanımını ve hedefe ulaşma yeteneğini artırabilirken, aynı zamanda istenmeyen etkileri minimize edebilir.

1.2.7. Hidrofobisite

İlaçların hidrofobisitesi, bir ilacın hidrofobik özelliklerini ve su ile etkileşimini belirlemek için kullanılan bir terimdir. Hidrofobiklik araştırmaları 19. yüzyıldan bu yana önemli ölçüde ilerlemiş olup ilaç tasarımı ve protein katlanmasına katkıda bulunmuştur (Sarkar & Kellogg, 2010). Hidrofobiklik, bir molekülün suya karşı olan iticiliğini ve suyla etkileşimini ifade eder. İlaçların hidrofobik özellikleri, ilacın farmakokinetik davranışını, hücre membranlarını geçme yeteneğini ve biyoyararlanımını etkileyebilir. Biyomoleküller ve ilaçlar arasındaki hidrofobik etkileşim, bağlanma afinitesi ve özgüllük için önemlidir, ilaç tasarımına yön verir ve güvenlik, etkililik ve farmakolojik özellikleri optimize eder (Lou & Martin, 2021). İdeal olarak, bir ilacın hidrofobikliği, hücre membranlarını geçme yeteneğini artıracak kadar yüksek olmalıdır, ancak aynı zamanda istenmeyen toksisiteye veya düşük su çözünürlüğüne neden olmayacak kadar da düşük olmalıdır. Nanotaşıyıcılarda hidrofobik iyon eşleşmesi, ilaç salınımını etkileyen sıvı kristal yapılar oluşturur, iç yapılar salınım hızlarını ve pH'a bağlı salınımı etkiler (Ristroph vd., 2021). Hidrofobik ilaçlar, dolaşımdaki lipoproteinlerle birleşerek kan sınırlamalarını aşabilir, bu da biyolojik aktivitelerini ve potansiyel toksisitelerini artırabilir (Wasan vd., 2008). İlaç adaylarının tasarımında, ilacın hidrofobik ve hidrofilik gruplarının dengeli bir şekilde ayarlanması ve ilacın hidrofobikliğinin optimize edilmesi önemlidir.

1.2.8. Bağlanma Eğilimi

Protein-ilaç arasındaki etkileşimin tahmini için Ki (inhibisyon sabiti), Kd (ayrışma sabiti) ve IC50 gibi gerekli parametreler bulunmaktadır. IC50 konsantrasyonu ise, belirli bir biyolojik veya biyokimyasal fonksiyonu yarı yarıya engellemek için

gereken inhibitör konsantrasyonunu ölçer (Caldwell vd., 2012). Ve bu çalışmada protein-ilaç arasındaki etkileşimi tahmin etmek için IC50 parametresi seçilmiştir.

IC50 değerleri, bir ilacın, çok sayıda hastalığın gelişiminde rol oynayan çeşitli enzimlere/biyolojik hedeflere karşı etkinliğini belirler (Thakur vd., 2022). Literatürde IC50 konsantrasyonu ile çalışmalar mevcuttur. Örneğin, Nevozhay (2014) çalışmasında, IC50 konsantrasyonu ile test ettiği ilacın kanser hücre popülasyonunun çoğalmasını teorik olarak mümkün olan etkinin %50'si veya ilacın pratikte elde edebileceği maksimum etkinin %50'si oranında engelleyen konsantrasyonu olduğunu belirtmiştir.

1.3. Problem Tanımı

Bilimsel araştırmaların ilerlemesi, farmakoloji ve ilaç keşfi alanında veri bilimi ve makine öğrenmesi tekniklerinin kullanımıyla yeni ufuklar açmaktadır. Bu çalışma, farmasötik alandaki önemli bir enzim olan asetilkolinesteraz (AChE) proteinine odaklanarak, veri madenciliği süreçlerini ve makine öğrenmesi algoritma ve yöntemlerini kullanarak ilaç moleküllerinin bağlanma eğilimine ait IC50 değerlerinin regresyon analizi ile tahmin edilmesi hedeflenmiştir. Enzimlerin biyolojik etkilerini anlamak ve ilaç geliştirme süreçlerini iyileştirmek için veri odaklı yöntemlerin kullanımı, ilaç tasarımı ve etkinliği üzerinde etkilidir.

1.4. Katkılar

Enzim-ilaç bağlanma eğilimini belirlemek için deneysel yöntemler zaman alıcı ve maliyetlidir. Bu nedenle, bu çalışma, deneysel verilerin analiz edilmesi için veri madenciliği tekniklerini benimseyerek verimliliği arttırmayı amaçlamaktadır. Bu çalışmanın ilaç keşfi literatüründe akıllı algoritmaların kullanılmasını içeren katkıları aşağıdaki gibi özetlenebilir:

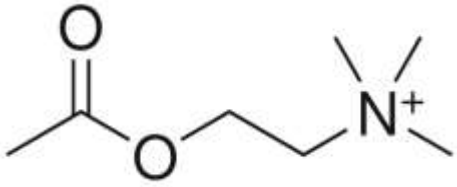
1. Moleküler parmak izlerinin kullanıldığı veri setlerinde farklı makine öğrenme yöntemleri kullanılarak performans değerlendirilmesi yapılmıştır.
2. Farklı uzunluklardaki moleküler parmak izlerinin bağlanma eğiliminin tahminindeki etkileri araştırılmıştır.
3. Verideki düşük varyanslı özellikler elendiğinde makine öğrenmesi algoritmalarının hız ve performansındaki değişimler gözlenmiştir.

Bu çalışmanın sonuçları, farmakoloji ve ilaç keşfi alanındaki uygulamalar için veri seti üzerindeki çalışmaların önemine ışık tutabilir. İlaç endüstrisindeki araştırmacılar için, veri madenciliği ve makine öğrenmesi tekniklerinin kullanımının, yeni ilaç adaylarının tasarımını hızlandırmak amacıyla nasıl kullanılabileceğini göstermektedir. Ayrıca bu çalışma, benzer biyolojik veri setleri üzerinde veri madenciliği yaklaşımlarının nasıl genelleştirilebileceğine dair geniş kapsamlı fikirler sunabilir. Elde edilen bulgular, ilaç keşfi süreçlerinin verimliliğini artırmak için yeni yöntemler geliştirmek isteyen araştırmacılar için önemli bir kaynak olabilir.

2. KAYNAK ARAŞTIRMASI

2.1. AChE Enzimi

Acetylcholinesterase (AChE) enzimi, esteraz ailesine ait bir tür serin hidrolitik enzimdir.



Şekil 2.1. Asetilkolinesteraz molekülü

Vücutta AChE'nin nin önemli bir rolü vardır; asetilkolin moleküllerini sinir hücreleri arasındaki bağlantı noktası olan sinaptik boşlukta hidroliz ederek sinyal iletimini sonlandırır. AChE çeşitli dokularda bulunur, ancak en yüksek düzeyde sinir uçlarında ve sinir-kas bağlantılarında bulunur. Bu bölgelerde, asetilkolinin hızla parçalanması, sinir sinyallerinin hızlı bir şekilde sonlanmasını sağlar. Bu, sinir sisteminin doğru çalışmasını sağlayarak, kasların kontrolünü sağlar. AChE ayrıca, asetilkolin düzeylerinin dengelenmesinde ve sinir hücrelerinin yeniden kullanıma hazır hale gelmesinde rol oynar. AChE birçok canlı türünde bulunan bir enzimdir. Memelilerde, diğer omurgalılarda, böceklerde, sürüngenlerde ve bazı bitkilerde de bulunur. Farklı çalışma alanlarında kullanılmıştır. Örneğin Karakuş vd. (2022) çalışmasında, sıçan beyin dokusundan AChE enziminin klonlanması ve bakteriyel bir konakta protein elde edilmesini amaçlamıştır.

AChE, birçok fizyolojik ve farmakolojik sürecin araştırmanın konusu olmuştur. Birçok zehir ve sinir gazı, AChE' yi inhibe ederek sinir iletimini etkiler. Bu inhibitörler, asetilkolinin parçalanmasını engeller ve sinir iletimini sürekli olarak uyarır, bu da kasların kontrolsüz kasılmasına ve ciddi sağlık sorunlarına yol açabilir. Bununla birlikte, AChE inhibitörleri, bazı tıbbi durumların tedavisinde de kullanılır. Örneğin, Alzheimer hastalığı tedavisinde, AChE inhibitörleri sinaptik asetilkolin seviyelerini artırarak bellek ve bilişsel fonksiyonlarda iyileşme sağlayabilir. AChE enzimi, normalde sinir sistemi işlevlerinin düzenlenmesinde önemli bir rol oynar ancak, bazı durumlarda AChE'nin aşırı aktivitesi veya yetersiz aktivitesi sağlık sorunlarına neden olabilir. Pope (1999) çalışmasında AChE inhibisyonunu, organofosfat toksik maddelere bağlı olarak çeşitli dokularda doza bağlı

olarak gözlemlemiş ve inhibisyon düzeylerinin biyolojik işlev bozukluğu sıklığıyla genel olarak uyumlu bir şekilde ilişkili olduğu sonucuna varmıştır. AChE inhibitörleri, AChE enzimini geçici veya kalıcı olarak inhibe ederek asetilkolinin parçalanmasını yavaşlatır veya durdurur. Bu da sinaptik asetilkolin seviyelerini artırır ve sinir iletimini düzenler. AChE inhibitörlerinin takibi için J. Chen vd. (2023) çalışmasında, asetilkolinin hidroliz reaksiyonu sırasında ortaya çıkan empedans değişimlerinin, asetilkolin miktarını tespit etmek için kullanılabileceğini göstermektedir. Ayrıca Akocak ve Lolak (2020) çalışmalarında, belirli karbonik hidrojen transfer enzimi inhibitörlerinin antioksidan özelliklerini ve kolinesteraz aktivitelerini değerlendirmek amacıyla çalışma gerçekleştirmiş, AChE inhibisyon profilinde genel olarak, zayıf DPPH, ABTS ve metal bağlayıcı aktivitesi gözlemiştir. AChE, Alzheimer hastalığı gibi durumlarda merkezi sinir sistemi üzerinde önemli bir rol oynamaktadır. Bu nedenle, daha etkili inhibitörlerin ve daha güvenli ilaç adaylarının belirlenmesine yönelik araştırmalar AChE'ye odaklanmaktadır. Bu ilaçlar, hastaların bellek, düşünme becerileri ve yaşam kalitesinde iyileşme sağlamak için kullanılır (Camadan & Akkemik 2022). Hormonlar üzerinde yıkıcı etkisi olduğu bilinen BPA (Bisphenol A) bileşiği üzerinde çalışma gerçekleştiren Heredia-García vd. (2023) çalışmasında BPA'nın AChE üzerindeki etkilerini, dolaylı olarak sinir sistemi fonksiyonlarına etki edebilecek hormonal değişikliklerle ilişkilendirmiştir.

Aynı zamanda, AChE inhibitörleri, zehirli gaz veya böcek ilaçları gibi AChE aktivitesini inhibe eden maddelerle zehirlenmelerin tedavisinde de kullanılabilir. Bu ilaçlar, zehirin etkisini azaltarak ve sinir iletimini düzenleyerek hastanın sağlığını korumaya yardımcı olur. AChE enzimi, sinir sistemi işlevinde önemli bir rol oynayan bir enzimdir. AChE'nin asetilkolin hidrolizi yoluyla sinir iletimini düzenlemesi, sinir sisteminin doğru çalışmasını sağlar. AChE, zararlı kimyasalların hedefi olabilirken, aynı zamanda tıbbi tedavilerde de kullanılabilir.

2.1.1. AChE Enzimi Çalışmaları

AChe enzimi çalışmalarının temelinde, Alzheimer hastalığının tedavisinde potansiyel AChE inhibitörleri üzerine yapılmış araştırmalar vardır. Makine öğrenme modelleri, çeşitli türlerde asetilkolinesteraz inhibisyon aktivitesini tahmin etmekte ve potansiyel toksinleri belirlemede yardımcı olmaktadır (Vignaux vd., 2023). Motor sinir kaynaklı faktörler farklı kaslarda AChE alt birimlerinin ekspresyonunu değiştirerek kas

fonksiyonunu etkileyebilir (Tsim vd., 2008). Kadir vd. (2008) çalışmasında, hafif Alzheimer hastalarında galantamin tedavisinin dikkat üzerinde etkili olduğu sonucuna varmıştır. Ayrıca galatamin tedavisi 3 hafta ile 12 ay arasında %30-40 oranında kortikal AChE inhibisyonu sağlamıştır. Bondžić vd. (2020) çalışmasında, volüminöz ve negatif yüklü moleküller tarafından asetilkolinesteraz inhibisyonundan sorumlu yeni bir allosterik bölgeyi ortaya çıkararak, yeni ilaç tasarım stratejileri ve daha verimli AChE modülatörleri için potansiyel sunmaktadır. AChE enzimi inhibitörleri üzerine yapılmış Makine Öğrenmesi çalışmaları sınıflandırma ve regresyon olarak ikiye ayrılmaktadır. Sandhu vd. (2022) çalışmasında, AChE enzimi ile etkileşimde olan 4140 molekülü BindingDB veri tabanından toplamıştır. Bu moleküllerden 1000 Nm'den küçük IC50 değerine (pIC50 > 6) sahip 1862 molekülü aktif diğer 2278 molekülü ise inaktif olmak üzere etiketlemiştir. Padel moleküler tanımlayıcı ve R dilinin kullanıldığı çalışmada özellik eleme yöntemleri kullanıldıktan sonra 179 tanımlayıcı ve 116 uzunluğunda moleküler parmak izi vektörü elde etmiştir. K-en yakın komşu, SVR ve rassal orman algoritmasının kullanıldığı çalışmada rassal orman tabanlı model ile Alzheimer hastalığı tedavisi için AChE inhibitörlerini %85,38 doğrulukla tahmin edebilmiştir. Destek vektör makinesi gibi makine öğrenme yöntemleri ise Alzheimer hastalığı ilaç keşfinde moleküler tanımlayıcılarla ilişkilendirilmiş asetilkolinesteraz inhibitörlerini doğru bir şekilde tahmin edebilir (Lv & Xue, 2010). Yücel (2022) çalışmasında moleküllerin asetilkolinesteraz inhibitörleri için aktif veya inaktif olduğunu tahmin etmek için Derin Sinir Ağları (DNN), Destek Vektör Makineleri (SVM) ve Aşırı Gradyan Artırma (XGBOOST) algoritmaları ile sınıflandırma uygulaması gerçekleştirmiştir. Alzheimer hastalığı ve multipl skleroz gibi nörodejeneratif hastalıklarla ilişkili olan AChE'nin rolüne dikkat çekmektedir. Çalışmada, AChE inhibitörlerine ilişkin deneysel IC50 değerleri ChEMBL veritabanından alınmış ve veriler pIC50 değerlerine dönüştürülmüştür. PIC50 değeri 7'den yüksek olan moleküller aktif olarak etiketlenmiştir. Eksik verilerin temizlenmesi ve tekrarlı verilerde düşük pIC50 değerinin dikkate alınması ile 5328 adet molekülün 2048 uzunluktaki parmak izleri üzerinde çalışılmıştır. Üç farklı makine öğrenimi modeli kullanılmıştır. 3 gizli katmanlı bir DNN modeli oluşturulmuş, ReLU aktivasyon fonksiyonu gizli katmanlarda, sigmoid fonksiyonu ise çıkış katmanında kullanılmıştır. XGBOOST algoritması n_estimators=300 ve max_depth=10 parametreleri ile optimize edilmiştir. SVM modeli için C=10 ve gamma=0.4 değerleri kullanılmıştır. DNN ve SVM modeli ile %93 doğruluk elde ederken XGBOOST yöntemi ile %87 doğruluk değerine ulaşılmıştır. Bu araştırmalar, Alzheimer hastalığının tedavisi

için potansiyel yeni ilaçların geliştirilmesine ve mevcut tedavi seçeneklerinin iyileştirilmesine katkıda bulunabilir.

AChE proteini ile regresyon çalışmaları da literatürde yer almaktadır. Örneğin, Lan vd. (2019) yaptıkları regresyon çalışmasında; şemsiye örnekleme yöntemi ile, potansiyel Alzheimer hastalığı inhibitörlerinin taranmasına yardımcı olarak ligandların AChE proteinine bağlanma afinitesini etkili bir şekilde tahmin eder. Moleküler bağlanma simülasyon ve modelleme tekniklerini kullandığı çalışmasında korelasyon katsayısını 0.94 olarak hesaplayarak AChE inhibitörlerinin etkinliğini ve bağlanma afinitelerini başarılı bir şekilde tahmin etmiştir.

Nguyen vd. (2022) çalışmasında ligandların hedef protein olarak belirlediği AChE'ye bağlanma serbest enerjisini yüksek korelasyonla tahmin edebilecek bir regresyon modeli eğitmeyi amaçlamıştır. Chembl veri tabanından elde edilen veri setine ait 600 veri eğitim veri seti için, 162 veri test veri seti için ayrılmıştır. Ayrılan veri setleri RDKit aracının hesapladığı 200 bit uzunluğunda moleküler parmak izlerinden oluşmaktadır. Yüksek ve düşük varyanslı parmak izlerinin elenmesi ile veri seti 123 bit uzunluğunda vektörlere indirgenmiştir. Regresyon modeli için Lineer Regresyon, Rassal Orman Regresyonu, XGBOOST ve Grafik Evrişimsel Ağ modellerini kullanmıştır. Bu modeller, AChE inhibitörlerinin bağlanma enerjilerini tahmin etmek için kullanılmıştır. En iyi tahmin sonuçlarını 0.72 Pearson R korelasyonu ve 1.580 RMSE ile Grafik Evrişimsel Ağ modeli elde etmiştir. Ayrıca, moleküler yerleştirme (docking) ve moleküler dinamik simülasyonları gibi atomistik hesaplamalarla ligandların AChE'ye bağlanma süreçleri incelenmiştir. Çalışmanın sonuçlarına göre, iki bileşik olan benzil trifluorometil keton ve trifluorometilstiril keton, AChE üzerinde güçlü inhibitör etkiler göstermiştir. Bu bileşiklerin IC50 değerleri sırasıyla 0.51 μM ve 0.33 μM olarak hesaplanmış ve bu değerler, mevcut bir Alzheimer tedavisinde kullanılan galantaminin IC50 değerinden (2.10 μM) oldukça düşük bulunmuştur.

Khedekar vd. (2022) çalışmasında, ChEMBL 20 veritabanından AChE ile etkileşen moleküllere ait 5103 benzersiz kanonik SMILES verisi toplayarak bir regresyon yöntemi geliştirmiştir. 500 tahmin edici parametreli Rassal Orman modeli, PaDEL parmak izi tanımlayıcılarını kullanarak AChE ile etkileşen moleküllerin pIC50 değerlerini tahmin etmek için kullanılmıştır. Bu QSAR modeli, Pearson korelasyon katsayısı (r = 0,93), kök ortalama kare hatası (RMSE = 1), ortalama kare hatası (MSE =

0,35) ve belirleme katsayısı (r-kare = 0,87) olmak üzere dört istatistiksel parametre kullanılarak değerlendirilmiştir. Bu çalışmanın sınırlaması, tüm veri noktalarının eğitim ve test kümelerine bölünmek yerine tek parça halinde modele uydurulmasıdır. Bu nedenle değerlendirme, yeni veri noktalarını keşfetme kapsamı sağlayan görünmeyen veriler üzerinde gerçekleştirilmez.

3. MATERYAL VE YÖNTEM

3.1. Veri madenciliği

Veri madenciliği, istatistik, makine öğrenimi ve veritabanı teknolojisinden yöntemler kullanarak büyük veri kümelerinde beklenmedik, değerli veya ilginç yapıları bulma bilimidir (Cuzzocrea vd., 2007). Büyük veri setlerinden anlamlı bilgileri keşfetmek ve anlamak için kullanılan disiplinlerarası bir alanı ifade eder. Bu, istatistiksel ve matematiksel tekniklerin, yapay zekâ, makine öğrenimi ve veritabanı yönetimi gibi alanlardan gelen yöntemlerle birleştirilmesini gerektirir. Büyük miktarda veri ve bilgi içinde desenleri tanımlayarak daha iyi kararlar almak için içgörüler sağlar, büyük miktarda veri ve bilgi içinde desenleri tanımlamak için çeşitli metodolojiler ve görevler kullanır (Ogunleye, 2021).

Veri madenciliği, işletmelerin veri tabanlarında gizli kalmış bilgileri ortaya çıkarmak için bir araçtır. Genellikle dört ana aşamada gerçekleştirilir: veri toplama, veri ön işleme, modelleme ve sonuçların yorumlanması. İlk olarak, çeşitli kaynaklardan veri toplanır ve uygun bir formata dönüştürülür. Daha sonra, veri ön işleme adımında, eksik veya gürültülü veriler temizlenir ve özellikler çıkarılır. Modelleme aşamasında, veri seti analiz edilir ve algoritmalar kullanılarak desenler ve ilişkiler keşfedilir. Elde edilen sonuçlar yorumlanır ve anlamlı bilgiler elde edilir. Bilgisayar bilimi ve istatistik kullanarak büyük veritabanlarından bilgi çıkarmak için gelişmekte olan bir alandır ve alana ilişkin istatistiksel temaları ve öğretileri vurgular (Glymour vd., 1997). Veri madenciliği birçok farklı alanda kullanılır. Örneğin, pazarlama alanında, müşteri davranışlarını anlamak ve hedef kitleye özel pazarlama stratejileri oluşturmak için, sağlık sektöründe, hastalık risklerini tahmin etmek veya tedavi sonuçlarını iyileştirmek için, finansal hizmetlerde, dolandırıcılığı tespit etmek veya yatırım stratejilerini geliştirmek için kullanılmaktadır. Hizmet endüstrisinde veri madenciliği, büyük veri kümelerini analiz etmeye, işlevleri iyileştirmeye ve büyüme fırsatlarını belirlemeye yardımcı olur (Olson, 2007). Tıp araştırmalarında veri analizi ve modelleme süreçlerini hızlandırır. Sağlık sektöründe klinik verilerden ve tıbbi literatürden elde edilen bilgilerle yeni tedavi yöntemlerinin keşfedilmesine katkıda bulunabilir. Özellikle ilaç sektöründe olumsuz ilaç reaksiyonlarını tespit etmek için büyük veri kümelerinde ilginç, beklenmedik veya değerli yapıları keşfetmeye yardımcı olur (Craig, 2007). İlaç endüstrisinde, veri madenciliği, yeni ilaçların geliştirilmesinde hız kazanılmasına ve maliyetlerin azaltılmasına yardımcı

olabilir. Klinik veritabanları ve biyomedikal literatürden büyük deneysel verilerle bilimsel hipotezler üretmeye ve klinik ve idari karar verme için yeni biyomedikal ve sağlık bilgisi keşfetmeye yardımcı olur (Yoo vd., 2012). Geleneksel farmakovijilans yöntemlerine potansiyel olarak faydalı bir ek, muhtemel olumsuz ilaç reaksiyonlarını belirlemeye yardımcı olur (Almenoff vd., 2005). Bilimsel literatürü madencilik için veri madenciliği araçları ve diğer kaynaklarla entegre etmek, ilaç keşfi ve geliştirilmesi için hayati öneme sahiptir (Agarwal & Searls, 2008).

3.1.1. Veri Setinin Hazırlanması

3.1.1.1. ChEMBL Veritabanı

ChEMBL, biyolojik aktivite verilerini içeren halka açık bir veritabanıdır ve özellikle ilaç keşfi ve biyomedikal araştırmalar için kullanılan kimyasal bileşiklerin, biyolojik hedeflerin bilgilerini sağlar. Cambridge'deki Avrupa Biyoinformatik Enstitüsü tarafından geliştirilen bu veritabanı, bilimsel araştırmalarda yaygın olarak kullanılmaktadır.

Molekül ChEMBL kimliği, ChEMBL veritabanında bulunan bir bileşiğin benzersiz kimlik numarasını ifade eder. ChEMBL dünya çapında biyolojik aktiflik verilerini içeren, ilaç araştırması ve kimyasal biyoloji alanında önemli bir kaynaktır. ChEMBL veritabanı, etkinlik odaklı bir tarama kütüphanesini verimli bir şekilde oluşturur ve fenotipik tarama sonuçlarından moleküler hedeflerin etkin bir şekilde geri izlenmesini sağlar (Mok & Brenk, 2011). ChEMBL, kimyasal biyoloji, öncü keşif ve ilaç keşfinde hedef seçimi için küçük molekül verilerinin bulunduğu çevrimiçi bir veritabanıdır (Kufareva vd., 2014). ChEMBL, yeni veri kaynakları, geliştirilmiş işlevsellik ve yeni erişim yöntemleri ile güncellenmiş olup, bunlar arasında yeni bir Kaynak Tanım Çerçevesi formatı bulunmaktadır (Bento vd., 2014). Molekül ChEMBL kimliği, bu veritabanında yer alan her bir bileşiğin özgün tanımlayıcısıdır ve genellikle bir harf ve ardından bir dizi sayıdan oluşur. Molekül ChEMBL kimliği, her bir bileşiğin kimyasal ve biyolojik özelliklerini birbirinden ayırt etmek için kullanılır. Verilerinin doğrudan biriktirilmesini, geliştirilmiş arama ve filtreleme yeteneklerini ve yeniden tasarlanmış bir web arayüzünü mümkün kılan önemli iyileştirmelerden geçmiştir (Mendez vd., 2019). Her bir molekül ChEMBL kimliği, ChEMBL veritabanındaki ilgili bileşik hakkında bir dizi bilgiyi içeren bir kayıtla ilişkilendirilir. Bu bilgiler arasında bileşiğin kimyasal yapısı, biyolojik aktivitesi, hedef proteinlerle etkileşimi ve biyolojik

test sonuçları gibi veriler yer alabilir. Molekül ChEMBL kimlikleri, ilaç tasarımı, kimyasal biyoloji araştırmaları, biyoinformatik analizler ve ilaç geliştirme süreçlerinde kullanılır. Araştırmacılar, bu kimlik numaralarını kullanarak ChEMBL veritabanında yer alan belirli bir bileşik veya bileşik grubuyla ilgili bilgilere erişebilirler. ChEMBL, ilaç keşfi için büyük ölçekli bir biyoaktivite veritabanı olup, 1 milyondan fazla bileşik ve 5200 protein hedefi için 5.4 milyon biyoaktivite ölçümü içerir (Gaulton vd., 2012). Molekül ChEMBL kimlikleri, aynı zamanda kimyasal yapay zeka ve makine öğrenimi modelleri için eğitim verileri olarak da kullanılabilir. Bu modeller, biyolojik aktivite tahmini, ilaç-tanıtma ve ilaç-ilaç etkileşimleri gibi önemli konuları incelemek için kullanılır. ChEMBL veritabanı artık ihmal edilen hastalık taramaları, bitki koruması, ilaç metabolizması ve patentlerden veriler içerirken, iyileştirmeler ve yeni özellikler de getirilmiştir (Gaulton vd., 2017).

3.1.1.2. Kanonik SMILES

Kanonik SMILES, bir bileşiğin yapısını tanımlamanın standart bir yoludur ve aynı molekül için farklı temsilasyonların tek bir, benzersiz bir formda ifade edilmesini sağlar. Kanonik SMILES, bir bileşiğin yapısını atomlar ve bağlar arasındaki ilişkileri belirterek ifade eder. Bu sistemde, her atom bir sembolle temsil edilir ve atomlar arasındaki bağlar, bunların arasına konulan sayılarla gösterilir. Bu sayılar, bağlanan atomların sırasını belirtir ve molekülün yapısını benzersiz bir şekilde tanımlar. Bu, moleküller arasında karşılaştırmalar yapmayı ve verileri tutmayı kolaylaştırır. Kanonik SMILES, bir bileşiğin yapısını temsil etmenin yanı sıra, kimyasal veri tabanlarında arama yapmak, yapay zeka tabanlı ilaç tasarımı ve yüksek verimli sanal tarama gibi birçok uygulama için de kullanılır. Bu sistem, büyük miktarda kimyasal bilgiyi depolamak, işlemek ve analiz etmek için temel bir araçtır. Birden fazla SMILES tabanlı arttırma, ilaç keşfi görevlerinde moleküler temsil ve tahmin performansını artırır (C. Li vd., 2022). C. K. Wu vd. (2021) çalışmasında BILSTM ile SMILES numaralandırması bir araya getirildiğinde, SMILES dizelerinden gizli özellik öğrenmeyi geliştirir. Kanonik SMILES'in oluşturulması, bir dizi kurallara ve algoritmaya dayanır. Bu kurallar, atomların sıralaması, bağların belirlenmesi ve izomerlerin tanımlanması gibi faktörleri içerir. Bu sayede, herhangi bir bileşiğin yapısı, birkaç temel kurala dayanarak benzersiz ve tutarlı bir şekilde ifade edilebilir. Örneğin, Arús-Pous vd. (2019) çalışmasında rastgele SMILES eğitimli LSTM hücreleri ile çalışmış ve moleküler üretim modellerinde daha büyük kimyasal alanlara genelleme yapmayı ve hedef kimyasal alanı daha iyi temsil etmeyi iyileştirmiştir. Ayrıca Liu vd.

(2019) çalışmasında, SMILES formatındaki ilaçlara yönelik ikiz Evrişimli Sinir Ağı, ilacın kanser hücre hatları üzerindeki etkilerini yüksek doğruluk ve kararlılıkla doğru bir şekilde tahmin ettiğini, ancak kör testlerde performansın düştüğünü gözlemlemiştir. SMILES tabanlı moleküler üretim modeli, ilaç keşfi ve moleküler tasarım optimizasyonunda umut vadeden bir uygulama gösterir ve yeni terapötik varlıkların oluşturulmasını sağlar (Kong vd., 2022).

3.1.1.3. Lipinski'nin Beş Kuralı

Lipinski'nin Beş Kuralı, ilaç tasarımı ve geliştirme sürecinde kullanılan bir kılavuzdur. Bu kural, bir bileşiğin oral biyoyararlanımını öngörmek için kullanılır ve ilaçların geçirgenlik ve etkinlik açısından ne kadar başarılı olacağını tahmin etmeye yardımcı olmaktadır. Lipinski kuralı, absorpsiyon, dağılım, metabolizma ve atılım faktörlerini dikkate alarak iyi biyoyararlanımı öngörür (Ivanović vd., 2020). Lipinski'nin Beş Kuralı, farmakokinetik özellikleri vurgulayarak ve ilaç benzerliği ve ilaç yapılabilirliği kavramlarını hedef belirleme ve seçme süreçlerine entegre ederek ilaç keşfi sürecinde devrim yaratmıştır (Keller vd., 2006). Fishburn (2013) beş kuralın, yeni biyolojiklerin davranışını tahmin etmenin bilgisayar tabanlı tahminlerde bir sonraki büyük atılım olabileceğini öngörmektedir.

Beş kural şunlardır:

1. Moleküler ağırlık 500 Da'dan fazla olmamalıdır.
2. Lipofilite (logP), 5'ten fazla olmamalıdır.
3. Hidrojen bağlama aktif grup sayısı 5'ten fazla olmamalıdır.
4. Hidrojen bağlama donör sayısı 10'dan fazla olmamalıdır.
5. Moleküler polarite için aşağıdaki kuralı kullanılır: 1.7 <logP < 2.5.

Bu kural, ilaç adaylarının farmakokinetik özelliklerini değerlendirmede kullanılır. Lipinski'nin Beş Kuralı, ilaçların sindirim sisteminden emilimini öngörmek için tasarlanmıştır. Bileşiklerin bu kurala uymaması, sindirim sisteminden yeterince emilimini engelleyebilir ve dolayısıyla ilaç adayının etkinliğini azaltabilir. Nendza ve Müller (2010) çalışmalarında, Lipinski''nin Beş Kuralı ve moleküler ağırlık eşikleri ile, gereksiz biyotestlere ihtiyaç duymadan düşük kaygılı endüstriyel kimyasalların %30 ila %40'ını etkili bir şekilde tanımlayabilmektedir. Lipinski'nin Beş Kuralı'nın bazı avantajları şunlardır:

- Lipinsk'nin Beş Kuralı, ilaç keşfi sürecinin erken aşamalarında potansiyel ilaç adaylarının seçiminde kullanıldığında, maliyet tasarrufu sağlayabilir. Bu kural, daha fazla zaman ve kaynak harcanmadan önce, potansiyel olarak uygun olmayan bileşikleri eleme konusunda rehberlik eder.
- Lipinski'nin Beş Kuralı'na uyan bileşiklerin, genellikle daha iyi geçirgenlik profiline sahip olduğu kabul edilir.
- Bu kuralın temel ilkeleri basit ve anlaşılır olduğundan, araştırmacılar ve ilaç tasarımcıları tarafından geniş çapta kullanılabilir. Bu nedenle, karmaşık bir analiz gerektirmez ve hızlı bir şekilde uygulanabilir.

Lipinski'nin Beş Kuralı, umut verici bileşikleri seçmek için yararlıdır, ancak dikkatli ve kriterlerle kullanılması, umut verici bileşiklerin dışlanmasını önlemek için önemlidir (Giménez vd., 2010).

3.1.1.4. Moleküler Parmak İzi ve Tanımlayıcılar

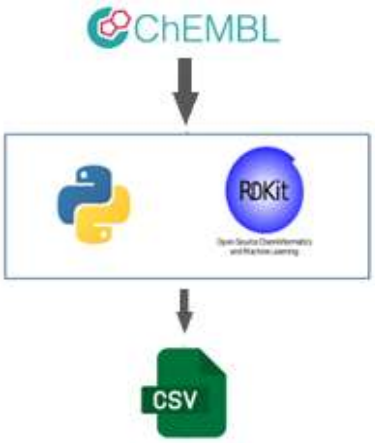
Moleküler parmak izleri, kimyasal bileşiklerin yapısını temsil etmek için kullanılan bir dizi bit dizisi veya vektörlerdir. Bu parmak izleri, kimyasal bileşiklerin yapısal özelliklerini tanımlamak ve karşılaştırmak için kullanılır. Moleküler parmak izleri, moleküler benzerlik arama, ilaç keşfi, sanal tarama ve yapı-aktivite ilişkilerinin analizi gibi birçok alanda kullanılır. Moleküler parmak izleri, kimyasal bileşiklerin benzerliklerini ve farklılıklarını analiz etmek için kullanılır. Veritabanı parmak izi, moleküler veritabanlarından önemli bilgilerini yakalar ve moleküler kütüphanelerin çeşitliliğini değerlendirmek ve kimyasal uzay karakterizasyonu ve sanal tarama yapmak için kullanılır (Fernández-De Gortari vd., 2017). Moleküler parmak izi araçları, biyolojik tarama sonuçlarının hızlı veri alışverişi, entegrasyon ve karşılaştırılmasını, keşif yapı-aktivite analizini ve hedef seçiciliği incelenmesini kolaylaştırır (Y. Wang vd., 2009).

Moleküler modelleme için açık kaynaklı PaDEL tanımlayıcı veya RDKit kütüphaneleri ile yapılan çalışmalar literatürde gözlendi. PaDEL-Descriptor, moleküler tanımlayıcılar ve parmak izleri hesaplamak için kullanılan açık kaynaklı bir yazılımdır. Harigua-Souiai vd. (2022) çalışmasında moleküler modelleme için RDKit kullandıktan sonra rastgele orman ve destek vektör makinesi modelleri ile moleküllerde anti-Leishmania etkilerini etkili bir şekilde tahmin eder ve her modelin ilk 10'unda yedi potansiyel ilacı belirler. Ayrıca RDKit, diğer kütüphaneler ile entegre çalışabilir. Örneğin, Dong vd. (2015) çalışmasında; ChemDes adını verdiği yazılımda, RDKit dahil olmak

üzere çok sayıda en son paketi entegre eder ve bu sayede ilaç keşfi süreçlerinde moleküler tanımlayıcılar ve parmak izleri hesaplamak için entegre bir web tabanlı platform ortaya çıkarmış olur. PaDEL tanımlayıcı da diğer kütüphaneler ile entegre edilerek yeni yazılımlar keşfedilebilir. Örneğin, Jiang vd. (2017) çalışmasında; DrugECs adını verdiği yeni tahmin sistemi ile, ilaç-hedef etkileşimlerini mevcut yöntemlerden daha hızlı ve verimli bir şekilde doğru tahmin eder. Ve Zhang vd. (2017) çalışmasında DrugRPE yöntemini, ilaç-hedef etkileşimlerini tahmin etmede diğer en iyi ilaç-hedef tahmincilerinden önemli ölçüde daha iyi performans gösterdiğini ve daha hızlı çalıştığını gözlemlemiştir.

RDKit, kimyasal bilim ve moleküler modelleme alanında kullanılan açık kaynaklı bir Python kütüphanesidir. Moleküler grafikler oluşturma, kimyasal özellikleri hesaplama, kimyasal veritabanları araştırma, ilaç tasarımı ve sanal tarama gibi birçok kimyasal hesaplama ve analiz işlemini gerçekleştirmek için kullanılır. RDKit'in hesaplamalı ilaç keşfindeki uygulamaları, yapı temelli ve ligand temelli ilaç tasarımı, sanal tarama teknikleri ve ilaç keşfi araştırmalarının ilerlemesine yardımcı olmak için deneysel rutinlerle entegrasyonu içerir (Macalino vd., 2015). Moleküler yapılar ve kimyasal bileşikler üzerinde çalışılırken RDKit kütüphanesi tercih edilir. Moleküler bağlanma, etkinlik merkezi modelleme, sıfırdan tasarım, moleküler benzerlik hesaplama ve dizi tabanlı sanal tarama gibi hesaplamalı ilaç keşfi yöntemleri, ilaç keşfi ve geliştirmesinde yaygın olarak uygulanmaktadır (Ou-Yang vd., 2012). RDKit kütüphanesi, açık kaynaklı bir proje olup, çeşitli akademik ve endüstriyel kuruluşlar tarafından kullanılmaktadır.

Veri setinin hazırlanması için ChEMBL veri tabanından AChE enzimi ile etkileşen moleküller toplandı. RDKit kütüphanesi ve Lipinski’nin 5 kuralı kullanılarak moleküler parmak izleri oluşturuldu. Veri setinin hazırlanması Şekil 3.1’de görülmektedir.



Şekil 3.1. Veri setinin hazırlanması

3.1.2. Veri Ön İşleme

Veri ön işleme, veri setlerinin analize hazırlanmasında kullanılır. Doğru bir veri ön işleme, analiz sürecinin doğruluğunu ve etkinliğini arttırmaktadır. Veri setlerindeki özniteliklerin anlamlı hale getirilmesi ve yeni özniteliklerin oluşturulması sürecine öznitelik çıkarımı denir. Bu aşama, daha etkili analizler yapılmasına olanak tanır. Öznitelik çıkarımı, veri setinin boyutunu azaltabilir ve analiz sürecini hızlandırabilir. Verilerin belirli bir aralığa veya dağılıma getirilmesi, analiz sonuçlarından daha tutarlı sonuçlar elde edilmesini sağlar. Farklı kaynaklardan elde edilen verilerin birleştirilmesi ve uyumlu bir formata getirilmesi veri entegrasyonu süreci ile gerçekleşir. Veri entegrasyonu veri bütünlüğünü sağlamak için önemlidir.

Veri ön işleme için temel teknikler, tahmin edici veri analizi çerçevesinde her bir ön işleme adımı için yaygın olarak kullanılan güncel algoritmaları içerir (Alexandropoulos vd., 2019). Veri ön işleme için kullanılan bir diğer yöntem ise standart sapma filtresidir. Standart sapma filtresi, veri setindeki anormal değerleri tespit etmek ve bu değerleri veri setinden çıkarmak için kullanılan bir yöntemdir. Matematiksel olarak, standart sapma aşağıdaki denklem ile hesaplanır:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} \qquad \textbf{(3.1.)}$$

Burada:

- x_i verinin i’ inci gözlemini,

- μ: veri setinin ortalamasını,
- N: veri setindeki toplam gözlem sayısını

ifade etmektedir. Veri setinde standart sapma filtresi uygulama 3 adımdan oluşmaktadır.

1. Verinin ortalaması (μ) ve standart sapması (σ) hesaplanır.
2. Veri setindeki anormal değerleri tanımlamak için belirli bir standart sapma eşiği kullanılır. Anormal değerlerin sınırları μ ± kσ aralığında tanımlanır.
3. Bu sınırlar dışında kalan veriler anormal olarak kabul edilir ve işleme dahil edilmez.

Veri ön işleme sürecindeki kararlar, sonraki analiz aşamalarının başarısını belirleyebilir. Veri akışları için veri ön işleme teknikleri, öğrenme hızını ve veri yapısı anlayışını artırabilir, ancak gelecekteki zorluklar için daha fazla araştırma ve geliştirme gerektirir (Ramírez-Gallego vd., 2017).

3.1.3. Makine Öğrenmesi ve Regresyon Yöntemleri

Makine öğrenimi modelleri algoritmalar tarafından temsil edilir. Bu algoritmalar, veriye dayalı olarak belirli bir amaca ulaşmak için optimize edilir. Üç temel makine öğrenimi türü vardır: denetimli, denetimsiz ve pekiştirmeli öğrenme (Scott vd., 2019).

- Denetimli Öğrenme: Bu yöntemde, makine öğrenimi modeline eğitim verisi olarak etiketlenmiş veri sağlanır. Model, bu verileri kullanarak girdi ve çıktı arasındaki ilişkiyi öğrenir ve yeni verileri tahmin etmek için bu ilişkiyi kullanır. Sınıflandırma ve regresyon görevlerinde kullanılır.
- Denetimsiz Öğrenme: Bu yöntemde, model etiketlenmemiş verilerle eğitilir. Model, veri içindeki örüntüleri veya yapılardaki doğal gruplamaları belirlemek için kendi kendine öğrenme yeteneğine dayanır. Kümeleme ve boyut azaltma görevlerinde kullanılır.
- Pekiştirmeli Öğrenme: Bu yöntemde, model bir çevreyle etkileşime girer ve belirli bir görevi en iyi şekilde gerçekleştirmek için ödüller veya cezalar alır. Model, deney ve hata yaparak en iyi davranışı öğrenir. Otomatik sürüş, oyun stratejileri ve robotik gibi alanlarda kullanılır.

Makine öğrenimi, deneyimle gelişen bilgisayar algoritmaları geliştirmeyi amaçlayarak, genom dizileme veri setleri gibi büyük, karmaşık veri setlerinin analizinde

insanlara yardımcı olmayı hedefler (Libbrecht & Noble, 2015). Makine öğrenimi, çok boyutlu veri kümelerinden tahmin modelleri oluşturmak için kullanılan veri analitik tekniklerinin bir koleksiyonudur ve modern biyolojik araştırmaları etkiler (Camacho vd., 2018). Örnek olarak destek vektör makineleri, karar ağaçları, XGBOOST, rassal orman regresyonu, PLS regresyon algoritmaları verilebilir. Makine öğrenimi modelleri eğitim, doğrulama ve test aşamalarından geçer. Eğitim verileri, modelin öğrenme sürecinde kullanılırken, doğrulama verileri modelin performansını değerlendirmek için kullanılır ve test verileri modelin gerçek dünya performansını ölçmek için kullanılır.

Makine öğrenimi ve derin öğrenme algoritmaları, büyük veri teknolojileri ile birlikte ilaç keşfi ve hedef tasarımının verimliliğini, etkinliğini ve kalitesini artırır (Patel vd., 2020). Makine öğrenimi, veri odaklı karar alma sürecini teşvik ederek ilaç keşfini ve geliştirmesini iyileştirebilir, ancak yorumlanabilirlik ve tekrarlanabilirlik gibi zorluklar daha fazla çalışma gerektirir. Makine öğreniminin ilaç keşfi alanında karşılaştığı zorluklar, makine öğrenmesi tarafından üretilen sonuçların yorumlanabilirliğinin ve tekrarlanabilirliğinin eksikliği olarak öne çıkar, bu da uygulama alanlarını sınırlayabilir (Vamathevan vd., 2019). Gelişmiş makine öğrenimi teknikleri, veri seyrekliği, yorumlanabilirlik ve otonom davranış gibi ilaç keşfindeki sınırlamaları aşabilir. İlaç keşfinde makine öğrenimi, yorumlanabilirlik eksikliği ve dağıtımdan sonraki yeniden eğitim ihtiyacı gibi sınırlamalarla karşı karşıyadır (Elbadawi vd., 2021). Tıp ve sağlık hizmetlerinde makine öğrenimi, özellikle karmaşık lineer olmayan modeller için model yorumlanabilirliği ve açıklanabilirlik açısından zorluklarla karşılaşır (Vellido, 2020). Bu zorluklar uzman desteğiyle aşılabilir.

3.1.3.1. Ridge

Ridge regresyonu, çok faktörlü verilerde karmaşık ilişkileri göstermek için iki boyutlu grafiksel bir prosedürdür ve olağan en küçük kare tahmininden daha iyi bir regresyon denklemi sağlar. Çoklu regresyonda bağımlı olmayan sorunlar için daha küçük ortalama kare hataya sahip önyargılı tahminler elde etmek için köşegen yönüne küçük pozitif miktarlar ekleyen bir tahmin prosedürüdür (Hoerl & Kennard, 2000). Ridge regresyonu, üç açıklayıcı değişken ile yapılan tüm araştırmalarda olağan en küçük kareler tarafından domine edilen bir k parametresine sahiptir (Al-Hassan & Mohammad Al-Hassan, 2008). Ridge regresyonu, doğrusal regresyon yöntemine bir düzenleme terimi ekleyerek çalışır ve bu sayede modelin karmaşıklığını azaltır ve aşırı uyumu önler. Bu

düzenleme terimi, katsayıların karelerinin toplamı olarak ifade edilir ve bu katsayıların büyüklüklerinin toplamının belirli bir eşik değerden küçük olması için optimize edilir. Düzenleme terimi, bazı katsayıları küçültür ancak sıfıra indirmez. Bu sayede, modeldeki tüm özelliklerin etkisini azaltırken, hiçbir özelliği tamamen atlamaz. Bu, ridge regresyonunun daha kapsamlı bir model oluşturmasını sağlar.

Ridge regresyonu, en küçük kare tahminlerini belirsiz olarak sıfıra doğru küçülterek destekleyen bir yöntemdir, ancak bağımlı olmayan verilerin zayıf olarak yanlış etiketlenmesi ve duruma özgü bilgiye dayanma gibi zayıflıklara sahiptir (Smith & Campbell, 2018). Bu nedenle, ridge regresyonu modeldeki katsayıları büyüklüklerine göre sınırlayarak modelin karmaşıklığını azaltır. Bağımlı bir değişken ile bazı açıklayıcı değerler arasında lineer bir ilişkiyi modelleme için istatistiksel bir yöntemdir (Y. R. Chen vd., 2018). Özellikle yüksek boyutlu veri setlerinde ve değişkenler arasında çoklu doğrusal bağlantıların olduğu durumlarda etkili bir şekilde kullanılır. Bu tür durumlarda, geleneksel regresyon yöntemleri performans kaybı yaşayabilirken, ridge regresyonu daha istikrarlı ve doğru sonuçlar verebilir. Ridge regresyonunun ana avantajlarından biri, modelin düzgünleştirilmesi ve aşırı uyumu azaltmasıdır.

3.1.3.2. PLS Regresyonu

PLS regresyon, çoklu değişkenler arasındaki ilişkiyi modellemek ve tahmin etmek için kullanılan bir regresyon tekniğidir. PLS, regresyon ve faktör analizi tekniklerinin bir kombinasyonunu içeren bir yöntemdir. Diğer regresyon yöntemlerinin zayıflıklarını ele alan bir yöntemdir (Geladi & Kowalski, 1986). Bir bağımlı değişkeni bir dizi açıklayıcı değişkenle ilişkilendiren ve belirgin açıklayıcı değişkenleri ve korunan PLS bileşenlerinin sayısını ayarlayarak olağan en küçük kareler kullanılan bir modeldir (Bastien vd., 2005). Tahmin edici değişkenlerin sonlu küme durumunun bir genişlemesi olup, yakınsama özellikleri kanıtlanmıştır (Preda & Saporta, 2005). PLS regresyonu, statik sistem modellemesinde girişler ile çıktılar arasındaki gizli ilişkiyi yakalamak için kullanılır (Y. Dong & Qin, 2018). Çoklu bağımsız değişkenlerin çoklu bağlantı, gereksizlik ve gürültüden etkilendiği çoklu yanıtlı regresyonu gerçekleştirmek için kullanılan çok değişkenli bir tekniktir (Stocchero vd., 2021).

PLS regresyonu, iki ana bileşen ile çalışır: bileşenler ve bileşen yükleri. Bileşenler, bağımsız değişkenler arasındaki ortak varyansı temsil ederken, bileşen yükleri, bağımsız değişkenlerin bağımlı değişkenle olan ilişkisini gösterir. PLS, bu

bileşenler ve yükler arasındaki ilişkiyi kullanarak tahminler yapar. Bağımlı değişkenleri bağımsız değişkenlerden en iyi tahmin edici güce sahip gizli değişkenleri çıkararak tahmin eder (Abdi, 2010). Bir veya daha fazla bağımlı değişkeni iki veya daha fazla bağımsız değişkene ilişkilendirir (Lorber vd., 1987). Özellikle örneklerden daha fazla değişken olduğunda çok değişkenli kalibrasyon için kullanır (Wakeling & Morris, 1993). PLS regresyon çoklu değişkenler arasındaki karmaşık ilişkileri modellemek için kimya, biyoloji, ekonomi, pazarlama ve mühendislik gibi çeşitli disiplinlerde yaygın olarak kullanılmaktadır.

3.1.3.3. Rassal Orman Regresyonu

Rassal orman regresyonu, birkaç rastgele karar ağacını bir araya getirerek ve tahminlerini ortalamayla birleştirerek genel amaçlı bir sınıflandırma ve regresyon yöntemidir (Biau & Scornet, 2016). Her bir karar ağacı, veri setinin farklı alt kümesi üzerinde eğitilir ve farklı özelliklerle çalışır. Bu, her bir ağacın birbirinden bağımsız olmasını sağlar ve modelin genelleme yeteneğini artırır. Rassal orman, yüksek boyutlu regresyon ve sınıflandırma için kullanılan bir makine öğrenme aracıdır ve yanıt değişkeninin koşullu ortalamasının doğru bir yaklaşımını sağlar (Meinshausen, 2006).

Rassal orman regresyonu, veri setinin rastgele seçilen alt uzaylarında büyüyen karar ağaçlarını kullanan bir modeldir ve bir tahminci topluluğu oluşturur (Biau & Fr, 2012). Her bir karar ağacı eğitilirken, rastgele seçilen özellikler üzerinde çalışması nedeniyle her bir ağacın farklı özellikler üzerinde eğitilerek geçerli bir model oluşturmasını sağlar. Ayrıca, rastgele özellik seçimi, modelin aşırı uyumu azaltmasına yardımcı olur. Rassal orman regresyonunun bir diğer önemli özelliği, toplu tahmin yapma yeteneğidir. Bu özellik, modelin daha doğru ve güvenilir tahminler yapmasını sağlar. Yüksek boyutlu veri ile sınıflandırma için bir araçtır ve aday tahmincileri değişken önem ölçümleri aracılığıyla sıralar (Scornet vd., 2015).

Rassal orman regresyonu, yanıt değişkeninin tam koşullu dağılımı hakkında bilgi sağlar ve yüksek boyutlu tahminci değişkenler için koşullu tahmin edebilir, bu da tahmin gücünde rekabetçi kılar (Meinshausen, 2006). Önemli değişkenleri bulmaya ve iyi bir sade tahmin modeli tasarlamaya yardımcı olabilir (Genuer vd., 2010). Sınıflandırma ve regresyonda yüksek doğruluk sağlar, az ayar gerektirirler ve yorumlanabilir çıktılar sağlar (Sega & Xiao, 2011).

3.1.3.4. XGBOOST

XGBOOST, makine öğrenimi alanında yaygın olarak kullanılan bir öğrenme algoritmasıdır. Regresyon ve sınıflandırma problemlerinde başarılı sonuçlar verir. XGBOOST, üstün tahmin performansına sahip bir gradyan artırma karar ağacı modelidir ve birçok sınıflandırma zorluğunda kullanılmaktadır (Sagi & Rokach, 2021). Bir bağımlı değişkenin diğer özellikler tarafından nasıl etkilendiğini veya tahmin edileceğini öğrenmek için kullanılabilir. Regresyon problemlerinde, bağımlı değişken sürekli bir sayısal değerdir ve bu değeri tahmin etmek için XGBOOST regresyonu kullanılır. Standart rastgele orman ve gradyan artırma makinelerinden daha iyi performans gösteren ağaç tabanlı bir toplu öğrenim tekniğidir ve habitat uygunluk modellemesinde kullanılır (Muñoz-Mas vd., 2019). Birden fazla karar ağacını bir araya getirerek bir tahmin modeli oluşturur. Bu ağaçlar, veri setinin farklı alt kümeleri üzerinde eğitilir ve birbirlerinin hatalarını düzelterek tahmin performansını artırırlar. Ayrıca XGBOOST, mineral potansiyel haritalama için optimum tahmin modelleri üretmede rastgele ormanlardan biraz daha iyi performans gösteren bir toplu öğrenim tekniğidir (Parsa, 2021). Veri dengesizliği durumlarında avantajları olan etkili bir toplu öğrenim algoritmasıdır (P. Zhang vd., 2022). XGBOOST regresyonunun temel avantajlarından biri, yüksek performansı ve genelleme yeteneğidir. Büyük veri setleri ile çalışırken ve karmaşık ilişkileri modellemek istendiğinde etkili bir şekilde çalışabilir.

XGBOOST, birçok makine öğrenme zorluğunda en son teknolojiyi başarıyla elde eder ve mevcut sistemlere kıyasla çok daha az kaynak kullanarak milyarlarca örneğin ötesine ölçeklenir (T. Chen & Guestrin, 2016). Eğitim hızı, genelleme performansı ve parametre kurulumunda avantajları olan güvenilir ve verimli bir makine öğrenme zorluğu çözücüdür (Bentéjac vd., 2019). Özellik uzayında karmaşık veri dağılımını aşarak yapılandırılmış verilerde sınıflandırma performansında daha yüksek doğruluk elde eder (J. Wu vd., 2021).

3.1.3.5. SVR

Support Vector Regression (SVR), Support Vector Machine (SVM) algoritmasının regresyon için uyarlanmış halidir. SVR, verileri sınıflandırmak yerine, bir dizi veri noktası arasındaki ilişkileri modelleyerek, sürekli bir çıktıya ulaşmayı amaçlar. SVM'de olduğu gibi, SVR de doğrusal olmayan verilerle başa çıkabilmek için çekirdek fonksiyonlarını kullanır.

SVR'nin temel amacı, verilerin çoğunu doğru tahmin edecek şekilde doğrusal olmayan bir fonksiyon bulmaktır. Bunu yaparken, belirli bir hata payı (epsilon) içinde kalan verileri dikkate alır. Epsilon, modelin ne kadar hata yapabileceğini belirler; yani tahmin edilen değer ile gerçek değer arasındaki fark epsilon değerini aşmazsa, model bu hatayı göz ardı eder. Bu sayede, modelin aşırı öğrenme yapması ve dolayısıyla genel performansının düşmesi engellenir.

SVR, ayrıca bir ceza parametresi olan C'yi kullanır. C parametresi, hatalara ne kadar tolerans gösterileceğini belirler. C değeri büyükse, model hatalara daha az tolerans gösterir ve tüm veri noktalarına uyan bir model oluşturmaya çalışır. Bu durumda, model daha karmaşık hale gelebilir ve aşırı öğrenmeye yatkın olabilir. C değeri küçükse, model daha fazla hatayı göz ardı edebilir, bu da modelin daha basit olmasına ve daha iyi genelleştirilmesine yol açabilir.

SVR'nin bir diğer önemli özelliği de çekirdek fonksiyonları kullanarak doğrusal olmayan veri kümelerini modelleyebilmesidir. Çekirdek fonksiyonları, veriyi daha yüksek boyutlu bir uzaya dönüştürerek, orada doğrusal olarak ayrılabilir hale getirir. Hangi çekirdek fonksiyonunun kullanılacağı, veri kümesine ve problemin doğasına bağlı olarak seçilir. Doğrusal SVR eğitim yöntemleri, bazı problemler için çekirdek SVR kadar iyi modeller üretebilir ve daha hızlı eğitim ve test olanağı sunabilir (Ho & Lin, 2012).

Literatürde SVR ile çalışmalar mevcuttur. Örneğin, Liu ve Zio (2016) çalışmalarında SVR için önerilen çevrimiçi uyarlanabilir öğrenme yaklaşımını uygulamış ve SVR için durağan olmayan koşullar için modelleri etkili bir şekilde değiştirerek hesaplama karmaşıklığını azaltır ve aşırı uyumu önler yorumunda bulunmuştur. Wu ve Huang (2019) çalışmalarında ise SVR öğrenme yönteminin, geleneksel yöntemlere kıyasla geniş bantlı varış yönü tahmin doğruluğunu iyileştirdiğini ve açı belirsizliğinin elde edilebildiği frekans aralığını genişlettiği sonucuna varmıştır.

3.1.3.6. Toplu Öğrenim Yöntemi

Toplu öğrenim yöntemi, birçok farklı modelin bir araya getirilerek daha güçlü ve kararlı bir model oluşturmak için kullanıldığı bir makine öğrenimi tekniğidir. Bu yöntem, birden fazla zayıf öğreniciyi bir araya getirerek daha güçlü bir öğrenici oluşturmayı amaçlar. Toplu öğrenim yöntemi, farklı öğrenme algoritmalarını veya aynı algoritmayı farklı alt örneklemelerle eğitmeyi içerebilir. Toplu öğrenim yöntemi, birçok makine

öğrenimi zorluğu için kabul edilen en son teknoloji çözümü olarak kabul edilebilir. Bu tür yöntemler, tek bir modelin tahmin performansını, birden fazla modeli eğiterek ve tahminlerini birleştirerek artırır (Sagi & Rokach, 2018). Bir toplu öğrenim yöntemi modelinde kullanılacak çeşitli öğreniciler seçilir veya oluşturulur. Bu öğreniciler, farklı algoritmaları veya aynı algoritmayı farklı parametrelerle kullanarak eğitilebilir. Veri setinden rastgele örnekler alınarak farklı öğreniciler için eğitim verisi oluşturulur. Bu örnekleme işlemi kümelenme veya artırma gibi tekniklerle yapılır. Kümelenme, yüksek varyanslı modellerle kullanılır. Kümelenme, rastgele alt örneklemler oluşturarak farklı öğrenicileri eğitmeyi ve sonuçlarını birleştirmeyi içerir. Her öğrenici, kendi alt örnekleminde eğitilir ve daha sonra tahminler birleştirilerek toplu öğrenim modelinin sonucunu oluşturulur. Bu yöntem, aşırı uyumu azaltır ve daha kararlı sonuçlar sağlar. Artırma, önceki öğrenicilerin hatalarına odaklanarak her bir öğreniciyi öncekinden daha iyi hale getirmeye çalışır. Her öğrenici, önceki öğrenicilerin hatalarını düzeltmeye odaklanarak eğitilir. AdaBoost ve gradyan artırma algoritmaları bulunmaktadır. Öğrenici eğitiminde her öğrenici, kendi eğitim verisi üzerinde eğitilir. Bu adımda, öğrenici modelin uygun parametreleri belirlenir ve veriye uyum sağlaması sağlanır. Eğitilen öğrenicilerin tahminleri bir araya getirilerek toplu öğrenim yöntemi modelinin tahmini oluşturulur. Bu birleştirme ağırlıklı oylama yöntemleriyle yapılır. Toplu öğrenim yöntemi, özellikle kümelenme ve artırma, geleneksel sinir ağlarının iflas tahmin görevlerinde performansını artırır (M. J. Kim & Kang, 2010). Oylama yönteminde, her öğrenici model kendi tahminini yapar ve bu tahminlerin çoğunluğu alınarak nihai tahmin belirlenir. Ağırlıklı oylama yönteminde ise, her bir öğrenici modelin tahmini, modelin performansına veya güvenilirliğine bağlı olarak belirlenen ağırlıklarla değerlendirilir. Daha güvenilir veya daha iyi performans gösteren modellerin tahminlerine daha yüksek ağırlık verilirken, zayıf veya hatalı modellerin tahminlerine daha düşük ağırlık verilebilir. Bu şekilde, daha güçlü ve dengeli bir tahmin elde edilir.

Toplu öğrenim yöntemi, makine öğrenimi alanında yaygın olarak kullanılan ve etkili bir tekniktir (Wyatt vd., 2005). Toplu öğrenim yöntemi, model çeşitliliğini temel öğrenici seçimi için bir uygunluk fonksiyonu olarak kullanarak öğrenici modellerin genelleme gücünü ve güvenilirliğini örnekleme ve optimizasyon teknikleriyle artırır (D. Wang & Alhamdoosh, 2013). Toplu öğrenim yöntemi, veri birleştirme, veri modelleme ve veri analizini birleştiren birleşik bir çerçevede birden fazla öğrenme algoritması

kullanarak, daha iyi bilgi keşfi ve tahmin performansı için makine öğreniminde tahmin performansını artırır (X. Dong vd., 2020).

Mienye ve Sun'un (2022) çalışmasındaki toplu öğrenim yöntemi modeli, rassal orman, uyarlanabilir boosting, gradyan boosting ve kategorik boosting gibi popüler algoritmaları ve kümelenme, artırma ve yığınlamayı içerir. Kao vd. (2021) çalışması sonucu elde ettiği EnsembleDLM ile, kimyasal bileşik ve protein dizileri kullanan bir derin öğrenme modelidir ve farklı biyolojik aktivite türleri ve protein sınıfları arasında ilaç-hedef etkileşimi tahmininde son teknoloji durumunu başarır. Pliakos ve Ven'in (2020) çalışmasında önerilen yöntem, yeniden yapılan ağlarda bi-küme ağaçlarını kullanarak ilaç-hedef etkileşimi tahmin doğruluğunu ve verimliliğini artırırken, ölçeklenebilirliği, yorumlanabilirliği ve indüktif ayarını korur.

Toplu öğrenim yönteminin önemli avantajı farklı öğrenme algoritmalarını veya aynı algoritmayı farklı alt örneklemlerle eğiterek daha güçlü ve kararlı bir model oluşturabilmesidir. Bu yöntem, farklı öğrenicilerin farklı hatalar yapma eğiliminde olması nedeniyle daha iyi genelleme yapar. Makine öğreniminde toplu öğrenim yöntemleri, kümelenme, arttırma ve rassal orman gibi birden fazla model kullanarak daha iyi performans elde eder (Ren vd., 2016). Toplu öğrenim yöntemi aynı zamanda aşırı uyumu azaltabilir, çünkü farklı öğrenicilerin farklı veri alt kümeleri üzerinde eğitilmesi ve tahminlerinin birleştirilmesi, modelin daha genel geçerli olmasını sağlar. Bu yöntem ayrıca daha kararlı sonuçlar sağlar, çünkü farklı öğrenicilerin tahminleri birleştirilerek bir ortalamaya ulaşılır. Toplu öğrenim yöntemi modelleri, karışım yöntemleri gibi, regresyon görevlerinde daha iyi tahmin performansı ve birleştirilmiş yorumlama sağlar (C. H. Chen vd., 2020).

3.1.4. Performans Ölçümü

3.1.4.1. R-kare

R-kare, istatistik ve regresyon analizinde kullanılan bir ölçüdür ve bir regresyon modelinin bağımsız değişkenler tarafından açıklanan varyansın yüzdesini ifade eder. R-kare, bir regresyon modelinin uygunluğunu ve açıklama gücünü değerlendirmek için kullanılır. R-kare, logit, probit, poisson, geometrik, gama ve üstel gibi doğrusal olmayan regresyon modelleri için uygunluk ölçüsüdür (Cameron & Windmeijer, 1996). R-kare, aşağıdaki denklem ile hesaplanır:

$$R^2 = 1 - \frac{\sum_{i-1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i-1}^{n}(y_i - \bar{y})^2} \quad \textbf{(3.2.)}$$

Burada:

- y_i: Gerçek değerleri,
- $\hat{y}_i$: Model tarafından tahmin edilen değerleri,
- $\acute{y}$: Gerçek değerlerin ortalamasını

ifade etmektedir. R-kare değeri, 0 ile 1 arasında bir değer alır. Değer 1 sayısına ne kadar yakınsa, modelin bağımsız değişkenler tarafından bağımlı değişkenin varyansını ne kadar iyi açıkladığını gösterir. Ayarlanmış R-kare tipi ölçüt, üstel dağılım modellerinde popülasyon değerinin yaklaşık olarak yansız bir kestiricisidir (Ricci, 2010). R-kare değeri 1'e eşit olamaz; çünkü bu, modelin hatasız olduğu anlamına gelir. R-kare değeri 0'a yaklaştıkça, modelin bağımsız değişkenler tarafından bağımlı değişkenin varyansını açıklama yeteneği azalır ve modelin uygunluğu kötüleşir. Negatif R-kare değerleri modelin beklenenden daha kötü tahminler yaptığını gösterir. En küçük kareler regresyonuna dayalı R-kare, kare hataların toplamını en aza indirir, ancak aykırı gözlemlere duyarlıdır (Saleh, 2014).

3.1.4.2. Mean Squared Error

Mean Squared Error **(**MSE), istatistik ve makine öğrenimi alanlarında kullanılan bir hata ölçüsüdür ve bir regresyon modelinin tahminlerinin gerçek değerlerden ne kadar uzak olduğunu ölçer. Regresyon analizinde MSE, regresyonun doğrusal mı yoksa doğrusal olmayan mı olduğunu gösteren ortalama karesel hatadır (Altman & Krzywinski, 2016). Regresyon analizinde MSE, tahmin edici hatadaki hatadır ve bu hata kareköküyle ölçülür (Namba, 2001). MSE, aşağıdaki denklem ile hesaplanır:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad \textbf{(3.3)}$$

Burada:

- n: Toplam veri noktası sayısını,

- y_i: Gerçek değerleri,
- $\hat{y}_i$: Model tarafından tahmin edilen değerleri

ifade etmektedir. Gupta vd. (2009) çalışmasındaki regresyon analizinde MSE'yi gözlemlenen verilerle hidrolojik modellerin kalibre edilmesi ve değerlendirilmesinde kullanılan bir kriter olarak tanımlamıştır. Regresyon analizinde MSE, her bir bireysel regresyon katsayısının tahmininden kaynaklanan hatadır (Namba, 2015). Regresyon analizinde MSE, regresyonun performansını gerçek değer öğelerinin dağılımı açısından değerlendirmek için kullanılan bir ölçümdür (Chicco vd., 2021). MSE hataların karelerini aldığı için negatif olamaz. Küçük MSE değerleri, modelin tahminlerinin gerçek değerlere yakın olduğunu ve modelin iyi bir uyum sağladığını gösterir. Büyük MSE değerleri, modelin tahminlerinin gerçek değerlerden uzak olduğunu ve modelin uygunluğunun düşük olduğunu gösterir.

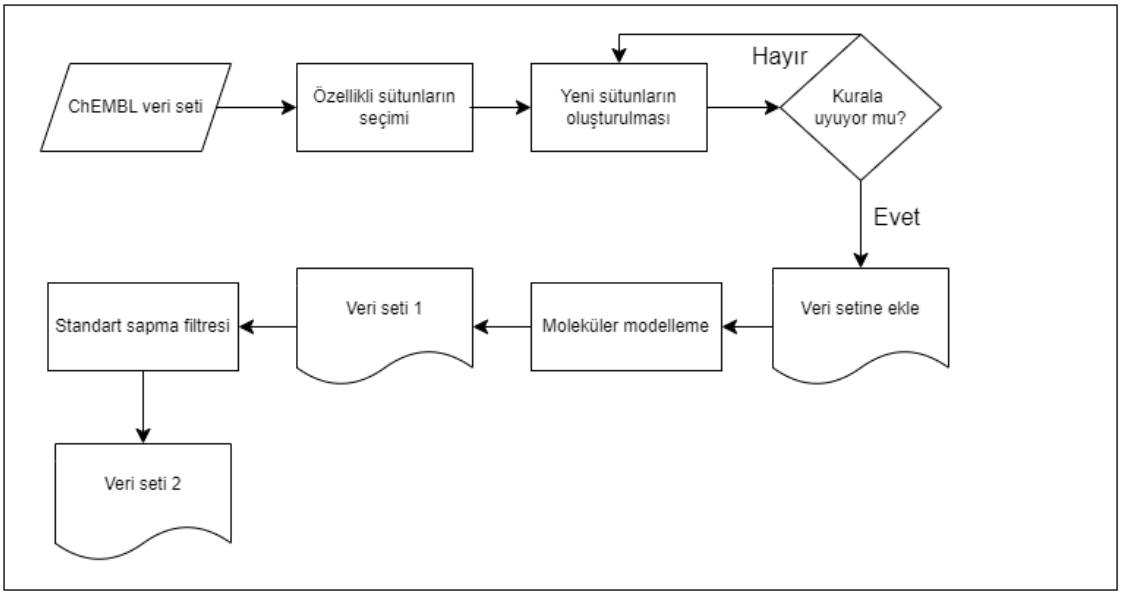
3.1.4.3. K-Katlı Çapraz Geçerleme

K-katlı çapraz geçerleme, makine öğrenimi modellerinin performansını değerlendirmek için kullanılan bir teknik olup, modelin genelleme yeteneğini test etmek amacıyla verinin farklı bölümler üzerinde nasıl çalıştığını ölçer. Bu yöntem, veri setinin eğitim ve test olarak bölünmesi sürecine dayanmaktadır, bu bölünme tek bir defa değil, K kez tekrarlanır. Çapraz geçerleme, test edilen model için gerçek tahmin hatasını değil, görülmemiş eğitim setlerine uyan modellerin ortalama tahmin hatasını tahmin eder (Bates vd., 2024).

Örneğin, bir 5 katlı çapraz geçerlemede, veri seti 5 eşit parçaya bölünür. Model ilk dört kat üzerinde eğitilir ve kalan bir kat üzerinde test edilir. Bu işlem, her kat bir kez test verisi olarak kullanıldığında tamamlanır ve sonuçlar kaydedilir. Bu sürecin sonunda, modelin performansını değerlendirmek için elde edilen sonuçların ortalaması alınır. Bu ortalama, modelin genel başarımı hakkında daha güvenilir bir tahmin sağlar. Model seçimi için çapraz geçerleme prosedürleri etkililik açısından farklılık gösterir ve sorunun belirli özelliklerine göre en iyi modeli seçmek için yönergeler sağlanır (Arlot & Celisse, 2010). Lei (2020) çalışmasında, çapraz geçerlemeye dayalı yeni bir istatistiksel araç oluşturarak, tutarlı değişken seçimi sağlamış ve çeşitli istatistiksel ve makine öğrenimi ortamlarında tahmin doğruluğu ile model yorumlanabilirliği arasında alternatif bir denge sağlamıştır.

3.2. Çalışmanın Akışı

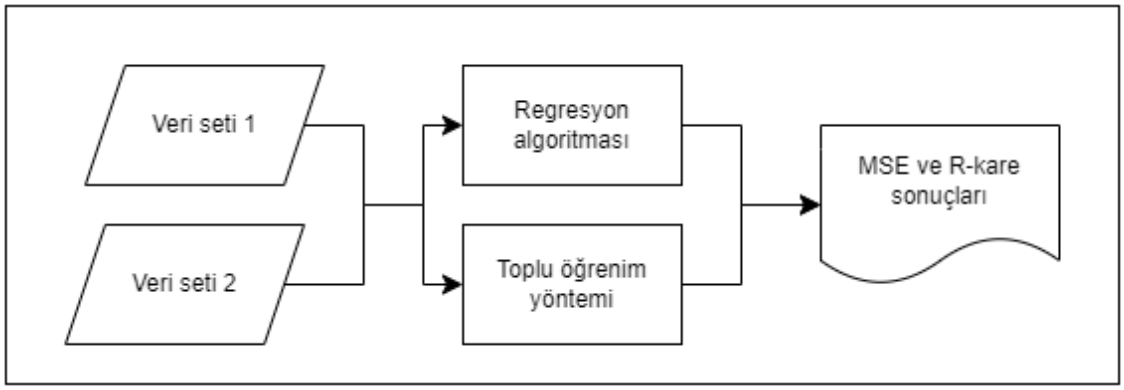
Öncelikle, geniş bir enzim-ilaç etkileşimi veri seti toplandı ve veri madenciliği çalışması uygulandı. Veri seti hazırlama aşamaları Şekil 3.2.'de verilmiştir.



Şekil 3.2. Veri madenciliği iş akış diyagramı

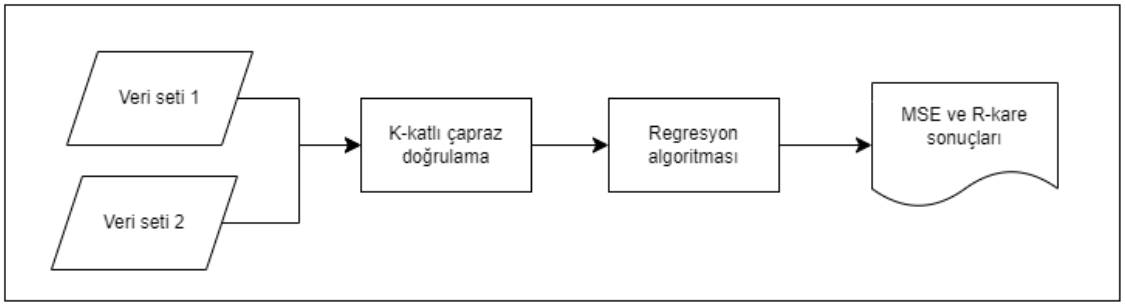
Makine öğrenmesi algoritmaları ile regresyon çalışması için uygun özelliklerin seçimi yapıldı. Özellik seçimi aşamasında, enzim aktivitesi üzerinde belirleyici olan özellikler belirlendi. Ardından, çeşitli makine öğrenmesi algoritmaları ve toplu öğrenim yöntemleri kullanılarak IC50 değerlerinin tahmin edilmesi için uygulamalar gerçekleştirildi. Algoritmaların performansını artırmak için k-katlı çapraz geçerleme ve topluluk öğrenimi teknikleri uygulandı.

Bu çalışmada, geliştirilen makine öğrenmesi algoritmalarının performansı kapsamlı bir şekilde değerlendirildi. Algoritma ve yöntemler arasında karşılaştırmalı analizler kullanılarak, hangi yöntemin en iyi tahmin yeteneğine sahip olduğu belirlendi. Algoritmaların eğitim ve test aşamalarındaki başarı metrikleri detaylı olarak incelendi ve sonuçlar istatistiksel olarak değerlendirildi. Makine öğrenimi için uygulanan adımlar Şekil 3.3.'te belirtilmiştir.



Şekil 3.3. Regresyon uygulaması 1 iş akış diyagramı

Regresyon eğitim sonuçlarının başarısını artırmak amacıyla k-katlı çapraz geçerleme uygulandı. Bu uygulama için akış diyagramı Şekil 3.4.'te verilmiştir.



Şekil 3.4. Regresyon uygulaması 2 iş akış diyagramı

4. ARAŞTIRMA BULGULARI VE TARTIŞMA

4.1. Veri

4.1.1. Veri Seti Hakkında

Enzim veri setine ChEMBL veri tabanından erişilmiştir. Bu veri tabanından AChE enzimine ait ChEMBL220 kimlik numarasına sahip veri seti seçilmiştir. Veri setinde enzime ait benzersiz ChEMBL veri tabanı kimlik numarası, veri geçerliliği ile ilgili yorumlar, standartlaştırılmış veri olarak adlandırılan IC50 değerinin türü ve birimi gibi enzim hakkında detaylı bilgiler bulunmaktadır. Enzim veri setindeki moleküllere ait özellikler ve açıklamaları Çizelge Ek-1.1.'de, ham veriye ait ilk 5 örnek ise Çizelge Ek-1.2.'de verilmiştir.

4.1.2. Veri Madenciliği Uygulaması

AChE veri setinde 7027 satır ve 43 öznitelikli sütun bulunmaktaydı. Bu veri seti için IC50 değerini temsil eden standard_value sütunu verileri ile regresyon çalışması yapılmıştır. Veri setinde, AChE enzimiyle etkileşen çeşitli bileşikler bulunmaktadır. Veri setindeki IC50 değerleri, farklı kimyasal bileşiklerin AChE'yi inhibe etme (durdurma) yeteneğini gösterir. Her bileşik farklı yapıya sahiptir ve bu bileşiklerin hangi formüllere sahip olduğu ChEMBL veri tabanındaki kimlikleriyle bulunabilir. IC50 değeri, her bir bileşiğin AChE'yi %50 oranında durdurabilmesi için ne kadarının gerekli olduğunu ifade etmektedir. Küçük bir IC50 değeri, daha az bileşikle enzimin durdurulabileceği anlamına gelir, bu da bileşiğin daha etkili olduğunu göstermektedir. Bu sütunun minimum değeri -324090 ve maksimum değeri 7,07946E+16 idi. Değer aralığı çok büyük olduğu için bu sütunun negatif logaritmik değerleri ile çalışılmaya karar verildi. Veri setinde pchembl_value sütununda verilen negatif logaritmik IC50 değerlerinin doğrulaması pIC50 = -log10(IC50) formülü ile yapıldı. Dönüşüm sonucunda sütundaki minimum değer 1.0, maksimum değer 11.22, ortalama değer ise 6.13 olarak gözlendi.

Veri madenciliği çalışması için veri setinden kanonik SMILES sütunu seçilmiştir. Kanonik SMILES, kimyasal bir bileşiğin benzersiz bir biçimde ifade edilmesi için kullanılan bir notasyondur. Elementin kimyasal sembolü ile temsil edilmiştir ve bağlar arasındaki ilişki çizgi ile gösterilmektedir. Kanonik SMILES sütunu üzerinde Bölüm 3.1.1.3.'te bahsedilen Lipinski'nin 5 kuralı uygulanarak 4 yeni sütun elde edilmiştir.

Çizelge 4.1.'de veri setine ait ilk 5 verinin Lipinski'nin 5 kuralı uygulaması ile elde edilen yeni değerleri verilmiştir.

Çizelge 4.1. Lipinski' nin 5 kuralı uygulaması ile elde edilen yeni sütunlar ve sütunlara ait ilk 5 veri

İndeks	**Moleküler Ağırlık**	**Lipofilite Değeri**	**Hidrojen Bağı Verici Atom Sayısı**	**Hidrojen Bağı Alıcı Atom Sayısı**
0	312.325	2.80320	0.0	6.0
1	376.913	4.55460	0.0	5.0
2	426.851	5.35740	0.0	5.0
3	404.845	4.70690	0.0	5.0
4	346.334	3.09530	0.0	6.0

Yapılan madencilik çalışması sonucunda moleküler parmak izi hesaplaması için oluşturulan veri seti 1 şu sütunlardan oluşmaktadır:

- Kanonik SMILES,
- Molekül ağırlığı,
- Lipofilite değeri,
- Hidrojen bağı verici atomların sayısı,
- Hidrojen bağı alıcı atomların sayısı

Moleküler parmak izi hesaplaması için Python programlama dili kullanılmıştır. Açık kaynaklı RDKit kütüphanesi ile molekül yapılar analiz edilmiştir. RDKit kütüphanesi, moleküller üzerinde çeşitli kimyasal ve fiziksel tanımlayıcıları hesaplayan bir fonksiyon tanımlanır. Bu tanımlayıcılar arasında moleküler parmak izi hesaplamaları, molekül ağırlığı, valans elektronlarının sayısı, topolojik polar yüzey alanı (TPSA), dönebilen bağların sayısı, hidrojen bağı verici ve alıcı atomların sayısı gibi önemli özellikler yer alır. Bu tanımlayıcılar, moleküllerin kimyasal reaktiviteleri, biyoyararlanımları, esneklikleri ve biyolojik hedeflerle etkileşimlerini anlamak için kullanılır. Kanonik SMILES bilgileri ile molekül dizisi oluşturulur ve bu moleküllerin parmak izleri çeşitli bitlik vektörler halinde hesaplanır. Hesaplanan parmak izleri ve diğer tanımlayıcılar bir veri çerçevesine eklenir ve moleküllerin biyolojik aktivitelerini ve kimyasal davranışlarını tahmin etmek amacıyla analiz edilir. Moleküler parmak izini hesaplarken aşağıda belirtilen parametreler dikkate alınır.

- Molekül içindeki atomların çevresindeki ikinci dereceden komşuları dahil edilir.

- Atom sayıları parmak izi hesaplamasına dahil edilir.
- Molekülün parmak izi bit vektörü olarak kayıt edilir.
- Molekülün kütlesi ve fiziksel özelliklerini anlamada moleküler ağırlık kullanılır.
- Molekülün biyolojik hedeflerle etkileşimini anlamak için moleküldeki hidrojen bağı verici ve alıcı atomların sayısını hesaplanır.
- Kimyasal reaktiviteyi anlamak için moleküldeki valans elektronlarının sayısı hesaplanır.
- Biyoyararlanım ve çözünürlük tahminlerinde kullanmak için molekülün polar yüzey alanı hesaplanır.
- Molekülün esnekliğini ve biyolojik sistemlerdeki davranışını anlamak için moleküldeki dönebilen bağların sayısını hesaplanır.

Moleküler parmak izi uzunluğu 100 olarak seçildiğinde elde edilen örnek bit vektörü Çizelge 4.2.'de verilmiştir.

Çizelge 4.2. NBits=100 için üretilen parmak izi vektörüne ait örnek parmak izi vektörü

Chem_0	**Chem_1**	**Chem_2**	**…**	**…**	**…**	**Chem_97**	**Chem_98**	**Chem_99**
1.0	0.0	1.0				0.0	1.0	0.0

Moleküler parmak izi hesaplaması yapıldıktan sonra elde edilen veri seti üzerinde standart sapma filtresi uygulanmıştır. Standart sapma filtresi ile veri seti üzerinde yapılan analizde, standart sapması 0.3'ün üzerinde olan özellikler dikkate alındı. Standart sapması bu eşiğin üzerinde olan özellikler ile veri setinin daha belirgin varyansına sahip özellikler seçilerek çalışma sonuçlarında iyileşme yapılması amaçlandı ve sonucunda veri seti 2 elde edildi. Çizelge 4.3.'te belirtilen her NBits değeri için veri seti 1 ve veri seti 2 oluşturuldu. Çalışma sonucunda 20 veri seti elde edildi ve regresyon analizinde kullanıldı.

Çizelge 4.3. Veri seti 1 NBits değerlerinin veri seti 2 için karşılık değerleri

Veri Seti 1 Sütun (NBits) Sayısı	100	600	1100	1600	2100	2600	3100	3600	4100	4600
Veri Seti 2 Sütun (NBits) Sayısı	97	119	101	98	100	89	88	95	91	84

Moleküler parmak izi, moleküler yapının özetlenmiş bir temsili olarak kullanılır. NBits parametresi, parmak izinin ne kadar ayrıntılı veya genel olacağını belirler. Örneğin, NBits=300 belirtildiğinde, oluşturulan parmak izi 300 bit uzunluğunda olacaktır. Bu, molekülün özelliklerini temsil etmek için 300 farklı yapı kullanılacağı anlamına gelir. Daha büyük NBits değerleri daha küçük NBits değerlerine kıyasla daha fazla hesaplama gücü gerektiren parmak izleri sağlamıştır. Çeşitli uzunlukta(NBits) vektör oluşturmanın temel amacı, moleküler yapıların çeşitli özelliklerini doğru bir şekilde temsil etmek için en uygun NBits değerini belirlemek ve moleküler parmak izi oluştururken en etkili ve verimli yaklaşımı belirlemektir. Farklı NBits değerlerinin kullanılmasıyla elde edilen parmak izleri arasındaki kalite ve ayrıntı düzeyi arasındaki ilişkiyi anlamak için sistematik denemeler gerçekleştirilmiştir. Bu çalışma ile, moleküler yapıların çeşitli analizlerinde ve özellikle ilaç keşfi gibi alanlarda kullanılan moleküler parmak izi yöntemlerinin optimize edilmesine katkıda bulunmayı amaçlanmaktadır.

4.2. Regresyon Analizi ve Tahmin Sonuçları

4.2.1. K-Katlı Çapraz Geçerleme ile Regresyon Analizi ve Tahmin Sonuçları

Bu çalışmada, veri madenciliği uygulamasından sonra elde edilen veri seti üzerinde 5 farklı makine öğrenimi algoritmalarının performansı değerlendirilmiş ve veri seti üzerindeki madencilik çalışmalarının tahmin sonuçlarını nasıl etkilediğini gözlemlemek amaçlanmıştır. Regresyon analizinde IC50 değerinin tahminlenmesi için aşağıdaki algoritmalar kullanılmıştır:

- Rassal Orman Regresyonu,
- XGBOOST Regresyonu,
- Ridge Regresyonu,

- SVR,
- PLS Regresyonu

İlk olarak, veri seti 1 üzerinde k-katlı çapraz geçerleme kullanılarak makine öğrenme algoritmaları ile regresyon uygulamaları gerçekleştirildi. Sonrasında veri seti 2 üzerinde k-katlı çapraz geçerleme kullanılarak makine öğrenme algoritmaları ile regresyon uygulamaları gerçekleştirildi. Bu regresyon uygulamasında, rassal orman regresyonu, SVR, XGBOOST regresyonu, ridge regresyonu ve PLS regresyonu algoritmaları kullanıldı. Sonuçlar Çizelge 4.4. ve Çizelge 4.5.'da karşılaştırıldı. Elde edilen en yüksek r-kare değeri ve en düşük MSE değeri kalın fontta belirtilmiştir.

Çizelge 4.4'e bakıldığında veri seti 1 üzerinde rassal orman regresyonu 4600 NBits uzunluğunda 0.7379 r-kare değeri ile verideki varyansı en iyi açıklayan yöntem olmuştur. Ayrıca 0.6287 hata oranıyla en düşük hata oranına da sahiptir. NBits sayısı azaldıkça korelasyonun azaldığı ve hatanın yükseldiği gözlenmiştir. En başarısız sonuçlar ise 0.3509 r-kare ve 1.5111 MSE ile 100 NBits uzunluğundaki vektörleri kullanan PLS Regresyonu yöntemi elde etmiştir.

Çizelge 4.4. K-katlı çapraz geçerleme uygulanan veri seti 1 için NBits sayısına göre performans karşılaştırılması

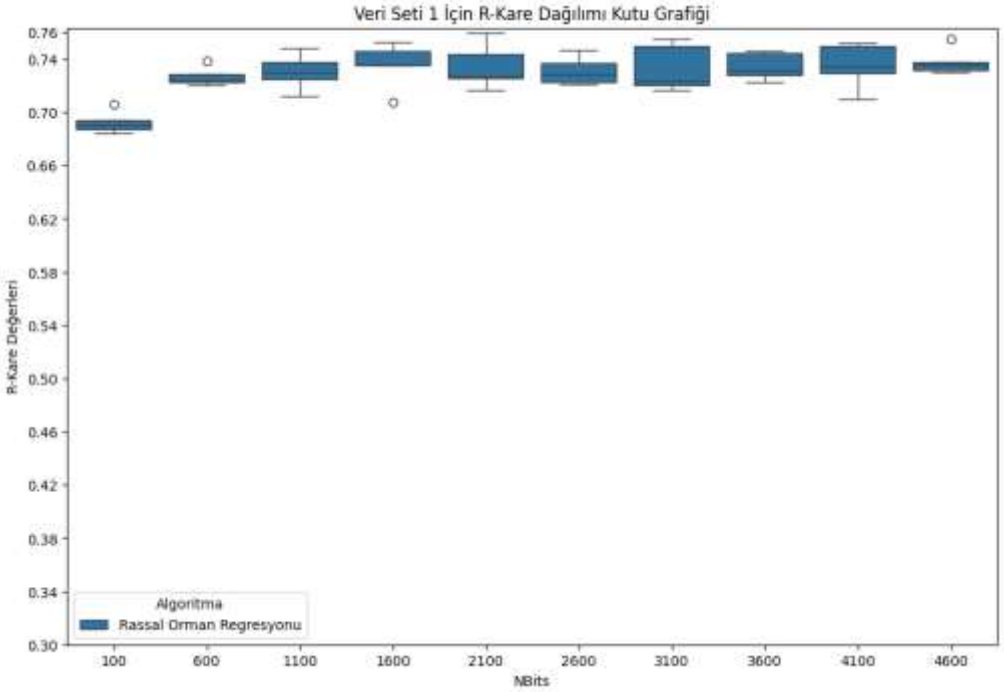
NBits	Rassal Orman Regresyonu		XGBOOST Regresyonu		Ridge Regresyonu		SVR		PLS Regresyonu	
	r-kare	MSE	r-kare	MSE	r-kare	MSE	r-kare	MSE	r-kare	MSE
4600	**0.7379**	**0.6287**	0.7233	0.6670	0.6684	0.7988	0.7229	0.6666	0.6027	0.9555
4100	0.7350	0.6370	0.7204	0.6729	0.6639	0.8075	0.7214	0.6701	0.5987	0.9646
3600	0.7345	0.6448	0.7235	0.6659	0.6546	0.8305	0.7217	0.6694	0.5940	0.9764
3100	0.7331	0.6417	0.7211	0.6720	0.6451	0.8551	0.7200	0.6735	0.5834	1.0023
2600	0.7311	0.6464	0.7195	0.6755	0.6211	0.9099	0.7187	0.6764	0.5751	1.0212
2100	0.7344	0.6384	0.7245	0.6636	0.6003	0.9611	0.7200	0.6733	0.5706	1.0326
1600	0.7352	0.6336	0.7166	0.6830	0.5946	0.9748	0.7165	0.6819	0.5551	1.0706
1100	0.7305	0.6448	0.7225	0.6680	0.5735	1.0253	0.7128	0.6908	0.5204	1.1536
600	0.7267	0.6664	0.7240	0.6646	0.5715	1.0305	0.7135	0.6893	0.4815	1.2481
100	0.6922	0.7376	0.6925	0.7403	0.3733	1.5077	0.6634	0.8102	0.3509	1.5111

Çizelge 4.5'ya bakıldığında veri seti 2 üzerinde rassal orman regresyonu en yüksek r-kare değerini 0.7130 ile 4600 NBits sayısında elde etmiştir. Aynı değeri 600 NBits sayısı ile XGBOOST regresyonu da aynı korelasyon değerine erişebilmiştir. Ayrıca en düşük hata oranına 0.6914 değeri ile 2100 ve 600 Nbits sayılarında XGBOOST regresyonu sahiptir. Veri seti 2'de 600 Nbits 0.3 standart sapma değeri ile filtrelenerek 119 uzunluğuna indirgenmiştir. Çizelge 4.4. incelendiğinde Veri seti 2'de en çok özelliğe sahip veridir. Bu sebeple diğerlerine oranla daha yüksek başarım elde ettiği düşünülmektedir.

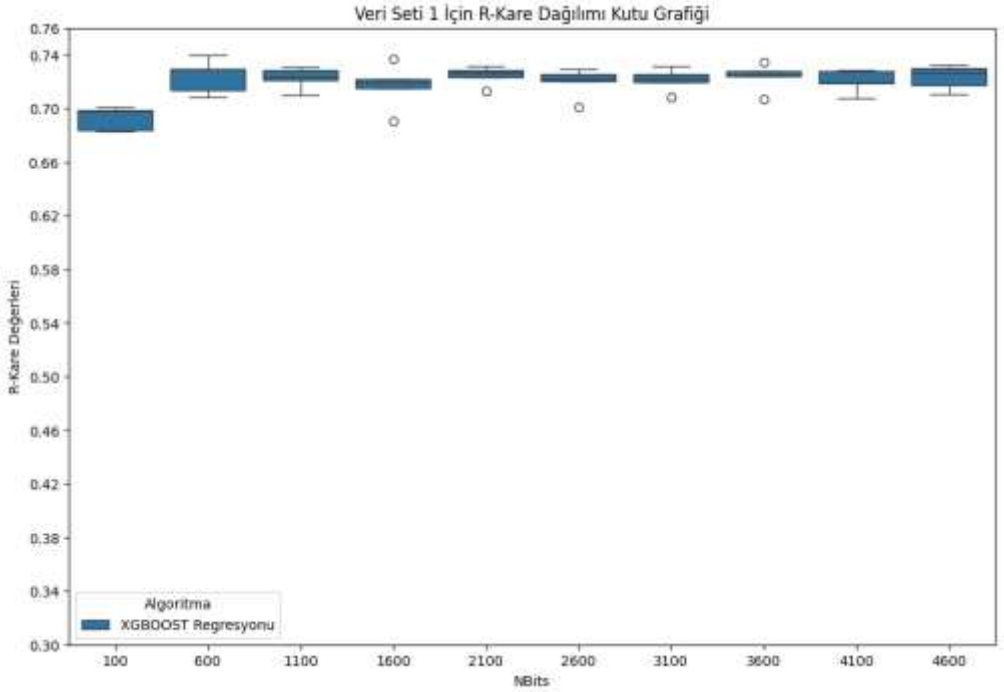
Çizelge 4.5. 5-katlı çapraz geçerleme uygulanan veri seti 2 için NBits sayısına göre performans karşılaştırılması

NBits	Rassal Orman Regresyonu		XGBOOST Regresyonu		Ridge Regresyonu		SVR		PLS Regresyonu	
	r-kare	MSE	r-kare	MSE	r-kare	MSE	r-kare	MSE	r-kare	MSE
4600	**0.7130**	0.6939	0.7044	0.7114	0.3926	1.4614	0.6294	0.8919	0.2841	1.7221
4100	0.7030	0.7058	0.6974	0.7271	0.3857	1.4775	0.6254	0.9015	0.2847	1.7203
3600	0.7067	0.7027	0.7035	0.7151	0.4068	1.4270	0.6442	0.8558	0.2894	1.7089
3100	0.7026	0.7200	0.7008	0.7202	0.3769	1.4989	0.6301	0.8903	0.2770	1.7389
2600	0.7082	0.7054	0.7009	0.7209	0.3861	1.4761	0.6329	0.8831	0.2802	1.7313
2100	0.7075	0.7007	0.7127	**0.6914**	0.4389	1.3504	0.6380	0.8708	0.3222	1.6302
1600	0.7074	0.7102	0.7096	0.7000	0.4118	1.4152	0.6472	0.8490	0.2978	1.6886
1100	0.7085	0.7039	0.7065	0.7065	0.4221	1.3906	0.6445	0.8555	0.3099	1.6600
600	0.7093	0.6996	**0.7130**	**0.6914**	0.4360	1.3563	0.6741	0.7841	0.3305	1.6100
100	0.6921	0.742 3	0.6847	0.7591	0.3696	1.5167	0.6661	0.8038	0.3473	1.5706

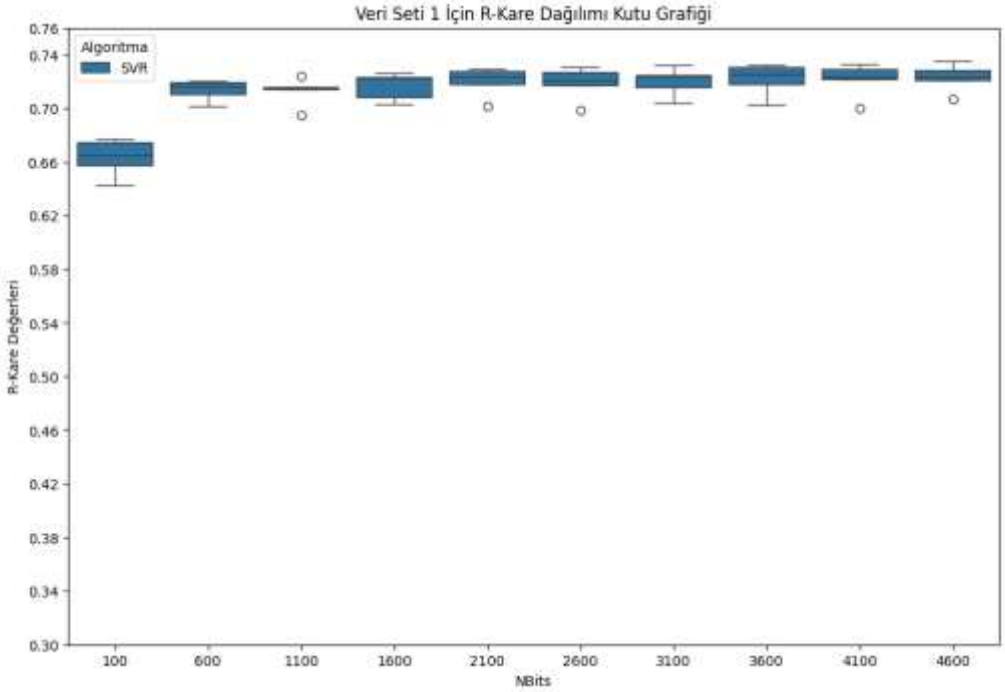
Veri seti 1 üzerinde 5-katlı çapraz geçerleme yöntemi uygulandı. Algoritmaların bu çalışmaya ait r-kare ve MSE değerleri Şekil 4.1., Şekil 4.2., Şekil 4.3., Şekil 4.4., Şekil 4.5., Şekil 4.6., Şekil 4.7., Şekil 4.8., Şekil 4.9., Şekil 4.10.'da verilmiştir. Bu şekiller incelendiğinde veri seti 1 için 5-katlı çapraz geçerleme uygulamasında rassal orman regresyonu, XGBOOST ve SVR regresyonu performanslarının ridge ve PLS regresyonuna göre daha az yayılım gösterdiği gözlendi. Bu durum ilk iki algoritmanın her bir çapraz doğrulama adımında benzer performans gösterdiğini ifade etmektedir.



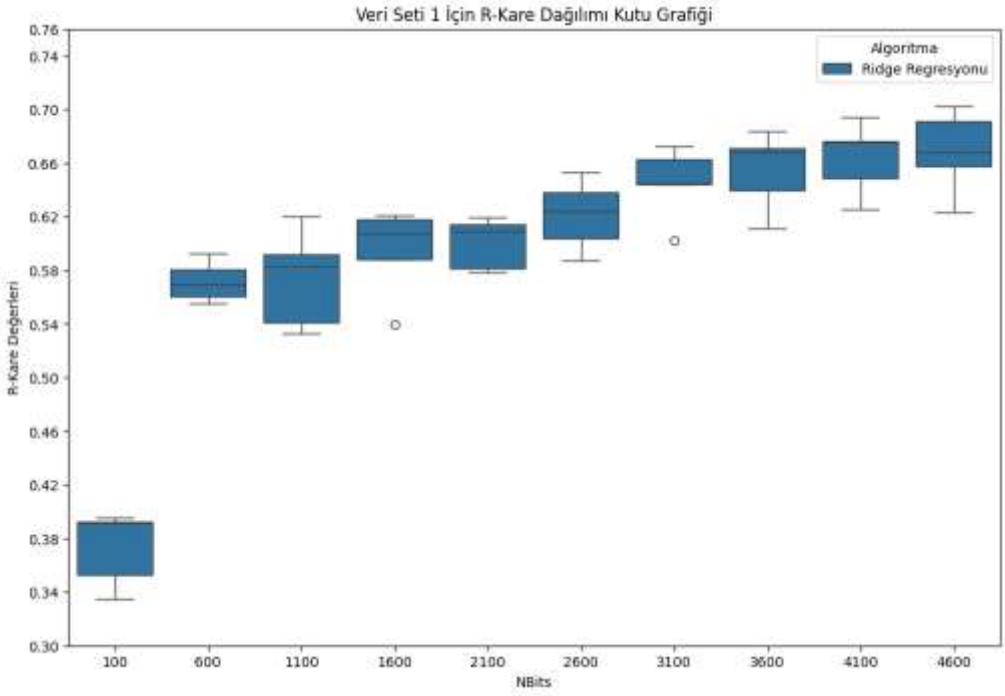
Şekil 4.1. Veri seti 1 üzerinde Rassal Orman Regresyonu algoritması için NBits sayısına göre r-kare dağılımı



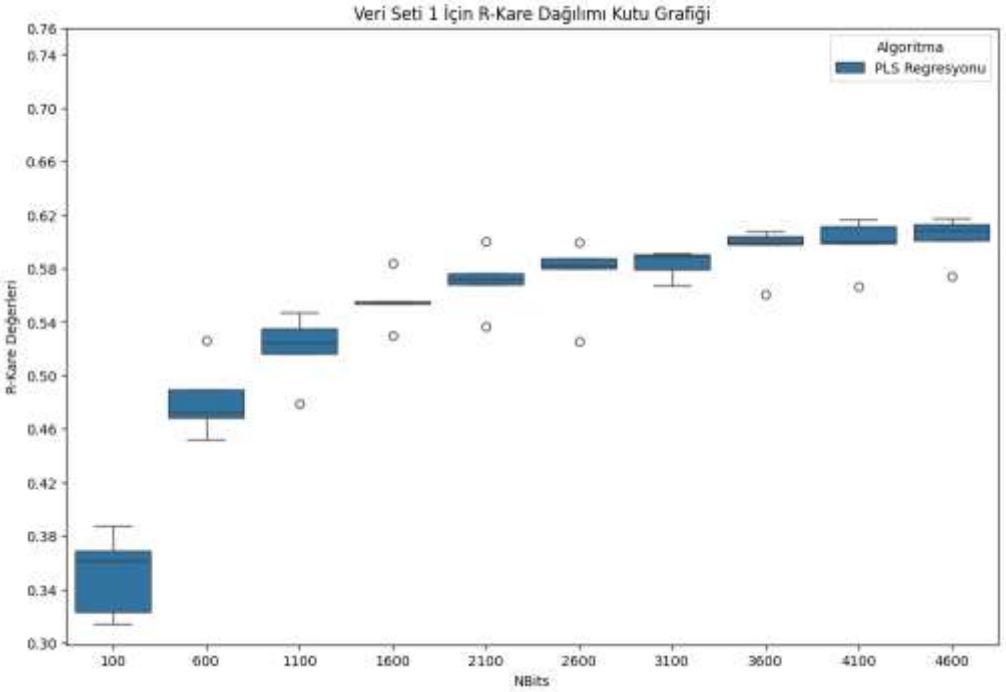
Şekil 4.2. Veri seti 1 üzerinde XGBOOST Regresyonu algoritması için NBits sayısına göre r-kare dağılımı



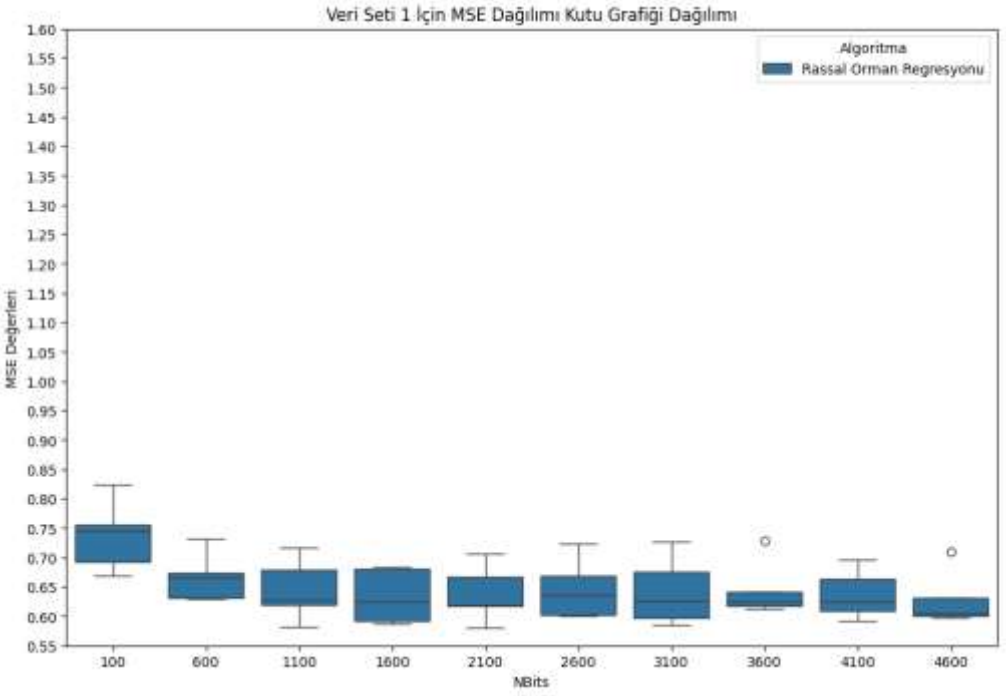
Şekil 4.3. Veri seti 1 üzerinde SVR algoritması için NBits sayısına göre r-kare dağılımı



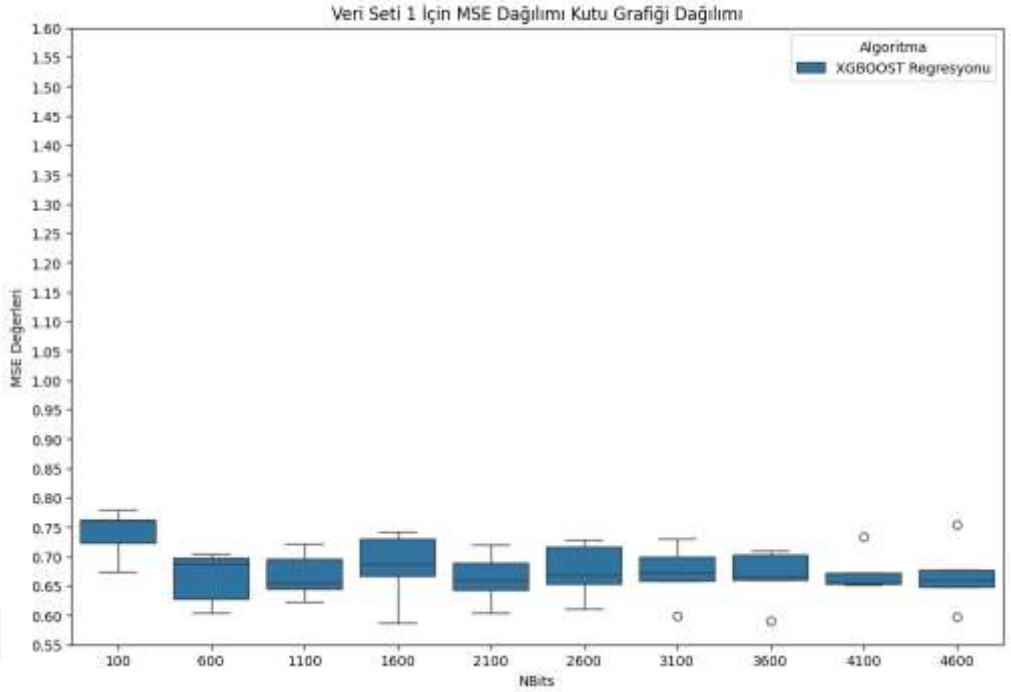
Şekil 4.4. Veri seti 1 üzerinde Ridge Regresyonu algoritması için NBits sayısına göre r-kare dağılımı



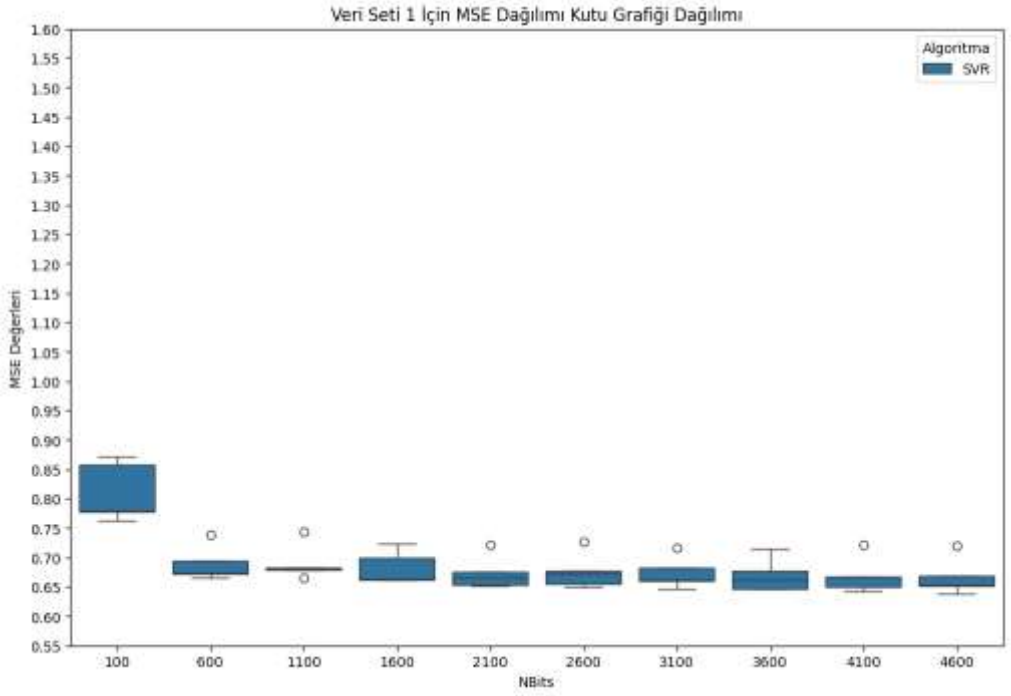
Şekil 4.5. Veri seti 1 üzerinde PLS Regresyonu algoritması için NBits sayısına göre r-kare dağılımı



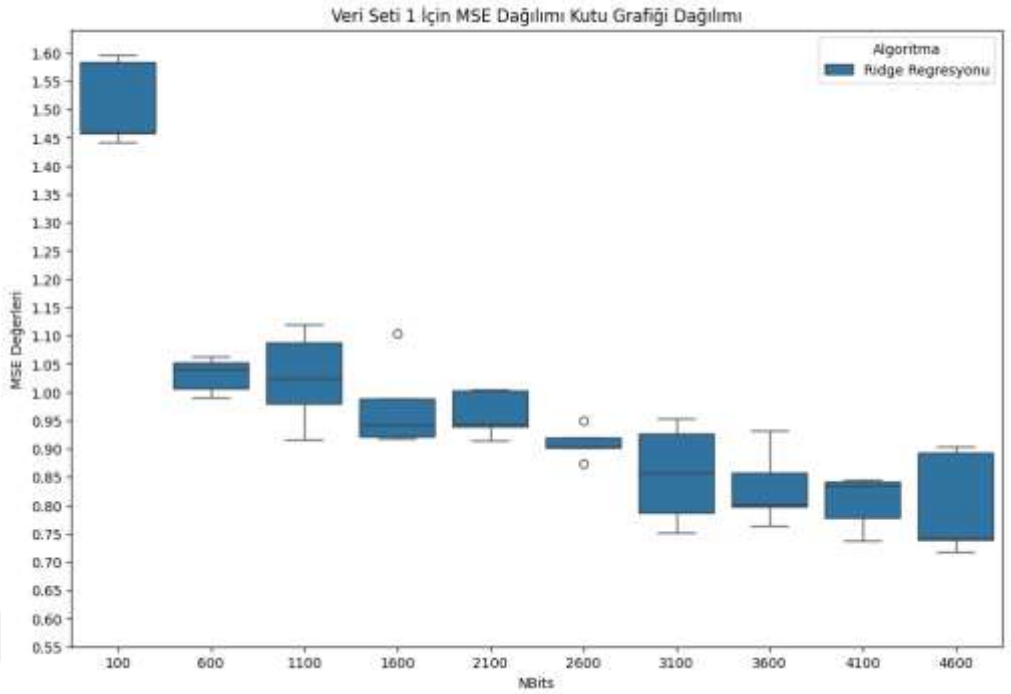
Şekil 4.6. Veri seti 1 üzerinde Rassal Orman Regresyonu algoritması için NBits sayısına göre MSE dağılımı



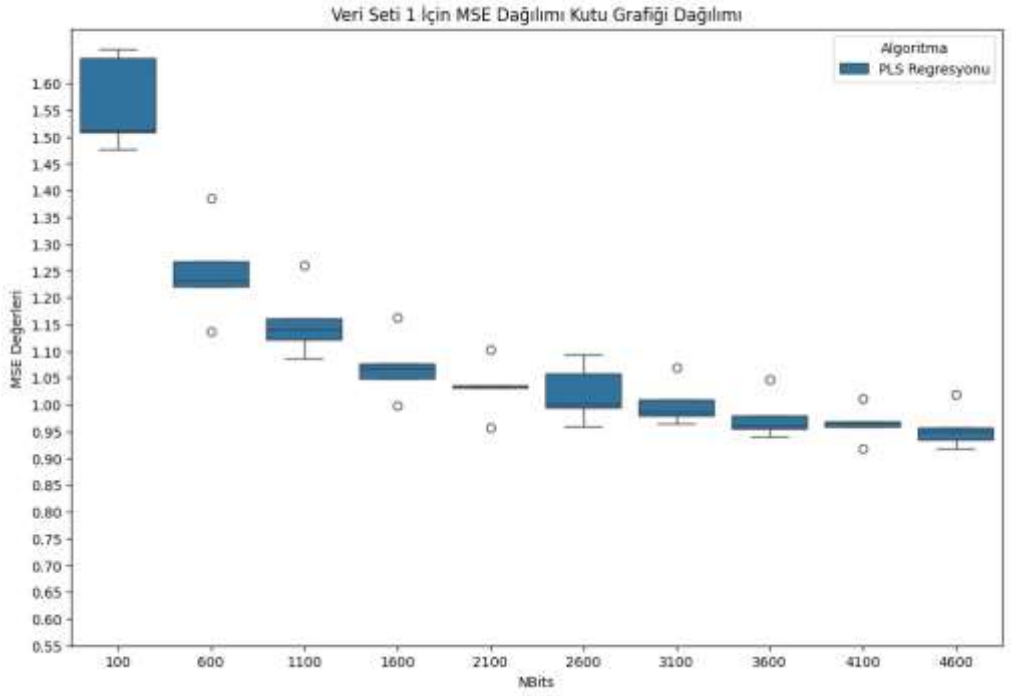
Şekil 4.7. Veri seti 1 üzerinde XGBOOST Regresyonu algoritması için NBits sayısına göre MSE dağılımı



Şekil 4.8. Veri seti 1 üzerinde SVR algoritması için NBits sayısına göre MSE dağılımı

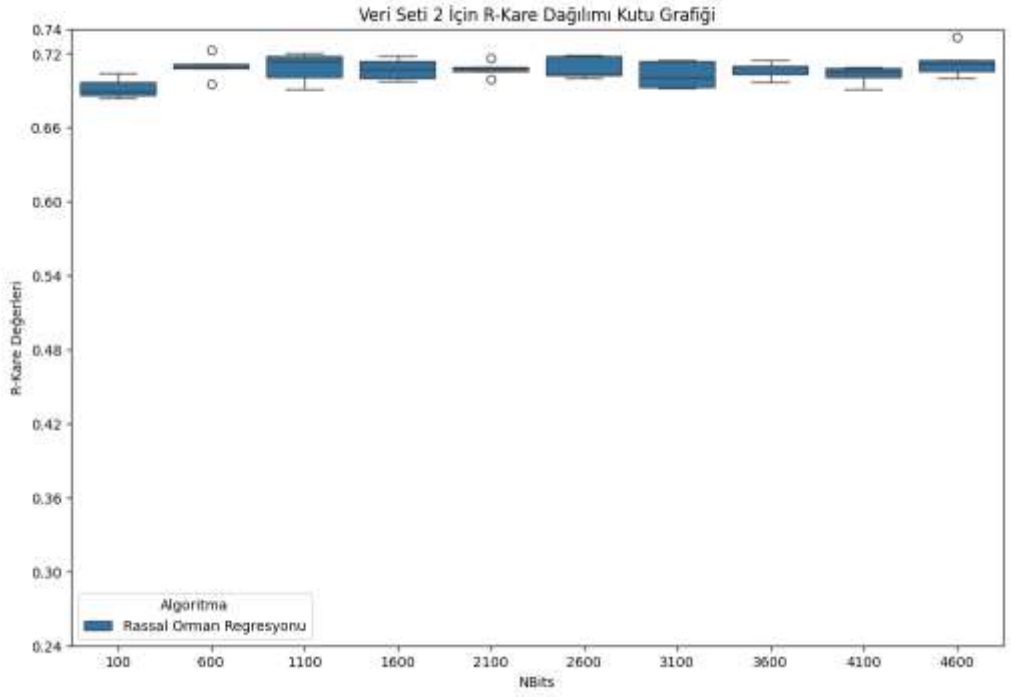


Şekil 4.9. Veri seti 1 üzerinde Ridge Regresyonu algoritması için NBits sayısına göre MSE dağılımı

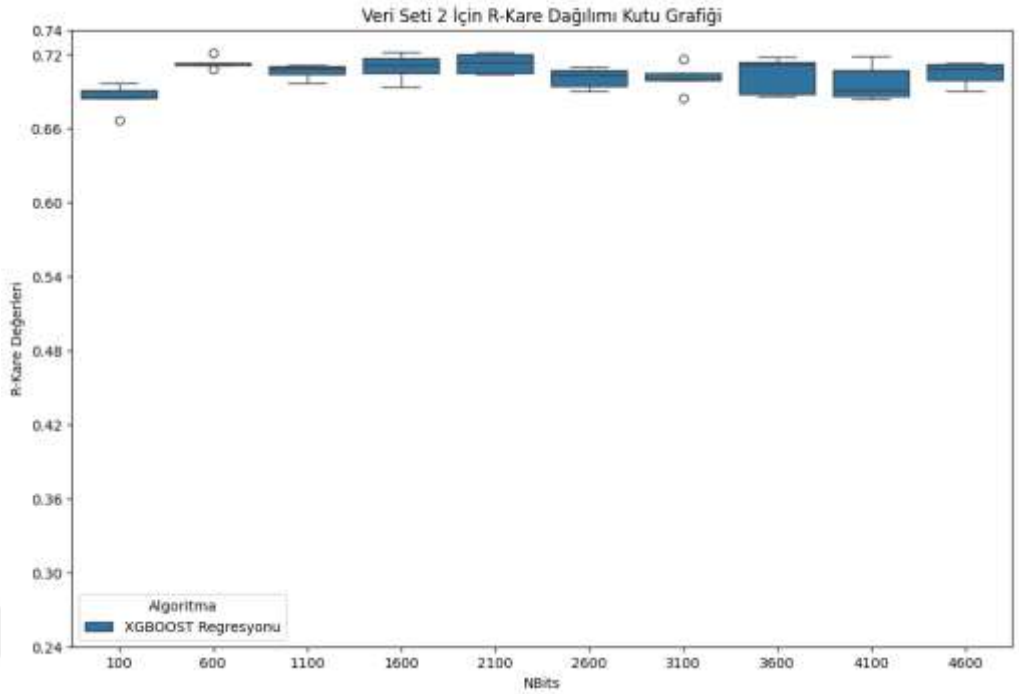


Şekil 4.10. Veri seti 1 üzerinde PLS Regresyonu algoritması için NBits sayısına göre MSE dağılımı

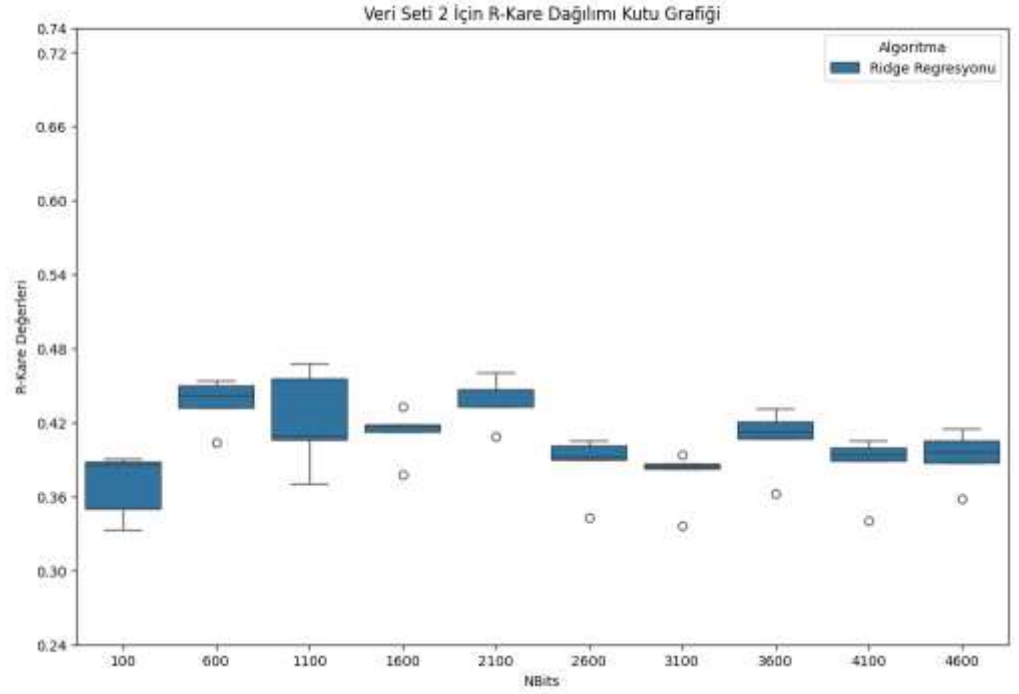
Veri seti 1'de Nbits sayısı arttıkça 5-katlı çapraz geçerleme performansının arttığı gözlendi. PLS ve ridge regresyonunda 100 Nbits sayısının performansının SVR, rassal orman, ve XGBOOST regresyonuna daha az başarı gösterdi. SVR, rassal orman ve XGBOOST regresyonları arasındaki farkın daha az olduğu gözlendi. Veri seti 2 üzerinde de 5-katlı çapraz geçerleme yöntemi uygulandı. Algoritmaların 5 katlı r-kare ve MSE değerleri Şekil 4.11., Şekil 4.12., Şekil 4.13., Şekil 4.14., Şekil 4.15., Şekil 4.16., Şekil 4.17., Şekil 4.18., Şekil 4.19., Şekil 4.20.'de verilmiştir. Bu şekiller incelendiğinde ise veri seti 2 için 5-katlı çapraz geçerleme uygulamasında rassal orman regresyonu, XGBOOST regresyonu ve SVR'ın performansının ridge ve PLS regresyonu performansından daha az yayılım gösterdiği gözlendi. Veri seti 2'de artan Nbits sayıları göz önüne alındığında r-karede artan ve MSE'de azalan bir trend gözlenmedi. Bunun sebebi, standart sapma filtresi uygulandığında Nbits sayılarının birbirine çok yakın değerler almış olması olarak yorumlanabilir.



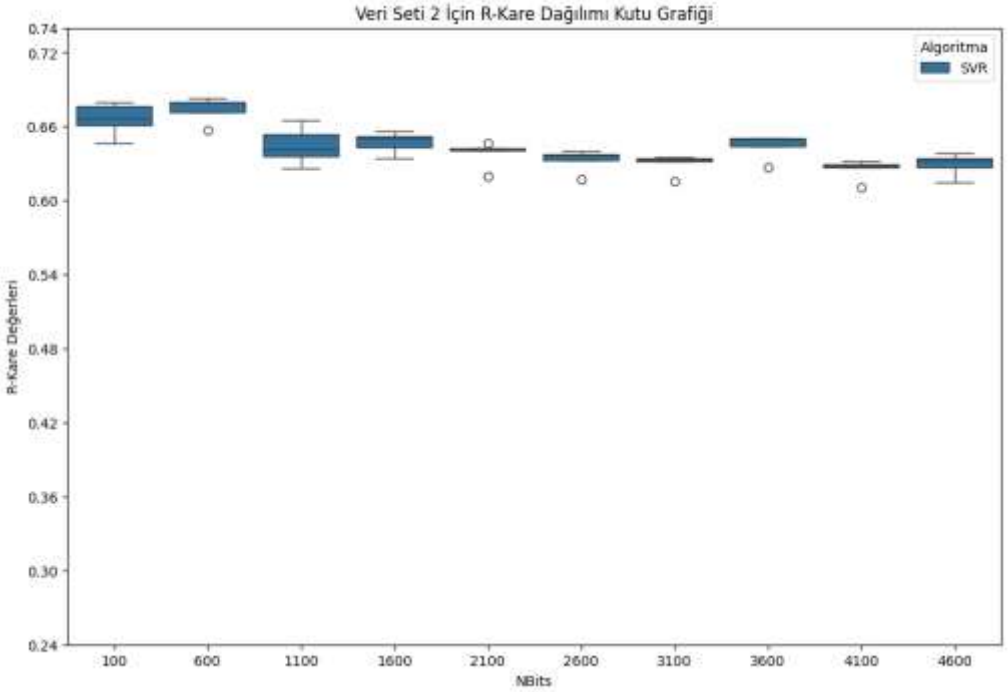
Şekil 4.11. Veri seti 2 üzerinde Rassal Orman Regresyonu algoritması için NBits sayısına göre r-kare dağılımı



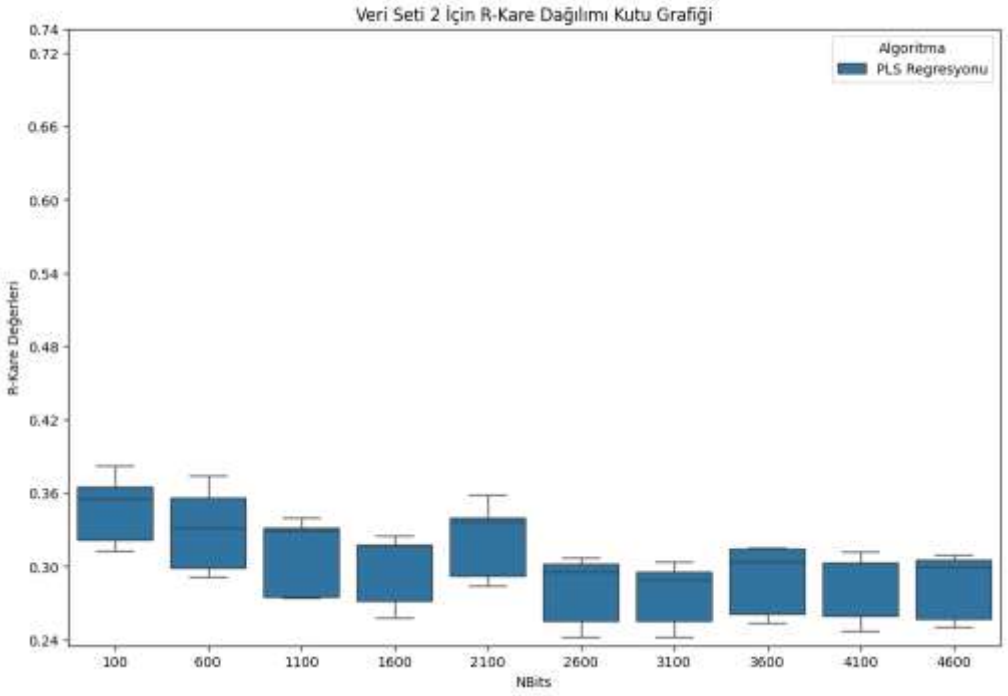
Şekil 4.12. Veri seti 2 üzerinde XGBOOST Regresyonu algoritması için NBits sayısına göre r-kare dağılımı



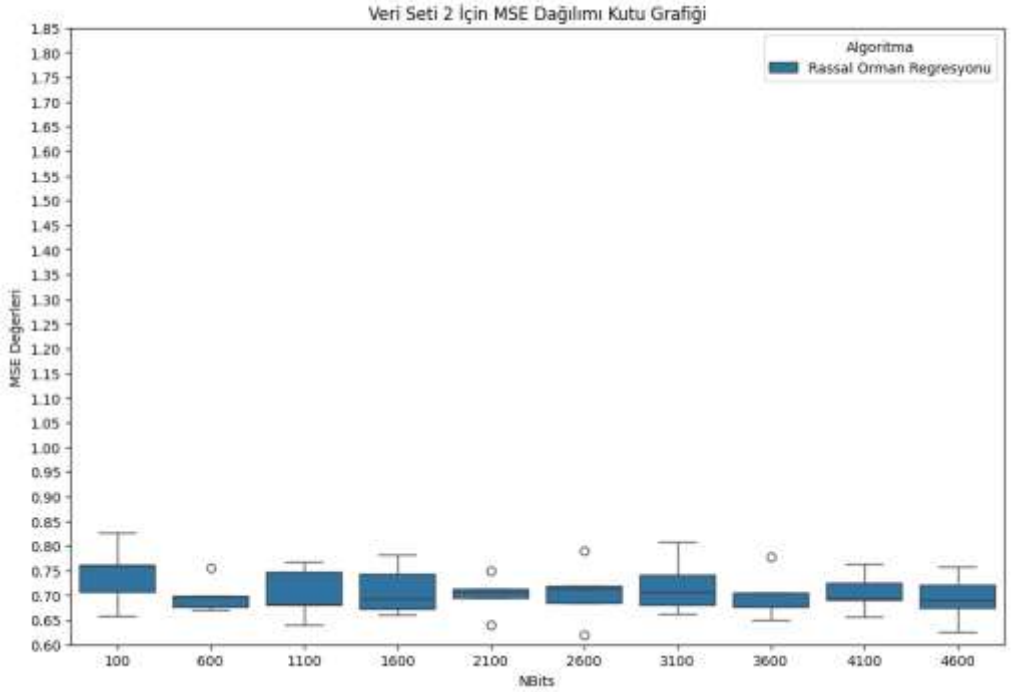
Şekil 4.13. Veri seti 2 üzerinde Ridge algoritması için NBits sayısına göre r-kare dağılımı



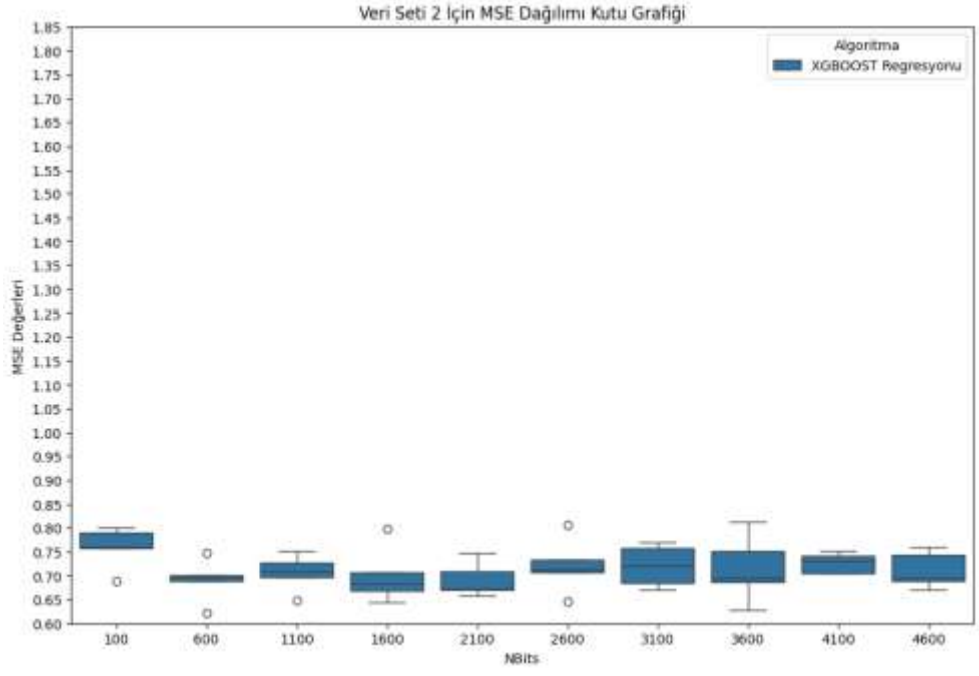
Şekil 4.14. Veri seti 2 üzerinde SVR algoritması için NBits sayısına göre r-kare dağılımı



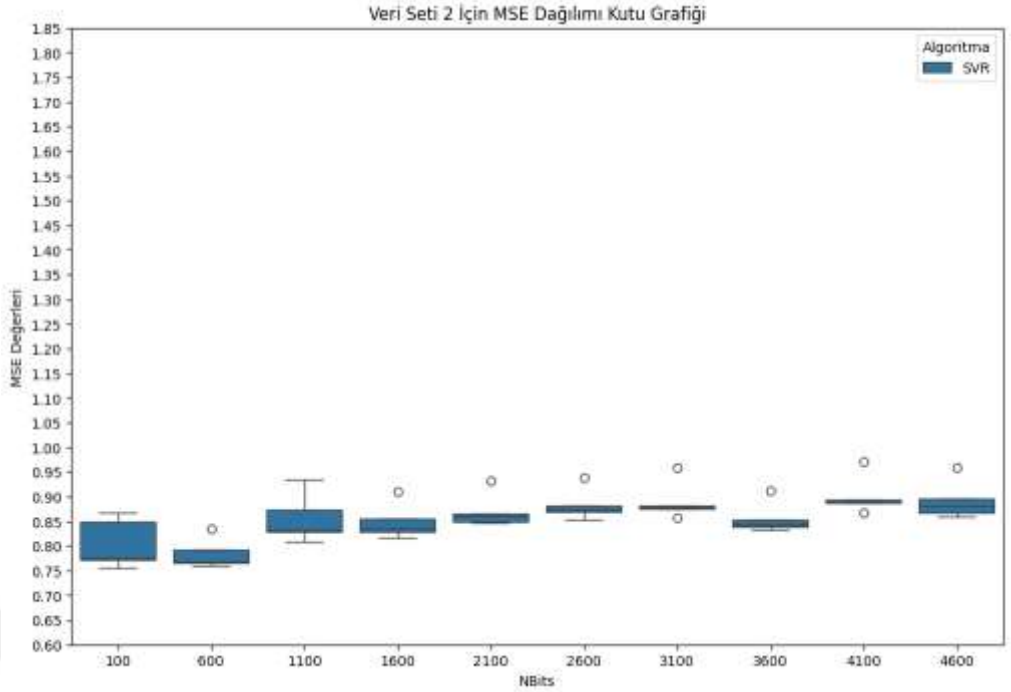
Şekil 4.15. Veri seti 2 üzerinde PLS Regresyonu algoritması için NBits sayısına göre r-kare dağılımı



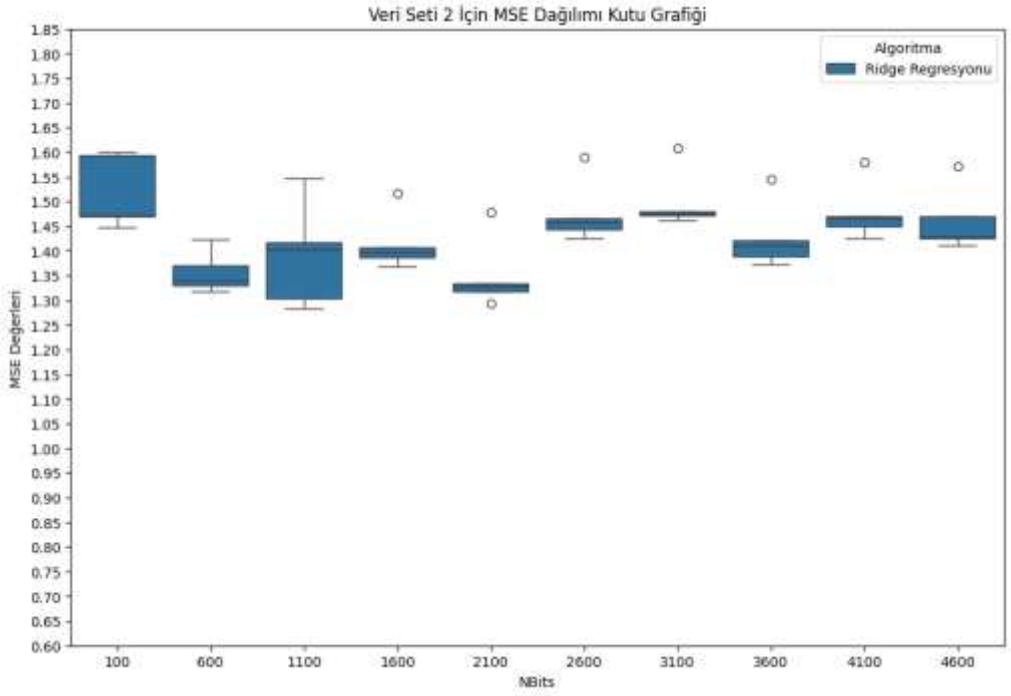
Şekil 4.16. Veri seti 2 üzerinde Rassal Orman Regresyonu algoritması için NBits sayısına göre MSE dağılımı



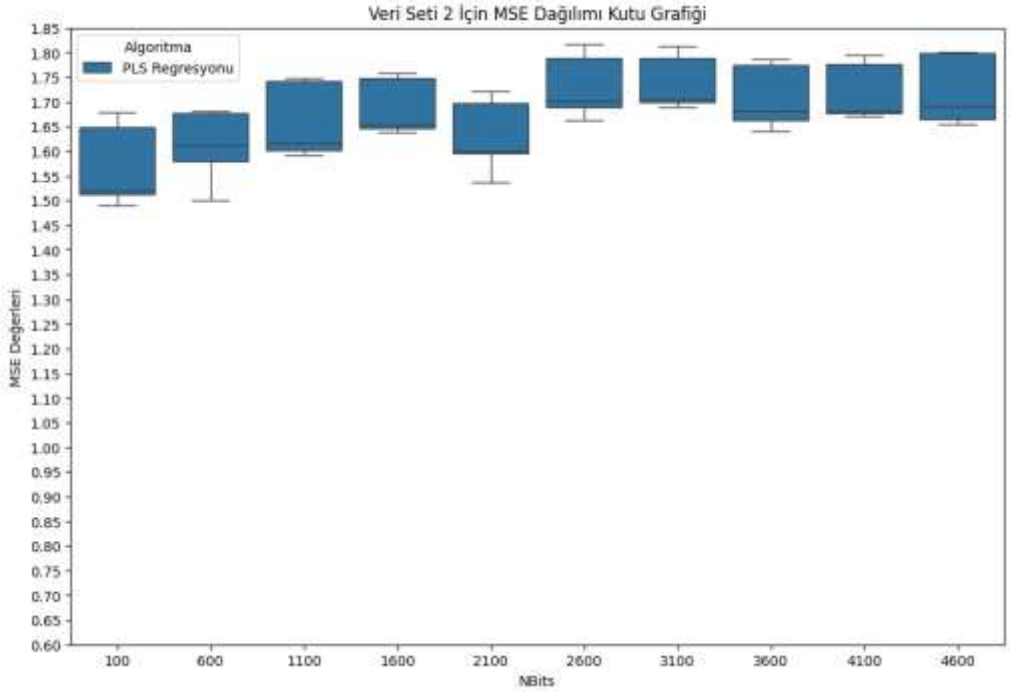
Şekil 4.17. Veri seti 2 üzerinde XGBOOST Regresyonu algoritması için NBits sayısına göre MSE dağılımı



Şekil 4.18. Veri seti 2 üzerinde SVR algoritması için NBits sayısına göre MSE dağılımı



Şekil 4.19. Veri seti 2 üzerinde Ridge Regresyonu algoritması için NBits sayısına göre MSE dağılımı

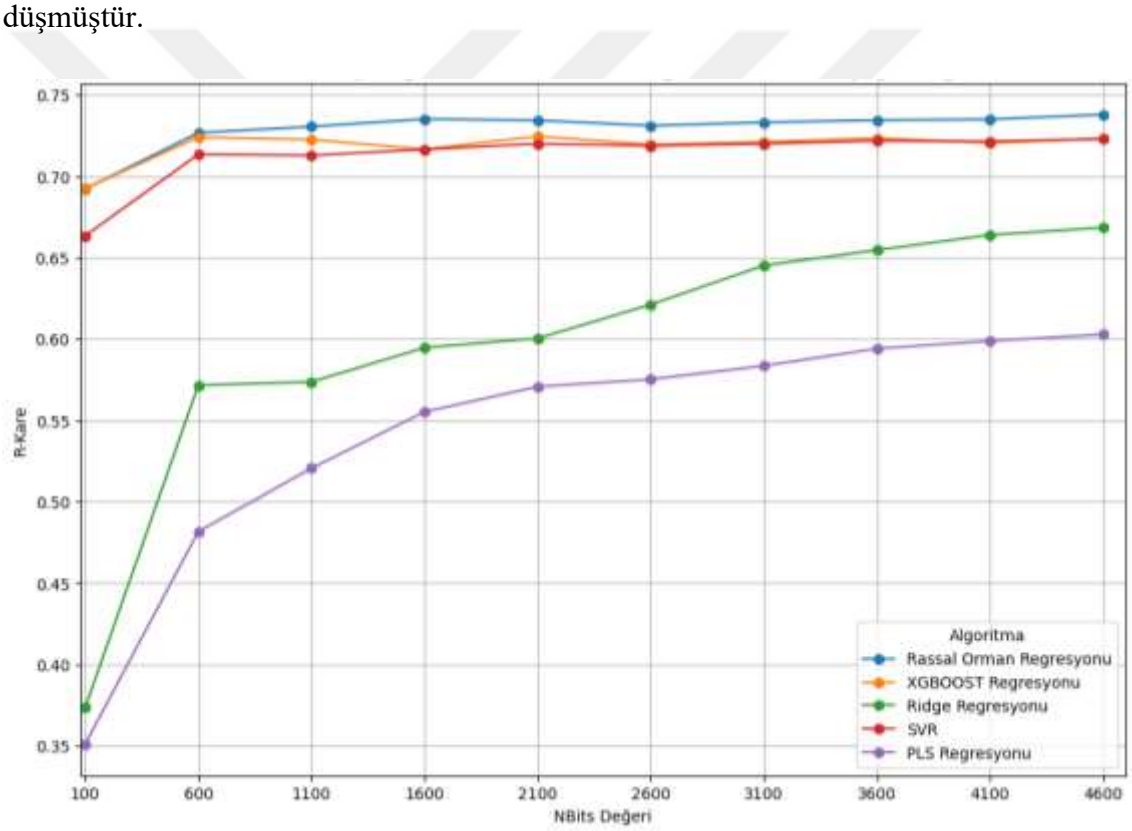


Şekil 4.20. Veri seti 2 üzerinde PLS Regresyonu algoritması için NBits sayısına göre MSE dağılımı

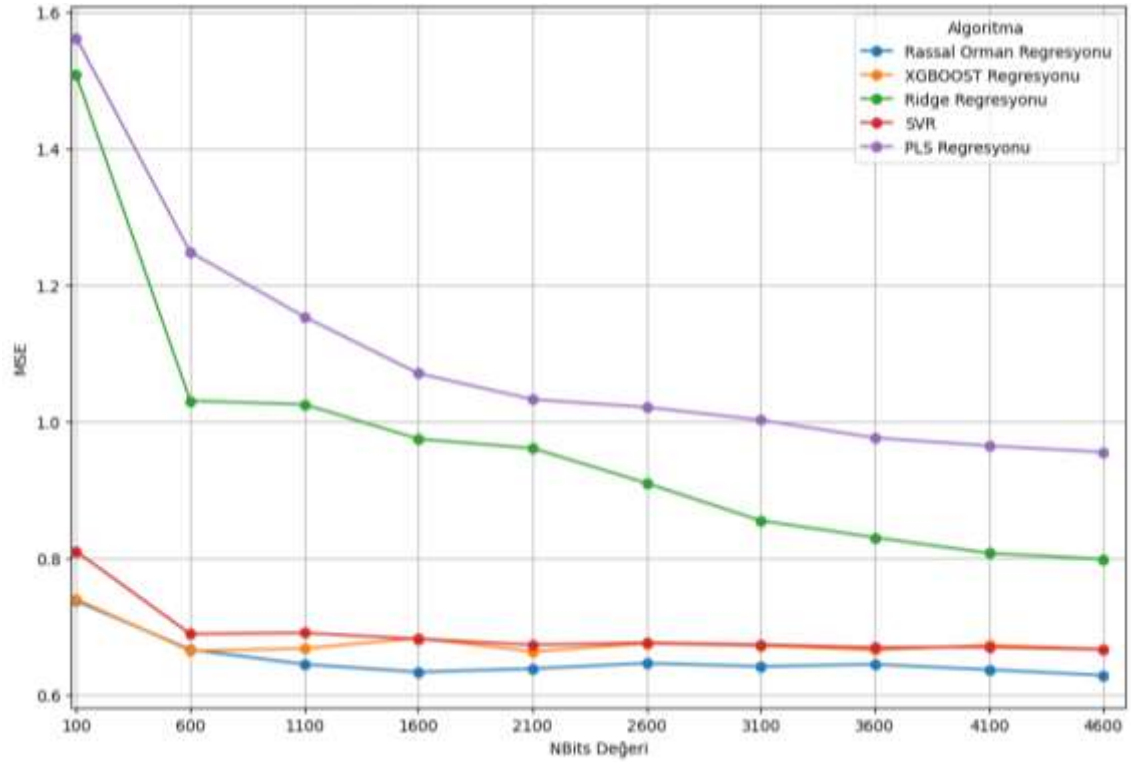
Çalışma kapsamında, iki veri setine ait sonuçların ortalama değerleri çizgi grafiği üzerinde görselleştirildi. Şekil 4.21. Veri seti 1 için 5-katlı çapraz geçerleme uygulanmış algoritmalara göre r-kare değerlerinde 5 farklı makine öğrenme algoritması için NBits sayısına göre veri seti 1 üzerinden elde edilen r-kare değerleri y ekseninde belirtilirken, Şekil 4.22. Veri seti 1 için 5-katlı çapraz geçerleme uygulanmış algoritmalara göre MSE değerlerinde 5 farklı makine öğrenme algoritması için NBits sayısına göre veri seti 1 üzerinden elde edilen MSE değerleri y ekseninde belirtilmiştir. X ekseninde veri seti için kullanılan NBits değeri artan sırada verilmiştir. Şekil 4.21. ve Şekil 4.22.'de rassal orman, XGBOOST ve SVR algoritmalarının eğilimlerinin benzer olduğu ve ridge ve PLS algoritmalarına göre daha başarılı performans gösterdiği görülmektedir. Aynı formatta Şekil 4.23. Veri seti 2 için 5-katlı çapraz geçerleme uygulanmış algoritmalara göre r-kare değerlerinde y ekseninde 5 farklı makine öğrenme algoritması için NBits sayısına göre veri seti 2'den elde edilen R-kare değerleri belirtilirken, Şekil 4.24. Veri seti 2 için 5-katlı çapraz geçerleme uygulanmış algoritmalara göre MSE değerleri' nde y ekseninde 5 farklı makine öğrenme algoritması için NBits sayısına göre veri seti 2'den elde edilen MSE değerleri belirtildi. Veri seti 2'nin oluşturulması aşamasında NBits sayısında azalma gözlendi. Bu azalma Çizelge 4.8. Veri seti 1 NBits değerlerinin veri seti 2 için karşılık değerleri'nde görülmektedir. Veri seti 1 ile karşılaştırılmasının sistematik olarak yapılması amacıyla Şekil 4.23. Veri seti 2 için 5-katlı çapraz geçerleme uygulanmış

algoritmalara göre r-kare değerleri ve Şekil 4.24. Veri seti 2 için 5-katlı çapraz geçerleme uygulanmış algoritmalara göre MSE değerlerinde NBits değerleri verilmiştir. Şekil 4.23. ve Şekil 4.24. incelendiğinde rassal orman, XGBOOST ve SVR algoritmalarının eğilimlerinin benzer olduğu ve ridge ve PLS algoritmalarına göre daha başarılı performans gösterdiği görülmektedir. Ridge ve PLS algoritmaları benzer eğilim göstermesine rağmen bazı noktalarda farklılık göstermiştir.

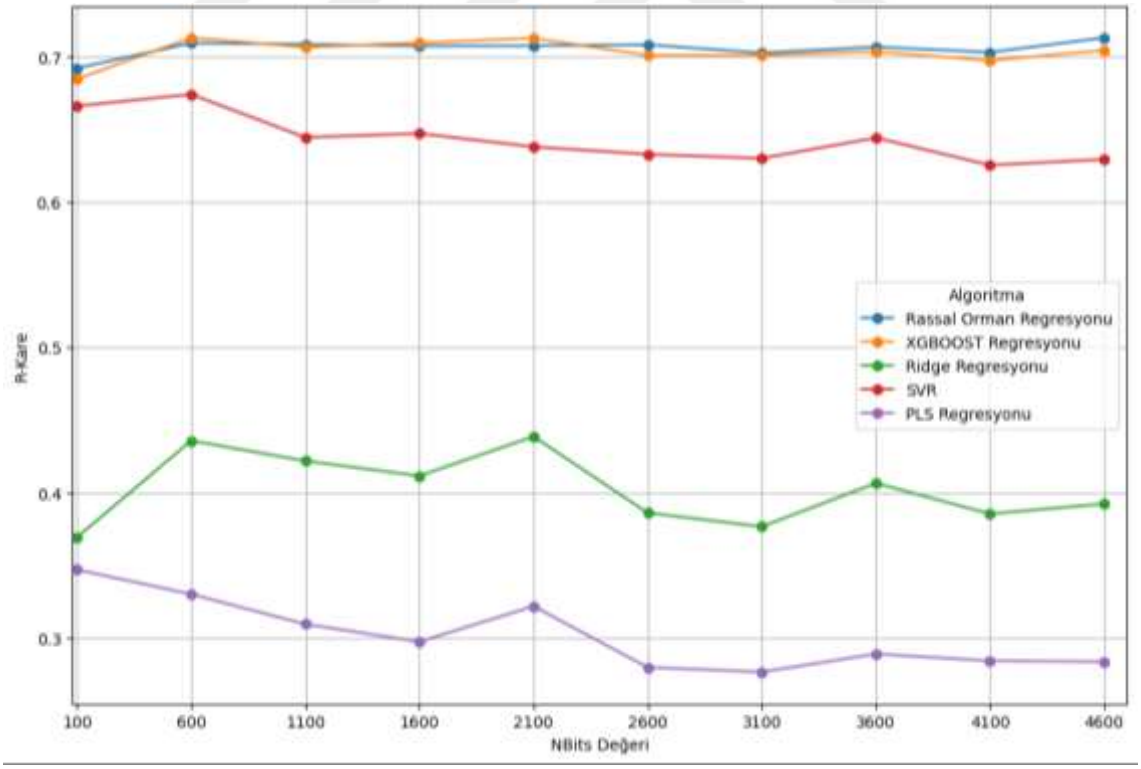
Rassal orman regresyonu, XGBOOST ve SVR için sonuçların, ridge ve PLS regresyonuna göre daha başarılı olduğu ve benzer eğilim gösterdiği gözlenmiştir. Grafikler, 5 algoritmanın da belirli bir trendi izlediğini ve bu trendin NBits değerine göre değiştiğini göstermektedir. NBits değeri arttıkça korelasyon artmış ve hata değerleri düşmüştür.



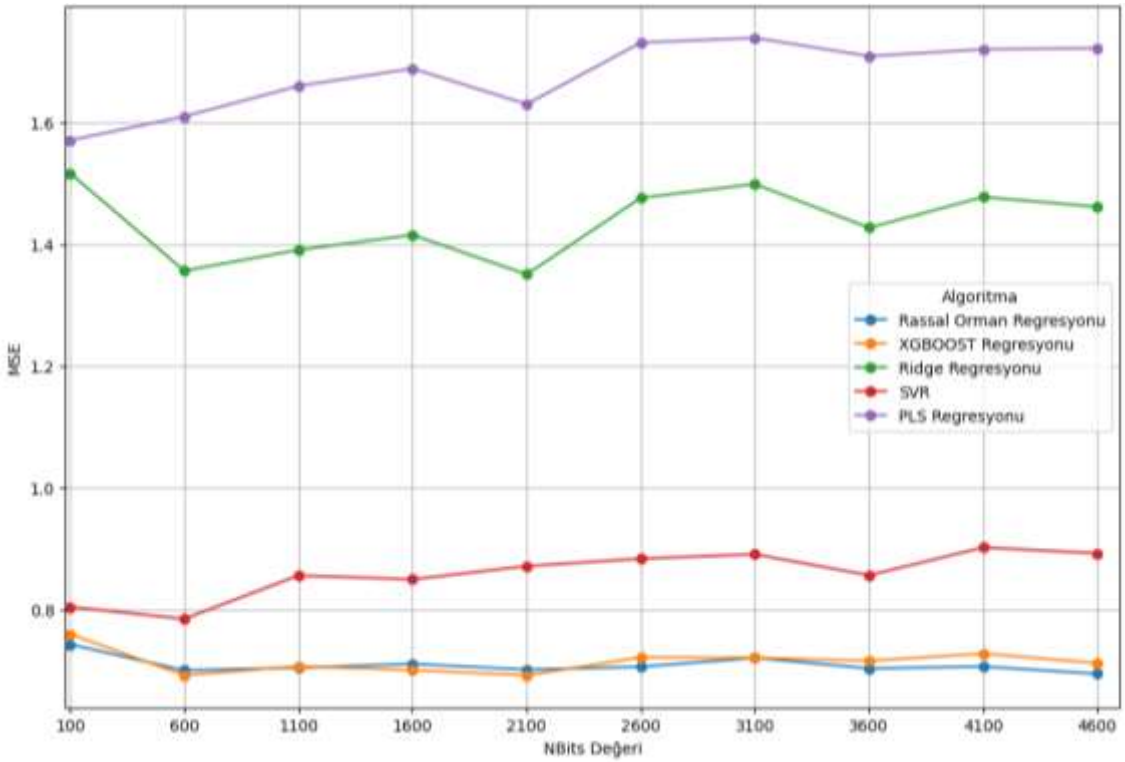
Şekil 4.21. Veri seti 1 için 5-katlı çapraz geçerleme uygulanmış algoritmalara göre r-kare değerleri



Şekil 4.22. Veri seti 1 için 5-katlı çapraz geçerleme uygulanmış algoritmalara göre MSE değerleri



Şekil 4.23. Veri seti 2 için 5-katlı çapraz geçerleme uygulanmış algoritmalara göre r-kare değerleri



Şekil 4.24. Veri seti 2 için 5-katlı çapraz geçerleme uygulanmış algoritmalara göre MSE değerleri

4.2.2. Toplu Öğrenim Yöntemi ile Tahmin Sonuçları

Algoritmaların birlikte nasıl bir performans göstereceğini belirlemek ve aslında daha iyi sonuçlar elde etmek amacıyla k-katlı çapraz geçerleme uygulamasından bağımsız olarak toplu öğrenim yöntemi de uygulandı. 5 farklı algoritma ile, yüzde 80 eğitim ve yüzde 20 test olarak ayrılan veri seti 1 ve veri seti 2 üzerinde regresyon analizleri yapıldı. Bu 5 algoritma aşağıda belirtilmiştir.

- Rassal orman regresyonu,
- XGBOOST regresyonu,
- Ridge regresyonu,
- SVR,
- PLS regresyonu

Bu çalışma, çeşitli makine öğrenimi algoritmalarının performansını değerlendirmek ve toplu öğrenme yöntemlerini kullanarak daha iyileştirmek amacıyla yapıldı. Tahmin sonuçları, r-kare ve MSE performans ölçütleri kullanılarak değerlendirildi ve algoritmalar başarılılık sırasına göre sıralandı. Ardından, toplu öğrenim yöntemi için XGBOOST, ridge ve PLS regresyonu algoritmaları tercih edildi. Toplu

öğrenim yönteminde ağırlıklı oylama yöntemi uygulandı. Uygulanan toplu öğrenim yönteminde XGBOOST regresyonu, ridge ve PLS Regresyon algoritmalarının ağırlık bilgisi Çizelge 4.6.'de belirtildi.

Çizelge 4.6. Toplu öğrenim yöntemi için ağırlık oranları

Algoritma	**Ağırlık Oranı**
XGBOOST Regresyonu	0.9
PLS Regresyon	0.05
Ridge Regresyonu	0.05

Toplu öğrenim yöntemi için r-kare ve MSE değerleri Çizelge 4.7. ve Çizelge 4.8.'de verilmiştir. Çizelge 4.7. Veri seti 1 için NBits sayısına göre toplu öğrenim yöntemi sonuçları'nda ve Çizelge 4.8. Veri seti 2 için NBits sayısına göre toplu öğrenim yöntemi sonuçları'nda r-kare ve MSE performanslarının bazı noktalarda farklılık gösterse de genel olarak NBits sayısına göre tutarlı sonuç verdiği gözlendi. Güçlü tahminleyici olarak belirlenen XGBoost, rassal orman ve SVR algoritmalarının birlikte eğitildiğinde toplu öğrenim yönteminin başarısında artış yaratmadığı gözlendi.

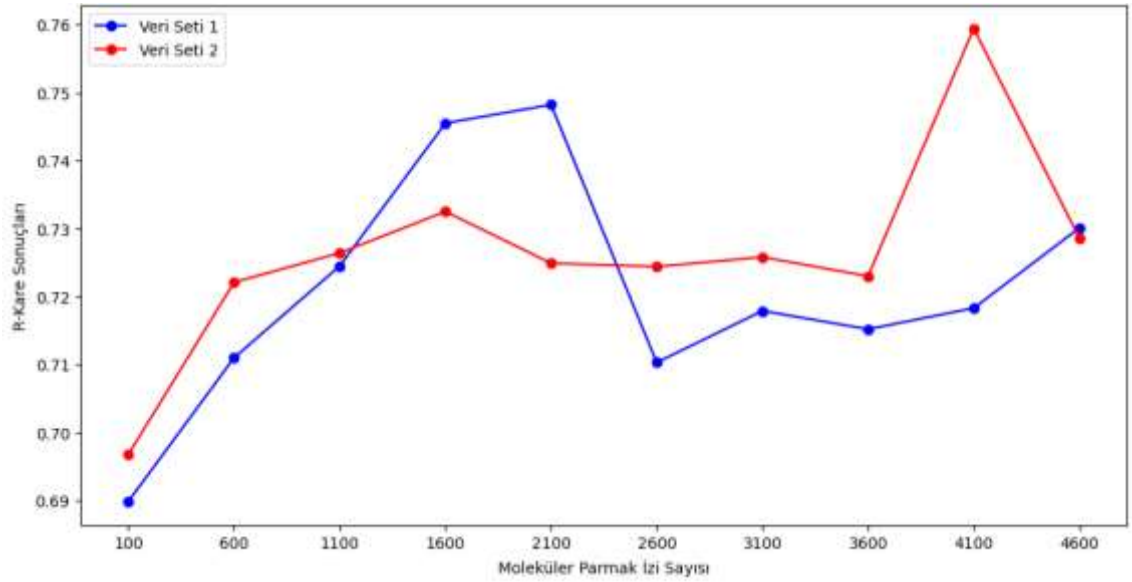
Çizelge 4.7. Veri seti 1 için NBits sayısına göre toplu öğrenim yöntemi sonuçları

Yöntem	**R-kare**	**MSE**	**NBits**
Toplu Öğrenim Yöntemi	0.6899	0.7700	100
	0.7110	0.6845	600
	0.7245	0.6668	1100
	0.7255	0.6371	1600
	0.7482	0.6305	2100
	0.7103	**0.6269**	2600
	0.7179	0.6788	3100
	0.7152	0.6817	3600
	0.7183	0.7403	4100
	0.7301	0.6392	4600

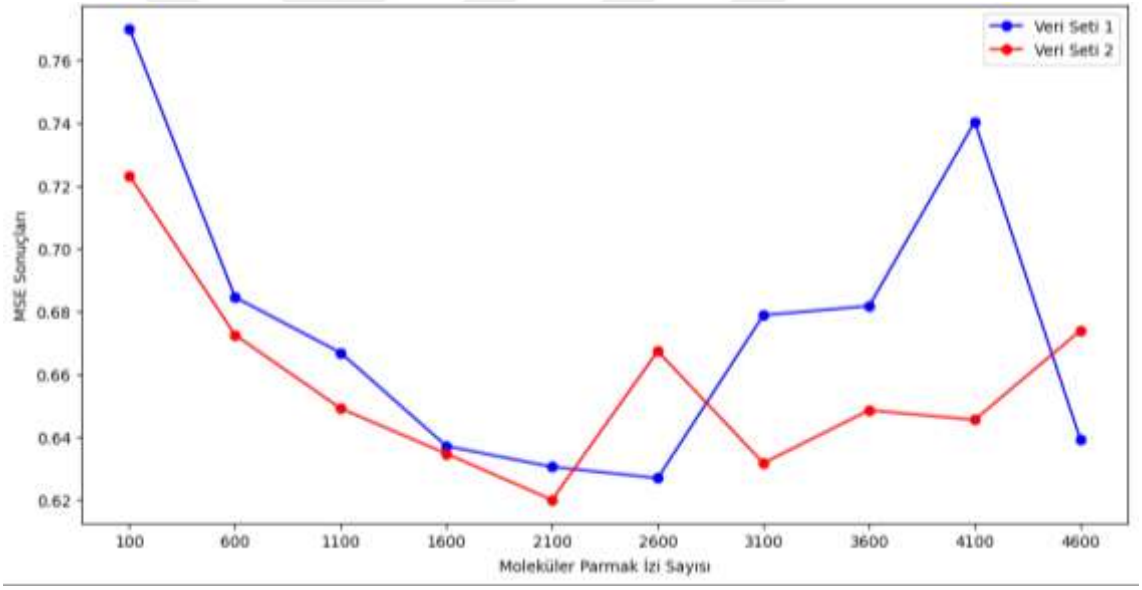
Çizelge 4.8. Veri seti 2 için NBits sayısına göre toplu öğrenim yöntemi sonuçları

Yöntem	**R-kare**	**MSE**	**NBits**	**Orjinal NBits Değeri**
Toplu Öğrenim Yöntemi	0.6968	0.7232	97	100
	0.7221	0.6725	119	600
	0.7264	0.6491	101	1100
	0.7325	0.6347	98	1600
	0.7249	**0.6200**	100	2100
	0.7244	0.6673	89	2600
	0.7258	0.6317	88	3100
	0.7230	0.6485	95	3600
	0.7594	0.6455	91	4100
	0.7286	0.6739	84	4600

Çalışma kapsamında, iki farklı sonucu aynı çizgi grafiği (Şekil 4.25) üzerinde görselleştirildi. Veri seti 1 ve veri seti 2 için NBits sayısına göre toplu öğrenim yöntemi r-kare sonuçlarında iki farklı veri setinden elde edilen r-kare değerleri y ekseninde belirtilirken, x ekseninde kullanılan NBits sayısı artan sırada verilmiştir. Veri seti 1 için en yüksek r-kare değerini 0.7482 ile 2100 NBits sayısı vermiştir. En düşük MSE değeri ise 2600 NBits sayısında 0.6269 olarak gözlenmiştir. Aynı formatta Şekil 4.26. Veri seti 1 ve veri seti 2 için NBits sayısına göre toplu öğrenim yöntemi MSE sonuçlarında y ekseninde iki farklı veri setinden elde edilen MSE değerleri verilmiştir. X ekseninde ise kullanılan NBits sayıları artan sırada verilmiştir. Veri seti 2 için en yüksek r-kare değerini 0.7594 ile 4100 NBits sayısı vermiştir. En düşük MSE değeri ise 2100 NBits sayısında 0.6200 olarak gözlenmiştir. Sonuçlar, her iki veri seti için benzer bir eğilim gösterirken, bazı noktalarda belirgin farklılıklar ortaya çıktı. Bu karşılaştırma ile farklı tahminleme yöntemlerinin sonuçlar üzerindeki etkileri değerlendirildi.



Şekil 4.25. Veri seti 1 ve veri seti 2 için NBits sayısına göre toplu öğrenim yöntemi r-kare sonuçları



Şekil 4.26. Veri seti 1 ve veri seti 2 için NBits sayısına göre toplu öğrenim yöntemi MSE sonuçları

Veri seti 1 ve veri seti 2 için NBits sayısına göre r-kare ve MSE değerlerinin karşılaştırıldığı Çizelge 4.9.'da verilmiştir.

Çizelge 4.9. Veri seti 1 ve veri seti 2 için toplu öğrenim yöntemi sonuçlarının NBits sayısına göre karşılaştırması

Veri Seti 1 Sütun Sayısı	**Veri Seti 2 Sütun Sayısı**	**Veri Seti 1 İçin R-kare Değeri**	**Veri Seti 2 İçin R-kare Değeri**	**Veri Seti 1 İçin MSE Değeri**	**Veri Seti 2 İçin MSE Değeri**
100	97	0.6899	0.6968	0.7700	0.7232
600	119	0.7110	0.7221	0.6845	0.6725
1100	101	0.7245	0.7264	0.6668	0.6491
1600	98	0.7255	0.7325	0.6371	0.6347
2100	100	**0.7482**	0.7249	0.6305	**0.6200**
2600	89	0.7103	0.7244	**0.6269**	0.6673
3100	88	0.7179	0.7258	0.6788	0.6317
3600	95	0.7152	0.7230	0.6817	0.6485
4100	91	0.7183	**0.7594**	0.7403	0.6455
4600	84	0.7301	0.7286	0.6392	0.6739

Çizelge 4.9. Veri seti 1 ve veri seti 2 için toplu öğrenim yöntemi sonuçlarının NBits sayısına göre karşılaştırması incelendiğinde sonuçlar, veri seti 2 üzerinde toplu öğrenim yönteminin 4100 NBits sayısı ile 0.7594 r-kare değeri, 2100 NBits sayısı ile 0.6200 MSE değerini elde ederek en başarılı tahmin sonuçlarını elde ettiğini gösterdi. Bu durum toplu öğrenim yönteminin veri setinin yüksek varyanslı sütunlarını k-katlı çapraz geçerleme yöntemine göre daha iyi açıkladığını göstermektedir.

4.3. Karşılaştırma

AChE enzimi ve potansiyel inhibitörlerinin tespiti ve bağlanma eğilimlerinin tahmini amacıyla yapılmış çalışmalar Kaynak Araştırması bölümünde özetlenmiştir. Bu çalışmalar arasında bu çalışmaya en benzeyenler Nguyen vd. (2022) ile Khedekar vd. (2022)'ye aittir. Nguyen vd. (2022), 762 adet molekülün 200 bitlik moleküler parmak izini kullanmış ve Grafik Evrişimsel Ağ modeli ile 0.72 Pearson R korelasyonu ve 1.580 RMSE sonuçlarını elde etmiştir. R korelasyonu, istatistikte yaklaşık olarak R-kare değerinin köküne tekabül ettiğinden kullandıkları veri modelinin pIC50 değerlerinin varyansının yaklaşık %50'sini açıkladığı söylenebilir. Bu çalışmada ise toplu öğrenim yöntemi kullanarak 2100 uzunluğundaki ham bit vektörleri kullanıldığında 0.7482 R-

kare, 4600 uzunluğundaki bit vektörlerine standart sapma filtresi uygulanıp uzunluk 91'e indirgendiğinde ise 0.7594 R-kare değerleri ile pIC50 değerlerinin varyansının %75'i açıklanmıştır. Khedekar vd. (2022) ise Chembl veri tabanından topladığı 5103 molekül üzerinde Rassal Orman Regresyonu modelini uygulayarak 0.87 R-kare ve 0.35 MSE değeri elde etmiştir. Ancak beyan ettikleri sonuçlar bütün verinin eğitim kümesi olarak kullanıldığı sonuçlar olduğundan görülmemiş veriler ile ilgili sonuçları içermemektedir. Karar ağacı tabanlı bir algoritma olan Rassal Orman Regresyonu aşırı öğrenmeye eğilimli bir yöntem olduğundan test sonuçları hakkında bir çıkarım yapılamamaktadır.

5. SONUÇLAR VE ÖNERİLER

5.1. Sonuçlar

Bu çalışmada, AChE enzimini inhibe eden bileşiklerin IC50 değerlerinin tahmin edilmesi amacıyla veri madenciliği ve makine öğrenme algoritmaları ve toplu öğrenim yöntemi kullanılmıştır. Veri seti üzerinde yapılan çalışmada, moleküllerin biyolojik aktivitelerini tahmin etmek için kanonik SMILES ve Lipinski'nin 5 kuralına göre yeni özellikler eklenmiştir. Lipinski'nin 5 kuralının uygulanmasıyla, molekül ağırlık, lipofilite değeri, hidrojen bağı verici ve alıcı atomların sayısı hesaplanmıştır ve bu hesaplama sonucunda elde edilen veriler moleküler parmak izi hesaplamasına dahil edilerek moleküler yapıların ilaç benzerliği ve geçirgenliği açısından değerlendirilmesi sağlanmıştır. Moleküler parmak izi hesaplamaları çeşitli bitlik vektörler (NBits = 100, 600, 1100, 1600, 2100, 2600, 3100, 3600, 4100, 4600) halinde hesaplanmıştır ve her biri için veri seti 1 oluşturulmuştur. Oluşturulan her veri seti üzerinde standart sapması 0.3'ün üzerinde olan sütunlar seçilerek veri seti 2 oluşturulmuştur. Bu işlem sonucunda NBits sayılarında azalma gözlenmiştir. Veri seti 2'nin NBits sayısı veri seti 1'e göre daha az olduğu için regresyon analizi çalışmalarında daha hızlı sonuç vermiştir. Sonuç olarak toplamda 20 veri seti edilmiştir. Oluşturulan bu veri setleri ile rassal orman, XGBOOST, ridge, SVR, PLS regresyonu makine öğrenme algoritmaları ile regresyon analizleri gerçekleştirilmiştir. Çalışma Python programlama dili kullanılarak gerçekleştirilmiştir. Açık kaynaklı RDKit kütüphanesi ve alt modülleri veri setini hazırlama çalışmalarında kullanılmıştır. Regresyon analizi için öncelikle k-katlı çapraz geçerleme 5 algoritma için uygulanmıştır. Çalışmada k=5 seçilmiştir. 5-katlı çapraz geçerleme uygulandığında aşağıdaki sonuçlar elde edilmiştir.

- Veri seti 1 üzerinde, en yüksek r-kare değeri 0.7379 ve en düşük MSE değeri 0.6287, rassal orman regresyonu algoritması ile 4600 NBits değerinde elde edilmiştir.
- Veri seti 2 üzerinde ise en yüksek r-kare değeri 0.7130 ile rassal orman regresyonu algoritması 4600 NBits değerinde ve aynı zamanda XGBOOST regresyonunun 2100 NBits değerinde elde edilmiştir. En düşük MSE değeri 0.6914 değeri ile XGBOOST regresyonunun 2100 ve 600 NBits değerlerinde elde edilmiştir.

Toplu öğrenim yöntemi kullanıldığında ise:

- Veri seti 1 için en yüksek r-kare değeri 0.7482 değeri 2100 NBits değerinde elde edilmiştir. En düşük MSE değeri 0.6269 değeri ile 2600 NBits değerinde elde edilmiştir.
- Veri seti 2 için en yüksek r-kare değeri 0.7594 değeri ile 4100 NBits değerinde elde edilmiştir. En düşük MSE değeri 0.6200 değeri ile 2100 NBits değerinde elde edilmiştir.

Sonuç olarak AChE enziminin potansiyel inhibitörler ile etkileşimini belirleyen bağlanma eğiliminin tahmininde elde edilen veri modeli pIC50 varyansının yaklaşık %75'ini açıklamış ve 0.62 civarında MSE ile gerçek değere yaklaşmıştır. Bu çalışmanın en büyük katkısı farklı uzunluklardaki moleküler parmak izi vektörlerinin pIC50 değerini tahminine olan etkisini ortaya koymak olmuştur. Parmak izi uzunluğu arttıkça tahmin başarısı artmıştır, ancak düşük varyanslı özelliklerin elenmesinin de tahmin performansına çok büyük etkisi olmasa da algoritmaların çalışma hızını arttırdığı gözlenmiştir. Ayrıca, parmak izlerinin yanı sıra Lipinski'nin 5 kuralını kullanarak hesaplanan özelliklerin parmak izi vektörlerine eklenmesi de literatürdeki benzer çalışmalardan daha iyi sonuçlar elde edilmesine yardımcı olmuştur.

5.2. Öneriler

Gelecek araştırmalar için, daha geniş ve çeşitli bir veri setinin kullanılması model performansının iyileştirilmesinde kritik bir rol oynayabilir. Bu, farklı kimyasal bileşiklerin ve moleküler yapıların daha iyi kapsanmasını sağlayabilir. Uygulama performansını artırmak için, moleküler yapısal özelliklerin yanı sıra, moleküllerin biyolojik etkileşimlerini daha iyi yakalayan derin öğrenme yöntemleri incelenebilir. Derin öğrenme teknikleri, moleküler etkileşimlerin daha ayrıntılı bir şekilde analiz edilmesine olanak tanıyabilir ve daha karmaşık ilişkileri daha iyi modelleyebilir. Bu tekniklerin moleküler tasarım ve ilaç keşfi süreçlerinde nasıl kullanılabileceği üzerine daha fazla araştırma yapılması gerekmektedir. Ayrıca, toplu öğrenim yöntemi ve diğer birleştirme tekniklerinde parametre optimizasyonu yapılması, uygulama performansını artırabilir. Bu teknikler, farklı makine öğrenimi algoritmalarını bir araya getirerek, her bir algoritmanın güçlü yönlerini birleştirerek daha güçlü bir tahminci oluşturabilir. Farklı toplu öğrenim yöntemleri ve birleştirme stratejilerinin değerlendirilmesi, uygulamanın genel performansını artırabilir. Son olarak, bu çalışmanın sonuçları, AChE inhibisyonunu hedefleyen ilaçların tasarımı ve geliştirilmesi alanında önemli bir ilerleme sağlayabilir.

Gelecekteki araştırmalar, bu alandaki bilgiyi genişleterek ve yeni teknikler geliştirerek, ilaç keşfi sürecine daha fazla katkı sağlayabilir. Bu çalışmanın tekniği ve bulguları, ilaç endüstrisinde yeni ilaçların geliştirilmesine yol gösterebilir.

KAYNAKLAR

Abdi, H. (2010). Partial least squares regression and projection on latent structure regression (PLS Regression). *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(1), 97–106. https://doi.org/10.1002/wics.51

Agarwal, P., & Searls, D. B. (2008). Literature mining in support of drug discovery. *Briefings in Bioinformatics*, *9*(6), 479–492. https://doi.org/10.1093/bib/bbn035

Akocak, S., & Lolak, N. (2020). Biological evaluation of aromatic bis-sulfonamide Schiff bases as antioxidant, acetylcholinesterase and butyrylcholinesterase inhibitors. *Cumhuriyet Science Journal*, *41*(2), 413–418. https://doi.org/10.17776/csj.595463

Alexandropoulos, S. A. N., Kotsiantis, S. B., & Vrahatis, M. N. (2019). Data preprocessing in predictive data mining. *Knowledge Engineering Review*, *34*. https://doi.org/10.1017/S026988891800036X

Al-Hassan, Y., & Mohammad Al-Hassan, Y. (2008). A Monte Carlo Evaluation of Some Ridge Estimators. In *J.J. Appl. Sci: Natural Sciences Series* (Vol. 10, Issue 2). https://www.researchgate.net/publication/260436784

Almenoff, J., Tonning, J. M., Gould, A. L., Szarfman, A., Hauben, M., Ouellet-Hellstrom, R., Ball, R., Hornbuckle, K., Walsh, L., Yee, C., Sacks, S. T., Yuen, N., Patadia, V., Blum, M., Johnston, M., Gerrits, C., Seifert, H., & Lacroix, K. (2005). Perspectives on the Use of Data Mining in Pharmacovigilance. In *Drug Safety* (Vol. 28, Issue 11).

Altman, N., & Krzywinski, M. (2016). Points of Significance: Regression diagnostics. In *Nature Methods* (Vol. 13, Issue 5, pp. 385–386). Nature Publishing Group. https://doi.org/10.1038/nmeth.3854

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, *4*, 40–79. https://doi.org/10.1214/09-SS054

Arús-Pous, J., Johansson, S. V., Prykhodko, O., Bjerrum, E. J., Tyrchan, C., Reymond, J. L., Chen, H., & Engkvist, O. (2019). Randomized SMILES strings improve the quality of molecular generative models. *Journal of Cheminformatics*, *11*(1). https://doi.org/10.1186/s13321-019-0393-0

Baron, S., Linton, S., & O'malley, M. A. (2023, December). On drugs. In The Journal of Medicine and Philosophy: A Forum for Bioethics and Philosophy of Medicine (Vol. 48, No. 6, pp. 551-564).

Bastien, P., Vinzi, V. E., & Tenenhaus, M. (2005). PLS generalised linear regression. *Computational Statistics and Data Analysis*, *48*(1), 17–46. https://doi.org/10.1016/j.csda.2004.02.005

Bates, S., Hastie, T., & Tibshirani, R. (2024). Cross-Validation: What Does It Estimate and How Well Does It Do It? *Journal of the American Statistical Association*, *119*(546), 1434–1445. https://doi.org/10.1080/01621459.2023.2197686

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2019). *A Comparative Analysis of XGBoost*. https://doi.org/10.1007/s10462-020-09896-5

Bento, A. P., Gaulton, A., Hersey, A., Bellis, L. J., Chambers, J., Davies, M., Krüger, F. A., Light, Y., Mak, L., McGlinchey, S., Nowotka, M., Papadatos, G., Santos, R., & Overington, J. P. (2014). The ChEMBL bioactivity database: An update. *Nucleic Acids Research*, *42*(D1). https://doi.org/10.1093/nar/gkt1031

Biau, G., & Fr, G. B. (2012). Analysis of a Random Forests Model. In *Journal of Machine Learning Research* (Vol. 13).

Biau, G., & Scornet, E. (2016). A random forest guided tour. *Test*, *25*(2), 197–227. https://doi.org/10.1007/s11749-016-0481-7

Bondžić, A. M., Lazarević-Pašti, T. D., Leskovac, A. R., Petrović, S., Čolović, M. B., Parac-Vogt, T. N., & Janjić, G. V. (2020). A new acetylcholinesterase allosteric site responsible for binding voluminous negatively charged molecules – the role in the mechanism of AChE inhibition. *European Journal of Pharmaceutical Sciences*, *151*. https://doi.org/10.1016/j.ejps.2020.105376

Caldwell, G. W., Yan, Z., Lang, W., & Masucci, J. A. (2012). The IC 50 Concept Revisited. In *Current Topics in Medicinal Chemistry* (Vol. 12).

Cameron, A. C., & Windmeijer, F. A. G. (1996). *An R-squared measure of goodness of fit for some common nonlinear regression models*.

CHAFFEY, N. (2003). Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. and Walter, P. Molecular biology of the cell. 4th edn. *Annals of Botany*, *91*(3), 401–401. https://doi.org/10.1093/aob/mcg023

Craig, I. D. (Iain D. (2007). *Object-oriented programming languages : interpretation*. Springer.

Cuzzocrea, A., Du, X., Kara, O., Liu, T., Ślęzak, D., Yang, X., Diniz, S., Barbosa, J., Chen, P., Filipe, J., Kotenko, I., Sivalingam, K. M., & Yuan, J. (2007). *Communications in Computer and Information Science 845 Commenced Publication in 2007 Founding and Former Series Editors: Editorial Board*. http://www.springer.com/series/7899

Camacho, D. M., Collins, K. M., Powers, R. K., Costello, J. C., & Collins, J. J. (2018). Next-Generation Machine Learning for Biological Networks. In *Cell* (Vol. 173, Issue 7, pp. 1581–1592). Cell Press. https://doi.org/10.1016/j.cell.2018.05.015

Camadan, Y., & Akkemik, E. (2022). Searching for New Natural Inhibitors of Acetylcholinesterase Enzyme. *Cumhuriyet Science Journal*, *43*(1), 66–71. https://doi.org/10.17776/csj.983869

Carvajal, F. J., & Inestrosa, N. C. (2011). Interactions of AChE with A? Aggregates in Alzheimer?s Brain: Therapeutic Relevance of IDN 5706. *Frontiers in Molecular Neuroscience*, *4*. https://doi.org/10.3389/fnmol.2011.00019

Chen, C. H., Tanaka, K., Kotera, M., & Funatsu, K. (2020). Comparison and improvement of the predictability and interpretability with ensemble learning models in QSPR applications. *Journal of Cheminformatics*, *12*(1). https://doi.org/10.1186/s13321-020-0417-9

Chen, J., Lin, K. C., Prasad, S., & Schmidtke, D. W. (2023). Label free impedance based acetylcholinesterase enzymatic biosensors for the detection of acetylcholine. *Biosensors and Bioelectronics*, *235*. https://doi.org/10.1016/j.bios.2023.115340

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-August-2016*, 785–794. https://doi.org/10.1145/2939672.2939785

Chen, Y. R., Rezapour, A., & Tzeng, W. G. (2018). Privacy-preserving ridge regression on distributed data. *Information Sciences*, *451–452*, 34–49. https://doi.org/10.1016/j.ins.2018.03.061

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, *7*, 1–24. https://doi.org/10.7717/PEERJ-CS.623

Davies, M. J. (2003). Singlet oxygen-mediated damage to proteins and its consequences. *Biochemical and Biophysical Research Communications*, *305*(3), 761–770. https://doi.org/10.1016/S0006-291X(03)00817-9

de Abreu Silva, M., Sette, C. D. A. B., Kiametis, A. S., Romeiro, L. A. S., & Gargano, R. (2019). Molecular modeling of cardanol-derived AChE inhibitors. *Chemical Physics Letters*, *731*. https://doi.org/10.1016/j.cplett.2019.07.019

De Vivo, M., Masetti, M., Bottegoni, G., & Cavalli, A. (2016). Role of Molecular Dynamics and Related Methods in Drug Discovery. In *Journal of Medicinal Chemistry* (Vol. 59, Issue 9, pp. 4035–4061). American Chemical Society. https://doi.org/10.1021/acs.jmedchem.5b01684

DiMasi, J. A., Hansen, R. W., & Grabowski, H. G. (2003). The price of innovation: New estimates of drug development costs. *Journal of Health Economics*, *22*(2), 151–185. https://doi.org/10.1016/S0167-6296(02)00126-1

Dong, J., Cao, D. S., Miao, H. Y., Liu, S., Deng, B. C., Yun, Y. H., Wang, N. N., Lu, A. P., Zeng, W. Bin, & Chen, A. F. (2015). ChemDes: An integrated web-based platform for molecular descriptor and fingerprint computation. *Journal of Cheminformatics*, *7*(1). https://doi.org/10.1186/s13321-015-0109-z

Du, X., Li, Y., Xia, Y. L., Ai, S. M., Liang, J., Sang, P., Ji, X. L., & Liu, S. Q. (2016). Insights into protein–ligand interactions: Mechanisms, models, and methods. In *International Journal of Molecular Sciences* (Vol. 17, Issue 2). MDPI AG. https://doi.org/10.3390/ijms17020144

Dong, X., Yu, Z., Cao, W., Shi, Y., & Ma, Q. (2020). A survey on ensemble learning. In *Frontiers of Computer Science* (Vol. 14, Issue 2, pp. 241–258). Higher Education Press. https://doi.org/10.1007/s11704-019-8208-z

Dong, Y., & Qin, S. J. (2018). Regression on dynamic PLS structures for supervised learning of dynamic data. *Journal of Process Control*, *68*, 64–72. https://doi.org/10.1016/j.jprocont.2018.04.006

Elbadawi, M., Gaisford, S., & Basit, A. W. (2021). Advanced machine-learning techniques in drug discovery. In *Drug Discovery Today* (Vol. 26, Issue 3, pp. 769–777). Elsevier Ltd. https://doi.org/10.1016/j.drudis.2020.12.003

Fishburn, C. S. (2013). A conversation with Chris Lipinski. *Science-Business EXchange*, *6*(46), 1309–1309. https://doi.org/10.1038/scibx.2013.1309

Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., & Overington, J. P. (2012). ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Research*, *40*(D1). https://doi.org/10.1093/nar/gkr777

Gaulton, A., Hersey, A., Nowotka, M. L., Patricia Bento, A., Chambers, J., Mendez, D., Mutowo, P., Atkinson, F., Bellis, L. J., Cibrian-Uhalte, E., Davies, M., Dedman, N., Karlsson, A., Magarinos, M. P., Overington, J. P., Papadatos, G., Smit, I., & Leach, A. R. (2017). The ChEMBL database in 2017. *Nucleic Acids Research*, *45*(D1), D945–D954. https://doi.org/10.1093/nar/gkw1074

Geladi, P., & Kowalski, B. R. (1986). PARTIAL LEAST-SQUARES REGRESSION: A TUTORIAL. In *Analytica Chimica Acta* (Vol. 186). Elsevier Science Publishers B.V.

Genuer, R., Poggi, J. M., & Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, *31*(14), 2225–2236. https://doi.org/10.1016/j.patrec.2010.03.014

Giménez, B. G., Santos, M. S., Ferrarini, M., & Dos Santos Fernandes, J. P. (2010). Evaluation of blockbuster drugs under the rule-of-five. *Pharmazie*, *65*(2), 148–152. https://doi.org/10.1691/ph.2010.9733

Glymour, C., Madigan, D., & Pregibon, D. (1997). Statistical Themes and Lessons for Data Mining. In *Data Mining and Knowledge Discovery* (Vol. 5, Issue 6). Kluwer Academic Publishers.

Goulian, M., & Simon, S. M. (2000). *Tracking Single Proteins within Cells*.

Greenfield, S. (2013). Discovering and targeting the basic mechanism of neurodegeneration: The role of peptides from the C-terminus of acetylcholinesterase: Non-hydrolytic effects of ache: The actions of peptides derived from the C-terminal and their relevance to neurodegeneration. *Chemico-Biological Interactions*, *203*(3), 543–546. https://doi.org/10.1016/j.cbi.2013.03.015

Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, *377*(1–2), 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003

Han, J., Lu, Z., Flach, A. L., Paton, R. S., Hammond, G. B., & Xu, B. (2015). Role of Hydrogen-Bonding Acceptors in Organo-Enamine Catalysis. *Chemistry - A European Journal*, *21*(33), 11687–11691. https://doi.org/10.1002/chem.201502407

Harigua-Souiai, E., Oualha, R., Souiai, O., Abdeljaoued-Tej, I., & Guizani, I. (2022). Applied Machine Learning Toward Drug Discovery Enhancement: Leishmaniases as a Case Study. *Bioinformatics and Biology Insights*, *16*. https://doi.org/10.1177/11779322221090349

Heredia-García, G., Elizalde-Velázquez, G. A., Gómez-Oliván, L. M., Islas-Flores, H., García-Medina, S., Galar-Martínez, M., & Dublán-García, O. (2023). Realistic concentrations of Bisphenol-A trigger a neurotoxic response in the brain of zebrafish: Oxidative stress, behavioral impairment, acetylcholinesterase inhibition, and gene expression disruption. *Chemosphere*, *330*. https://doi.org/10.1016/j.chemosphere.2023.138729

Hoerl, A. E., & Kennard, R. W. (2000). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *42*(1), 80–86. https://doi.org/10.1080/00401706.2000.10485983

Ivanović, V., Rančić, M., Arsić, B., & Pavlović, A. (2020). Lipinski's rule of five, famous extensions and famous exceptions. In *POPULAR SCIENTIFIC ARTICLE* (Vol. 3, Issue 1).

Kadir, A., Darreh-Shori, T., Almkvist, O., Wall, A., Grut, M., Strandberg, B., Ringheim, A., B. Eriksson, Blomquist, G., Långström, B., & Nordberg, A. (2008). PET imaging of the in vivo brain acetylcholinesterase activity and nicotine binding in galantamine-treated patients with AD. *Neurobiology of Aging*, *29*(8), 1204–1217. https://doi.org/10.1016/j.neurobiolaging.2007.02.020

Karakuş, A., CEYLAN, H., & ERDOĞAN, O. (2022). Cloning and Expression of Rat Brain Acetylcholinesterase Enzyme in Escherichia coli. *Journal of the Institute of Science and Technology*, 287–296. https://doi.org/10.21597/jist.962268

Keiser, M. J., Roth, B. L., Armbruster, B. N., Ernsberger, P., Irwin, J. J., & Shoichet, B. K. (2007). Relating protein pharmacology by ligand chemistry. *Nature Biotechnology*, *25*(2), 197–206. https://doi.org/10.1038/nbt1284

Keller, T. H., Pichota, A., & Yin, Z. (2006). A practical view of "druggability." In *Current Opinion in Chemical Biology* (Vol. 10, Issue 4, pp. 357–361). https://doi.org/10.1016/j.cbpa.2006.06.014

Kim, M. J., & Kang, D. K. (2010). Ensemble with neural networks for bankruptcy prediction. *Expert Systems with Applications*, *37*(4), 3373–3379. https://doi.org/10.1016/j.eswa.2009.10.012

Kong, W., Hu, Y., Zhang, J., & Tan, Q. (2022). Application of SMILES-based molecular generative model in new drug design.

Kufareva, I., Chen, Y.-C., Ilatovskiy, A. V, & Abagyan, R. (2014). *Compound activity prediction using models of binding pockets or ligand properties in 3D*.

Khedekar, S. A., Mhatre, N. S., & Mendhe, R. (2022). Prediction of pIC50 Values for the Acetylcholinesterase (AChE) using QSAR Model.

Lan, N. T., Vu, K. B., Dao Ngoc, M. K., Tran, P. T., Hiep, D. M., Tung, N. T., & Ngo, S. T. (2019). Prediction of AChE-ligand affinity using the umbrella sampling simulation. Journal of molecular graphics & modelling, 93, 107441. https://doi.org/10.1016/j.jmgm.2019.107441

Layer, P. G., Klaczinski, J., Salfelder, A., Sperling, L. E., Thangaraj, G., Tuschl, C., & Vogel-Höpker, A. (2013). Cholinesterases in development: AChE as a firewall to inhibit cell proliferation and support differentiation. *Chemico-Biological Interactions*, *203*(1), 269–276. https://doi.org/10.1016/j.cbi.2012.09.014

Lei, J. (2020). Cross-Validation With Confidence. *Journal of the American Statistical Association*, *115*(532), 1978–1997. https://doi.org/10.1080/01621459.2019.1672556

Li, C., Feng, J., Liu, S., & Yao, J. (2022). A Novel Molecular Representation Learning for Molecular Property Prediction with a Multiple SMILES-Based Augmentation. *Computational Intelligence and Neuroscience*, *2022*. https://doi.org/10.1155/2022/8464452

Li, Q., & Liny, N. (2010). The Bayesian elastic net. *Bayesian Analysis*, *5*(1), 151–170. https://doi.org/10.1214/10-BA506

Libbrecht, M. W., & Noble, W. S. (2015). Machine learning applications in genetics and genomics. In *Nature Reviews Genetics* (Vol. 16, Issue 6, pp. 321–332). Nature Publishing Group. https://doi.org/10.1038/nrg3920

Liu, P., Li, H., Li, S., & Leung, K. S. (2019). Improving prediction of phenotypic drug response on cancer cell lines using deep convolutional network. *BMC Bioinformatics*, *20*(1). https://doi.org/10.1186/s12859-019-2910-6

Lorber, A., Wangen, L. E., & Kowalski, B. R. (1987). A theoretical foundation for the PLS algorithm. *Journal of Chemometrics*, *1*(1), 19–31. https://doi.org/10.1002/cem.1180010105

Lou, L. L., & Martin, J. C. (2021). Selected thoughts on hydrophobicity in drug design. *Molecules*, *26*(4). https://doi.org/10.3390/molecules26040875

Lv, W., & Xue, Y. (2010). Prediction of acetylcholinesterase inhibitors and characterization of correlative molecular descriptors by machine learning methods. *European Journal of Medicinal Chemistry*, *45*(3), 1167–1172. https://doi.org/10.1016/j.ejmech.2009.12.038

Macalino, S. J. Y., Gosu, V., Hong, S., & Choi, S. (2015). Role of computer-aided drug design in modern drug discovery. In *Archives of Pharmacal Research* (Vol. 38, Issue 9, pp. 1686–1701). Pharmaceutical Society of Korea. https://doi.org/10.1007/s12272-015-0640-5

Martel, S., Gillerat, F., Carosati, E., Maiarelli, D., Tetko, I. V., Mannhold, R., & Carrupt, P. A. (2013). Large, chemically diverse dataset of log P measurements for benchmarking studies. *European Journal of Pharmaceutical Sciences*, *48*(1–2), 21–29. https://doi.org/10.1016/j.ejps.2012.10.019

Meinshausen, N. (2006). Quantile Regression Forests. In *Journal of Machine Learning Research* (Vol. 7).

Mendez, D., Gaulton, A., Bento, A. P., Chambers, J., De Veij, M., Félix, E., Magariños, M. P., Mosquera, J. F., Mutowo, P., Nowotka, M., Gordillo-Marañón, M., Hunter, F., Junco, L., Mugumbate, G., Rodriguez-Lopez, M., Atkinson, F., Bosc, N., Radoux, C. J., Segura-Cabrera, A., … Leach, A. R. (2019). ChEMBL: Towards direct deposition of bioassay data. *Nucleic Acids Research*, *47*(D1), D930–D940. https://doi.org/10.1093/nar/gky1075

Mienye, I. D., & Sun, Y. (2022). A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects. In *IEEE Access* (Vol. 10, pp. 99129–99149). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ACCESS.2022.3207287

Mok, N. Y., & Brenk, R. (2011). Mining the ChEMBL database: An efficient chemoinformatics workflow for assembling an ion channel-focused screening library. *Journal of Chemical Information and Modeling*, *51*(10), 2449–2454. https://doi.org/10.1021/ci200260t

Muñoz-Mas, R., Gil-Martínez, E., Oliva-Paterna, F. J., Belda, E. J., & Martínez-Capel, F. (2019). Tree-based ensembles unveil the microhabitat suitability for the invasive bleak (Alburnus alburnus L.) and pumpkinseed (Lepomis gibbosus L.): Introducing XGBoost to eco-informatics. *Ecological Informatics*, *53*. https://doi.org/10.1016/j.ecoinf.2019.100974

Najmanovich, R., Kuttner, J., Sobolev, V., & Edelman, M. (2000). Side-chain flexibility in proteins upon ligand binding. *Proteins: Structure, Function and Genetics*, *39*(3), 261–268. https://doi.org/10.1002/(SICI)1097-0134(20000515)39:3<261::AID-PROT90>3.0.CO;2-4

Namba, A. (2001). Statistical Papers MSE performance of the 2SHI estimator in a regression model with multivariate t error terms. In *Statistical Papers* (Vol. 42).

Namba, A. (2015). MSE dominance of the positive-part shrinkage estimator when each individual regression coefficient is estimated. *Statistical Papers*, *56*(2), 379–390. https://doi.org/10.1007/s00362-014-0586-6

Nendza, M., & Müller, M. (2010). Screening for low aquatic bioaccumulation (1): Lipinski's "rule of 5" and molecular size. *SAR and QSAR in Environmental Research*, *21*(5–6), 495–512. https://doi.org/10.1080/1062936X.2010.502295

Nevozhay, D. (2014). Cheburator software for automatically calculating drug inhibitory concentrations from in vitroscreening assays. *PLoS ONE*, *9*(9). https://doi.org/10.1371/journal.pone.0106186

Nguyen, T. H., Tran, P. T., Pham, N. Q. A., Hoang, V. H., Hiep, D. M., & Ngo, S. T. (2022). Identifying Possible AChE Inhibitors from Drug-like Molecules via Machine Learning and Experimental Studies. *ACS Omega*, *7*(24), 20673–20682. https://doi.org/10.1021/acsomega.2c00908

Ogunleye, J. O. (2021). *The Concept of Data Mining*. www.intechopen.com

Olson, D. L. (2007). Data mining in business services. In *Service Business* (Vol. 1, Issue 3, pp. 181–193). https://doi.org/10.1007/s11628-006-0014-7

Ou-Yang, S. S., Lu, J. Y., Kong, X. Q., Liang, Z. J., Luo, C., & Jiang, H. (2012). Computational drug discovery. In *Acta Pharmacologica Sinica* (Vol. 33, Issue 9, pp. 1131–1140). https://doi.org/10.1038/aps.2012.109

Parsa, M. (2021). A data augmentation approach to XGboost-based mineral potential mapping: An example of carbonate-hosted Zn–Pb mineral systems of Western Iran. *Journal of Geochemical Exploration*, *228*. https://doi.org/10.1016/j.gexplo.2021.106811

Patel, L., Shukla, T., Huang, X., Ussery, D. W., & Wang, S. (2020). Machine Learning Methods in Drug Discovery. *Molecules*, *25*(22). https://doi.org/10.3390/MOLECULES25225277

Pliakos, K., & Vens, C. (2020). Drug-target interaction prediction with tree-ensemble learning and output space reconstruction. *BMC Bioinformatics*, *21*(1), 1V. https://doi.org/10.1186/s12859-020-3379-z

Pope, C. N. (1999). Organophosphorus pesticides: Do they all have the same mechanism of toxicity. *Journal of Toxicology and Environmental Health - Part B: Critical Reviews*, *2*(2), 161–181. https://doi.org/10.1080/109374099281205

Poulin, P., & Theil, F. P. (2000). A priori prediction of tissue: Plasma partition coefficients of drugs to facilitate the use of physiologically-based pharmacokinetic models in drug discovery. *Journal of Pharmaceutical Sciences*, *89*(1), 16–35. https://doi.org/10.1002/(SICI)1520-6017(200001)89:1<16::AID-JPS3>3.0.CO;2-E

Preda, C., & Saporta, G. (2005). PLS regression on a stochastic process. *Computational Statistics and Data Analysis*, *48*(1), 149–158. https://doi.org/10.1016/j.csda.2003.10.003

Ramírez-Gallego, S., Krawczyk, B., García, S., Woźniak, M., & Herrera, F. (2017). A survey on data preprocessing for data stream mining: Current status and future directions. *Neurocomputing*, *239*, 39–57. https://doi.org/10.1016/j.neucom.2017.01.078

Ramsden, J. J. (1993). Partition coefficients of drugs in bilayer lipid membranes. *Experientia*, *49*, 688-692.

Ren, Y., Zhang, L., & Suganthan, P. N. (2016). Ensemble Classification and Regression-Recent Developments, Applications and Future Directions [Review Article]. In *IEEE Computational Intelligence Magazine* (Vol. 11, Issue 1, pp. 41–53). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/MCI.2015.2471235

Ricci, L. (2010). Adjusted R-squared type measure for exponential dispersion models. *Statistics and Probability Letters*, *80*(17–18), 1365–1368. https://doi.org/10.1016/j.spl.2010.04.019

Ristroph, K., Salim, M., Wilson, B. K., Clulow, A. J., Boyd, B. J., & Prud'homme, R. K. (2021). Internal liquid crystal structures in nanocarriers containing drug hydrophobic ion pairs dictate drug release. *Journal of Colloid and Interface Science*, *582*, 815–824. https://doi.org/10.1016/j.jcis.2020.08.045

Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. In *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* (Vol. 8, Issue 4). Wiley-Blackwell. https://doi.org/10.1002/widm.1249

Sagi, O., & Rokach, L. (2021). Approximating XGBoost with an interpretable decision tree. *Information Sciences*, *572*, 522–542. https://doi.org/10.1016/j.ins.2021.05.055

Saleh, S. (2014). MODEL SELECTION VIA ROBUST VERSION OF R-SQUARED. *Journal of Mathematics and Statistics*, *10*(3), 414–420. https://doi.org/10.3844/jmssp.2014.414.420

Sandhu, H., Kumar, R. N., & Garg, P. (2022). Machine learning-based modeling to predict inhibitors of acetylcholinesterase. *Molecular Diversity*, *26*(1), 331–340. https://doi.org/10.1007/s11030-021-10223-5

Sarkar, A., & Kellogg, G. E. (2010). Hydrophobicity-shake flasks, protein folding and drug discovery. *Current topics in medicinal chemistry*, *10*(1), 67-83.

Schmidt, M.F. (2022). Proteins as Drug Targets. In: Chemical Biology. Springer, Berlin, Heidelberg.

Scornet, E., Biau, G., & Vert, J. P. (2015). Consistency of random forests. *Annals of Statistics*, *43*(4), 1716–1741. https://doi.org/10.1214/15-AOS1321

Scott, I. A., Cook, D., Coiera, E. W., & Richards, B. (2019). Machine learning in clinical practice: prospects and pitfalls. *Medical Journal of Australia*, *211*(5), 203–205. https://doi.org/10.5694/mja2.50294

Sega, M., & Xiao, Y. (2011). Multivariate random forests. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*(1), 80–87. https://doi.org/10.1002/widm.12

Steiner, T., & Koellner, G. (2001). Hydrogen bonds with π-acceptors in proteins: Frequencies and role in stabilizing local 3D structures. *Journal of Molecular Biology*, *305*(3), 535–557. https://doi.org/10.1006/jmbi.2000.4301

Singh, N., Vayer, P., Tanwar, S., Poyet, J.-L., Tsaioun, K., & Villoutreix, B. O. (2023). Drug discovery and development: introduction to the general public and patient groups. *Frontiers in Drug Discovery*, *3*. https://doi.org/10.3389/fddsv.2023.1201419

Smith, G., & Campbell, F. (2018). *A Critique of Some Ridge Regression Methods*. https://about.jstor.org/terms

Tian, J., Wu, N., Chu, X., & Fan, Y. (2010). Predicting changes in protein thermostability brought about by single- or multi-site mutations. *BMC Bioinformatics*, *11*. https://doi.org/10.1186/1471-2105-11-370

Tsantili-Kakoulidou, A., & Demopoulos, V. J. (2021). Drug-like Properties and Fraction Lipophilicity Index as a combined metric. *ADMET and DMPK*, *9*(3), 177–190. https://doi.org/10.5599/admet.1022

Thakur, A., Kumar, A., Sharma, V., & Mehta, V. (2022). *PIC50: An open source tool for interconversion of PIC 50 values and IC 50 for efficient data representation and analysis*. https://doi.org/10.1101/2022.10.15.512366

Tsim, K. W. K., Choi, R. C. Y., Xie, H. Q., Zhu, J. T. T., Leung, K. W., Lau, F. T. C., Chu, G. K. Y., Chen, V. P., Mok, M. K. W., Cheung, A. W. H., & Bi, C. W. C. (2008). Transcriptional control of different subunits of AChE in muscles: Signals triggered by the motor nerve-derived factors. *Chemico-Biological Interactions*, *175*(1–3), 58–63. https://doi.org/10.1016/j.cbi.2008.04.014

Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., & Zhao, S. (2019). Applications of machine learning in drug discovery and development. In *Nature Reviews Drug Discovery* (Vol. 18, Issue 6, pp. 463–477). Nature Publishing Group. https://doi.org/10.1038/s41573-019-0024-5

Vellido, A. (2020). The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Computing and Applications*, *32*(24), 18069–18083. https://doi.org/10.1007/s00521-019-04051-w

Vignaux, P. A., Lane, T. R., Urbina, F., Gerlach, J., Puhl, A. C., Snyder, S. H., & Ekins, S. (2023). Validation of Acetylcholinesterase Inhibition Machine Learning Models for Multiple Species. *Chemical Research in Toxicology*, *36*(2), 188–201. https://doi.org/10.1021/acs.chemrestox.2c00283

Yağmuroğlu, O., & Emir Diltemiz, S. (2020). Development of acetylcholinesterase immobilized CMD (Carboxymethyldextran) chip-based sensor for the detection of nerve agent simulant parathion. *Cumhuriyet Science Journal*, *41*(4), 815–825. https://doi.org/10.17776/csj.725122

Yoo, I., Alafaireet, P., Marinov, M., Pena-Hernandez, K., Gopidi, R., Chang, J. F., & Hua, L. (2012). Data mining in healthcare and biomedicine: A survey of the literature. *Journal of Medical Systems*, *36*(4), 2431–2448. https://doi.org/10.1007/s10916-011-9710-5

Yucel, M. A. (2022). Comparing machine learning models for acetylcholine esterase inhibitors. *Bioorganic and Medicinal Chemistry Reports*, *2*, 20–27. https://doi.org/10.25135/bmcr.29.22.06.2483

Zagidullin, B., Wang, Z., Guan, Y., Pitkänen, E., & Tang, J. (2021). Comparative analysis of molecular fingerprints in prediction of drug combination effects. *Briefings in Bioinformatics*, *22*(6). https://doi.org/10.1093/bib/bbab291

Zhang, P., Jia, Y., & Shang, Y. (2022). Research and application of XGBoost in imbalanced data. *International Journal of Distributed Sensor Networks*, *18*(6). https://doi.org/10.1177/15501329221106935

Zhou, A., Hu, J., Wang, L., Zhong, G., Pan, J., Wu, Z., & Hui, A. (2015). Combined 3D-QSAR, molecular docking, and molecular dynamics study of tacrine derivatives as potential acetylcholinesterase (AChE) inhibitors of Alzheimer's disease. *Journal of Molecular Modeling*, *21*(10). https://doi.org/10.1007/s00894-015-2797-8

Wakeling, I. N., & Morris, J. J. (1993). A test of significance for partial least squares regression. *Journal of Chemometrics*, *7*(4), 291–304. https://doi.org/10.1002/cem.1180070407

Wang, D., & Alhamdoosh, M. (2013). Evolutionary extreme learning machine ensembles with size control. *Neurocomputing*, *102*, 98–110. https://doi.org/10.1016/j.neucom.2011.12.046

Wang, Y., Wang, H., & Chen, H.-Z. (2016). Send Orders for Reprints to reprints@benthamscience.ae AChE Inhibition-based Multi-target-directed Ligands, a Novel Pharmacological Approach for the Symptomatic and Disease-modifying Therapy of Alzheimer's Disease. In *Current Neuropharmacology* (Vol. 14).

Wang, Y., Xiao, J., Suzek, T. O., Zhang, J., Wang, J., & Bryant, S. H. (2009). PubChem: A public information system for analyzing bioactivities of small molecules. In *Nucleic Acids Research* (Vol. 37, Issue SUPPL. 2). https://doi.org/10.1093/nar/gkp456

Wasan, K. M., Brocks, D. R., Lee, S. D., Sachs-Barrable, K., & Thornton, S. J. (2008). Impact of lipoproteins on the biological activity and disposition of hydrophobic drugs: Implications for drug discovery. In *Nature Reviews Drug Discovery* (Vol. 7, Issue 1, pp. 84–99). https://doi.org/10.1038/nrd2353

Wu, L. L., & Huang, Z. T. (2019). Coherent SVR Learning for Wideband Direction-of-Arrival Estimation. *IEEE Signal Processing Letters*, *26*(4), 642–646. https://doi.org/10.1109/LSP.2019.2901641

Wu, C. K., Zhang, X. C., Yang, Z. J., Lu, A. P., Hou, T. J., & Cao, D. S. (2021). Learning to SMILES: BAN-based strategies to improve latent representation learning from molecules. *Briefings in Bioinformatics*, *22*(6). https://doi.org/10.1093/bib/bbab327

Wu, J., Li, Y., & Ma, Y. (2021). Comparison of XGBoost and the Neural Network model on the class-balanced datasets. *2021 IEEE 3rd International Conference on Frontiers*

Technology of Information and Computer, *ICFTIC 2021*, 457–461. https://doi.org/10.1109/ICFTIC54370.2021.9647373

Wyatt, J. L., Tino, P., & Brown, G. (2005). Managing Diversity in Regression Ensembles Managing Diversity in Regression Ensembles Peter Ti ˇ no. In *Journal of Machine Learning Research* (Vol. 6). https://www.researchgate.net/publication/220320750

EKLER

EK-1 VERİ SETİNE AİT EK BİLGİ

Çizelge Ek-1.1. ChEMBL220 veri setine ait sütunlar ve sütun veri tipleri

Sütun Numarası	**Sütun Adı**	**Veri Tipi**	**İçerik**
1	Activity Comment	string	Aktivite ile ilgili ek yorumları veya notları içerir
2	Activity Id	integer	Aktiviteye ait benzersiz kimlik numarası
3	Activity Properties	list	Aktivite ile ilgili ek özellikleri veya meta verileri içeren liste
4	Assay ChEMBL Id	string	Deneyin benzersiz ChEMBL kimlik numarası
5	Assay Description	string	Deneyin açıklaması
6	Assay Type	string	Deneyin türü. Örneğin, biyokimyasal (B) olarak belirtilmiş
7	Bao Endpoint	string	BioAssay Ontology (BAO) tarafından atanmış özel kod
8	Bao Format	string	Deney formatını belirten BAO kodu
9	Bao Label	string	Deney formatının etiketi. Bu veri seti için “tek “protein format” belirtilmiş
10	Canonical SMILES	string	Molekülün standartlaştırılmış SMILES (Simplified Molecular Input Line Entry System) dizisi, molekülün yapısını temsil eder
11	Data Validity Comment	string	Veri geçerliliği ile ilgili yorumlar
12	Data Validity Description	string	Veri geçerliliği açıklaması
13	Document Chembl Id	string	İlgili dokümanın benzersiz ChEMBL kimlik numarası

14	Document Journal	string	İlgili dokümanın yayımlandığı dergi
15	Document Year	float	İlgili dokümanın yayımlandığı yıl
16	Ligand Efficiency	json	Bağlayıcı etkinlik indeksi
17	Molecule ChEMBL Id	string	Molekülün benzersiz ChEMBL kimlik numarası.
18	Molecule Pref Name	string	Molekülün tercih edilen adı.
19	Parent Molecule ChEMBL Id	string	Ana molekülün benzersiz ChEMBL kimlik numarası
20	PChEMBL Value	float	Aktivite değerlerinin negatif logaritması
21	Potential Duplicate	boolean	Potansiyel olarak tekrarlayan kayıtları belirten alan
22	Qudt Units	string	QUDT (Quantities, Units, Dimensions and Data Types) tarafından tanımlanan birimler
23	Record Id	integer	Kaydın benzersiz kimlik numarası
24	Relation	string	Değerin diğer değerlerle ilişkisini belirten alan (örneğin, '=', '<', '>')
25	Src Id	integer	Kaynağın kimlik numarası
26	Standard Flag	boolean	Standartlaştırılmış veri olup olmadığını belirten bayrak
27	Standard Relation	string	Standartlaştırılmış değerlerin diğer değerlerle ilişkisini belirten alan
28	Standard Text Value	string	Standartlaştırılmış metin değeri
29	Standard Type	string	Standartlaştırılmış veri türü
30	Standard Units	string	Standartlaştırılmış veri birimleri
31	standard_upper_value	string	

32	Standard Value	float	Standartlaştırılmış veri değeri
33	Target ChEMBL Id	string	Hedef organizmanın kimlik numarası
34	Target Organism	string	Hedef organizma, bu veri seti için "Homo sapiens"
35	Target Pref Name	string	Hedefin tercih edilen adı, bu veri setinde "Acetylcholinesterase"
36	Target Tax Id	integer	Hedef organizmanın taksonomik kimlik numarası
37	Text Value	integer	Metin değeri
38	Toid	integer	Tanımlayıcı kimlik numarası
39	Type	string	Veri türü, bu veri seti için "IC50"
40	Units	string	Verinin birimleri, bu veri seti için "uM"
41	Uo Units	string	UO (Units Ontology) tarafından tanımlanan birimler
42	Upper Value	string	Üst değer
43	Value	float	Aktivite değeri

Çizelge Ek-1.2. Orijinal veri setine ait ilk 5 veri

İndeks No	Activity Comment	Activity Id	Activity Properties	Assay Chembl Id	Assay Description	Assay Type	Bao Endpoint	Bao Format	Bao Label	Canonical Smiles
0	None	33969	[]	CHEMBL643384	Inhibitory concentration against acetylcholine...	B	BAO_0000190	BAO_00357	single protein format	CCOc1nn(-c2ccccc(OCc3ccccc3)c2)c(=O)o1
1	None	37563	[]	CHEMBL643384	Inhibitory concentration against acetylcholine...	B	BAO_0000190	BAO_00357	single protein format	O=C(N1CCCCC1)n1nc(-c2ccc(Cl)cc2)nc1SCC1CC1
2	None	37565	[]	CHEMBL643384	Inhibitory concentration against acetylcholine...	B	BAO_0000190	BAO_00357	single protein format	CN(C(=O)n1nc(-c2ccc(Cl)cc2)nc1SCC(F)(F)F)c1ccccc1
3	None	38902	[]	CHEMBL643384	Inhibitory concentration against acetylcholine...	B	BAO_0000190	BAO_00357	single protein format	O=C(N1CCCCC1)n1nc(-c2ccc(Cl)cc2)nc1SCC(F)(F)F
4	None	41170	[]	CHEMBL643384	Inhibitory concentration against acetylcholine...	B	BAO_0000190	BAO_00357	single protein format	CSc1nc(-c2ccc(OC(F)(F)F)cc2)nn1C(=O)N(C)C

İndeks No	Data Validity Comment	Data Validity Description	Document Chembl Id	Document Journal	Document Year	Ligand Efficiency	Molecule Chembl Id	Molecule Pref Name	Parent Molecule Chembl Id	Pchembl Value	Potential Duplicate
0	None	None	CHEMBL1148382	J. Med. Chem.	2004.0	{'bei': '19.61', 'le': '0.36', 'lle': '3.32', ...	CHEMBL133897	None	CHEMBL133897	6.12	FALSE
1	None	None	CHEMBL1148382	J. Med. Chem.	2004.0	{'bei': '18.57', 'le': '0.38', 'lle': '2.45', ...	CHEMBL336398	None	CHEMBL336398	7.00	FALSE
2	None	None	CHEMBL1148382	J. Med. Chem.	2004.0	None	CHEMBL131588	None	CHEMBL131588	None	FALSE
3	None	None	CHEMBL1148382	J. Med. Chem.	2004.0	{'bei': '16.11', 'le': '0.34', 'lle': '1.81', ...	CHEMBL130628	None	CHEMBL130628	6.52	FALSE
4	None	None	CHEMBL1148382	J. Med. Chem.	2004.0	{'bei': '17.60', 'le': '0.36', 'lle': '3.00', ...	CHEMBL130478	None	CHEMBL130478	6.10	FALSE

İndeks No	Qudt Units	Record Id	Relation	Src Id	Standard Flag	Standard Relation	Standard Text Value	Standard Type	Standard Units	Standard Upper Value
0	http://www.openphacts.org/units/Nanomolar	252547	=	1	TRUE	=	None	IC50	nM	None
1	http://www.openphacts.org/units/Nanomolar	252533	=	1	TRUE	=	None	IC50	nM	None
2	http://www.openphacts.org/units/Nanomolar	252530	>	1	TRUE	>	None	IC50	nM	None
3	http://www.openphacts.org/units/Nanomolar	252534	=	1	TRUE	=	None	IC50	nM	None
4	http://www.openphacts.org/units/Nanomolar	252552	=	1	TRUE	=	None	IC50	nM	None

İndeks No	Standard Value	Target Chembl Id	Target Organism	Target Pref Name	Target Tax Id	Text Value	Toid	Type	Units	Uo Units	Upper Value	Value
0	750.0	CHEMBL220	Homo sapiens	Acetylcholinesterase	9606	None	None	IC50	uM	UO_0000065	None	0.75
1	100.0	CHEMBL220	Homo sapiens	Acetylcholinesterase	9606	None	None	IC50	uM	UO_0000065	None	0.1
2	50000.0	CHEMBL220	Homo sapiens	Acetylcholinesterase	9606	None	None	IC50	uM	UO_0000065	None	50.0
3	300.0	CHEMBL220	Homo sapiens	Acetylcholinesterase	9606	None	None	IC50	uM	UO_0000065	None	0.3
4	800.0	CHEMBL220	Homo sapiens	Acetylcholinesterase	9606	None	None	IC50	uM	UO_0000065	None	0.8