# GDSC VERİLERİNİ KULLANARAK YAPAY ÖĞRENME YÖNTEMLERİ İLE AKCİĞER KANSERİ İÇİN HEDEF İLAÇ VE YOLAK TAHMİNİ

Abdullah TERCAN[1], Gıyasettin ÖZCAN[2*]

[1] Bursa Uludağ Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği, Bursa, ORCID No : http://orcid.org/0000-0002-7922-1249

[2] Bursa Uludağ Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği, Bursa, ORCID No : http://orcid.org/0000-0002-1166-5919

**Anahtar Kelimeler**

*Yapay öğrenme, GDSC2 veri kümesi, Hedef ilaç tahmini, Hedef yolak tahmini CTDBase veri kümesi*

**Öz**

*Bu çalışmada literatürde yer alan ve uluslararası alanda öneme sahip olan GDSC veri kümesinde yer alan akciğer kanseri verileri toplanmış, ve bu veriler üzerinde yapay öğrenme yöntemleri kullanarak tahmin yapmak hedeflenmiştir. Bu amaçla ilaç dozunun yarılanma süresine bağlı hedef ilaç ve hedef yolak tahminleri yapılmıştır. Elde edilen bu iki tahminin yine literatürde yer alan CTDBase isimli bir veri kümesinden hastalık tahmini için kullanılması amaçlanmıştır. Böylece ilaçların doz kullanım bilgilerinin hangi hastalıkla ilişkili olabileceği sayısal verilerden tahmin edilmeye çalışılmıştır. Yapılan tahmin işlemi makine öğrenmesi algoritmaları kullanılarak yapılmıştır. Bu süreçte Python programlama dili ile kodlama yapılmış ve bu dilin makine öğrenmesi araçlarından faydalanılmıştır. Her biri on kere tekrarlanan deney sonuçlarına göre oluşturulan makine öğrenmesi modellerinin GDSC veri kümesinde verimli tahmin performansına ulaştığı sonucuna varılmıştır. Özellikle, Light GBM, SVC and kNN algoritmaları analiz edilmiştir. Deney sonuçlarına göre geliştirilen LightGBM ve SVC modellerinin doğruluk oranı %84'ün üstündedir. Bu sonuçlar makine öğrenmesi algoritmalarının kanser ilaç verilerine ait bilinmeyen anlamlı örüntüleri ortaya çıkarma potansiyeli olduğunu göstermektedir.*

# PREDICTION OF TARGET DRUGS AND TRADITIONS FOR LUNG CANCER WITH MACHINE LEARNING METHODS USING GDSC DATA

**Keywords**

*Machine learning, GDSC2 dataset, Lung adenocarcinoma, Drug-target prediction, Target pathway prediction CTDBase dataset*

**Abstract**

*In this study, lung cancer data is collected from literally cited GDSC dataset, and aimed to make predictions on the data using machine learning algorithms. For this purpose, target drug and target pathway estimates were made depending on the half-life of the drug dose. These two predictions are aimed to be used for disease prediction from a dataset called CTDBase, which is also cited in literature. Thus, it can be possible to predict relation between disease and the dose usage information of drugs. The estimation process was made using machine learning algorithms. In this process, coding was done with the Python programming language and its machine learning tools of this language were used. Ten times repeated test results of each experiments denote that our machine learning models achieved efficient prediction performance on GDSC dataset. Particularly, Light GBM, SVC and kNN algorithms were analyzed. Accuracy rates of Light GBM and kNN were no less than 84%. These results show that machine learning algorithms have the potential to reveal unknown significant patterns in cancer drug data.*

Araştırma Makalesi — Research Article

Başvuru Tarihi : 06.02.2023 — Submission Date : 06.02.2023

Kabul Tarihi : 08.05.2023 — Accepted Date : 08.05.2023

* Sorumlu yazar: gozcan@uludag.edu.tr

https://doi.org/10.31796/ogummf.1248489

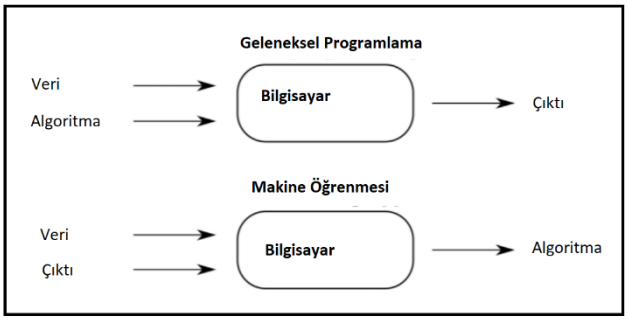
## 1. Giriş

Makine öğrenmesi veriden öğrenerek algoritma geliştirme esasına dayanır. Başka bir deyişle yapılan eylem, veriden öğrenme yaparak bir model oluşturmak ve bu modeli aynı özellikteki yeni verilerin özelliklerinin tahmin edilmesinde kullanmaktır. Bu nedenle klasik algoritma mimarisinden farklıdır.

Şekil 1'de gösterildiği üzere Geleneksel Programlamada bilgisayar aldığı girdi değişkenlerine bağlı olarak önceden tanımlı komutları icra ederek çıktı değerlerini üretir. Algoritma komutları rastgele değişken

kullanmadığı takdirde sonuçları deterministik bir biçimde hesaplar.

Makine öğrenmesi ise girdilerin konumu açısından farklı bir tasarıma sahiptir. Yine şekil 1'deki gösterildiği üzere hem veri, hem de çıktı makine öğrenmesi modeline girdi olarak sunulmaktadır. Bu modelin çıktısı ise bir algoritmadır. Yani algoritmanın nasıl olacağını girdiler ve çıktılar belirlemektedir.




Şekil 1. Geleneksel Programlama ve Makine Öğrenmesi

Makine öğrenmesi algoritmaları 1960'lı yıllardan bu yana birçok alanda veriden tahmin yapabilmek için kullanılmıştır (Alpaydın, 2020). İlk zamanlarda bir sinir hücresinin potansiyel eşdeğerinin benzetimi makine öğrenmesi için araç olarak görülmüştür (McCulloch ve Pitts, 1943).

Makine öğrenmesi yöntemleri istatistiksel yöntemleri temel almaktadır. Örneğin doğrusal regresyon temelli model girdi özellikleri arasında doğrusal bir fonksiyon bulmaya çalışır. Elde edilen fonksiyon katsayıları, yeni girdi sunulduğunda çıktıyı hesaplamanın kolayca yapılmasını sağlamaktadır (Hastie, Tibshirani ve Friedman, 2009).

Makine Öğrenmesi çalışmalarında Destek Vektörleri kullanmak ta alternatif öğrenme olanağı sunmaktadır. Bu yöntemin amacı bağımlı değişkenleri sınıflarına göre ayırabilmek için çok boyutlu düzlemler tanımlamaktır (Noble, 2016).

Literatürde, çok sayıda farklı makine öğrenmesi algoritması mevcuttur. SVM (Boser, Guyon ve Vapnik, 1992), kNN (Fix ve Hodges, 1951), Gradient Boosting (Ke ve diğ., 2017) yakın tarihte kullanılan algoritmalardan üçüdür. Diğer algoritmalara dair kapsamlı bilgi literatürde mevcuttur (Alpaydın, 2020). Algoritmaların doğruluk başarımı oranı verilerin özelliğine göre değişmektedir.

Makine öğrenmesi algoritmalarında bağımlı değişken ayrık ya da sürekli olabilmektedir. Bağımlı değişkenlerin alabileceği sınırlı sayıda ayrık değer olması sınıflandırma problem olarak tanımlanmaktadır. Öte yandan bağımlı değişkenin sürekli olması durumunda yaklaşımlar regresyon olarak çözülmek zorundadır.

En yakın *k* komşu algoritması, kNN, sınıflandırma yapmak amacıyla geliştirilmiştir. Bu algoritmada vektör elemanı olarak tanımlanan veriler, komşularının çoğul oyuna bağlı olarak sınıflandırma yapmaktadır. Başka bir deyişle, en yakın *k* komşusu arasında çoğunluk oyun belirlediği sınıfa atanır.

Algoritmada ideal *k* değeri veriye bağlı olarak değişmektedir. Ancak büyük k değerleri gürültünün etkisini azaltmaktadır. Bu nedenle overfitting ile karşılaşılması büyük *k* değerleri için daha düşük ihtimaldir. kNN algoritmasının optimizasyonu amacıyla geliştirilen yöntemlerden birisi de Komşu bileşen analizidir. Bu amaçla Goldberger, Hinton, Roweis ve Salakhutdinov (2004) kNN sınıflandırmada kullanmak üzere mesafe ölçüm yöntemi geliştirmişlerdir.

Sağlık alanında toplanan verilerin işlenmesi sürecinde Makine öğrenmesi algoritmaları önem arz etmektedir. Yakın geçmişte birçok alanda olduğu gibi akciğer kanseri konusunda elde edilen veriler makine öğrenmesi algoritmaları ile işlenebilmektedir (Huang, Chang, Hsu, Huang, ve Ng, 2016). Elde edilen veri analiz sonuçları önleyici veya tedavi edici araçların ortaya çıkmasına yardımcı olmaktadır (Qureshi ve diğ, 2022).

Kanser ilaç direnci çalışmaları, tedavi edici araçlar geliştirmeyi hedefler. Bu amaçla insan dokusunu oluşturan proteinlerin ilaçlarla etkileşimleri araştırılmaktadır. Bunun yanı sıra sağlıklı ve kanserli hücrelerde proteinlerin etkileşiminin bütününü içeren kimyasal tepkimeler(yolaklar) araştırılmaktadır (Alison ve diğ., 2014). Yolak, hücre içinde meydana gelen bir dizi kimyasal tepkimedir. Hücrenin sağlıklı işleyişi için yolakların uyumlu çalışması gerektiği biyoloji camiasında anlaşılmıştır (Alison, Papachristodoulou, Despo, Elliott ve Elliott, 2014).

Bu amaçla yazarlar GDSC veri tabanı kullanıcılara sunmuştur. Zira gen mutasyonları sonucu ortaya çıkan hastalıkları geri döndürebilmek için etkili ilaçların keşfi gereklidir. Başka bir deyişle ilaçlardaki etken kimyasalların genlerle etkileşiminin en iyi şekilde anlaşılması gereklidir. Bir diğer önemli tanım ise yolaktır.

Bu çalışmada GDSC ilaç direnç veri tabanından (Yang ve diğ.., 2013) akciğer kanseri ilaç veriler toplanmıştır. Toplanan veriler üzerinde makine öğrenmesi modelleri geliştirilerek hedef ilaç ve hedef yolak tahminleri yapılmıştır. Bu iki özelliğin birlikte doğru tahmin edilmesi hastalık ilaç ilişkisinin daha iyi anlaşılmasına yardımcı olabilecektir.

## 2. Sağlıkta Yapay Zeka Bilimsel Yayın Taraması

Gelişen tıbbi cihaz teknolojileri nedeniyle üretilen tıbbi veri miktarında artış gözlenmektedir. Öte yandan hastalıkların ön işaretini belirleyen biomarker yakalama, hastalık teşhisinin bilgisayar yardımı ile doğru tahmin

edilmesi, tedavi için doğru ilaç kombinasyonunun belirlenmesi gibi süreçler büyük veri nedeniyle bilgisayar desteğini zorunlu kılmıştır. Bu gelişim süreci Sağlıkta yapay zekâ konusunda araştırmalarının hızla artmasına neden olmuştur (Yu, Beam ve Kohane, 2018).

Yapay zekâ sağlık biliminin bir çok alanında araç olarak kullanılmaktadır. Literatürde Oftalmoloji (Atwany, Sahyoun ve Yaqub, 2022), Radyoloji (Erickson, Korfiatis, Akkus ve Kline, 2017), Genel Cerrahi, Dermatoloji (Özcan ve Yazici, 2022), Onkoloji (Ali ve Aittokallio, 2019), ve Genetik (Libbrecht ve Noble, 2015) alanlarında çok sayıda çalışma veya inceleme örneği mevcuttur. Elde edilen veriler analiz edilerek hastalık ve tedavi öngörülerinin yapay zekâ ile yapılması hedeflenmektedir.

Literatür incelendiğinde, özellikle görüntü verilerinde derin öğrenme yöntemlerinin tercih edildiği görülmektedir (Shen, Wu ve Suk, 2017). Resim dosyalarının çok boyutlu olması ve resim dosya sayısının çok fazla olması daha verimli bir öğrenme olanağı sağlamaktadır. Derin Öğrenme convolution, max pooling, fully connected ve softmax katmanlarının doğru kombinasyon ve sırada kullanılması yüksek tahmin başarımını sunabilmektedir. (Bengio, 2008). Bu fikirden yola çıkarak (Qiu, Lee, Kim, Yoon, ve Kang, 2021) kanser hücre hat verileri üzerindeki çalışmasında hem derin öğrenme, hem de makine öğrenme algoritmalarını kullanarak tahmin yapmıştır.

Resim ya da video verilerinin aksine, metinsel verilerin işlenmesinde makine öğrenmesi yöntemleri öne çıkabilmektedir. Zira metinsel veriler, derin öğrenme katmanlarında ağırlıkların doğru belirlenmesini sağlayacak kadar büyük miktarda veri içermeyebilir (Tan ve diğ., 2020).

Makine öğrenmesi algoritmaları doğru parametrelerin bilinmesi ya da öngörülmesi durumunda verim sunabilmektedir. SVM (Boser, Guyon ve Vapnik, 1992), kNN (Fix ve Hodges, 1951) ve türevleri sağlıkta yapay zeka alanında kullanılmıştır (Rafique ve diğ., 2021). Öte yandan (Menden ve diğ., 2013) Doğrusal Regresyon, Lasso, yapay sinir ağları ve Rastgele Orman yaklaşımlarını kullanmıştır. Öte yandan Gao ve diğ. (2021) akciğer kanserinde Cisplatin direncini makine öğrenmesi yöntemleri ile tahmin etmiştir. Bu amaçla Destek Vektör Makinelerini kullanmış ve Diferansiyel Gen İfadesi Analizi yapmıştır.

Yakın geçmişte biyolojik alanda büyük miktarda veri açık kaynak olarak ortaya çıkmıştır. Sunulan açık kaynaklı veri tabanları hakkında genel bilgi Özcan ve Yazici (2021) tarafından derlenmiştir.

Kanser alanında sunulan yaygın kullanıma sahip açık kaynaklı temel kaynak TCGA ve Cosmic olarak bilinmektedir. Öte yandan kanserde ilaç direnci konusunda çalışmalar önem arz etmektedir. Bu alanda GDSC (Yang ve diğ., 2013), CTDBase ilaç ve ilaç direnci konusunda bilinen önemli kaynaklardan ikisidir. Her iki veri tabanında yer alan özelliklerin bazılar ortak olmakla birlikte bazı diğer özellikleri farklıdır. Bir veri tabanında elde edilen sonucun diğer veri tabanında girdi olarak sunulma potansiyeli vardır.

Literatürde GDSC ve CTBase veri tabanlarını kullanarak sunulan makine öğrenmesi çalışmaları mevcuttur. Xia ve diğ. (2022) kansere ilaç cevaplarının biyolojik açıdan makine öğrenmesi yöntemleri ile analizini yapmıştır. Öte yandan Tang, Powell ve Gottlieb (2022) ise Moleküler yolaklarda ilaç cevabını tahmin eden bir model sunmuştur. Paltun ve diğ. (2021) ise ilaç kombinasyonlarının etkisini makine öğrenmesi ile araştırılmıştır. Diğer bir çalışma ise ilaç tepkilerini tahmin etmek için Derin Öğrenme modelini sunmuştur (Kuenzi, 2020). Öte yandan Raises ve diğ. (2022) DrugnomeAI framework'u geliştirirken CTbase veri tabanını kullanmıştır.

Sunulan açık kaynak verilerinin Sağlıkta Yapay Zeka konusunda yeni araştırmalara olanak sağladığı anlaşılmaktadır. Öte yandan bu veri kümelerinin boyutlarının büyüklüğü ve karmaşıklığı makine öğrenmesi algoritmalarının kullanımı gereksinimi doğurmaktadır.

## 3. Yöntem

Bu bölümde ilk önce araştırmaya konu olan veri kümesi kaynağı, verinin toplanması ve modellenmesi açıklanacaktır. Ardından verinin makine öğrenme gereksinimleri tanımlanacaktır. Son olarak gereksinim duyulan örüntüleri bulan makine öğrenmesi modeli bulma yöntemi açıklanacaktır.

### 3.1. Veri Kümesi

Yapılan çalışma kapsamında kullanılan veriler GDSC veri tabanının alt kümesinden oluşan GDSC-2 veri setidir. Yapılan çalışma bu iki veri setinin içerdiği bilgileri aynı amaç altında işleyip anlamlı sonuçlar elde etmek üzerinedir.

Çalışmanın amacı GDSC2 veri kümesinden faydalanarak Drug Target ve Target Pathway değerlerini tahmin etmektir. Daha sonra elde edilen tahminlerin CTDBase ya da benzeri veri tabanlarında kullanılma potansiyellerini sunmaktır.

#### 3.1.1 GDSC-2 Veri Kümesi

Yakın zamanda sunulan açık kaynaklı veri kümelerinden birisi Genomics of Drug Sensitivity in Cancer, GDSC'dir (Yang ve diğ., 2013). Bu kaynakta insan kanser hücreleri yüzlerce kimyasal denenerek izlenmiş ve sonuçları sunulmuştur. Sitede ilaç etki verileri ve genomik işaretçiler bulunmaktadır.

GDSC çok sayıda kanser hastalığı için veri barındıran bir sitedir. Bu veri seti içerisinde kanser hastalığı tedavisi için kullanılan ilaçların doz bilgileri bulunmaktadır. Bunun yanı sıra veri seti içerisinde bulunan ilacın hangi gene etki ettiği bilgisi ve ilacın hangi yol ile genlere etki ettiği bilgisi de yer almaktadır. Yapılan bu çalışma kapsamında GDSC-2 veri setinde akciğer kanserinin tedavisine ait ilaçların bulunduğu alt grup kullanılmıştır.

Veri setinde bulunan bu bilgiler ışığında GDSC-2 veri setinin bu çalışmadaki yeri, herhangi bir doz bilgisinin girilmesine karşın makine öğrenmesi algoritmaları yardımıyla girilen doz bilgilerinin hangi gene, hangi yolak ile etki ettiğinin tahmin edilmesidir.

Veri indirme işlemi 28 Ekin 2021'de yapılmıştır. Sitenin Download sekmesinde ANOVA alt sekmesi seçilmiştir. Bu tarih itibarı ile çalışma kapsamında kullanılan GDSC-2 veri seti yaklaşık 2.975 satırdan oluşmaktadır. Sırası önemli olmaksızın kullanılan GDSC-2 veri setinin sütun isimleri şu şekildedir;

- drug_name,
- drug_target,
- target_pathway,
- feature_ic50_t_pval,
- feature_delta_mean_ic50,
- feature_pos_ic50_var,
- feature_neg_ic50_var,
- feature_pval,
- fdr.

Üst tarafta anlatılanların yanı sıra drug_id gibi bilgiler de GDSC-2 veri seti içerisinde bulunmaktadır. Fakat bu çalışma kapsamında bu tip bilgiler işlevsiz görüldüğü için veri setinden çıkartılmıştır.

### 3.1.2 CTDBase Verisi

CTDBase çevresel etkenlerin insan sağlığına etkisini inceleyen bir veri setidir (Davis ve diğ., 2019). Bu veri tabanında kimsayal gen, protein, hastalık ile ilişkilerini içeren veriler bulunmaktadır.

Araştırma amacıyla veri kaynağının sitesine erişilerek bu sitede yer alan CTD_diseases_pathways.tsv dosyası 9 Mart 2022 tarihinde indirilmiştir. Kullanılan CTDBase veri setinde hastalık ismi, hastalığın tedavisi için etki edilmesi gereken gen ve gene nasıl etki edileceği bilgisi bulunmaktadır. Bu çalışma kapsamında kullanılan veri seti 599.032 satırdan oluşmaktadır.

Anlatılan bu bilgilerin CTDBase veri setinde bulunan isimleri aşağıdaki gibidir;

- DiseaseName,
- PathwayName,
- InferenceGeneSymbol.

Yukarıda anlatılanların yanı sıra DiseaseID dahil çok sayıda özelik CTDBase veri seti içerisinde bulunmaktadır. Ancak bu çalışmada bu özellikler kullanılmamıştır.

Kullanılan bu veriler sayesinde GDSC-2 veri seti baz alınarak makine öğrenmesi yardımıyla tahmin edilen drug target ve target pathway bilgilerinin hangi hastalıkların tedavisinde kullanılabileceği tahmin edilmesi amaçlanmıştır.

### 3.2. GDSC-2 Verisinde makine öğrenmesi

Bölüm 3.2'de belirtildiği üzere GDSC-2 veri tabanında yer alan hücre örneklerinde DrugTarget ve TargetPathway ve diğer ilaç özellikleri aynı anda yer almaktadır. Diğer ilaç özelliklerine bağlı olarak DrugTarget ve TargetPathway özelliklerinin öngörülmesi bu çalışmada amaçlanmaktadır. Dolayısıyla GDSC-2 veri tabanı analizinde bu iki özellik bağımlı değişken olarak belirlenmiştir.

Birbiriyle ilişkili olması muhtemel olan DrugTarget ve TargetPathway özelliklerinin bilinmesi önemlidir. Zira DrugTarget ve TargetPathway özelliklerinin bilinmesi durumunda CTBase veri tabanında girdi olacak ve bağımsız değişken olarak kullanılabilecektir. Bu sayede CTDBase veri tabanı kullanılarak gen, ilaç, yolak, hastalık ilişkisi incelenebilecek ve ilacın hangi gene ve hastalığa etki edebileceği konusunda yeni fikirler sunulması mümkün olacaktır.

GDSC-2 Veri tabanından faydalanarak DrugTarget ve TargetPathway tahmini için iki aşamalı bir süreç yürütülmüştür. Birinci aşamada veri kümesinde yer alan TargetPathway değişkeni veri kümesi dışına alınarak makine öğrenmesinin DrugTarget değişkenini tahmini yapması amaçlanmıştır. Daha sonra deneye katılan bağımlı değişkenler yer değiştirilerek diğer bağımlı değişken olan TargetPathway değişkeni tahmin edilmiştir.

Makine öğrenmesi algoritmalarının GDSC-2 veri setinde bağımsız değişken olarak kullandığı özellikler şunlardır:

- feature_ic50_t_pval,
- feature_delta_mean_ic50,
- feature_pos_ic50_var,
- feature_neg_ic50_var,
- feature_pval,
- fdr.

DrugTarget ve TargetPathway değişkenlerinin aynı anda tahmin edilmiş olması bir diğer veri tabanı olan CTDBase verileri için faydalı bir kaynak durumundadır. Zira bu

veri kaynağı ile Gen, Hastalık, TargetPathway ve DrugTarget ilişkisi incelenebilecektir. Bu nedenle DrugTarget ve TargetPathway değerlerinin aynı anda tahmin edilmesi gereklidir ve aynı tabloda yer alması hedeflenmiştir. Elde edilecek bulgu bize tahmini yapılan ilacın hangi hastalıkların tedavisinde kullanılabileceği konusunda bir fikir vermektedir.

### 3.3. Makine Öğrenmesi Algoritmaları

Makine Öğrenmesi algoritmalarını uygulamak için Python programlama dili ve bu dilde tanımlanmış scikit-learn kütüphanesi kullanılmıştır. Bunun yanı sıra Light GBM algoritması için esas kaynağı kullanılmıştır (Ke, 2017). Dil ve kütüphane tercihi yapılırken literatürde en yaygın kullanıma sahip olma koşulu dikkate alınmıştır.

GDSC2 veri kümesinde DrugTarget ve TargetPathway özelliklerini tahmin etmek için farklı makine öğrenmesi algoritmaları denenmesi gerektiğine kanaat getirilmiştir. Denenen algoritmalardan en yüksek doğruluk oranına sahip olan model aranmıştır. Denenen algoritmalar şunlardır:

- Support Vector Classification,

- LightGBM Gradient Boosting

- K Nearest Neighbour.

Kullanılan algoritmalar hazır kütüphane yardımıyla çalışma kapsamında denenmiştir. Bu algoritmalar target pathway değişkeninin tahmini için öncelikli olarak denenmiştir.

### 3.4. Deneylerin icra akışı

Deneylerin icra akışı Şekil 2'de gösterilmiştir. Makine öğrenmesi Uygulama aşamasında GDSC-2 verilerinin rastgele belirlenen %80'i eğitim için ayrılmıştır. Kalan % 20 veri ise test için kullanılmıştır. Test verilerinin bağımlı değişkenleri makine öğrenmesi yöntemleri ile kıyaslanarak doğruluk oranları belirlenmiştir.

Her makine öğrenmesi yöntemi için uygulama 10 kez çalıştırılmıştır Her program çalışmasında GDCS_2 veri setinde eğitim-test için rastgele seçimler farklı olacaktır. Elde edilen 10 makine öğrenmesi model çıktısının doğruluk ortalaması o deneyin nihai performansı olarak belirlenmiştir. Böylece makine öğrenmesi modellerinde bias ve aşırı öğrenmeye karşı önlem planlanmıştır.

Her makine öğrenmesi modeli ile tahmin edilen bağımlı değişkenler DrugTarget ve TargetPathway olarak tanımlanmıştır. DrugTarget tahmin edilirken adaletli olması için makine öğrenmesi TargetPathway değişken değerini dikkate almamıştır. Sonraki deneyde ise TargetPathway tahmin edilirken DrugTarget değişken değeri dikkate alınmamıştır.

Makine öğrenme model oluşturma ve test tahminin 10 deney modeli ortalamasının hesaplanması ve her algoritmanın iki farklı tahminde bulunması nedeniyle her makine öğrenmesi algoritması toplamda 20 kez çalıştırılmıştır. Üç farklı makine öğrenmesi algoritması için toplamda 60 program icrası, model oluşumu mevcuttur.

Model oluşturmak için algoritmaların optimum parametreleri araştırılmıştır. En iyi sonucu veren makine öğrenmesi modelleri için bulunan parametreler Tablo 1'te sunulmuştur.

Tablo 1.Makine Öğrenme Algoritmaları ve Parametreleri

| kNN | SVC | LightGbm |
|---|---|---|
| n_neighbors = 1 | c = default (1.0) | learning_rate=default ( 0.1) |
| weights = default ('uniform') | kernel = default ('rbf') | num_leaves=default ( 31) |
| algorithm = default ('auto') | degree = default (3) | early_stopping_rounds= default( 0 ) |
| leaf_size = default (30) | gamma = 'auto' | num_iterations= default (100) |
| p = default (2) | coef0 = default (0.0) | |
| metric = default ('minkowski') | | |

### 4. Bulgular

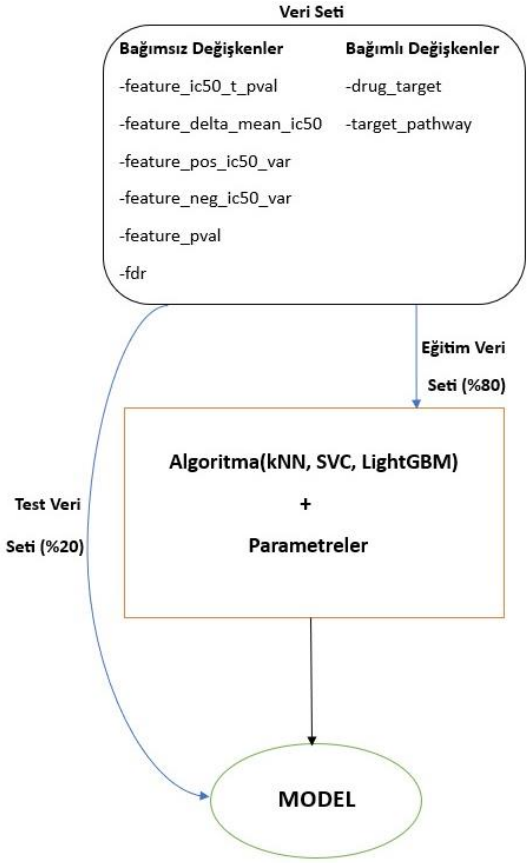
Çalışmada üç farklı algoritma için ayrı ayrı deney sonuçları elde edilmiştir. Bunun yanı sıra kNN algoritmasında tercih edilmesi gereken ideal *k* değeri araştırılmıştır. Ön deney çalışmalarında ideal k değerinin 1 olduğu sonucuna varılmıştır.

GDSC2 veri kümesinden iki adet bağımlı özellik tahmini yapılması amaçlanmıştır. Bunlar DrugTarget ve TargetPathway özellikleridir. Bu nedenle algoritmalar her özellik için ayrı ayrı çalıştırılmıştır. CTDBase veri kümesine doğru girdi sağlanabilmesi için her iki bağımlı değişkenin de yüksek doğruluk oranına sahip olması hedeflenmiştir.

Algoritmaların doğruluk değerini belirlemek için aşağıdaki yöntem kullanılmaktadır:

Veri kümesinde yer alan satırların % 80'I eğitim, kalan %20'si test amacıyla kullanılmıştır. Eğitim kümesindeki veriler makine öğrenmesi uygulamasına girdi olarak verilmiştir.

Makine öğrenmesi tahminleri deneyle edildiğinde doğruluk performansı ölçümü gereklidir. Çalışmada doğru tahmin sayısının tüm tahmin sayısına oranı analiz için kullanılmıştır. Modelin ürettiği küme tahminlerin test veri kümesinde yer alan küme ile uyumlu olması doğru tahmin, aksi takdirde yanlış tahmin olarak tanımlanmıştır. Plan doğrultusunda deneyler ve bulgular aşağıda açıklanmıştır.




Şekil 2. GDSC-2 Model İş akışı

Birinci deneyde üç algoritmanın Hedef Yolak Tahmin değişkeni için doğruluk oranı hesaplanmıştır. Tablo 2'de gösterilen bulgulara göre sırasıyla kNN ve LightGBM modelleri tatminkar sonuç vermiştir:

Tablo 2. Hedef Yolak Tahmin ortalaması

| Algoritma | Doğruluk Oranı (%) |
|---|---|
| SVC | 51.27 |
| LightGBM | 86.20 |
| kNN | 89.96 |

İkinci deneyde ise Algoritmalar İlaç Tahmin deneylerine tabi tutulmuştur. Tablo 3'de ise deney sonuçları sunulmuştur. İkinci deneyde de kNN ve LightGBM model ortalamaları başarılı sonuç vermiştir. Burada belirtmek gerekir ki, kNN ve LightGBM modeli performansını arttırmak için bir çok parametre denenmiştir.

Tablo 3. İlaç Tahmin ortalaması

| Algoritma | Doğruluk Oranı (%) |
|---|---|
| SVC | 55.24 |
| LightGBM | 84.60 |
| kNN | 88.42 |

Yapılan bu çalışmalar sonucunda target pathway değişkeninin tahmininde hem LightGBN hem de kNN model tahminlerinin çok daha iyi sonuç verdiği görülmüştür. Bu deney esnasında *k*'nın 1 olan değeri test edilmiştir.

Çalışmada kNN algoritmasının başarılı sonuçlar vermesi, veri kümesinde yer alan özelliklerin kendi aralarında uyumlu özelliklere sahip olduğu yargısını ortaya çıkarmaktadır. Zira boyut sayısı arttığında SVC'in doğruluk performansının da artması beklenir. Öte yandan 6 boyutlu ve 2975 vektör elemanı olan veri boyutu kNN'in daha iyi sonuç vermesi ile sonuçlanmıştır.

Tablo 3'te elde edilen sonuçlara göre vektör elemanları arasında ideal kümeleme yapılırken sadece en yakın bir komşuyu hesaba katmak daha verimli örüntü yakalamaktadır. Bu durumda her vektör elemanının özgün karakterini dikkate almak gerektiği kanaatine varılmaktadır.

kNN algoritmasının yanı sıra LightGBM algoritmasını baz alan model çalışmaları da kNN modellerine yakın performanslar elde etmiştir. Gradient Boosting temelli bir algoritma olan LightGBM literatürde iyi performans göstermektedir (Ke ve diğ., 2017). Hem karar ağaçlarını hem de zayıf öğrenme yaklaşımını kullanmaktadır. Öte yandan algoritmanın başarısında öğrenme oranı, ağaç derinliği ve yaprak sayısı parametrelerinin doğru belirlenmesi gerekmektedir.

## 4. Sonuçlar

Kanserin oluşum süreci, genetik faktörlerin kansere etki şekli, ilaçların etki düzeyi, ilaçların hassasiyet düzeyi gibi konular yakın zamanda araştırmacıların ilgisini çekmektedir. Bunun sonucunda açık kaynaklı veri tabanları ortaya çıkmış ve araştırmacıların kullanımına sunulmuştur. CTDBase ve GDSC açık kaynaklı veri setlerinden ikisini oluşturmaktadır.

Gen-hastalık-ilaç üçlüsü arasındaki etkileşiminin iyi şekilde anlaşılması kanser tedavisinde doğru ilaç ve dozaj kullanımına destek sağlayacaktır. İlaç dirençlerinin ve yolaklarının daha iyi anlaşılabilmesi için açık kaynak veri sunulan GDSC, makine öğrenmesi çalışmaları için değerli kaynak sunmaktadır.

Hastalık-Yolak verilerinin doğru öngörülmesi koşulu ile ilgili genin doğru tahmin edilmesi hem ilaç geliştirme hem de ilaç yeniden konumlandırma konu başlıklarında önem arz etmektedir. Bu amaçlar doğrultusunda ilaç özelliklerinin tahmini için makine öğrenmesi algoritmaları üzerinde çalışılmıştır.

Çalışmanın gelişim sürecinde yer alan ilk ana konu GDSC veri tabanından akciğer kanser verilerini toplamaktır. Elde edilen veriler üzerinde Hedef ilaç ve hedef yolak tahminlerinin diğer özelliklere göre belirlenmesidir. Çalışmada GDSC veri setinde yer alan hedef ilaç, ilaç yolağının doğru tahmin edilebilmesi için makine

öğrenmesi tabanlı yöntemler araştırılmıştır. Bu özelliklerin doğru tahmin edilmesi durumunda CTDBase veri seti kullanılarak İlişkili hastalığın belirlenmesine yardımcı olacaktır. Böylece ilaçlara ait bazı özellikler ile hastalık ilişkisi daha iyi anlaşılabilecektir.

Problemin çözümü için farklı makine öğrenmesi algoritmaları denenmiştir. Rastgeleliği en aza indirebilmek için deneyler 10'ar kez tekrar edilmiştir. Süreç içerisinde Hedef ilaç ve yolak tahminlerini doğru biçimde tahmin eden doğru parametreler araştırılmıştır.

Geliştirilen modeller arasında hem LightGBM hem de kNN doğru tahmin performansları sunmuştur. Gen-hastalık-ilaç etkileşiminin daha iyi anlaşılması hastalara doğru ilacın doğru zamanda ve doğru dozajda belirlenmesi çalışmasına katkıda bulunması konusunda önem arz etmektedir.

Sunulan açık kaynaklı genomik veriler geçen yirmi yılda üstel olarak artmıştır. Bu durum, veriden öğrenme olanaklarının artmasına yardımcı olmaktadır. Bunun sonucunda gen-hastalık-ilaç ilişkisini daha iyi tanımlayan makine öğrenmesi yöntemlerinin gelişmesine yardımcı olacaktır.

## Etik standartlara uygunluk

Bu çalışmada kullanılan verilerin tamamı açık kaynaklıdır ve bilimsel dergilerde yayımlanmıştır. İlgili açık veri kaynakları atıflanmış ve Referanslara eklenmiştir.

## Araştırmacıların Katkısı

Bu çalışmada; Abdullah TERCAN, kodların hazırlanması, veri toplama ve etiketleme, kavramsal tasarım, bulguların elde edilmesi, değerlendirilmesi, makalenin yazılması; Gıyasettin ÖZCAN literatür tarama, kavramsal tasarım, makalenin yazılması, araştırmanın bilimsel danışmanlığının yürütülmesi.

## Çıkar Çatışması

Yazarlar tarafından herhangi bir çıkar çatışması beyan edilmemiştir. Bu çalışma Abdullah Tercan'ın, Doç. Dr. Gıyasettin ÖZCAN danışmanlığında tasarladığı lisans bitirme projesinin geliştirilmiş halidir.

## Kaynaklar


Ali, M., & Aittokallio, T. (2019). Machine learning and feature selection for drug response prediction in precision oncology applications. *Biophysical reviews*, 11(1), 31-39.

Alison, S., Papachristodoulou, D.K., Despo, K., Elliott, W.H., & Elliott, D.C. (2014). *Biochemistry and molecular biology* (Fifth ed.). Oxford. ISBN 978-0-19-960949-9. OCLC 862091499.

Alpaydin, E. (2020) Introduction to machine learning. 4th ed. MİT press.

Atwany, M. Z., Sahyoun, A. H., & Yaqub, M. (2022). Deep learning techniques for diabetic retinopathy classification: A survey. *IEEE Access*. , 10, 28642-28655.

Bengio, Y. (2008) Learning deep architectures for AI. *Foundations and Trends in Machine Learning* 2(1): 1–127.

Boser, B.E., Guyon, I.M. & Vapnik, V. (1992). "A training algorithm for optimal margin classifiers". Proceedings of the fifth annual workshop on Computational learning theory – COLT '92. p. 144.

Brent M. K., Park J., Fong, S.H., Sanchez, K.S., Lee, J., Kreisberg, J.F., Jianzhu, M., & Ideker, T. (2020). Predicting Drug Response and Synergy Using a Deep Learning Model of Human Cancer Cells. *Cancer Cell*, Volume 38, Issue 5, Pages 672-684.e6, ISSN 1535-6108, https://doi.org/10.1016/j.ccell.2020.09.014.

Callahan, A., & Shah, N. H. (2017). Machine learning in healthcare. In Key Advances in Clinical Informatics (pp. 279-291). Academic Press.

Davis, A. P., Grondin, C. J., Johnson, R. J., Sciaky, D., McMorran, R., Wiegers, J., ... & Mattingly, C. J. (2019). The comparative toxicogenomics database: update 2019. Nucleic acids research, 47(D1), D948-D954.

Erickson, B. J., Korfiatis, P., Akkus, Z., & Kline, T. L. (2017). Machine learning for medical imaging. *Radiographics*, 37(2), 505-515. https://doi.org/10.1148/rg.2017160130

Fix, E., & Hodges, J. L. (1951). Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties (PDF Report). *USAF School of Aviation Medicine, Randolph Field, Texas*.

Gao, Y., Lyu, Q., Luo, P., Li, M., Zhou, R., Zhang, J., & Lyu, Q. (2021). Applications of Machine Learning to Predict Cisplatin Resistance in Lung Cancer. *International Journal of General Medicine*, 14, 5911.

Goldberger, J., Hinton, G. E., Roweis, S., & Salakhutdinov, R. R. (2004). Neighbourhood components analysis. *Advances in neural information processing systems*, 17.

Grossman,, R.L., Heath, A.P., Ferretti, V., Varmus, H.E., Lowy, D.R., Kibbe, W.A., & Staudt, L.M., (2016). Toward a Shared Vision for Cancer Genomic Data. *N. Engl. J. Med.,* 375, 1109–1112.

Hamilton, D., Pacheco, R., Myers, B., & Peltzer, B. (2020). kNN vs. SVM: A comparison of algorithms. In: Hood, Sharon M.; Drury, Stacy; Steelman, Toddi; Steffens, Ron, eds. . *Proceedings of the Fire Continuum-Preparing for the future of wildland fire*.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2, pp. 1-758). New York: Springer.

Huang, C. H., Chang, P. M. H., Hsu, C. W., Huang, C. Y. F., & Ng, K. L. (2016). Drug repositioning for non-small cell lung cancer by using machine learning algorithms and topological graph theory. *BMC bioinformatics* (Vol. 17, No. 1, pp. 13-26). BioMed Central.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30.

Libbrecht, M. W., & Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, 16(6), 321-332.

Kuenzi, B.M., Park, J., Fong, S.H., Sanchez, K.S., Lee, J., Kreisberg, J.F., et al. (2020). Predicting Drug Response and Synergy Using a Deep Learning Model of Human Cancer Cells. *Cancer Cell*, 38:672–84.

McCulloch, & W., Pitts, W. (1943). A Logical Calculus of Ideas Immanent in Nervous Activity. *Bulletin of Mathematical Biophysics*, 5: 115–133

Menden, M. P., Iorio, F., Garnett, M., McDermott, U., Benes, & C. H., Ballester, P. J., & Saez-Rodriguez, J. (2013). Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. *PLoS one*, 8(4), e61318.

Noble, W. S. (2006). What is a support vector machine. *Nature biotechnology*, 24(12), 1565-1567.

Özcan G, ve Yazici S. (2022). Açık Erişimli veri kaynakları ve veri analizi. Türsen Ü, editör. *Dermatolojide Yapay Zekâ.* 1. Baskı. Ankara: Türkiye Klinikleri. p.9-15.

Paltun, B.G., Kaski, S., & Mamitsuka, H., (2021). Machine learning approaches for drug combination therapies, *Briefings in Bioinformatics*, Volume 22, Issue 6, November, https://doi.org/10.1093/bib/bbab293

Rafique, R., Islam, S. R., & Kazi, J. U. (2021). Machine learning in the prediction of cancer therapy. *Computational and Structural Biotechnology Journal*, 19, 4003-4017.

Raies, A., Tulodziecka, E., Stainer, J., Middleton, L., Dhindsa, R. S., Hill, P., ... & Vitsios, D. (2022). DrugnomeAI is an ensemble machine-learning framework for predicting druggability of candidate drug targets. *Communications Biology*, 5(1), 1291.

Qiu, K., Lee, J., Kim, H., Yoon, S., & Kang, K. (2021). Machine learning based anti-cancer drug response prediction and search for predictor genes using cancer cell line gene expression. *Genomics & informatics,* 19(1).

Qureshi, R., Basit, S. A., Shamsi, J. A., Fan, X., Nawaz, M., Yan, H., & Alam, T. (2022). Machine learning based personalized drug response prediction for lung cancer patients. *Scientific Reports*, 12(1), 18935.

Shen, D., Wu, G., & Suk, H. I. (2017). Deep learning in medical image analysis. *Annual review of biomedical engineering,* 19, 221.

Tan, X., Yu, Y., Duan, K., Zhang, J., Sun, P., & Sun, H. (2020). Current advances and limitations of deep learning in anticancer drug sensitivity prediction. *Current Topics in Medicinal Chemistry*, 20(21), 1858-1867.

Tang, Y.C., Powell, R.T. & Gottlieb, A. (2022). Molecular pathways enhance drug response prediction using transfer learning from cell lines to tumors and patient-derived xenografts. *Sci Rep,* 16109. https://doi.org/10.1038/s41598-022-20646-1

Tate J.G., Bamford, S., Jubb, H.C., Sondka. Z., Beare, D.M., Bindal. N., et al. (2019). COSMIC: the Catalogue of Somatic Mutations ın Cancer. *Nucleic Acids Research*, 47(D1):D941-D7. doi: 10.1093/nar/gky1015.

Xia, F., Allen, J., Balaprakash, P., Brettin, T., Garcia-Cardona, C., Clyde, A., ... & Stevens, R. (2022). A cross-study analysis of drug response prediction in cancer cell lines. *Briefings in bioinformatics*, 23(1), bbab356.

Yang, W., Soares, J., Greninger, P., Edelman, E.J., Lightfoot, H., Forbes, S., et al. (2013) Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res*. 41 (Database issue):D955-61

Yu, K. H., Beam, A. L., & Kohane, I. S. (2018). Artificial intelligence in healthcare. *Nature biomedical engineering*, 2(10), 719-731.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* 67(2), 301-320.