MAKİNE ÖĞRENME YAKLAŞIMLARININ BİYOİNFORMATİKTE İLAÇ GELİŞTİRME PROBLEMİNDE KULLANILMASI

USING MACHINE LEARNING APPROACHES IN DRUG DEVELOPMENT PROBLEM IN BIOINFORMATICS

TUĞÇE SEMERCİ

PROF. DR. ÇAĞDAŞ HAKAN ALADAĞ

Tez Danışmanı

Hacettepe Üniversitesi

Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliğinin

İstatistik Anabilim Dalı için Öngördüğü

YÜKSEK LİSANS TEZİ olarak hazırlanmıştır.

2023

ÖZET

MAKİNE ÖĞRENME YAKLAŞIMLARININ BİYOİNFORMATİKTE İLAÇ GELİŞTİRME PROBLEMİNDE KULLANILMASI

Tuğçe SEMERCİ

Yüksek Lisans, İstatistik Bölümü

Tez Danışmanı: Prof. Dr. Çağdaş Hakan ALADAĞ

Haziran 2023, 46 sayfa

İlaç araştırma ve geliştirme sürecinin odak noktasında insan vardır. Hastanın, hastalığını yenmesine yardım etmek ve yaşam kalitesini iyileştirmek amaçlanır. İlaç geliştirme sürecinde yenilikçi ilaçların etkin, güvenilir ve mümkün olan en kısa sürede hastaların kullanımına sunulacak tedaviler olması amaçlanır. Ancak bir ilacın keşfedilerek tıbbın hizmetine sunulması zaman alıcı ve yüksek maliyet gerektirir. Son yıllarda ise bilişim teknolojilerinin gelişmesi ve biyoinformatik tabanlı uygulamalar sayesinde bu sürecin daha az maliyetle ve hızlı bir şekilde klinik aşamaya geçilmesinde gelişme sağlanmıştır. Bu tez çalışmada Tip-2 diyabet tedavisi için, DPP-4 inhibitörleri kullanılarak ve makine öğrenme yaklaşımları yardımıyla ilaç adayı olabilecek moleküllerin tespit edilebilmesi amaçlanmıştır. ChEMBL veri tabanından elde edilen veriler 10 adet makine öğrenimi algoritmalarıyla ve yapay sinir ağı modeliyle analiz edilmiştir. Modellerin performanslarının karşılaştırılmasında Hata Kareler Ortalaması Karekök (HKOK) ölçütleri ile değerlendirilmiştir. Uygulama sonucunda, en iyi öngörüleri üreten makine öğrenme yaklaşımlarının Rastgele Orman ve tek tabakalı ileri beslemeli sinir ağı olduğu görülmüştür. Bu iki yöntemin birbirlerine yakın öngörü

sonuçları verdiği gözlemlenmiştir. Modellerin performanslarının değerlendirilmesinde, literatürde en yaygın ölçüt olan kök ortalama kare hatası değerine göre, Rastgele Orman modeli daha yüksek performans gösterdiği için optimum model olarak seçilmiştir. Yapılan bu çalışma sonuçlarına göre, Tip-2 diyabet tedavisi için ilaç adayı olabilecek moleküllerin tespit edilmesinde Rastgele Orman yaklaşımı kullanmanın iyi sonuçlar ürettiği görülmüştür.

Anahtar sözcükler: ilaç keşfi, makine öğrenmesi, yapay sinir ağları, dipeptidil peptidaz-4 inhibitörleri, QSAR

ABSTRACT

USING MACHINE LEARNING APPROACHES IN DRUG DEVELOPMENT PROBLEM IN BIOINFORMATICS

TUĞÇE SEMERCİ

Master's Degree Thesis, Department of Statistics

Supervisor: Prof. Dr. Çağdaş Hakan ALADAĞ

June 2023, 46 pages

Humans are at the center of the drug research and development process. It is aimed to help the patient overcome his illness and improve his quality of life. In the drug development process, innovative drugs are aimed to be effective, reliable and treatments that will be offered to patients as soon as possible. However, the discovery of a drug and putting it into the service of medicine requires time consuming and high cost. In recent years, thanks to the development of information technologies and bioinformatics-based applications, progress has been made in moving this process to the clinical stage with less cost and quickly. In this thesis, it is aimed to detect molecules that can be drug candidates for the treatment of Type-2 diabetes by using DPP-4 inhibitors and with the help of machine learning approaches. The data obtained from the ChEMBL database were analyzed with 10 machine learning algorithms and artificial neural network model. In comparison of the performances of the models, the Root Mean Square Error (RMSE) criteria were evaluated. As a result of the application, it has been seen that the machine learning approaches that produce the best predictions are Random Forest and a single layer feedforward neural network. It has been observed that these two methods give predictive results close to each other.

In the evaluation of the performances of the models, the Random Forest model was chosen as the optimum model because it showed higher performance than the root mean square error value, which is the most common criterion in the literature. According to the results of this study, it has been seen that using the Random Forest approach produces good results in detecting molecules that can be drug candidates for the treatment of Type-2 diabetes.

Keywords: drug discovery, machine learning, artificial neural networks, dipeptidyl peptidase-4 inhibitors, QSAR

TEŞEKKÜR

Bu çalışmanın gerçekleştirilmesinde, değerli bilgi, birikim ve tecrübelerini benimle paylaşan, kendisine her danıştığımda beni sabırla dinleyip çözüm üreten, bu süreçte karşılaştığım her zorlukta yanımda olan, yoluma ışık tutan değerli danışmanım, hocam Sayın Prof. Dr. Çağdaş Hakan ALADAĞ'a (Hacettepe Üniversitesi Fen Fakültesi İstatistik Bölümü); akademik bilgi ve deneyimlerini benimle paylaşan ve bana yol gösteren, ilgisini, güler yüzünü ve samimiyetini benden esirgemeyen değerli hocam Sayın Prof. Dr. Birdal ŞENOĞLU'na (Ankara Üniversitesi Fen Fakültesi İstatistik Bölümü); eğitimim ve tez çalışmam süresince varlıklarıyla, maddi, manevi destekleriyle bugünlere gelmemde en büyük emeği harcayan, sabır ve metanetle her zaman yanımda olan daima da yanımda olacaklarını bildiğim annem Hatice SEMERCİ, babam Cengiz SEMERCİ, ablam Kübra SEMERCİ, kardeşim Sefer SEMERCİ'ye; bu süreçte yanımda olan tüm arkadaşlarıma ve özellikle manevi desteğiyle her zaman yanımda olan değerli arkadaşım Aylin AZİZOĞLU'na tüm kalbimle teşekkür ederim.

Tuğçe SEMERCİ

Ankara, Haziran 2023

İÇİNDEKİLER

TABLOLAR

ŞEKİLLER

SİMGELER VE KISALTMALAR

AChE	Asetilkolinesteraz
Bagging	Bootstrap Aggregating
CADD	Computer Aided Drug Design
Docking	Moleküler Kenetlenme
DPP-4	Dipeptidil-peptidaz IV
DTI	Drug target interaction
IC50	Yarı maksimal inhibitör konsantrasyonu
K-EYK	K-en yakın komşu
LightGBM	Light Gradient Boosting Machine
MAE	Mean absolute error
MLR	Çoklu Doğrusal Regresyon
MSE	Mean squared error
NB	Naive Bayes
PCA	Temel Bileşen Analizi
pIC50	IC50 değerinin negatif logaritma sonucu
R^2	Belirtme katsayısı
RF	Random forest
RMSE	Root mean squared error
SMILES	Simplified molecular input line entry
SVM	Support Vector Machine
VS	Virtual Screening
YSA	Yapay Sinir Ağları
XGBoost	Extreme Gradient Boosting
QSAR	Quantitative Structure-Activity Relationship

1. GİRİŞ

Dünya üzerinde bulunan yüzlerce hastalığın her biri için, özel bir tedavi yöntemi uygulanmaktadır. Fakat dünya üzerinde bu kadar çok fazla hastalık varken etkin bir ilaç molekülünün keşfi oldukça yüksek maliyet, emek ve zaman almaktadır. Biyoloji, moleküler biyoloji ve aynı zamanda diğer temel bilimlerde çalışan araştırmacılar her gün çok sayıda yeni çalışma yayınlamaktadır. Özellikle son yıllarda teknolojinin gelişmesiyle yapılan araştırmaların sonucunda çok yüksek miktarlarda veri üretilmeye başlandı. Biyoloji günümüzde sadece laboratuvarda çalışılan bir bilim alanı olmaktan çıkmış ve bilgi teknolojisiyle iç içe çalışan bir bilim dalı haline gelmiştir (Duman, 2022). Biyolojik problemlerin çözülmesi için bilişim teknolojilerinin kullanılmaya başlanmasıyla biyoinformatik bilim dalı ortaya çıkmıştır. Farklı tiplerde üretilen veriler veri tabanlarında depolanır. Nükleotitler, proteinler, küçük moleküller vb. bu farklı veri tabanları ilişkilendirilmeye çalışılarak aradaki biyolojik ilişkiler açıklanmaya çalışılır. Bu alanda çalışan uzmanlar, hücreleri oluşturan sistemin içerisindeki bileşenlerin birbirleriyle olan ilişkilerini açıklamaya çalışarak dokuların, organların ve sistemlerin nasıl çalıştığını anlamaya çalışırlar. Böylece bu biyolojik mekanizmalar ilişkilendirilerek bir insanın tamamını anlayıp, öğrenilen bu bilgilerin farklı popülasyonlar arasındaki farklı hastalıklara yatkınlıklar ve bunların arasındaki ilişki nedir, nasıl tedavi edilir ya da nasıl erken teşhis edilir? Sorularına cevap aranır. Teknolojideki güncel ilerlemelerin biyolojik veriyi daha erişilebilir ve ucuz kılmasıyla beraber ilaç geliştirme problemi için de makine öğrenme yaklaşımları kullanılmaya başlanmıştır. Makine öğrenme yaklaşımları ile moleküler özellik verisi içindeki ilişkiler belirlenebilmekte ve deneysel verisi olmayan benzer özelliğe sahip başka ilaç adayı moleküller için tahmin yapılabilmektedir. Bu yaklaşımla, hastalar için daha doğru tedavi seçenekleri daha az süre ve maliyet ile gerçekleştirilebilir.

1.1. Araştırmanın Amacı ve Önemi

İlaç araştırma ve geliştirme sürecinde uzmanlar, hastalıkla ilişkili hedef biyomolekülü belirlendikten sonra ilgili hedef biyomolekül ile etkileşime girecek olan moleküllerin hangilerinin daha önemli olduğunu araştırırlar. Bu süreçte önemli olan, herhangi bir molekülün değil sadece hedeflenen biyomolekül ile etkileşime girecek moleküllerin bulunmasıdır. Aksi takdirde yan etkiler ortaya çıkar. Örneğin, bir hastalık tedavisi için alınan bir ilacın yüksek seviyede baş ağrısı veya mide bulantısı yapması durumunda,

aslında bu ilaç hedeflenen biyomolekül dışındaki biyomoleküllerle de etkileşip onlarında fonksiyonunu değiştirdiği için bazı yan etkiler ortaya çıkmıştır. Makine öğrenme yaklaşımları ile, bir hastalığa sebep olan mekanizmalar (reseptör, enzim gibi biyomoleküler) hedeflenerek, bir ilaç için önemli olabilecek uygun küçük moleküllerin seçilmesine daha az emek, zaman ve maliyetle yardımcı olmak amaçlanır. Bir başka ifadeyle, makine öğrenme yaklaşımlarının kullanılmasıyla, ilaç geliştirme sürecinde sarf edilen çaba azaltılmaya, maliyetler düşürülmeye ve oluşabilecek yan etkiler ortadan kaldırılmaya çalışılır. Bunun sonucunda, ilgilenilen hastalığın tedavisi için güvenilir bir ilaç daha optimal bir süreçle geliştirilebilir. Belirtilen süreç insan yaşam kalitesinin artırılmasında hayati bir rol oynar.

1.2. İlaç Geliştirme Problemi

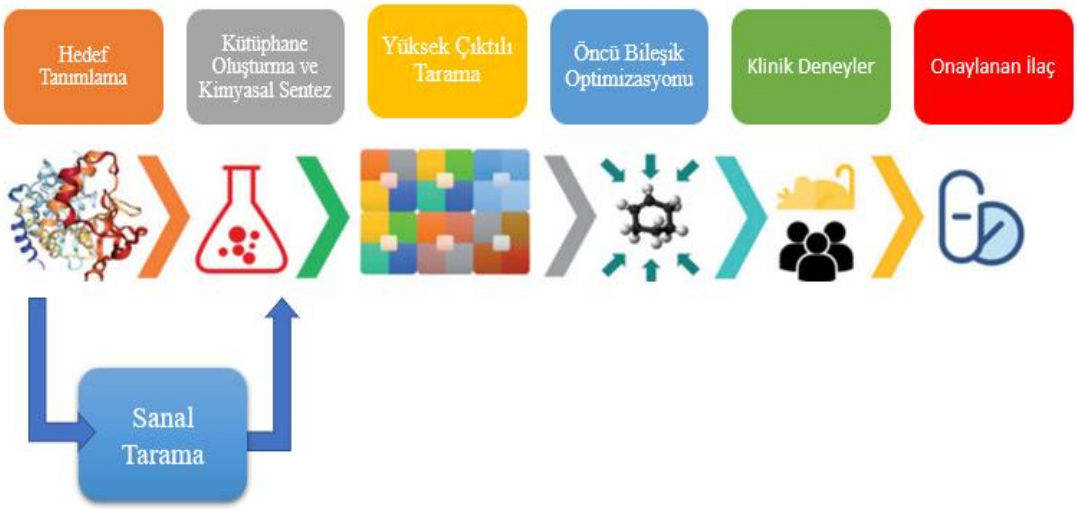
Bir hastalık için ucuz ve etkili ilaçların geliştirilmesi, insanlığın üzerinde yoğunlaştığı problemlerden biridir. İlaç geliştirme probleminin ilk aşaması genellikle ilaç keşfi olarak kabul edilir. İlaç keşif aşaması, belirli bir hastalığa sebep olan mekanizma hedeflenerek uygun ilaç adaylarının ortaya çıkarılmasını sağlayan keşif adımıdır. Ancak bir ilacın keşfedilerek tıbbın hizmetine sunulması zaman alıcı ve yüksek maliyet gerektirmesinin yanı sıra, geliştirilen yeni ilaçların onaylanma oranları da gittikçe düşmektedir.

Günümüzde artan verilerin, sistematik bir şekilde toplanması, depolanması, erişilebilir olması ve üzerinde değişiklikler yapılabilmesi amacıyla veri tabanları oluşturulmasına büyük önem verilmeye başlanmıştır. Bunun sonucunda, biyoinformatikle (hesaplamalı biyoloji ile) uğraşan araştırmacılar uluslararası açık erişim veri tabanlarındaki verilere rahatça erişebilir ve veri üzerinde analizler gerçekleştirerek yeni bilgiler üretebilmektedirler. Bu çok çeşitli ve değişik ölçekteki verilerin entegre edilip belirli hedefe göre kullanılması önem taşıyan bir konudur. Belirtilen entegrasyonun sağlanması temel bilimlerle veya klinik bilimlerle uğraşan araştırmacıların hesaplamalı bilimlerden elde edilen sonuçları kendi çalışmalarına uygulayarak bir sonraki projelerini planlamalarına yardımcı olur. Böylece maliyetleri ve sarf edilen çaba düşürülerek, yeni ilaç keşifleri daha kısa bir sürede gerçekleştirilmeye çalışılır.

1.2.1. İlaç Geliştirme Süreci

İlaç geliştirme probleminde bir hastalığın tedavisi için biyolojik mekanizmalar anlaşılmaya çalışılır. Bu amaçla belirlenen hastalık için hedef biyomolekül uzmanlar tarafından doğrulanır ve en çok bağlanma ihtimali olan moleküller bir araya getirilip bir kütüphane oluşturulur. Daha sonra moleküllerin aktif bölgelerine uyum sağlayıp sağlamayacağı deneysel olarak hedef biyomoleküle karşı denenir. Bunun sonucunda etkinliği en yüksek olanlar alınıp optimize edilerek etkinlik maksimum seviyeye çıkarılır. Bir sonraki aşamada etkisi yüksek olan etkileşimler hayvan deneyleri ve klinik araştırmalarda denenir. Elde edilen sonuçlar istenilen seviyede ise ilaç onaylanır ve marketlerdeki yerini alır. Belirtilen kütüphane oluşturma, deneme ve optimize etme süreçlerini daha verimli hale getirmek için makine öğrenmesi yaklaşımlardan faydalanılır. Bahsedilen süreçler Şekil 1.1'de gösterilmektedir.

Şekil 1.1. İlaç Geliştirme Süreci (Rifaioglu ve ark., 2019)

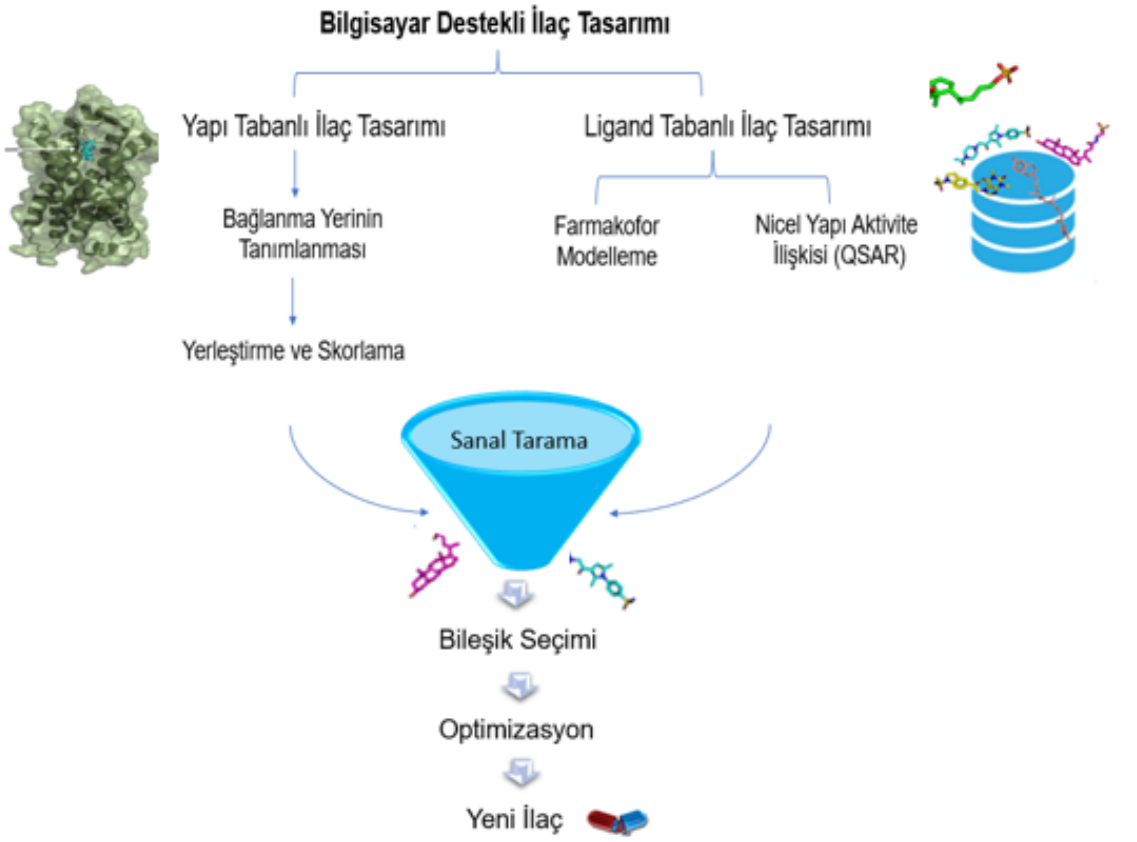


1.3. Bilgisayar Destekli İlaç Tasarımı

İlaç geliştirmeye yönelik yeni akılcı yaklaşımlar geliştirerek moleküllerin yapı ve etkinlik arasındaki ilişkilerinin aydınlatılmasına olanak sağlayan *Bilgisayar Destekli (Rasyonel) İlaç Tasarımı (Computer Aided Drug Design -CADD)* önemli bir bilimsel çalışma alanıdır. CADD tanımı Şekil 1.2'de gösterilmektedir. CADD, birbiriyle örtüşen biyokimyasal özelliklere sahip molekül ya da ilaçları oluşturan yapıları tanımlayarak bu saptanan bulguları geliştirmek ve analizlerini yapmak amacıyla hesaplamalı tekniklere başvurmaktadır (Özcan & Yöşili , 2022). CADD, küçük moleküllerin (ligand) hedef biyomoleküle nasıl bağlanacağının tahmin edilmesine ve

yaklaşık bağlanma afinitesinin tahmin edilmesine katkı sağlamak amacıyla ligand temelli ve yapı temelli olmak üzere ikiye ayrılmaktadır. İki yaklaşımda da öncü moleküllerin ortaya çıkarılması, optimize edilmesi ve klinik çalışmalara geçilmesi amaçlanır.

Şekil 1.2. Bilgisayar Destekli İlaç Tasarımı (Macalino ve ark., 2015)



1.3.1. Yapı Tabanlı İlaç Tasarımı

Yapı tabanlı ilaç tasarımı, 3D yapısı bilinen bir protein hedefine aday ligandların yerleştirilmesi ve ardından aday ligandın yüksek bağlanma değeri ile proteine bağlanma olasılığını tahmin etmek için bir skorlama fonksiyonu uygulanmasını içerir. Yapı tabanlı ilaç tasarımında, CADD yönteminin temel amacı, aday ligandın ilgili hedefe bağlanıp bağlanmayacağı veya ne kadar güçlü bağlanacağına dair tahminde bulunmaktadır (Jyothirmayee, 2014). Yapı tabanlı ilaç tasarımı, hedef biyomolekülün bağlanma bölgesiyle etkileştiği düşünülen bileşikleri araştırmak için genellikle moleküler docking yöntemini kullanmaktadır (Guedes R. ve ark., 2016).

Moleküler Kenetlenme (Docking), üç boyutlu yapısı bilinen reseptör ile etkileşime girebilecek molekülün (protein/ligant, protein/protein) en düşük enerjili bağlanma modlarının bulunması metodudur. Docking, aday ligandın uygun konformasyonu ile reseptör arasında anahtar-kilit ilişkisine benzer bir uyum olması durumudur (Lengauer, T., Rarey, M., 1996). Enerji bakımından en çok tercih edilen konformasyonel bağlanma modunu bulma amacıyla biyomolekülün (protein) aktif bölgesinde küçük molekülün (ligand) çevrilmesiyle, dönmesiyle ve bükülmesiyle docking, bir nevi tanıma prosesi gerçekleştirmekte ve bağlanma enerjisi olan protein-ligand afinitesi puanlama fonksiyonu olan algoritmayla hesaplanmaktadır (Guedes R. ve diğerleri, 2016). İlgili reseptörün sabit tutulduğu, fakat bağlanacak olan aday ligandın tercihe göre sabit tutularak veya sabit tutulmayarak gerçekleştirilen 'rijit docking', ilgili reseptör ve aday ligandın indüklenmiş uyum hipotezine göre en olabilecek bağlanma modunun gerçekleştiği 'indüklenmiş uyum docking' gibi çeşitli docking metotları mevcuttur (Olgac A., 2017).

1.3.2. Ligand Tabanlı İlaç Tasarımı

Ligand tabanlı ilaç tasarımı, bir hedefe karşı etkinliği bilinen bileşiklerin bilgisayar ortamında ligand bilgilerinin kaydedildiği veri tabanlarını kullanarak yeni bileşiklerin keşfedilmesinde ve geliştirilmesinde aynı zamanda önceden bilinen bileşiklerin etkilerinin nedenlerinin incelenmesinde kullanılabilmektedir (Petterson I. ve ark., 2010). Benzer bileşiklerin kimyasal yapılarından yola çıkarak benzer biyolojik tepki ve etkileşime sahip oldukları düşünülerek geliştirilmiş bir yaklaşımdır. Hedef biyomolekülün 3-boyutlu yapı bilgisi mevcut olmadığı durumlarda etkinliği daha önceden bilinen ligand yapı bilgileri ile yeni öncül ilaç molekülleri tasarlanmaktadır. Ligand tabanlı ilaç tasarımı için popüler bir yaklaşım, hedef biyomolekülün bilinen kimyasal yapı bilgileri kullanılarak, bir aday bileşiğin yapısı ile biyolojik aktivitesi (örneğin; Ki, Kd, IC50) arasında matematiksel bir korelasyon kurulması ve yeni ilaç adayı bileşiklerin tahmininin yapılmasıdır. Bu matematiksel olarak ifade edilebilen ve istatistiksel bir yaklaşım olan nicel yapı-aktivite ilişkisi (Quantitative Structure-Activity Relationship, QSAR) olarak adlandırılır. QSAR modelleriyle, biyolojik aktiviteleri bilinen benzer bileşiklerin kimyasal yapılarından yola çıkarak yeni ya da test edilmemiş kimyasalların biyolojik aktiviteleri (ya da özellik, reaktivite gibi), belirlenebilir (Ağca, 2014).

Sanal Tarama

İlaç geliştirme sürecinde kullanılan önemli yaklaşımlardan biri Sanal taramadır (Virtual Screening –VS). Sanal tarama metodu ile bilgisayar ortamında oldukça fazla sayıda ve birçok farklı sentezlenmiş / sentezlenmemiş bileşiğe sahip kimyasal veri bankasındaki ligandların belirili bir biyolojik hedefe karşı aktifliklerini belirleyebilmek için kullanılan bir tarama metodudur (Olgac A., 2017). Kısaca sanal tarama, moleküllerin ilgili hedef biyomoleküle karşı taranması, laboratuvar ortamında değil de bilgisayarda taranması olarak da ifade edebilir.

Sanal tarama metodu kullanılarak ilaç geliştirme problemi ele alınırsa, hedef biyomolekülün klinik çalışmalarda belirlenip doğrulanmasının ardından elde edilen bilgiler veri analizini gerçekleştirecek uzmanlara sunulur. Bunun sonucunda da veri analizini gerçekleştiren uzmanlar belirlenen hedef biyomolekül ile etkileşime girebilecek molekülleri sanal tarama metoduyla bulmaya çalışırlar. Elde edilen bu potansiyel ilaç adayı moleküllerin hedef biyomolekül ile uyum sağladığına kanaat getirilirse deneysel çalışmaları da yapılarak, etkisi yüksek ve yan etkisi düşük olduğu gözlemlendiği takdirde ilaç raflardaki yerini alır. Böylece örneğin 1 milyonun üzerinde molekül için zaman ve para harcanmak yerine, sadece 10.000 molekül için zaman ve para harcanarak istenilen sonuca ulaşılmış olur.

1.4. Çözümlenen Problem Tanımı

Dünyada en hızlı büyüyen kronik hastalıklardan birisi olan Tip-2 diyabet, pankreasın yeterli miktarda insülin hormonu üretmemesi veya ürettiği insulin hormonunun etkili bir şekilde kullanılamaması durumunda gelişen bir hastalıktır. Literatürde Tip-2 diyabet tedavisi için kullanılan yöntemlerden birisi olan ve dokularda yaygın biçimde bulunan DPP-4 enzimleri hedeflenir. Besin alınımı ile bağırsaktan inkretin hormonlar salgılanır. Bunlar GIP ve GLP-1 hormonlarıdır. GIP' in diyabette etkisi az olduğu için Tip-2 diyabet tedavisinde göz ardı edilebilir. GLP-1, pankreas da bulunan GLP-1 reseptörü üzerinde etki eder ve insülin salgılanmasına yardımcı olur. Ancak GLP-1'in, DPP-4 enzimi tarafından parçalanması durumunda kandaki glikoz (kan şekeri) seviyesi düşürülemez ve beraberinde birçok hastalığa da sebep olur. Tip-2 diyabeti tedavi etmek için DPP-4 enzimi hedeflenerek GLP-1' in yapısı bozulmadan GLP-1 reseptörüne bağlanması sağlanabilir. DPP-4 inhibitörleri, DPP-4'ü inhibe eder ve

böylece GLP-1'in yapısı bozulmaz. Pankreas üzerinde etkisi artar ve daha fazla insülin salgılanır. Bu tez çalışmasında, Tip-2 diyabet tedavisi için DPP-4 inhibitörleri kullanılarak yüksek etkiye sahip uygun ilaç adayı moleküllerin tahmini güncel çeşitli makine öğrenme yöntemleri kullanılarak gerçekleştirilmiştir.

1.5. Literatürde İlgili Çalışmalar

Veri tiplerinin ve modelleme yaklaşımlarının farklılığından dolayı literatürde ilaç-hedef etkileşim (Drug Target Interaction-DTI) problemi için geliştirilen yöntem sayısı fazladır. Bu kısımda pIC50 değerlerinin tahmini veya DPP-4 İnhibitörleri kullanılarak yapılan daha önceki çalışmalar göz önüne alınarak seçilmiştir.

Simeon ve ark. (2016) Alzheimer hastalığı tedavisi için kullanılan Asetilkolinesteraz (AChE), biliş ve hafıza için gerekli olan nörotransmiter asetilkolinin parçalanmasını katalize eden bir enzimdir. ChEMBL veri tabanından IC50 metriğine göre bileşikler elde edilmiş ve biyoaktivite kökenleri hakkında fikir edinmek için nicel yapı-aktivite ilişkisi (QSAR) çalışmasında kullanılmıştır. AChE inhibitörleri, 12 farklı parmak izi tanımlayıcısı tarafından tanımlanmış ve QSAR tahmin modelleri oluşturulmuştur. Modellerin değerlendirilmesi 10 kat çapraz ve Y-scrambling test doğrulama yöntemleri kullanılmıştır.

DPP-IV inhibitörlerinin iskeleleri yapısal olarak çeşitlidir. Bu sebeple geleneksel QSAR yaklaşımları kullanılarak yeni yapıların tahmin edilmesi zordur. Cai ve ark. (2017) yaptıkları çalışmada Naive Bayes (NB) ve özyinelemeli bölümleme yöntemlerini içeren makine öğrenimi yaklaşımlarıyla DPP-IV inhibitörlerini tahmin etmek için yeni bir strateji sunulmuştur. Tanımlayıcı olarak moleküler özelliklere ve topolojik parmak izlerine göre optimize edilmiş bilinen 1307 DPP-IV inhibitörünü temel alan 247 makine öğrenimi modeli oluşturulmuştur. Optimize edilmiş modellerin genel tahmin doğruluğu % 80' den fazla olduğu gösterilmiştir.

Ghamali ve arkadaşları (2017) yaptıkları çalışmada temel bileşenler analizi (PCA), çoklu doğrusal regresyon (MLR) ve yapay sinir ağları (YSA) gibi istatistiksel yöntemler kullanarak fenollerin ve tiofenollerin P. phosphorum'a toksitesinin QSAR

modellerini tahmin etmektedir. Çalışmada oluşturulan iki modelin değerlendirilmesi sonucunda YSA modeli MLR modeline kıyasla daha iyi performans göstermiştir.

Rastgele parametre seçimi, derin sinir ağının (Deep Neural Network-DNN) düşük performansına yol açtığı için Ghasemi ve arkadaşları (2018) tarafından DNN' leri başlatmak için derin inanç ağı (deep belief network -DBN) uygulanmıştır. DBN, tüm numuneler için hesaplama logaritmik olabilirlik gradyanı (log likelihood gradient) gerektiren enerji tabanlı bir yöntem olan bazı kısıtlanmış Boltzmann makinesi yığınlarından oluşur. Bu gradyanı çözmek için üç farklı örnekleme yaklaşımı önerilmiş ve DNN' e göre daha iyi performans sergilediği gösterilmiştir.

Haris ve arkadaşları (2020) yaptığı bu çalışmada, Tip-2 diyabet tedavisi için moleküler yapı ve biyolojik aktivite değerleri arasındaki ilişkiye dayanarak DPP-4 inhibitörlerini tahmin edebilen bir yöntem geliştirilmiştir. Tanımlayıcı olarak topolojik parmak izleri kullanılarak çıkarılan 1185 aktif bileşik ve 588 aktif olmayan kullanılmıştır. Veri setinde bir sınıf dengesizliği olduğundan, SMOTE tekniğini kullanılmış ve özellik seçme yöntemi olarak CatBoost kullanılmıştır. Sonuç olarak, ECFP_6 ile birleştirilen DNN yöntemi ve özelliğin önem değeri oranı %90 olan özellik seçimi kullanılarak doğruluk değeri 0.906 olduğu gösterilmiştir.

Rifaioglu ve arkadaşları (2020) tarafından yayımlanan çalışmada, evrişimli sinir ağları kullanarak ilaç hedef etkileşim tahmin sistemi olan DEEPScreen önerilmiştir. DEEPScreen sisteminde girdi olarak geleneksel tanımlayıcılar yerine ilaç adayı bileşiklerin 2 boyutlu temsilleri kullanılarak ilaç hedef etkileşim tahmini yapılmıştır. DEEpScreen sistemi, 704 hedef protein için eğitilmiş ve hiper parametre optimizasyonu yapılmıştır. Önerilen yaklaşımın etkinliğini göstermek ve moleküler yerleştirme analizi ve literatüre dayalı doğrulama yoluyla seçilen yeni tahminleri doğrulamak için DEEPScreen'in performansını çok sayıda kıyaslama veri seti ile karşılaştırılmış ve deneysel olarak gösterilmiştir.

Gelecekteki araştırmalar için, QSAR modelleri için modelin performansını artırmak üzere çeşitli diğer hibrit derin öğrenme yöntemleri literatüürde denenmiştir. Ulfa ve arkadaşları (2021) çalışmada Tip-2 diyabet tedavisi için moleküler yapı ve biyolojik aktivite değerleri arasındaki ilişkiye dayanarak DPP-4 inhibitörlerini tahmin edebilen

bir yöntem geliştirilmiştir. DPP-4 inhibitörünü ikili sınıflandırma olarak (aktif ve aktif olmayan) bileşiklere göre tahmin etmek için Conv1D-LSTM modeli olan iki derin öğrenme yaklaşımının bir kombinasyonunu önerilmiştir.

Hermansyah ve arkadaşları (2021) yaptıkları araştırmada DPP-8 ve DPP-9 enzimlerine karşı seçici olan DPP-4-inhibitörleri ile uygun bileşikleri belirlemek için yapı-aktivite ilişkisi (QSAR) modeli oluşturulmuştur. 4 farklı sınıflandırma algoritması (DL, XGBoost, RF, SVR) ve 5 farklı regresyon algoritması (DL, XGBoost, MLR, RF, SVR) tahmin sonuçlarının karşılaştırılmıştır. Sınıflandırma ve regresyon modellerinden elde edilen tahminler IC50 değeri pIC50 değerine dönüştürülmüş ve ara bileşikler çıkarılmıştır. Regresyon modeli için karar destek vektörleri ve sınıflandırma modeli için RandomForest algoritması en iyi model olarak seçilmiştir.

Wan ve arkadaşları (2019) geliştirdiği DeepCPI, büyük miktarda etiketlenmemiş veriden, ilaç hedefleri ve bileşikler arasındaki etkileşimleri tahmin etmek için derin öğrenme yöntemini kullanan bir algoritmadır. DeepCPI tarafından başlatılan sanal tarama ile küçük moleküllü bileşiklerin, çeşitli ilaç hedefleri üzerinde gerçekleştirilen deneysel sonuçlar sunulmuştur. Bu sonuçlar, DeepCPI'nın yüksek doğruluk oranları sağladığını, büyük veri setlerinde etkili bir şekilde çalışabildiğini ve bu nedenle ilaç keşfine yardımcı olabileceğini göstermektedir.

2. İLAÇ GELİŞTİRME PROBLEMİNDE MAKİNE ÖĞRENME YAKLAŞIMLARI

Makine öğrenmesi, bilgisayarın bir olay ile ilgili özellikleri deneyim yolu ile öğrenerek daha önce karşılaşılmayan benzer olaylar için karar verebilmesi veya çözüm üretebilmesidir (Kemal, 2020). Kısaca insan faktörü minimize edilerek, oluşturulan modelin bilgisayara problem hakkında sunulan veriden kendi kendine öğrenmesi olarak ifade edilebilir. Denetimli Makine Öğrenimi, etiketli veri (hangi verinin hangi bilgiye karşılık geldiği bilinen) ile eğitilen, ortak özellikleri öğrenebilen, genelleştirebilen ve gelecekteki yeni veriler hakkında tahminlerde bulunabilen bir makine öğrenmesi modelidir.

Son yıllarda yapılan araştırmaların artmasıyla birlikte artan biyolojik verinin analizinde klasik yöntemlerin yetersiz kaldığı söylenebilir. Buna karşın yapay sinir ağları gibi modern makine öğrenme yöntemleri, içlerindeki gizli örüntüleri bulmak ve doğru tahminler yapmak için çok büyük veri kümelerini etkin biçimde analiz edebilmeyi vaat etmektedir (Angermueller ve ark., 2016).

İlaç geliştirme probleminin ilk aşaması olan ilaç keşfinde makine öğrenme yaklaşımlarının uygulama alanlarından biri ilaç-hedef etkileşim tahminidir. İlaç-hedef etkileşimlerinin tahmin edilebilmesi için literatürde çeşitli yaklaşımlar kullanılmıştır. Bugüne kadar yapılan çalışmalar incelenmiş ve ilaç-hedef etkileşimlerinin tespitinde, doğrusal regresyon, destek vektör makineleri, K-en yakın komşu, karar ağaçları, rastgele orman, torbalama yöntemi , Adaboost, gradyan artırma, aşırı gradyan artırma, hafif gradyan artırma ,ileri beslemeli sinir ağları ve derin öğrenme modelleri gibi yöntemlerin güncel çalışmalarda kullanıldığı görülmüştür. Aşağıda belirtilen yaklaşımlardan kısaca bahsedilmiştir.

2.1. Doğrusal Regresyon

Regresyon modeli, bir fonksiyon aracılığıyla bağımlı değişkeni bağımsız değişken veya değişkenlerle ilişkilendirir (Luu ve ark.,2021). Doğrusal regresyon (Linear Regression, LR), iki sürekli değişken arasındaki ilişkiyi açığa çıkarmak amacıyla kullanılan istatistiksel bir yöntemdir.

y bağımlı değişkeni ile doğrusal bir ilişkiye sahip x bağımsız değişkeni lineer regresyon modeli olarak tanımlanmaktadır. Bu model aşağıdaki denklem ile ifade edilir.

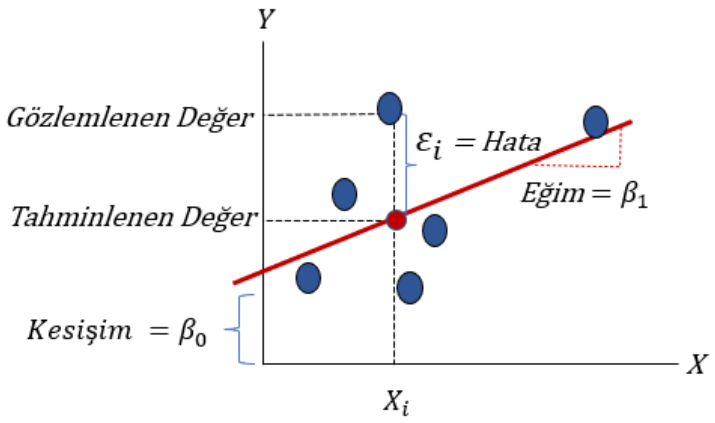
$$y = \beta_0 + \beta_1 X + \varepsilon$$

β: Bilinmeyen regresyon parametreleri (Regresyon Katsayıları)

ε: Hata Terimi

β_0 kesim noktası ve β_1 ise regresyon doğrusunun eğimini veren regresyon katsayısı olarak tanımlanır. Buradaki β_1 katsayısı x' teki bir birime karşı y'deki değişimin ölçüsünü gösterir. Şekil 2.1'de basit doğrusal regresyon grafiği gösterilmiştir.

Şekil 2.1. Doğrusal Regresyon Grafiği



Doğrusal regresyon yönteminde bağımlı değişkeni açıklamak için birden fazla bağımsız değişken de kullanılabilir. Bu modele ise çoklu doğrusal regresyon modeli denir. Bu model ise aşağıdaki denklem ile ifade edilir.

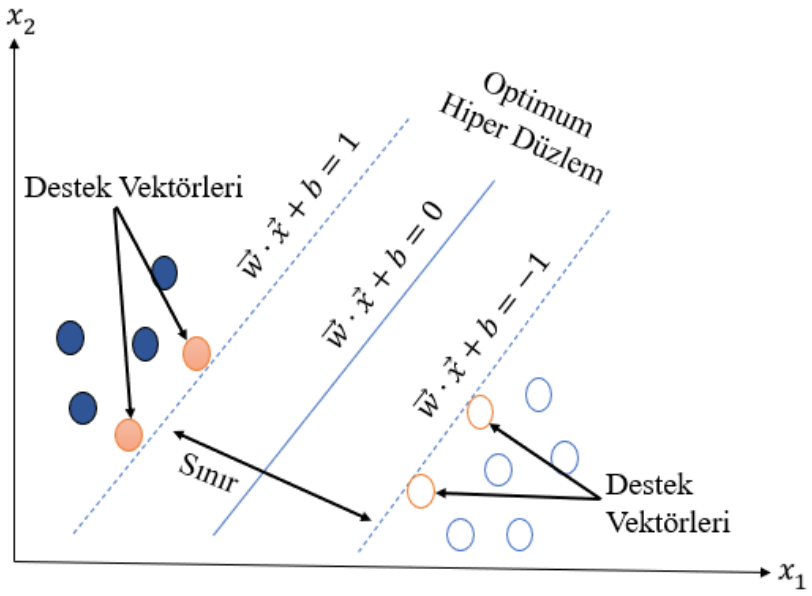
$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon$$

2.2. Destek Vektör Makineleri

Destek vektör makineleri (Support Vector Machine -SVM) girdi olarak verilen veri kümesini doğrusal vektörler yardımıyla öğrenmeyi gerçekleştiren, regresyon ve sınıflandırma için kullanıma elverişli olan denetimli bir makine öğrenme algoritmasıdır. Yöntem model karmaşıklığı üzerinde esnek bir kontrol sağladığı için doğrusal olmayan problemleri çözmede çok küçük eğitim setlerinde de başarılı sonuçlar vermektedir (Deng ve ark., 2018).

Farklı sınıflara ait hiper-düzlem üzerindeki vektörlere "destek vektörleri" denir. SVM algoritmasının amacı, bu destek vektörleri için oluşturulan hiper-düzlemler arasında optimal sınıra sahip sadece bir hiper-düzlem bulmaktır. Şekil 2.2'de destek vektör makinelerinin sınıflar arasında nasıl ayrım yaptığı grafiksel olarak gösterilmiştir.

Şekil 2.2. Destek Vektör Makineleri



Destek vektör regresyonunda amaçlanan eğitim verisi xi değeri ile önceden belirlenen hata toleransından (ε) küçük olan bir uzaklıktaki yi tahmin değerini hesaplayabilen f(x) fonksiyonu oluşturmaktır (Li ve ark., 2009). Bu fonksiyon aşağıda ifade edildiği gibidir.

$$y = f(x) = W.\emptyset(x) + b$$

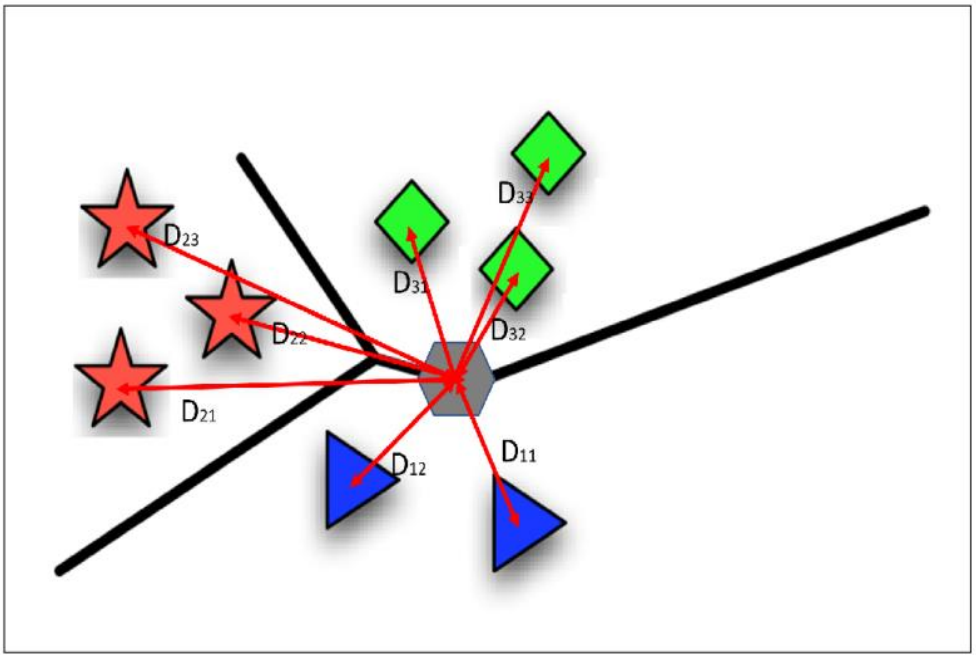
2.3. K-En Yakın Komşu

K-en yakın komşu (K-EYK) girdi olarak verilen veri kümesindeki K örneğin, benzerliklerini veya yakınlıklarını kullanarak öğrenmeyi gerçekleştiren, regresyon ve sınıflandırma için kullanıma elverişli olan denetimli bir makine öğrenme algoritmasıdır. Yeni test verisini sınıflandırmak için, veri setindeki her bir mesafesi ölçülür. Mesafe hesaplamada en çok kullanılan uzaklık ölçüsü, Euclidean uzaklığıdır (Hu ve ark., 2016). Aşağıdaki şekilde hesaplanır;

$$D(x,y) = \sqrt{\sum_{i=1}^{k} |xi - yi|^2}$$

Mesafeler hesaplandıktan sonra büyükten küçüğe sıralanır ve en küçük uzaklık değerine sahip gözlemler seçilir. Böylece k gözlemin içinde en çok tekrar eden sınıf tahmini sınıf değeri olarak atanır. Yeni gelen bir veri için en yakın komşu gösterimi Şekil 2.3'te gösterilmiştir.

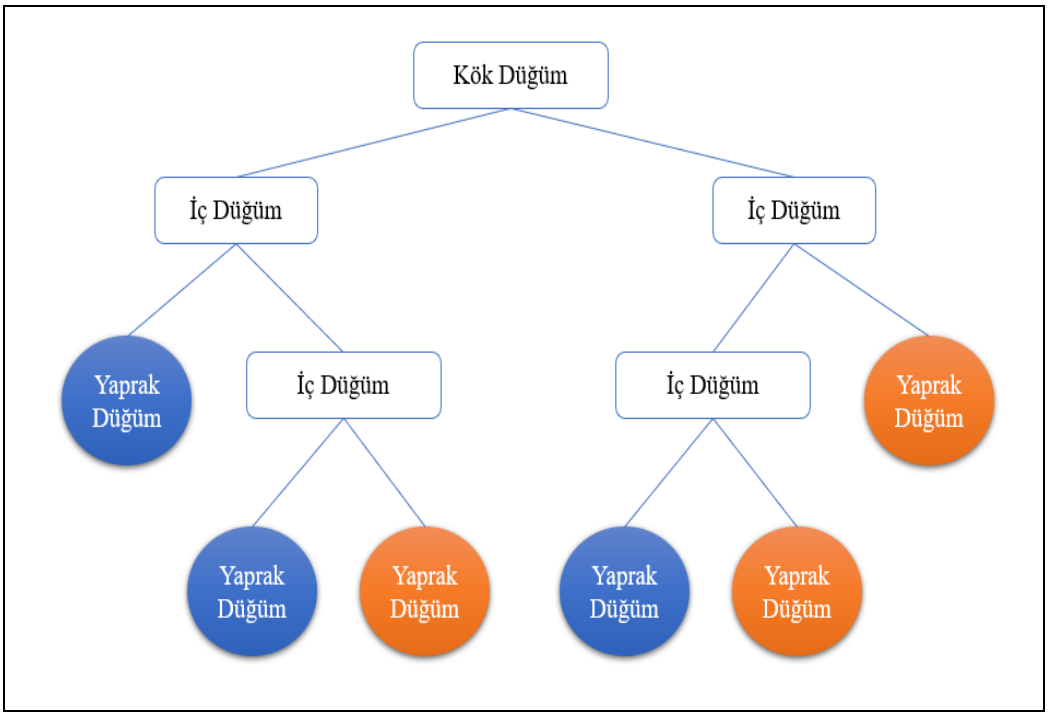
Şekil 2.3. Yeni Gri Altıgen Veri Noktasının En Yakın Komşulara Olan Mesafelerinin Gösterimi (Erdoğan,2019).



2.4. Karar Ağaçları

Karar Ağaçları (Decision Tree) yöntemi, ağaç benzeri bir yapı gösterdiği için kolay anlaşılabilir bir mantığa sahiptir. Karar ağaçları, bir ağacın yapısından etkilenerek geliştirilen kök, düğümler, dallar ve yapraklardan oluşan denetimli bir makine öğrenmesi algoritmasıdır. Her bir düğüm veri kümesindeki belirli bir özelliği ve her yaprak sınıf tahminini (düğümün sonucunu) temsil eder. Belirli koşullara dayalı olarak ilgili probleme ait tüm olası çözümlerin elde edildiği bir grafiksel gösterimdir. Karar ağaçları, bağımlı değişkendeki farklılıkları maksimize edecek şekilde veriyi kök düğümden yaprak düğüme doğru anlamlı bir şekilde bölmeye çalışır. Bu yaprak düğüme gelen yeni veri sınıflandırma problemi ise aynı etikete ve regresyon problemi için benzer sayısal sonuçlara sahiptir (Alpaydin,2010). Bu karar ağacı diyagramı Şekil 2.4'te gösterilmektedir.

Şekil 2.4. Karar Ağacı Diyagramı

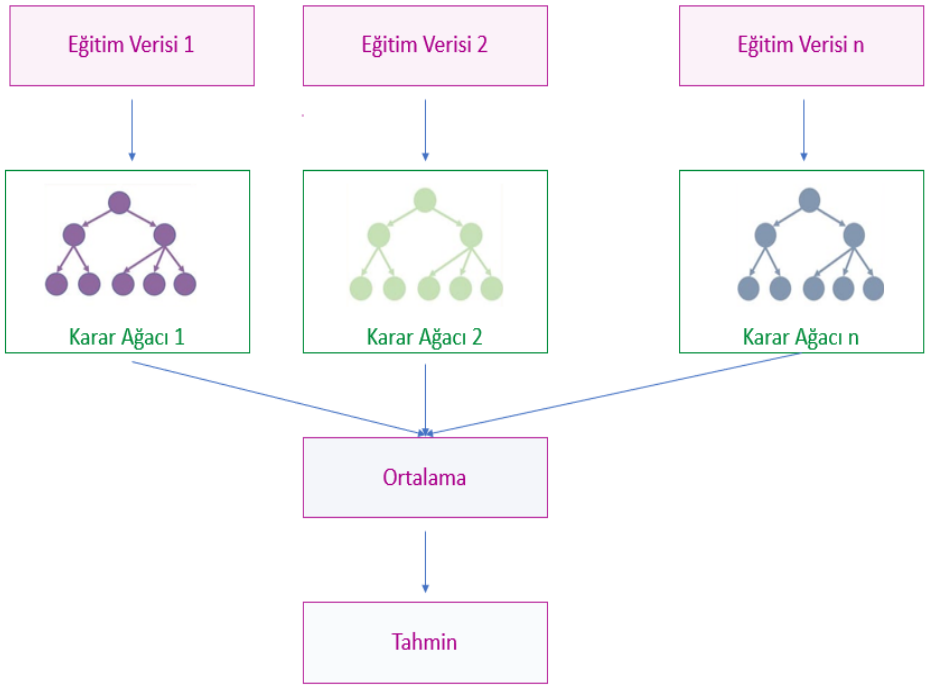


2.5. Torbalama Yöntemi

Torbalama yöntemi (Bootstrap Aggregating-Bagging), sınıflandırma ve regresyon modelleri için uygulanabilen aşırı öğrenmeye karşı güçlü olan, doğru sınıflandırma oranını arttıran, varyans düşürücü etkisi olan ve veri setinde kayıp verilerin yer aldığı durumlarda başarılı sınıflandırma yapabilen bir topluluk öğrenme yöntemidir (Akman,2010).

Bagging regresyon modelinin temel amacı, her bir alt modelin kendi hatalı tahminlerini birbirine dengeleyerek daha doğru ve kararlı bir tahmin sağlamaktır. Orijinal eğitim verileri farklı alt kümelere ayrılır ve bu alt kümeleri kullanarak karar ağacı modelleri oluşturulur. Her alt küme üzerinde ayrı ayrı regresyon modelleri eğitilir ve her bir modelin tahminleri alınır. Son olarak, tahminlerin ortalaması veya medyanı alınarak birleştirilir ve böylece nihai bir sonuç üretilir. Farklı eğitim setlerinin seçilmesi ve model çeşitliliği sayesinde karar farklılıkları elde ederek overfitting (aşırı uyum) sorununu azaltır. Şekil 2.5'te Bagging regresyon modeli gösterilmiştir.

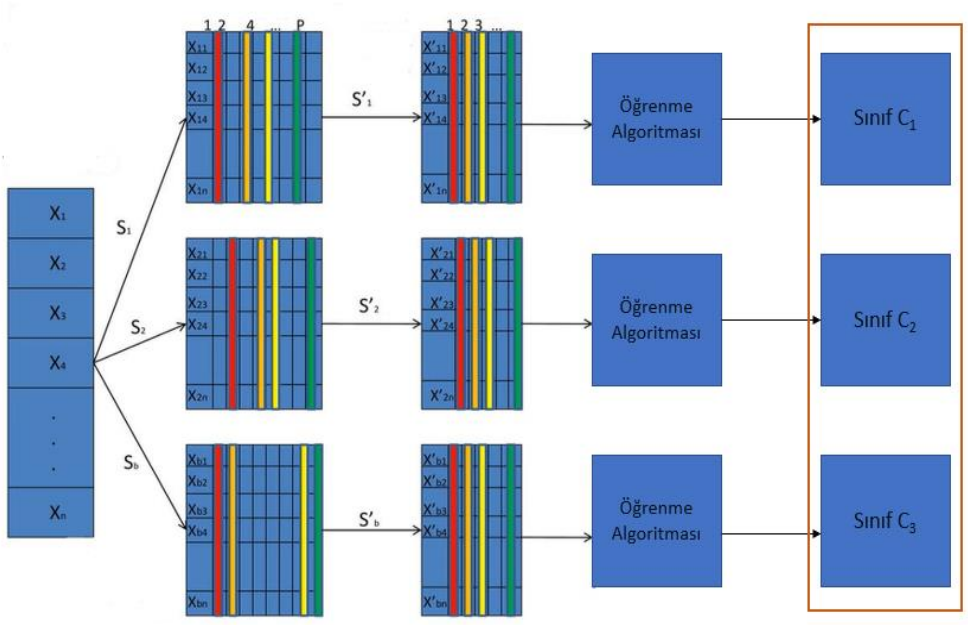
Şekil 2.5. Bagging Yöntemi



2.6. Rastgele Orman

Rastgele Orman (Random Forest-RF) yönteminde, her bir karar ağacı, orijinal veri setinden seçilen rastgele örneklemle ve her düğümde tüm değişkenlerden belirlenen sayıda rastgele değişken seçilmesi ile geliştirilmiş tahmin başarısı yüksek olan bir makine öğrenmesi algoritmasıdır (Akman,2010). Topluluk (ensemble) metoduna dayanan RF algoritmasında, orijinal veri kümesinden birden çok alt küme oluşturulur. Bu alt kümelerin her birinden bir model oluşturulur. Nihai tahmin, tüm modellerden gelen tahminleri birleştirerek, en çok atanan sınıf değerini sınıflandırma tahmini olarak belirler. Şekil 2.6'da rastgele orman algoritmasının rastgele alt küme seçimi ve sınıflandırması gösterilmektedir.

Şekil 2.6. RF Yönteminde Uygulanan Rastgele Alt Küme Seçimi Aşamaları (Panov Ve Džeroski, 2007).



Birden fazla karar ağacı ile çalıştığı için yüksek doğruluk sağlar ve overfitting (aşırı uyum) riskini azaltır. Ayrıca her bir karar ağacı, yalnızca kendi alt kümesindeki verilere dayanarak tahmin yaptığı için eksik veriler ve aykırı değerlere başa çıkmada dayanıklıdır.

2.7. AdaBoost

Adaboost algoritması, 1997 yılında Freund ve Schapire tarafından önerilen, çeşitli zayıf öğrenicilerin bir araya getirilmesiyle karma ve daha güçlü bir eğitim modeli sunmayı amaçlanmaktadır. AdaBoost yönteminde her bir zayıf öğrenici, önceki zayıf öğrenici tarafından yanlış bir şekilde sınıflandırılmış eğitim örneklerine odaklanır ve onları iyi bir şekilde sınıflandırmayı amaçlayan topluluk öğrenme ve yükseltme algoritmalarından biridir. (Kalaycı,2018). AdaBoost yöntemi T farklı zayıf öğreniciyi eğitmekte, daha sonra da bu öğrenicilerin sonuçlarını alıp ağırlıklı oy çokluğuna dayanarak ilgili örneğin sınıfına karar vermektedir (Schapire, 2013). Buna ilişkin formül aşağıda gösterilmiştir.

$$H(x) = \sum_{1}^{T} \alpha_t \, h_t(x)$$

T: Eğitilen zayıf öğrenici sayısı

h: Her bir sınıf sonucunu

α : İlgili sınıflandırıcının ağırlığını

H: Karar verilen sınıf

2.8. Gradyan Artırma

Gradyan artırma (Gradinet Boosting), ilk olarak 1999 yılında Friedman ve arkadaşları tarafından geliştirilmiş olan, günümüzde sınıflandırma ve regresyon problemlerinde yaygın olarak kullanılan bir makine öğrenmesi algoritmasıdır. Gradyan artırma algoritması, gradyan artırma yöntemi ile zayıf öğrenicileri güçlü öğreniciye dönüştürme algoritmasıdır (Feng ve ark., 2018). Topluluk öğrenme ve yükseltme algoritmalarından biri olan gradyan artırma regresyon modelinde ilk olarak regresyon ağacı oluşturulur. Her bir ağacın dalı için tahminler ile hedef değer arasındaki fark yani hata oranı hesaplanır ve bu farklar için de bir aşağıda ifade edildiği gibi bir kayıp fonksiyonunu oluşturulur.

$$\text{MSE}=\sum(y_i - y_i^p)^2$$

Hesaplanan hata oranları yeni gözlem verileri olarak kullanılır ve yeni ağaçlar oluşturularak bu hata oranlarının düşürülmesi amaçlanır. Gradyan tahminler öğrenme oranına göre iterasyonlarda güncellenir ve kayıp fonksiyonun minimum değerlere yaklaşıncaya kadar bu adımlar tekrarlanır. Aşağıdaki denklemler de y_i i. hedef değer, y_i^p i. tahmin değeri ve α ise öğrenme oranı olarak ifade edilmektedir.

$$y_i^p = y_i^p + \alpha * \delta \sum \frac{(y_i - y_i^p)^2}{\delta y_i^p}$$

$$y_i^p = y_i^p - \alpha * 2 * \sum(y_i - y_i^p)^2$$

Sonuç olarak, tahmin değerlerinin gerçek değerlere yakın olacak şekilde güncellemeler yapılarak en iyi model oluşturulur.

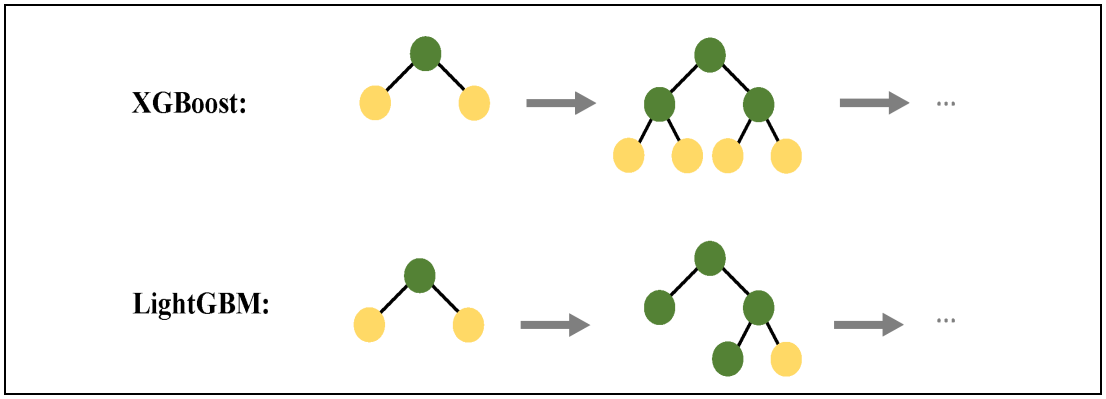
2.9. Aşırı Gradyan Artırma

Aşırı gradyan artırma (Extreme Gradient Boosting -XGBoost) algoritması 2016 yılında Washington Üniversitesi'nde Tianqi Chen ve Carlos Guestrin tarafından bir araştırma projesi olarak geliştirildi (Kurt ve Cedimoğlu,2020). Gradyan Artırma algoritmasının optimize edilmesiyle geliştirilen hızlı ve yüksek performanslı bir makine öğrenme algoritmasıdır. Algoritma ilk olarak maksimum derinlik değerini kullanarak ağacın derinliğini belirlemektedir. Ağaç aşağı yönde fazla derin ise aşırı öğrenmeyi ve yanlılığı azaltmak için budama gerçekleştirilir. Gradient Artırma algoritması, kayıp fonksiyonun hesaplanmasında birinci dereceden fonksiyon kullanırken, XGBoost bu hesaplamaları ikinci dereceden fonksiyonlar kullanarak gerçekleştirir (Kelle ve Hüseyin,2022). En önemli özelliklerinden bir tanesi ise veri kümesindeki gözlem noktalarını ağırlıklarını kullanarak verileri ağaçlara ayırırken doğru noktayı ayırmasıdır (Muratlar, 2020). XGBoost algoritması overfitting (aşırı uyum) engellemek için Gradyan Artırma algoritmasından farklı olarak, ağaçların karmaşıklığını kontrol etmek ve gereksiz dallanmaları engellemek için düzenlileştirme (regularization) teknikleri kullanır. Ek olarak, budama ve eksik değerlerle çalışabilmesi ve hızlı hesaplama gibi özellikleri ile Gradient Artırma algoritmasından ayrılır ve daha iyi bir performans sunar. Günümüzde gereksiz maillerin tespiti, reklam eşleştirme sistemleri, dolandırıcılık tespit sistemleri, fizikte anomali olay tespit sistemleri gibi alanlarda bu yöntem başarılı bir şekilde kullanılmaktadır (T. Chen ve Guestrin, 2016).

2.10. Hafif Gradyan Artırma Makineleri

Hafif gradyan artırma makineleri (Light Gradient Boosting Machine -LightGBM), karar ağacı algoritmasına dayanan sınıflandırma ve regresyon problemlerinde yaygın olarak kullanılan bir makine öğrenme algoritmasıdır. Yüksek düzeyde optimize edilmiş histogram tabanlı bir karar ağacıdır. XGBoost algoritmasına benzerdir ancak temek fark ağaçların yapısıyla ilgilidir. Bu algoritma karar ağacı algoritmalarına dayandığı için, ağacı en uygun şekilde yaprak bilgisine böler. Bu özelliği ile diğer artırma algoritmalarından farklıdır çünkü ağacı yaprak bazında değil, derinlik bazında veya seviye bazında bölmektedir (İlya ve ark.2021). Bu yaklaşım, büyüklüğü daha az düğüm içeren daha derin ağaçlar oluşturmayı sağlar. Böylece, daha az bellek kullanılır ve hızlı hesaplamalar yapılır. Bu şekilde mevcut artırma algoritmalarından herhangi biri tarafından elde edilebilecek çok daha iyi doğruluk değerlerine ulaşılmaktadır (Khandelval, 2017). Ayrıca LightGBM algoritması, XGBoost daha hızlı ve büyük veri kümelerinde daha başarılı bir performansa ulaşır. Şekil 2.7'de XGBoost ve LightGBM arasındaki farkı göstermektedir.

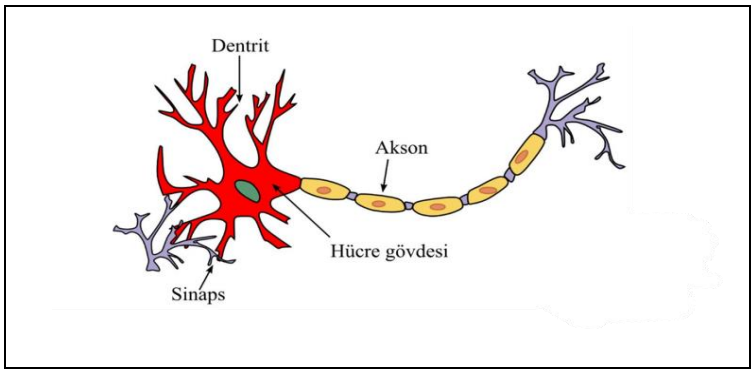
Şekil 2.7. XGBoost ve LightGBM (Rezazadeh, 2020)



2.11. Yapay Sinir Ağları

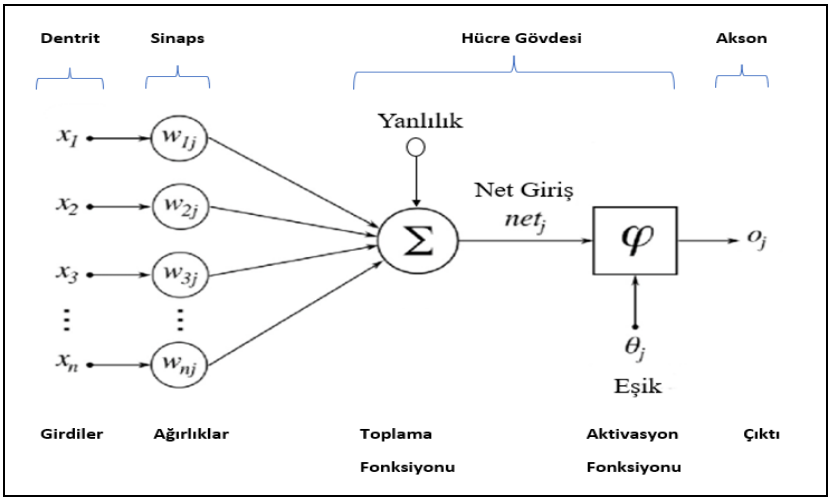
Yapay sinir ağları (YSA), biyolojik sinir ağlarını taklit eden sentetik ağlardır (Eğrioğlu ve ark., 2009). Sinir sisteminin en temel fonksiyonel birimi nöronlar, birbirleri ile sinapslar vasıtası ile iletişim kurarlar. Bir sinir hücresini; dentritler, hücre gövdesi ve aksonlar olmak üzere üç ana kısımdan oluştuğunu söyleyebiliriz. Dentritlere gelen sinyaller hücre gövdesinde işlenir ve işledikleri bilgileri aksonlar yolu ile diğer hücrelere gönderirler. Şekil 2.8'de bir sinir hücresi biyolojik olarak gösterilmiştir.

Şekil 2.8. Bir Sinir Hücresinin Biyolojik Gösterimi (Yeşilkanat Ve Ark., 2014)



Aynı biyolojik sinir ağlarında olduğu gibi yapay sinir ağlarında da giriş sinyallerini aldıkları, bu sinyalleri toplayıp işledikleri ve çıktıları ilettikleri bölümleri bulunmaktadır. YSA, biyolojik sinir hücrelerine benzer yapıdadır ancak yapı ve yetenek açısından çok daha farklıdır (Zurada,1992). Model verilerini işleyen ve tahmin yapabilen değişkenlere parametre denir. YSA temel parametreleri, ağırlıklar ve yan değerlerden (bias) oluşur. Şekil 2.9'da yapay sinir ağının matematiksel modeli gösterilmiştir.

Şekil 2.9. Yapay Sinir Ağının Matematiksel Modeli



Bir yapay sinir hücresi beş bölümden oluşmaktadır;

1.Girdiler (x): Nöronlara gelen verilerdir.

2.Ağırlıklar (w): Yapay sinir ağlarını nöronları birbirlerine bağlayan bağlantıların değerlerine ağırlık değerleri denir ve girdi değerleri ile çarpılarak biyolojik sinir hücrelerinde olduğu gibi işlenmek için hücre gövdesine gönderilir.

3.Toplama Fonksiyonu (Birleştirme Fonksiyonu): Hücre gövdesine ağırlıklarla çarpılarak gelen girdileri birleştiren bir fonksiyondur. Çeşitli toplama fonksiyonları vardır. (Örneğin; Toplam, çarpım, maksimum, minimum)

4.Aktivasyon Fonksiyonu: Bu aşamaya gelen veriler üzerinde işlem yaparak net çıktı değerlerini oluşturur.

5.Çıktılar: Aktivasyon fonksiyonundan çıkan değer hücrenin çıktı değeri olmaktadır.

Yapay sinir ağının matematiksel modelini özetlemek gerekirse; dentritlerden gelen girdi değerleri (x_0) ve ağırlıklar (w_0) çarpıldıktan sonra (w_0x_0) bir toplama fonksiyonu ile toplanır ardından bir bias (b) ile toplandıktan sonra aktivasyon fonksiyonu tarafından işlenir ve çıktı katmanına aktarılır. Kısaca Yapay Sinir Ağlarının amacı; modelin en iyi değerini vereceği w (ağırlık parametresi) ve b (bias değeri) optimum parametreleri bulmaktır.

Yapay sinir ağının temel bileşenleri;

- Aktivasyon Fonksiyonu,
- Mimari Yapısı,
- Öğrenme Algoritmaları

2.11.1. Aktivasyon Fonksiyonu

Yapay sinir ağında girdi verilerini işleyen ve net çıktı değerlerini oluşturan fonksiyondur. Aktivasyon fonksiyonu girdi ve çıktı birimleri arasındaki eğrisel eşleşmeyi sağlar (Erilli ve ark., 2010). Çeşitli aktivasyon fonksiyonları mevcuttur. Bu aktivasyon fonksiyonlarından bazıları aşağıda verilmiştir.

Düzleştilmiş Doğrusal Birim (Rectified Linear Units) (ReLU))

Gizli katmanlarda yapılan matematiksel işlemler sayesinde doğrusal yapıda olan ağı doğrusal olmayan yapıya dönüştürmek için ReLU aktivasyon fonksiyonu kullanılır (Özkan ve ark., 2017). Negatif sayılarla işlem yapmaz. $f(x)$ = max (0, x)'dir. Yani

girdi değeri; 0'dan küçük ise 0, büyük ise x değerini alarak bağımlı değişken ile doğrusal bir ilişkiye sahip olduğu söylenebilir.

Sigmoid Fonksiyon (Sigmoid Function)

(-∞,∞) aralığındaki bağımsız değişkenleri [0,1] aralığına dönüştüren bir fonksiyondur. Genellikle ikili sınıflandırma problemlerinde tercih edilir ve çıktı tabakasında kullanılır.

Softmax Fonksiyonu (Softmax Function)

Softmax aktivasyon fonksiyonu, Sigmoid aktivasyon fonksiyonuna benzer şekilde her bir sınıfın olasılık değerini temsil ettiği için her bir çıktı değerini 0 ile 1 aralığına dönüştürerek sınıflandırma yapmaktadır. Sigmoid fonksiyonundan farkı çoklu sınıflandırma problemlerinde tercih edilmesidir.

Hiperbolik Tanjant (Hyberbolic Tangent) (tanh(z))

Hiperbolik Tanjant aktivasyon fonksiyonu, ikili sınıflandırma problemleri için kullanılır ve Sigmoid fonksiyonu ile oldukça benzerdir. Ancak temel fark bu fonksiyon [-1,1] değer aralığında değerler alır.

2.11.2. Mimari Yapısı

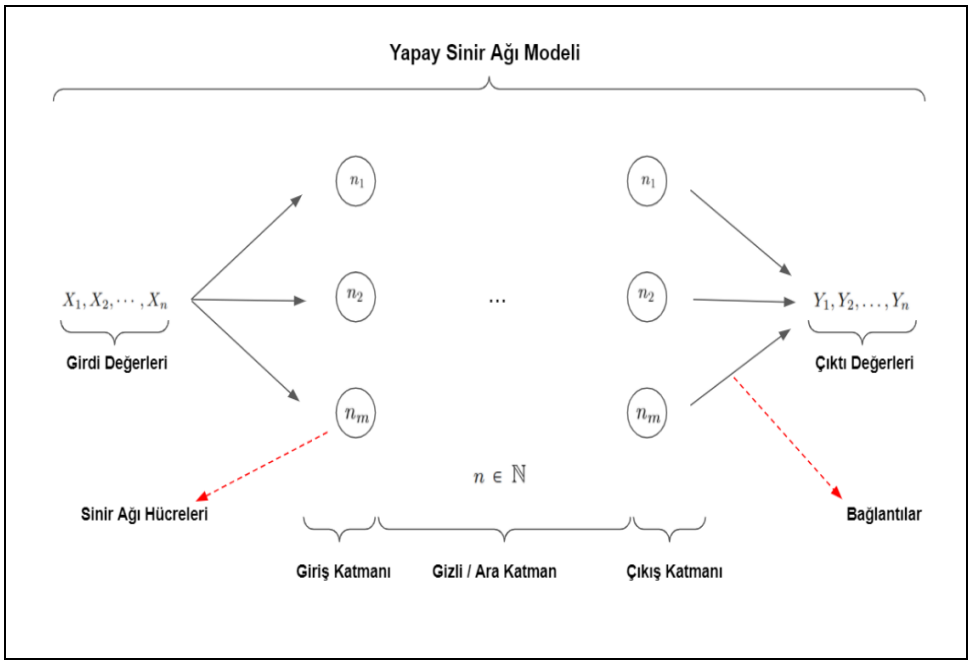
Biyolojik sinir ağlarına benzer şekilde yapay sinir ağlarında da gelen bilgiler, ağa girdi tabakasından iletilir, gizli tabakalarda işlenir ve çıktı tabakasına gönderilir. Üretilen çıktı ağ bağlantıları yardımıyla diğer hücrelere aktarılır. Yapay sinir ağları yapılarına göre iki farklı gruba ayrılır bunlar; İleri Beslemeli Yapay Sinir Ağları ve Geri Beslemeli Yapay Sinir Ağları.

İleri beslemeli yapay sinir ağında nöronlar genellikle girdi, gizli ve çıktı tabakaları olmak üzere üç kısımdan oluşur. Sinyaller, tek yönlü ve ileri doğru şekilde girdi tabakasından çıktı tabakasına iletilir.

Geri beslemeli yapay sinir ağında ise çıktı ve gizli tabakalardaki çıktıların, girdi birimlerine veya önceki gizli katmanlarla geri beslendiği bir ağ yapısıdır. Böylece, sinyaller çift yönlü (hem ileri yönde hem de geri yönde) aktarılmış olur.

Özet olarak Yapay Sinir Ağı modelini genelleyerek temsili olarak görselleştirecek olursak Şekil 2.10' da gösterilmiştir.

Şekil 2.10. İleri Yönlü Beslemeli Yapay Sinir Ağı (Öztürk,2021)



2.11.3. Öğrenme Algoritması

Yapay sinir ağlarında öğrenme, ağırlıkların en iyi değerinin bulunması işlemidir. Öğrenme işlemi, ağırlıkların ve yan değerlerinin güncellenmesi ile modeli en iyi temsil edebilecek değerlerin bulunmasıyla gerçekleşir. Yapay sinir ağlarında en iyi ağırlıkların bulunması bir optimizasyon problemi olarak düşünülebilir ve model eğitiminde kullanılan optimizasyon algoritması da öğrenme algoritması olarak tanımlanır. Bu optimizasyon probleminin çözümü için geliştirilmiş çeşitli öğrenme algoritmaları vardır. Bunlardan en çok tercih edilenleri geri yayılım, gradyan iniş, RMSProp ve Adam öğrenme algoritmalarıdır.

Modelin en iyi parametrelerini bulmak için kullanılan ikiden fazla gizli tabakaya sahip ağ yapısı ***derin öğrenme modeli*** olarak adlandırılır. Son yıllarda derin sinir ağları pek çok veri yapısının (ses, video, metin gibi) işlenebilmesini mümkün kılmakla birlikte oldukça başarılı sonuçlar ortaya koymuştur (LeCun ve ark., 2015).

3. TİP 2 DİYABET TEDAVİSİ İÇİN UYGULAMA

Tezin bu bölümünde Tip-2 Diyabet tedavisi için ilaç adayı olabilecek moleküllerin belirlenmesinde kullanılan makine öğrenmesi algoritmalarının uygulanması ve çıkan sonuçlar hakkında ayrıntılı bilgi sunulmaktadır.

3.1. ChEMBL Veri Tabanı

ChEMBL, genomik bilgiler kullanılarak yeni etkin ilaçların keşfedilmesine yardımcı olmak amacıyla, ilaç benzeri küçük moleküllerin bir araya getirildiği veri tabanıdır. ChEMBL veri tabanı, biyolojik etkinlik, kimyasal yapı ve ilaç hedefleri gibi çeşitli kimyasal ve biyolojik bilgileri içeren kapsamlı bilgiler sağlar. Akademi, endüstri ve bir dizi kimyasal biyoloji alanını kapsayan küresel kullanıcı kitlesi ile yaygın olarak kullanılan kimyasal bir bilgi kaynağıdır. Açık erişim kaynağı olarak kullanılması, araştırmacıların verilere kolay bir şekilde ulaşmasını, verilerin iyileştirilmesi ve standardizasyonunun gerçekleştirilmesine yardımcı olur. Böylece veri tabanındaki veri hacmi büyür, verilerin kalitesi artar ve araştırmacıların güvenilir analizler yapmasına olanak sağlayarak çalışmalarını destekler. Bu tez çalışmasında kullanılan veriler ChEMBL veri tabanından alınmıştır. ChEMBL'ın en son sürümü olan (v32) sürümü kullanılmıştır. Python'da bulunan "chembl_webresource_client" kütüphanesi ile ChEMBL verilerine erişimi mümkündür. Hem erişimi açısından hem güvenilir modeller geliştirmek açısından gerçekleştirilen tez çalışmasında Python programlama dili kullanılmıştır.

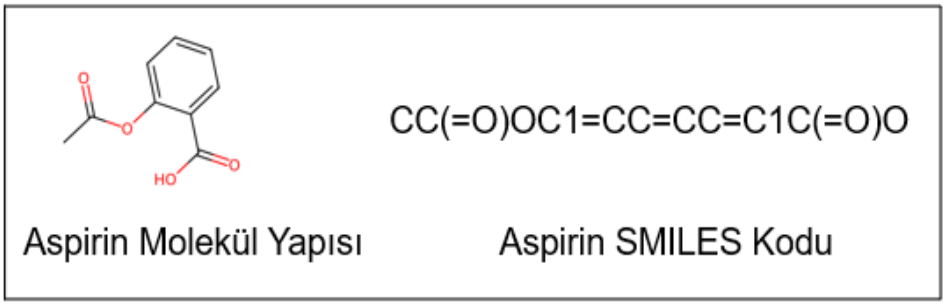
Tez çalışmasında Tip-2 diyabet tedavisi için kimyasal veri tabanından (ChEMBL) elde edilen 5050 bileşikten oluşan dipeptidil-peptidaz IV (DPP-4) inhibitörleri kullanılmıştır. Çeşitli makine öğrenme algoritmaları kullanılarak biyoaktivite tahmini yapılmış ve performansları değerlendirilmiştir. Çalışmada uygulanan adımlar Şekil 3.1'de gösterilmiştir.

Şekil 3.1. Çalışmanın Mimari Yapısı

3.2. Basitleştirilmiş Moleküler Girdi Hattı Giriş Sistemi (SMILES)

Literatürde en yaygın kullanılan notasyon Basitleştirilmiş Moleküler Girdi Hattı Giriş Sistemi (SMILES), ilaç molekül yapılarını (molekül üzerindeki atomları ve bağları niteleyen) ifade etmek için kullanılan belli karakterlerden oluşan bir satır notasyonudur. Tez çalışmasında da SMILES notasyonu kullanılmıştır. Örnek olarak Aspirin molekülüne ait SMILES dizisi Şekil 3.2'de gösterilmiştir.

Şekil 3.2. Aspirin Molekülüne ait SMILES Dizisinin Gösterimi



SMILES dizilerini makine öğrenmesi algoritmalarına girdi olarak verilebilmesi için, SMILES dizileri birer sayısal vektöre dönüştürülerek ilaç parmak izleri çıkartılır (Özcan, 2022).

3.3. Veri Ön İşleme

Veri ön işlememizin amacı ham veri setinden yararlı veriler elde ederek daha iyi sonuçlar elde etmektir. ChEMBL veri tabanından elde edilen veriler IC50 değerlerinden ve Basitleştirilmiş Moleküler Giriş Hattı Giriş Sisteminden Simplified Molecular Input Line Entry System (SMILES) oluşmaktadır. Veriler aşağıda verilen kriterlere göre filtrelenmiştir:

(1) Sadece insana ait DPP-4 inhibitörü seçilir;
(2) Biyolojik aktivite değeri IC50 olan seçilir;
(3) Yinelenen ve SMILES dizi olmayan bileşikler çıkarılır.

IC50 değerleri enzim aktivitesini %50 oranında engelleyebilen inhibitör konsantrasyon ölçümüdür. Kullanılan DPP-4 inhibitörlerinin IC50 değerlerine ilişkin etkinlik sınıfları ve sayıları Tablo 3.1'de gösterilmiştir.

Tablo 3.1. IC50 Değerlerini Sınıflandırma

Değer	Kategori	Sınıf	Bileşik Sayısı
IC50	IC50 < 50	Aktif	1692
IC50	50 < IC50 < 500	Gri	1100
IC50	IC50 > 500	Aktif Olmayan	807

Gri bileşik olarak tanımlanan 1100 bileşik veriden çıkarılmıştır. Kalan 2499 bileşik, regresyon modeli için kullanılmıştır.

IC50 değerlerindeki bu büyük veri aralıklarını daha küçük ve daha anlamlı değerlere dönüştürmek amacıyla ve farklı bileşiklerin inhibisyon potansiyelini daha doğru bir şekilde karşılaştırmak için her bir moleküler yapı için aşağıdaki pIC50 dönüşümü yapılmıştır.

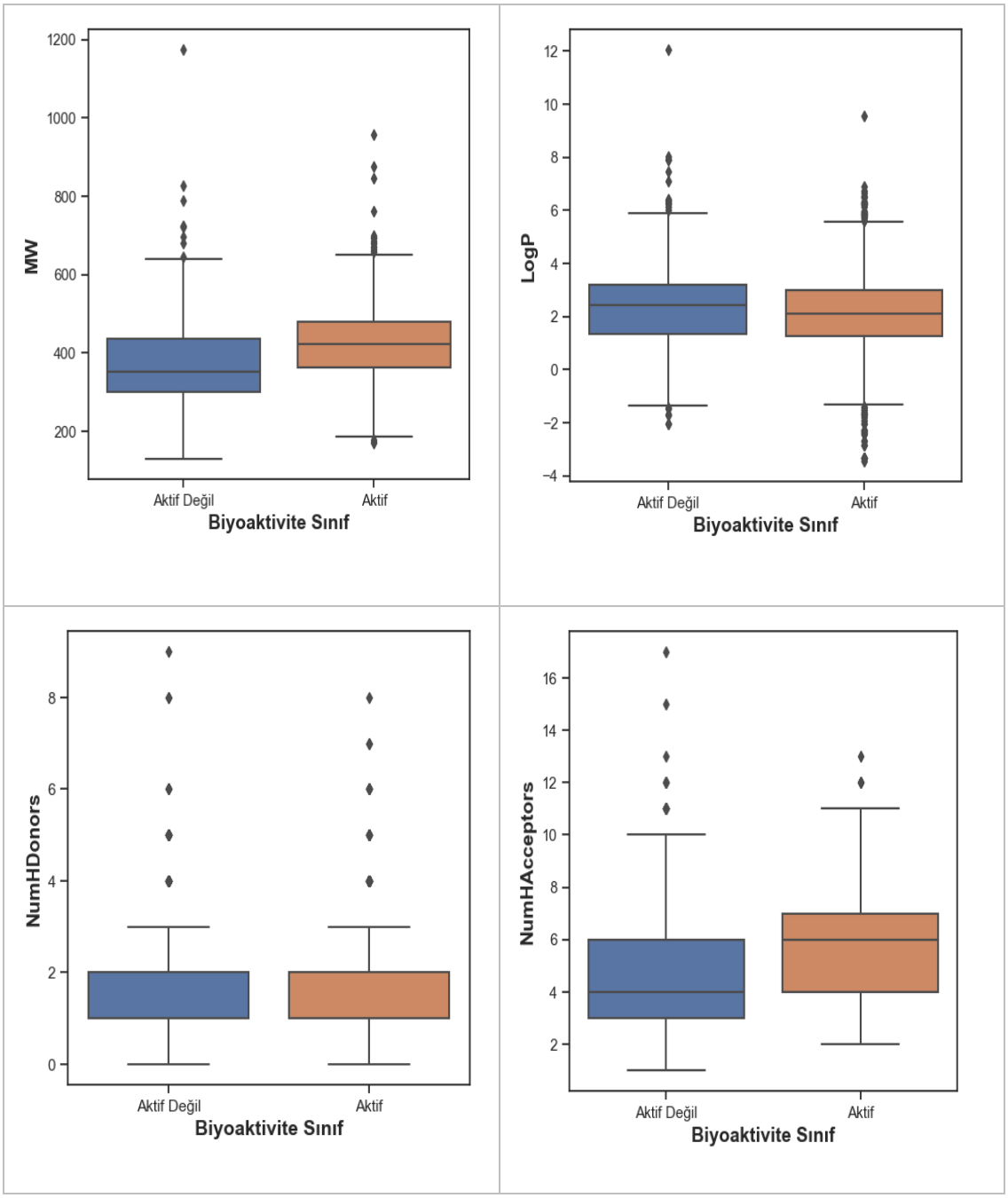
$$pIC50 = -log\ (IC50*10^{-9})$$

Bir molekülün ilaç adayı olabilmesi için aşağıdaki özelliklerin sağlanmış olması gerekmektedir.

1) Moleküler ağırlık (MW)<500 g/mol,
2) Oktanol-su dağılım katsayısının logaritması (LogP)<5,
3) Hidrojen bağı vericisi sayısı<5,
4) Hidrojen bağı alıcı sayısı<10

Bu özellikler Lipsinki'nin "beş kuralı" olarak adlandırılmaktadır. Lipinski'nin bu özelliklerini taşıyan bir molekül bir ilaç adayı olarak görülmesi için gerekli koşulları taşımış olur (Lipinski ve diğ., 2001). Şekil 3.3'te Lipsinki'nin "beş kuralı" kullanılarak DPP-4inibitörlerinin kutu grafiği gösterilmiştir.

Şekil 3.3. Lipsinki'nin "Beş Kuralı" Kullanılarak DPP-4 İnhibitörlerinin Kutu Grafiği

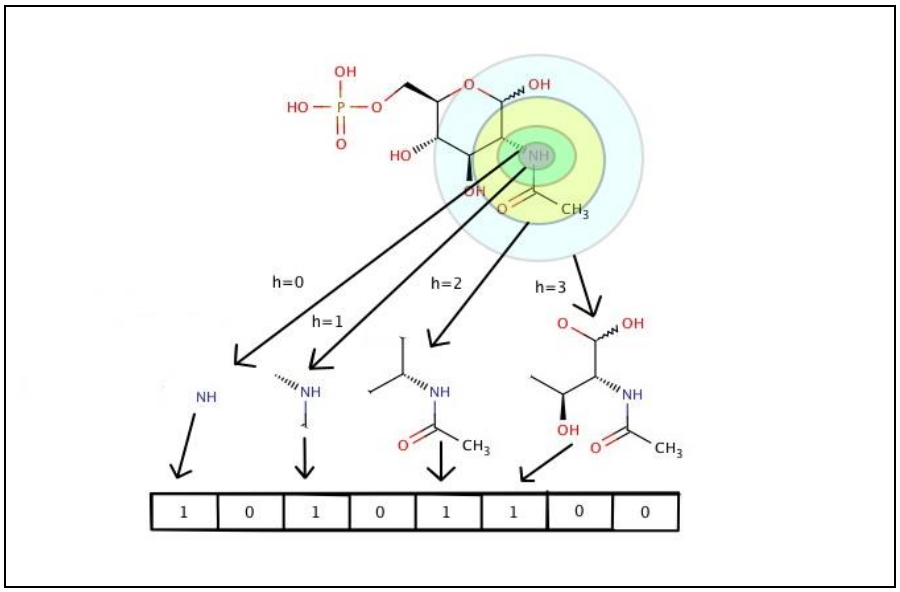


DPP-4 inhibitörleri için aktiflerin aktif olmayanlara kıyasla LogP değerleri daha düşük olduğu ve MW değerleri daha yüksek olduğu gözlemlenmiştir.

3.4. Moleküler Tanımlayıcıların Hesaplanması

Molekülün atomlar ve aralarındaki bağlardan oluştuğu bilinmektedir. Ancak bu bilgileri makine öğrenmesi modeline girdi olarak verilebilmesi için öznitelik vektörüne (molekülün önemli özelliklerini yansıttığı vektör) çevrilmesi gerekir. Moleküllerin yapıları, moleküler temsilleri ve tanımlayıcıları tarafından oluşan bu biyolojik veri (metinsel veri) nicel (nümerik veri) olarak modele girdi olarak verilmelidir. Bir molekülün yapısal özelliklerini temsil etmek için kullanılan matematiksel tanımlayıcı moleküler parmak izi vektörleri olarak tanımlanır. Şekil 3.4'te moleküler parmak izi vektörü oluşturma yöntemi gösterilmiştir.

Şekil 3.4. Moleküller İçin İkili Vektör Oluşturma Yöntemi (Jadhav, 2019)



PaDEL yazılımı, bu moleküler değişkenleri hesaplamak için geliştirilen ve Python'da RDKit kütüphanesi ile kullanılabilen açık kaynak kodlu bir yazılımdır. Bu çalışmada, PaDEL yazılımı kullanılarak 881 özellikten oluşan PubChem parmak izleri kullanılmıştır. Böylece veri sayısal bir forma dönüştürülmüştür.

3.5. Veri Analizi için Kullanılan Kütüphaneler

Çalışmada, tahmin modeli oluşturmak için hangi regresyon modelinin veri setinde daha iyi performans gösterdiğini belirlemek amacıyla Lazy Prediction kütüphanesini kullanılmıştır. Bu kütüphane yardımıyla Python programlama dilinde, Sklearn kütüphanesinde bulunan modellerin tamamı tek satır kod ile çalıştırılır ve modelin değerlendirme metrikleri değerlendirilir. Böylece modelin eğitim aşamasında çeşitli modelleri denemekle zaman harcamadan hızlı bir şekilde tahmin yapmaya olanak sağlar.

Bir yapay sinir ağı modeli oluşturulurken, yapay sinir ağının temel bileşenleri göz önünde bulundurulmalıdır. Sinir ağının model parametrelerini kontrol eden ve model performansını etkileyen ayarlanabilir parametrelere hiperparametreler denir. Yapay sinir ağında kullanılan hiperparametreler, nöron sayısı, tabaka sayısı, aktivasyon türü ve öğrenme oranı gibi değerler örnek olarak verilebilir. Bu hiperparametreler, modelin eğitim verilerini nasıl işleyeceğini, ağırlıkları nasıl güncelleyeceğini ve ne kadar süreyle eğitim yapacağını belirleyerek modelin iyi bir performans göstermesini sağlayabilir. Fakat YSA mimari yapısı açısından kaç tane tabaka kullanılacağı ya da bir tabakada kaç tane nöron bulunacağına karar vermek güçtür. Yanlış ayarlanmış hiperparametreler, modelin eğitiminde sorunlara ve düşük performansa yol açabilir. Bu sebeple bu tez çalışmasında Python da yapay sinir ağı modeli oluşturmak için kullanılan Keras kütüphanesinden faydalanılmıştır. En iyi hiperparametreleri otomatik olarak elde etmek üzere Keras Tuner kütüphanesi kullanılmıştır. Keras Tuner, farklı hiperparametre kombinasyonlarını denemek için bir arama algoritması kullanır. Kullanılan bu arama algoritması, belirli bir hiperparametre uzayında farklı değerleri deneyerek en iyi performansı sağlayan modeli bulmaya çalışır. Yapılan çalışmada Random Search algoritması kullanılmıştır.

3.6. Değerlendirme Metrikleri

Eğitilen makine öğrenimi algoritmalarının test seti performanslarını değerlendirmek ve karşılaştırmak amacıyla 4 farklı ölçüt kullanılmıştır. Kullanılan bu ölçütlerin formülleri ve değişkenlerinin tanımları aşağıda açıklanmıştır.

Aşağıdaki denklemde y; bağımlı değişkeni pIC50 değeri (biyolojik aktivite değeri), x ise; bağımsız değişken moleküler tanımlayıcı olacaktır.

$$y = mx + c$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} xi \qquad i = 1,2,3,\dots,n$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} yi \qquad i = 1,2,3,\dots,n$$

$$\tilde{y}i = Tahmin\ değeri \qquad i = 1,2,3,\dots,n$$

Belirtme Katsayısı (Coefficient of Determination, R^2)

Bağımsız değişkenlerin bağımlı değişkenin yüzde kaçını açıkladığını ifade eden değerdir.

$$R^2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(\tilde{y}_i - \bar{y})^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad \text{veya} \quad R^2 = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad \text{ifade edilir.}$$

$0 < R^2 < 1$ aralığında alır. R^2 değeri 0'a yaklaştıkça değişkenler arasında korelasyonun azaldığını, değer 1'e yaklaştıkça değişkenler arasındaki korelasyonun arttığı anlamına gelir. Kısaca, R^2 değeri 1'e yaklaştığında bağımsız değişkenlerin bağımlı değişkeni iyi açıkladığını söylenebilir.

Ortalama Mutlak Hata (Mean Absolute Error, MAE)

Gerçek değerler ile tahmin edilen değerler arasındaki farkların mutlak değerinin ortalamasıdır.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\tilde{y}i - yi|$$

yi: Gerçek değer

$\tilde{y}i$: Tahmin değeri

n: Toplam veri noktası sayısı

Ortalama Kare Hata (Mean Squared Error, MSE)

Gerçek değerler ile tahmin edilen değerler arasındaki farkın karelerinin toplamının ortalamasıdır.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(yi - \tilde{y}i)^2$$

yi: Gerçek değer

$\tilde{y}i$: Tahmin değeri

n: Toplam veri noktası sayısı

Kök Ortalama Kare Hatası (Root Mean Squared Error, RMSE)

MSE'nin kareköküdür ve MSE' e göre daha yaygın kullanılır.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}i)^2}$$

yi: Gerçek değer

$\tilde{y}i$: Tahmin değeri

n: Toplam veri noktası sayısı

4. BULGULAR, YORUMLAR VE TARTIŞMA

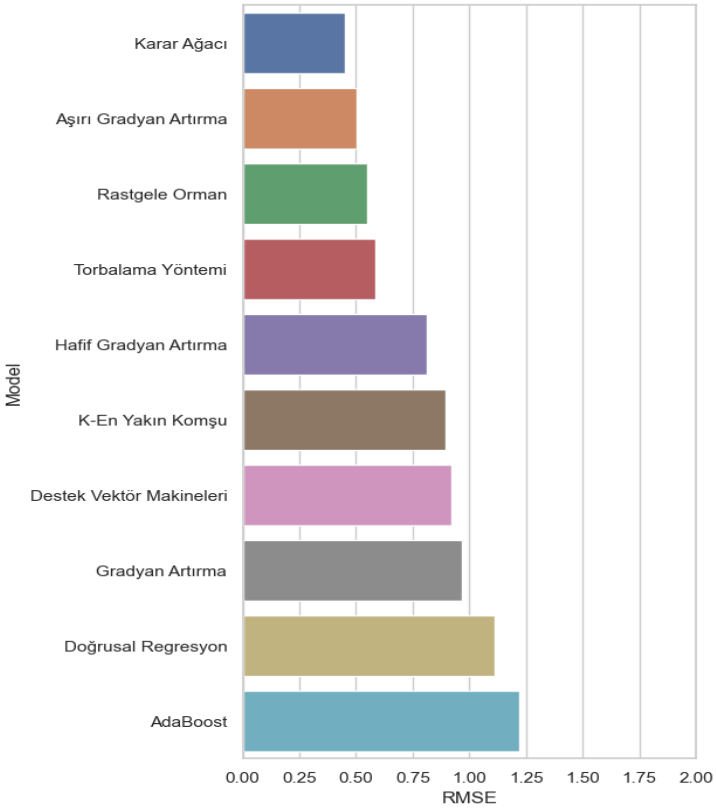
İlk olarak veriler pIC50 değer tahmini yapma amacıyla eğitilmiştir. Bu eğitim için verileri AdaBoost, Bagging, Doğrusal Regresyon, Gradyan Artırma, Karar Ağaçları, Karar Destek Vektörleri, K-En Yakın Komşuluk, LighGBM, Rastgele Orman, XGBoost ve Yapay Sinir Ağları uygulanmıştır. Uygulamada, 881 özellikten oluşan moleküler tanımlayıcıdan düşük varyanslı olan özellikler çıkarılmış ve 132 tane moleküler tanımlayıcı kullanılmıştır. Bir sonraki aşamada verinin %80'lik bir kısmı eğitim, %20'lik kısmı test için ayrılmıştır. Aşağıdaki Tablo 4.1'de analizde kullanılan veriye ait bilgi yer almaktadır.

Tablo 4.1. Analizde Kullanılan Veri

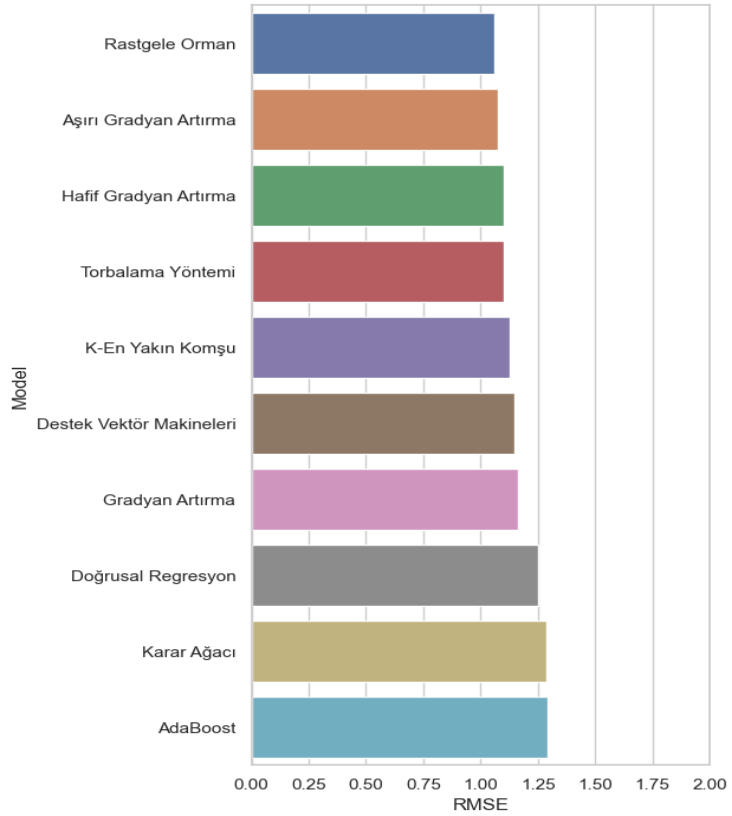
Aktivite	Bileşik Sayısı	Özellik Sayısı	Eğitim Verisi	Test Verisi
DPP-4 İnhibitörleri	2499	132	1999	500

Bu tez çalışmasında, regresyon modeli için Lazy Regressor kullanılarak elde edilen sonuçlar ve literatürde ilaç keşif çalışmalarında sıklıkla kullanılan 10 makine öğrenmesi modelleri seçilmiştir. Bunlar; AdaBoost, Bagging, Doğrusal Regresyon, Gradyan Artırma, Karar Ağaçları, Karar Destek Vektörleri, K-En Yakın Komşuluk, LighGBM, Rastgele Orman, XGBoost makine öğrenme modelleridir. Uygulanan makine öğrenmesi yöntemlerinden eğitim verisindeki pIC50 değerlerini tahmin etmedeki performansları Şekil 4.1' de grafik ile gösterilmiştir. Bu grafikte en başarılı yöntemin eğitim verisi tahmini için Karar Ağaçları Algoritması olduğu görülmektedir. Test verisindeki pIC50 değerlerini tahmin etmedeki performansları ise Şekil 4.2' de grafik ile gösterilmiştir. Bu grafikte ise en başarılı yöntemin eğitim verisi tahmini için Rastgele Orman Algoritması olduğu görülmektedir. Sonuç olarak RF, çok sayıdaki karar ağacının birleşmesiyle ortaya çıkmaktadır ancak karar ağaçlarından temel farkı, farklı alt kümeler oluşturularak karar farklılıkları elde edilir. Gözlem ve değişken bazında rassallık sağlandığı için aşırı öğrenme problemi azalır ve tahmin başarısı artar. Bu sonuçlar doğrultusunda RF algoritmasının hiç görmediği veriyi tahmin etmedeki başarısından dolayı en iyi model olarak seçilir.

Şekil 4.1. Makine Öğrenme Algoritmalarının Eğitim Verisindeki Tahmin Performansları



Şekil 4.2. Makine Öğrenmesi Algoritmalarının Test Verisindeki Tahmin Performansları



Tablo 4.2. Makine Öğrenmesi Algoritmalarının Eğitim ve Test Verisindeki Tahmin Performans Değerleri

Model	Test Verisi	Eğitim Verisi
	RMSE	RMSE
Rastgele Orman	1,06	0,55
Aşırı Gradyan Artırma	1,08	0,50
Hafif Gradyan Artırma	1,10	0,81
Torbalama Yöntemi	1,10	0,58
K-En Yakın Komşu	1,13	0,90
Destek Vektör Makineleri	1,15	0,92
Gradyan Artırma	1,17	0,97
Doğrusal Regresyon	1,25	1,11
Karar Ağacı	1,29	0,45
AdaBoost	1,29	1,22

Tablo 4.2'de modelin hiç görmediği veriler olan test verisinde en düşük RMSE değerine sahip olan Rastgele Orman modelinin performans değeri 1,06 olarak uygulanan diğer makine öğrenmesi modellerinden daha iyi bir performans sergilediği gösterilmektedir. Karar Ağacı modeli eğitim verisinde iyi bir performans göstermesine rağmen test verisinde kötü bir performans sergilediğini aşırı öğrenme yapmış olabileceğini söyleyebiliriz. Bu sebeple hem eğitim verisinde hem test verisinde iyi performans gösteren Rastgele Orman modeli, yapay sinir ağları ile performansını kıyaslayacağımız model olarak seçilmiştir.

Bu tez çalışmasında, Tablo 4.3'teki arama uzayı ve belirlenen arama uzayında rastgele seçilen noktaları kullanarak arama yapan RandomSearch arama algoritması kullanılarak en iyi hiperparametreler seçilmiştir. ReLU aktivasyon fonksiyonu ve Adam optimizasyon algoritması kullanılarak 1 ve 2 gizli tabakalı ileri beslemeli sinir ağı modeli için öğrenme oranları 0,01; 0,001 ve 0,0001 olan, her tabakada 1'den

256’ya kadar nöron denenerek en iyi model seçilmiştir. Arama uzayı Tablo 4.3’te ve elde edilen en iyi 10 model Tablo 4.4’te gösterilmiştir.

Tablo 4.3. Hiperparametre Uzayı

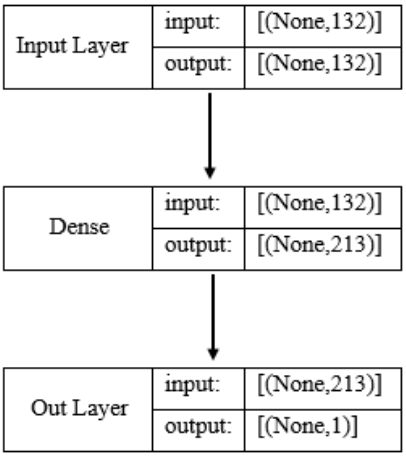
HİPERPARAMETRE	**ARAMA DEĞERLERİ**
Gizli Katman Sayısı	[1-2]
Katmandaki Nöron Sayısı	[1-256]
Öğrenme Katsayısı	[0,01; 0,001; 0,0001]
Aktivasyon Fonksiyonu	ReLU
En İyileme	Adam

Tablo 4.4. Keras Tuner İle Oluşturulan En İyi 10 Model

Model	Gizli Tabaka Sayısı	Gizli Tabakadaki Nöron Sayısı	Öğrenme Oranı	MSE
1	1	(213)	0,001	0,996
2	1	(247)	0,001	1,006
3	1	(228)	0,001	1,008
4	1	(167)	0,001	1,017
5	2	(136,24)	0,001	1,040
6	2	(152,252)	0,001	1,043
7	2	(148,226)	0,001	1,044
8	2	(169,244)	0,0001	1,070
9	1	103	0,01	1,099
10	2	(179,204)	0,001	1,117

Tablo 4.4’ te görüldüğü üzere 1 ve 2 gizli tabakaya sahip modeller arasında çok önemli bir farkın olmadığı ancak model karmaşıklığı açısından ve yapay sinir ağı modelinin performans ölçütü olan MSE değerinin daha düşük olması nedeniyle tek tabakalı ileri beslemeli yapay sinir ağı model seçilmiştir. Keras Tuner ile oluşturulan Yapay Sinir Ağı yapısı; 1 girdi tabakası, 1 çıktı tabakası ve 1 gizli tabaka olmak üzere toplam 3 tabakalı bir yapıya sahiptir. Giriş katmanı 132 adet girdi bilgisini alarak gizli tabakaya iletir. Gizli tabakada 213 adet nöron ve çıktı tabakasında 1 adet çıkış nöronuna sahiptir. Şekil 4.3’te kullanılan örnek bir tek tabakalı ileri beslemeli yapay sinir ağı modeli gösterilmiştir.

Şekil 4.3. Yapay Sinir Ağı Modeli



Tablo 4.5’te eğitilen tek tabakalı ileri beslemeli yapay sinir ağı modelinin eğitim ve test verisindeki performansları gösterilmiştir.

Tablo 4.5. Kullanılan Tek Tabakalı İleri Beslemeli Yapay Sinir Ağı Modelinin Eğitim Ve Test Verisindeki Performans Değerleri

Veri	**MSE**	**RMSE**
Eğitim	0,71	0,84
Test	1,11	1,05

Eğitim verisi ile eğitilmiş Rastgele Orman ve Yapay sinir ağı modellerinin test verisine uygulandığındaki tahmin değerleri ve gerçek değerleri Tablo 4.6’ da gösterilmiştir.

Tablo 4.6. Test Verisi İçin pIC50 Tahmin Değerleri İle Gerçek Değerlerinin Karşılaştırması

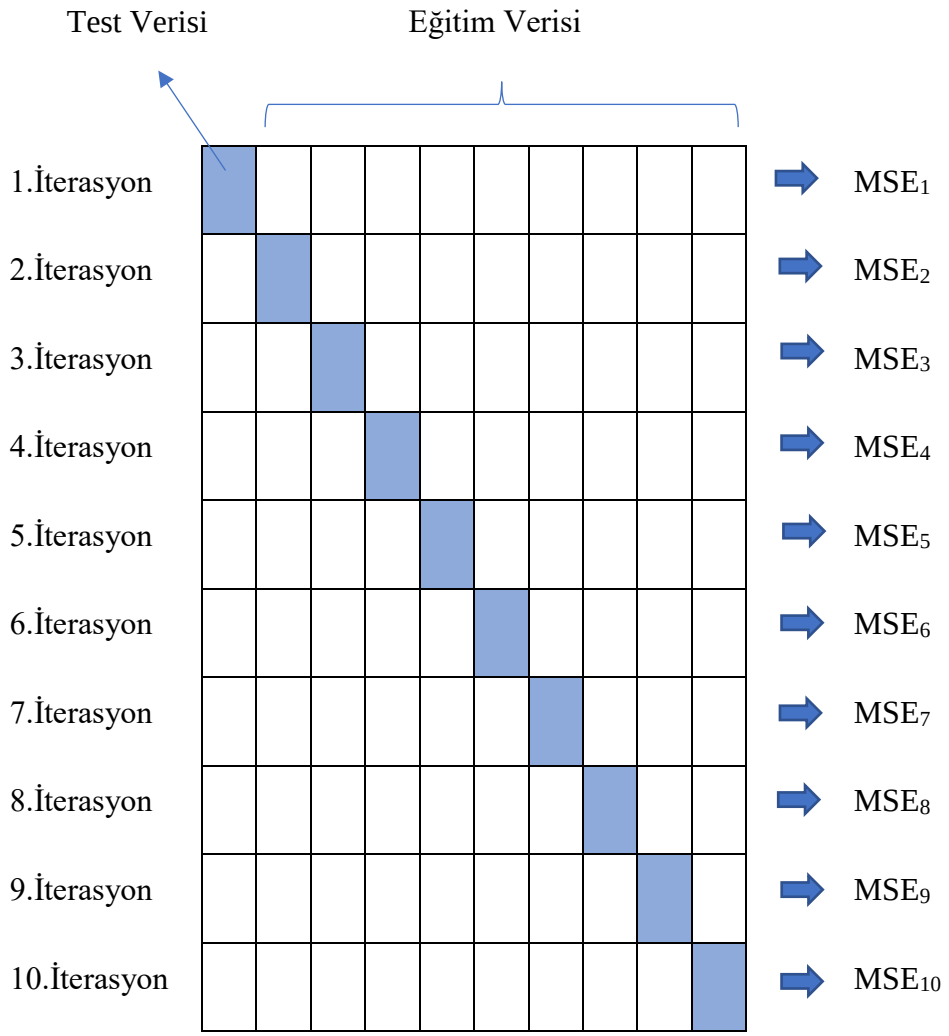
Bileşik No	Rastgele Orman	Yapay Sinir Ağı	pIC50
1	8,09	8,20	7,81
2	7,58	8,74	7,69
3	6,01	6,44	7,49
4	8,55	9,04	7,43
5	7,81	8,02	7,95
6	6,14	6,56	7,76
7	6,18	6,06	7,51
8	6,87	6,65	6,22
9	7,91	7,87	7,92
10	6,43	6,49	8,09
11	5,42	5,64	5,43
12	8,17	8,49	7,31
13	8,32	8,97	8,75
14	7,28	7,95	5,76
15	8,52	8,13	8,36
16	8,03	8,11	8,52
17	8,16	8,55	7,51
18	7,46	7,79	7,41
19	6,52	6,83	5,73
20	4,71	5,10	4,08
RMSE	**1,00**	**1,05**	
MAE	**0,71**	**0,80**	

Sonuç olarak Rastgele Orman modeli en düşük hata oranına sahip olup optimum çözüme en yakın model olarak belirlenmiştir.

Çapraz Doğrulama

Diğer adı Cross-Validation metod, regresyon analizi sonucunda korelasyon denklemindeki doğruluk ve güvenirliği saptamak için kullanılmıştır. Bu tez çalışmasında performansı test etmek amacıyla 10 katlı çapraz doğulama testi yapılmıştır. Veri setini 10 tane parçaya ayırdıktan sonra 9 tanesiyle model eğitilir ve 1 tanesi ile model değerlendirilir. En son çıkan doğrulama skoru 10 tane doğrulama skorunun ortalamasıdır. Şekil 4.4'te çapraz doğrulama için eğitim ve test verisi seçimi gösterilmiştir.

Şekil 4.4. 10 Kat Çapraz Doğrulama İçin Eğitim Ve Test Verisi Seçimi



%80 eğitim ve %20 test şeklinde ayırdığımız veriyi çapraz doğrulama işleminde yine aynı testi farklı eğitim ve test kümeleriyle 10 kez tekrarladık. Çıkan MSE değerlerinin

ortalamasını alarak modelin gerçek performansı hesaplanmıştır. Elde edilen sonuçlar Tablo 4.7'de gösterilmiştir.

Tablo 4.7. Rastgele Orman 10-Kat Çapraz Doğrulama Sonuçları

İterasyon	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Ortalama**
MSE	1,42	1,21	1,33	1,50	1,37	0,74	1,25	1,86	1,19	0,94	1,28

Tablo 4.7'de görüldüğü gibi kurulan Rastgele Orman modelinde 1,28 gibi bir MSE değeri elde edilmiştir. Modelin genelleştirme performansına göre başarılı olduğunu söylenebilir. Bir sonraki aşamada Rastgele Orman modeli tüm veri setine uygulanmıştır. Elde edilen sonuçlar Tablo 4.8'de sunulmuştur.

Tablo 4.8. Tüm Veri Seti İçin Rastgele Orman Değerlendirme Metrikleri

VERİ	**MSE**	**RMSE**
Model	0,32	0,56

Tablo 4.8'e göre, tüm veriye Rastgele Orman modeli uygulandığında RMSE performans ölçütü değeri 0,56 olduğu görülmektedir. Uygulanan makine öğrenme yaklaşımlarından elde edilen sonuçlara göre, Tip 2 Diyabet tedavisinde kullanılan DPP-4 inhibitörlerinin moleküler tanımlayıcıları kullanılarak pIC50 biyoaktivite değerini en iyi tahmin eden modelin Rasgele Orman modeli olduğu gösterilmiştir.

5. SONUÇ VE ÖNERİLER

Başarılı bir ilacın geliştirilebilmesi için pek çok şartın birlikte sağlanması gerekmektedir. Bu nedenle ilaç geliştirme süreci oldukça karmaşık ve zor bir süreç haline gelmektedir. Teknolojinin gelişmesiyle birlikte ilaç adayı moleküllerin, makine öğrenme yaklaşımları kullanarak sanal olarak taranması sağlanmaktadır. Böylece harcanması gereken zaman ve maliyet azaltılmaya çalışılmaktadır.

Bu tez çalışmasında, Tip-2 diyabet hastalığı için ilaç adayı moleküllerinin biyolojik aktivite değerini (pIC50) tahmin etmek amacıyla farklı makine öğrenme yöntemleri kullanılmıştır. İlaç tanımlayıcı özelliğe sahip ilaç parmak izlerinin oluşturulmasında SMILES dizileri kullanılmış ve PubChem parmak izleri ile veri sayısal forma dönüştürülmüştür. Optimum sonuca ulaşabilmek için verilerimize 10 farklı makine öğrenme modeli ve yapay sinir ağı yaklaşımı uygulayarak tahmin modelleri oluşturulmuştur. Makine öğrenmesi modellerinden Rastgele Orman modeli 1,06 RMSE değeri ile en iyi model olarak belirlenmiştir. En iyi parametreleri otomatik olarak belirlenen yapay sinir ağı modeli ve Rastgele Orman modeli eğitim verisi üzerinden eğitilerek, modellerin test verisi için verdikleri tahminler karşılaştırılmıştır. Test verisi performansları; RMSE değeri 1,00 olan Rastgele Orman ve RMSE değeri 1,05 olan yapay sinir ağı modelinin birbirlerine yakın sonuçlar verdiği gözlemlenmiştir. RMSE performans ölçütüne göre, kullanılan modeller arasında en iyi performansa sahip olan Rastgele Orman modeli belirlenmiş ve 10 kat çapraz doğrulama uygulanmıştır. Yapılan çalışmanın sonuçlarına göre, Rastgele Orman yönteminin ilaç keşif çalışmalarında kullanılabilecek iyi bir makine öğrenme yöntemi olduğu gözlemlenmiştir. İleride yapılacak çalışmalar için, mevcut biyoaktivite verilerinin miktarının artması durumunda, bu modellerin performansının artmasının bekleneceği de unutulmamalıdır.

6. KAYNAKLAR

Ağca, Fatma. *Piridoimidazolon Türevlerinin (Braf İnhibitörleri) Elektron Konformasyonel Ec-Ga Metodu İle 4d-Qsar Çalışması*. Yüksek Lisans Tezi, Erciyes Üniversitesi, 2014

Akman, M. (2010). *Veri madenciliğine genel bakış ve Rastgele Ormans yönteminin incelenmesi: sağlık alanında bir uygulama* (Master's thesis, Sağlık Bilimleri Enstitüsü).

Alpaydin, E., *Introduction to Machine Learning*, MIT Press, 2010.

Angermueller, C., Pärnamaa, T., Parts, L., & Stegle, O. (2016). Deep learning for computational biology. Molecular Systems Biology, 12(7), 878.

Cai, J., Li, C., Liu, Z., Du, J., Ye, J., Gu, Q., & Xu, J. (2017). Predicting DPP-IV inhibitors with machine learning approaches. Journal of computer-aided molecular design, 31, 393-402.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

Deng, H., Fannon, D.J., & Eckelman, M.J. (2018). Predictive modeling for US commercial building energy use: A comparison of existing statistical and machine learning algorithms using CBECS microdata. Energy and Buildings, 163, 34-43.

Dhaliwal, S. S., Nahid, A. A., & Abbas, R. (2018). Effective intrusion detection system using XGBoost. *Information*, *9*(7), 149.

Duman, Y. E. (2022, 11 28). ACADEMIA. academia.edu: https://www.academia.edu/36792514/Biyoinformatik_Ve_Ilac_Kesfi adresinden alındı.

Egrioglu, E., Aladag, C. H., Yolcu, U., Uslu, V. R., & Basaran, M. A. (2009). A new approach based on artificial neural networks for high order multivariate fuzzy time series. Expert Systems with Applications, 36(7), 10589-10594.

Erdoğan, S. (2019). Non-contact breathing abnormality detection using machine learning (Master's thesis, Biyo-Medikal Mühendislik Enstitüsü).

Erilli, N. A., Eğrioğlu, E., Yolcu, U., Aladağ, Ç. H., & Uslu, V. R. (2010). TÜRKİYE'DE ENFLASYONUN İLERİ VE GERİ BESLEMELİ YAPAY SİNİR AĞLARININ MELEZ YAKLAŞIMI İLE ÖNGÖRÜSÜ. *Doğuş Üniversitesi Dergisi*, *11*(1), 42-55.

Feng, Z., Xu, C., & Tao, D. (2018, September). Historical Gradient Boosting Machine. In *GCAI* (pp. 68-80).

Guedes, R. A., Serra, P., Salvador, J. A., & Guedes, R. C. (2016). Computational approaches for the discovery of human proteasome inhibitors: An overview. Molecules, 21(7), 927.

Ghamali, M., Chtita, S., Ousaa, A., Elidrissi, B., Bouachrine, M., & Lakhlifi, T. (2017). QSAR analysis of the toxicity of phenols and thiophenols using MLR and ANN. Journal of Taibah University for Science, 11(1), 1-10.

Ghasemi, F., Mehridehnavi, A., Fassihi, A., & Pérez-Sánchez, H. (2018). Deep neural network in QSAR studies using deep belief network. Applied soft computing, 62, 251-258.

Hamzah, H., Bustamam, A., Yanuar, A., & Sarwinda, D. (2020, October). Predicting the molecular structure relationship and the biological activity of dpp-4 inhibitor using deep neural network with Catboost method as feature selection. In 2020 International Conference on Advanced Computer Science and Information Systems (ICACSIS) (pp. 101-108). IEEE.

Hermansyah, O., Bustamam, A., & Yanuar, A. (2021). Virtual screening of dipeptidyl peptidase-4 inhibitors using quantitative structure–activity relationship-based artificial intelligence and molecular docking of hit compounds. Computational Biology and Chemistry, 95, 107597.

Hu, L. Y., Huang, M. W., Ke, S. W., & Tsai, C. F. (2016). The distance function effect on k-nearest neighbor classification for medical datasets. SpringerPlus, 5(1), 1-9.

Hughes, J. P., Rees, S., Kalindjian, S. B., & Philpott, K. L. (2011). Principles of early drug discovery. British journal of pharmacology, 162(6), 1239-1249.

İlya, K. U. Ş., Keser, S. B., & YOLAÇAN, E. (2021). Saldırı tespit sistemlerinde topluluk öğrenme yöntemlerinin kıyaslanması. *Avrupa Bilim ve Teknoloji Dergisi*, (31), 725-734.

Jadhav, A.(2019,February26). *Applications of graph neural networks*. Medium. https-medium-com-aishwaryajadhav-applications-of-graph-neural-networks-1420576be574

KALAYCI, T. E. (2018). Kimlik hırsızı web sitelerinin sınıflandırılması için makine öğrenmesi yöntemlerinin karşılaştırılması. *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi*, *24*(5), 870-878.

KELLE, A. C., & Hüseyin, Y. Ü. C. E. (2022). MQTT Trafiğinde DoS Saldırılarının Makine Öğrenmesi ile Sınıflandırılması ve Modelin SHAP ile Yorumlanması. *Journal of Materials and Mechatronics: A*, *3*(1), 50-62.

Khandelwal, P. (2017). Which algorithm takes the crown: Light GBM vs XGBOOST?. Erişim adresi: https://www.analyticsvidhya.com/blog/2017/06/whichalgorithm-takes-the-crown-light-gbm-vs-xgboost/

KURT, A., BULDU, B., & CEDİMOĞLU, İ. H. XGBOOST VE RASTGELE ORMAN ALGORITMALARININ AĞ TABANLI SALDIRI TESPITINE YÖNELIK PERFORMANSLARININ KARŞILAŞTIRILMASI.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436-444.

Luu, Q. H., Lau, M. F., Ng, S. P., & Chen, T. Y. (2021). Testing multiple linear regression systems with metamorphic testing. Journal of Systems and Software, 182, 111062.

Macalino, S. J. Y., Gosu, V., Hong, S., & Choi, S. (2015). Role of computer-aided drug design in modern drug discovery. Archives of pharmacal research, 38, 1686-1701.

Muratlar, E. R. (2020). XGBoost Nasıl Çalışır? Neden İyi Performans Gösterir. Erişim adresi: https://www.veribilimiokulu.com/xgboost-nasil-calisir/

Olgac A. (2017). *Sanal tarama ve farmakolojik değerlendirme yoluyla yeni 5-lipoksijenaz aktive edici protein (flap) inhibitörü bileşiklerin keşfi.* [Doktora Tezi, Gazi Üniversitesi Sağlık Bilimleri Enstitüsü, Ankara]. Gazi Üniversitesi Akademik Veri Yönetim Sistemi. https://avesis.gazi.edu.tr/yonetilen-tez/865cf59f-1681-4b05-848c-e59aa0daf474/sanal-tarama-ve-farmakolojik-degerlendirme-yoluyla-yeni-5-lipoksijenaz-aktive-edici-protein-flap-inhibitoru-bilesiklerin-kesfi

OLĞAÇ, A., CAROTTI, A., Garscha, U., Werz, O., Macchiarulo, A., & BANOĞLU, E. SANAL TARAMA VE FARMAKOLOJİK DEĞERLENDİRME YOLUYLA YENİ 5-LİPOKSİJENAZ AKTİVE EDİCİ PROTEİN (FLAP) İNHİBİTÖRÜ BİLEŞİKLERİN KEŞFİ.

Özcan, E. N., & Yöşili, S. (2022, 06 23). bioinforange: https://www.bioinforange.com/bioinforeviews/biyoinformatik/ilac-kesfine-biyoenformatik-yaklasim-bilgisayar-destekli-ilac-tasarimi/ adresinden alındı.

Özkan, İ. N. İ. K., & Ülker, E. (2017). Derin öğrenme ve görüntü analizinde kullanılan derin öğrenme modelleri. *Gaziosmanpaşa Bilimsel Araştırma Dergisi*, *6*(3), 85-104.

Öztürk,M. (2021,Ocak). *Python ile Sınıflandırma Analizleri – Yapay Sinir Ağları (YSA).* miracozturk.https://miracozturk.com/python-ile-siniflandirma-analizleri-yapay-sinir-aglari-ysa/

Panov, P., & Džeroski, S. (2007). Combining bagging and random subspaces to create better ensembles. In Advances in Intelligent Data Analysis VII: 7th International Symposium on Intelligent Data Analysis, IDA 2007, Ljubljana, Slovenia, September 6-8, 2007. Proceedings 7 (pp. 118-129). Springer Berlin Heidelberg.

Petterson, I., Balle, T., & Liljefors, T. (2010). Ligand based drug design. Textbook of Drug Design and Discovery, 43-57.

Rezazadeh, A. (2020). A generalized flow for B2B sales predictive modeling: An azure machine-learning approach. *Forecasting*, *2*(3), 267-283.

Rifaioglu, A. S., Atas, H., Martin, M. J., Cetin-Atalay, R., Atalay, V., & Doğan, T. (2019). Recent applications of deep learning and machine intelligence on in silico drug discovery: methods, tools and databases. Briefings in bioinformatics, 20(5), 1878-1912.

Rifaioglu, A. S., Nalbat, E., Atalay, V., Martin, M. J., Cetin-Atalay, R., & Doğan, T. (2020). DEEPScreen: high performance drug–target interaction prediction with convolutional neural networks using 2-D structural compound representations. Chemical science, 11(9), 2531-2557.

Schapire, R. E. (2013). Explaining adaboost. Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik, 37-52.

Simeon, S., Anuwongcharoen, N., Shoombuatong, W., Malik, A. A., Prachayasittikul, V., Wikberg, J. E., & Nantasenamat, C. (2016). Probing the origins of human acetylcholinesterase inhibition via QSAR modeling and molecular docking. PeerJ, 4, e2322.

Ulfa, A., Bustamam, A., Yanuar, A., Amalia, R., & Anki, P. (2021, April). Model QSAR Classification Using Conv1D-LSTM of Dipeptidyl Peptidase-4 Inhibitors. In 2021 International Conference on Artificial Intelligence and Mechatronics Systems (AIMS) (pp. 1-6). IEEE.

van der Kamp M.W., Shaw K.E., Woods C.J., Mulholland A.J. (2008). Biomolecular simulation and modelling: status, progress and prospects. Journal of the Royal Society, Interface / the Royal Society. 5 Suppl 3(July), S173--90.

Wan, F., Zhu, Y., Hu, H., Dai, A., Cai, X., Chen, L., ... & Zeng, J. (2019). DeepCPI: a deep learning-based framework for large-scale in silico drug screening. *Genomics, proteomics & bioinformatics*, *17*(5), 478-495.

Yeşilkanat, C., Kobya, Y., TAŞKIN, H., & Çevik, U. (2014). Yapay Sinir ağları yöntemi ile Artvin ilinde ölçülen gama doz oranlarının ara değer modellemesi ve haritalanması. *Cumhuriyet Üniversitesi Fen Edebiyat Fakültesi Fen Bilimleri Dergisi*, *35*(4), 36-52.

Zurada, J. (1992). *Introduction to artificial neural systems*. West Publishing Co..