Abbasoğulları, E. G., Günay, F. B. (2024). Talasemi Hastalığı Tahmini İçin Farklı Makine Öğrenmesi Yöntemlerinin Kullanılması ve Karşılaştırılması. *Karadeniz Fen Bilimleri Dergisi*, 14(4), 1990-2007.

Karadeniz Fen Bilimleri Dergisi, 14(4), 1990-2007, 2024. DOI: 10.31466/kfbd.1512278

KFBD

Karadeniz Fen Bilimleri Dergisi
The Black Sea Journal of Sciences
ISSN (Online): 2564-7377 https://dergipark.org.tr/tr/pub/kfbd

Araştırma Makalesi / Research Article

Talasemi Hastalığı Tahmini İçin Farklı Makine Öğrenmesi Yöntemlerinin Kullanılması ve Karşılaştırılması

Ece Gülşah ABBASOĞULLARI[1*], Faruk Baturalp GÜNAY[2]

Öz

Talasemi, insan vücudunda az miktarda hemoglobin ve kırmızı kan hücresine neden olan kalıtsal bir hastalıktır. Bu hastalık tedavi edilemediği gibi bazı hastalarda ömür boyu kan nakli gerektirmektedir. Hastalığın erken teşhis edilmesi büyük önem taşımaktadır. Çalışmanın amacı makine öğrenmesi sınıflandırma yöntemleri kullanarak talasemi hastalığı tahmini yapmaktır. Çalışmada kullanılan veriler Erzurum Atatürk Üniversitesi Araştırma Hastanesine gelen hastalardan oluşmaktadır. Çalışma, python dili ile Jupyter Notebook ortamında sınıflandırma yöntemleri kullanılarak gerçekleştirilmiştir. Çalışmada, Naive Bayes (NB), K-En Yakın Komşu (KNN), Destek Vektör Makineleri (SVM), Lojistik Regresyon (LR), Rastgele Orman (RF) ve Karar Ağaçları (DT) gibi farklı sınıflandırma yöntemlerin karşılaştırılması yapılmıştır. Bu sınıflandırma yöntemleri kullanılarak en iyi tahmin sonucuna ulaşmaya çalışılmıştır. Veri seti %70 eğitim ve %30 test aşamasında kullanmak için ayrılmıştır. Bu aşamalarda oluşan sapmaların önüne geçmek için k kat çapraz doğrulama (k fold cross validation) yöntemi uygulanmıştır. Sınıflandırma yöntemlerinin performans değerlendirmesinde kesinlik (precision), duyarlılık (recall), f1-skoru (f1 score), doğruluk (accuracy), işlem karakteristik eğrisi (ROC-AUC), log loss (logaritmik kayıp) gibi performans metriklerine bakılmıştır. Çalışma sonucunda, yöntem uygulanmadan kurulan modeller içerisinde KNN yöntemi ile en başarılı doğruluk değeri %94,14 olarak, k katlı çapraz doğrulama yöntemi kullanıldıktan sonra kurulan modeller içerisinde ise RF yöntemi ile en başarılı doğruluk değeri %93,92 olarak elde edilmiştir.

Anahtar Kelimeler: Makine Öğrenmesi, Sınıflandırma, Talasemi, K Katlı Çapraz Doğrulama.

Using and Comparing Different Machine Learning Methods for Thalassemia Disease Prediction

Abstract

Thalassemia is an inherited disease that causes a low amount of hemoglobin and red blood cells in the human body. This disease cannot be treated and some patients require lifelong blood transfusions. Early diagnosis of the disease is of great importance. The aim of this study is to predict thalassemia disease using machine learning classification methods. The data used in this study consists of patients coming to Erzurum Atatürk University Research Hospital. This study was carried out using classification methods in the Jupyter Notebook environment with the Python language. In this study, different classification methods such as Naive Bayes (NB), K-Nearest Neighbor (KNN), Support Vector Machines (SVM), Logistic Regression (LR), Random Forest (RF) and Decision Trees (DT) were compared. Using these classification methods, the best estimation result was tried to be achieved. The dataset was divided into 70% for training and 30% for testing. To prevent deviations in these stages, k fold cross validation (k fold cross validation) method was applied. In the performance evaluation of classification methods, performance metrics such as precision (precision), recall (recall), f1 score (f1 score), accuracy (accuracy), operating characteristic curve (ROC-AUC), log loss (logarithmic loss) were examined. As a result of this study, the most successful accuracy value was obtained as 94.14% with the KNN method among the models established without applying any method, and the most successful accuracy value was obtained as 93.92% with the RF method among the models established after using the k-fold cross-validation method.

Keywords: Machine Learning, Artificial Intelligence, Classification, Thalassemia, K Fold Cross Validation.

[1,2]Atatürk Üniversitesi, Bilgisayar Mühendisliği, Erzurum, Türkiye, ecegulsahabbasogullari@gmail.com baturalp@atauni.edu.tr

[*]Sorumlu Yazar/Corresponding Author
Geliş/Received: 08.07.2024 **Kabul/Accepted:** 09.09.2024 **Yayın/Published:** 15.12.2024

1. Giriş

Talasemi, hemoglobin sentezinde kalıtsal bozuklukların neden olduğu dünya genelinde çok yaygın bir gen hastalığıdır. Akdeniz bölgelerinde yaygın olarak görülmektedir (Mohammed & Al-Tuwaijari, 2021). Dünya Sağlık Örgütü raporlarına bakıldığında, dünya üzerinde %5,1 talasemi hastalığı ve 266 milyon talasemi taşıyıcısı olduğu görülmektedir. Talasemi hastalığı Türkiye'de yaklaşık olarak 1.300.000 talasemi taşıyıcısı ve 4500 bireyin de talasemi hastası olduğu görülmektedir (Karaaziz & Okyayuz, 2020). Talasemi hastalığı tanısı için genetik testler yaygın olarak kullanılmaktadır. Genetik testlerin yanı sıra çeşitli tarama yöntemleri hastalığın teşhis edilmesinde ilk adım olarak tercih edilmektedir. Talasemi hastalığında geleneksel tarama yöntemlerinden kan yayması, tam kan sayımı, ozmotik kırılganlık testi, hemoglobin elektrofarezi gibi yöntemler uzun yıllar boyunca hastalığın teşhis edilmesinde klinik uygulamalarda kullanılmaktadır (Gao & Liu, 2022). Talasemi hastaları için erken ve doğru tarama yapmak oldukça önemlidir. Son yıllarda Yapay Zekâ (YZ) teknolojileri birçok alanda özellikle tıp alanında çeşitli hastalıkların teşhis, tedavi ve tanı aşamalarında büyük bir başarıyla kullanılmaktadır (Akgül et al., 2024). YZ uygulamaları denildiğinde akla ilk olarak Makine Öğrenmesi (MÖ) yöntemleri gelmektedir. MÖ, verilerin daha verimli bir şekilde kullanılmasını makinelere öğretmek amacıyla kullanılmaktadır. Veriler görüntülendikten sonra elde edilen bilgiler yorumlanamayabilir. Bu bilgileri yorumlayabilmek için MÖ yöntemleri kullanılmaktadır. Bir sorunu çözmek için en iyi ve tek tip bir algoritma yoktur. Kullanılan MÖ algoritmasının türü, problemin türüne, veri setindeki değişkenlerin sayısına ve probleme uygun modelin yapısına bağlıdır (Mahesh, 2020).

Talasemi ve benzer birçok hastalığın teşhisinde MÖ yöntemleri kullanılarak önemli katkılar sağlanmaktadır. Bu çalışma doğrultusunda literatür taraması yapılmıştır. Ferih ve arkadaşları (Ferih et al., 2023) yaptıkları bir çalışmada, MÖ tekniklerini kullanarak talasemi hastalığı tahmini için modelleme yapmışlardır. Tam kan sayımı testi ile talasemi tanısına yönelik parametreler açıklanmıştır. Çalışmada model olarak, K-En Yakın Komşu (KNN), Naive Bayes (NB), Karar Ağacı (DT) ve Çok Katmanlı Algılayıcı (MLP) kullanılmıştır. En başarılı sonuca MLP ile ulaşılmıştır. MCH, MCHC ve RDW gibi parametreler kullanılarak, talasemi ve diğer kan hastalıklarının teşhisinde en iyi yöntem olduğu gösterilmiştir. Liu ve arkadaşları (Liu, 2024) yaptıkları bir çalışmada, talasemi hastaların Akdeniz anemisine sahip olup olmadığını tahmin etmek için iki model önerilmiştir. Bunlar Temel Bileşen Analiziyle Lojistik regresyon (PCA-LR) ve kısmi en küçük kareler yöntemi (PLS) modelleri olmuştur. Çalışma sonucunda %92,5 doğruluk oranı ile PLS modeli en başarılı sonuca ulaşmıştır. Devanath ve arkadaşları (Devanath et al., 2022) yaptıkları bir çalışmada, Yapay zekânın (YZ) önemli bir parçası olan MÖ yöntemleri kullanılarak talasemi tahmini için modeller oluşturulmuştur. Modellerin geliştirilmesi amacıyla 297 talasemi pozitif ve 297 talasemi negatif

kişiden elde edilen veriler kullanılmıştır. Uygulanan modeller; KNN, LR, Destek Vektör Makineleri (), NB, Rastgele Orman (RF), Uyarlanabilir Güçlendirme (ADA), Xgboost, DT, MLP ve Gradient Boosting sınıflandırıcısıdır. Bu çalışma sonucunda, ADA modelinin %100 doğruluk oranı ile en iyi sonucu verdiği görülmüştür. Phirom ve arkadaşları (Phirom et al., 2022) yaptıkları bir çalışmada, α+-talasemi tahmini için kırmızı kan hücresi (RBC) parametrelerini kullanan bir model ile teşhis performansını, tek bir RBC parametresi kullanan geleneksel bir yöntem ile karşılaştırmayı amaçlamışlardır. Hemoglobin H hastalığı bulunan ve fetüs açısından risk altında olan çiftler üzerinde çalışma yapılmıştır. Doğum öncesi talasemi hastalarından elde edilen yaş, cinsiyet, kırmızı kan hücresi (RBC) parametreleri (Hb, Hct, MCV, MCH, MCHC, RDW ve RBC sayısı) ile MÖ tabanlı yöntemler kullanılmıştır. Çalışmada, CNN, SVM, MLP, RF, PLS, LR, ET, LGBM, XGB, DT, KNN yöntemleri ile 10 kat çapraz doğrulama yöntemi kullanılmıştır. Modelde en iyi sonuç %100 doğruluk oranı ile MLP ile elde edilmiştir. Laengsri ve arkadaşları (Laengsri et al., 2019), talasemi ile demir eksikliği anemisini ayırt etmek için web tabanlı bir tahminleme aracı oluşturmayı amaçlamışlardır. Çalışmada KNN, Yapay Sinir Ağı, MLP, SVM ve DT modelleri uygulanmıştır. Uygulanan modeller arasında %99,12 doğruluk oranı ile MLP en iyi sonucu vermiştir. Yağmur ve arkadaşları (Yağmur et al., 2023), bir sınıflandırma yapmak amacıyla NN yöntemini kullanmışlardır. Çalışmada Öğrenmeli Vektör Kuantalama (LVQ), Rekabetçi Katman Sinir Ağı (CLNN), Örüntü Tanıma Yapay Sinir Ağı (PRNN), Kendiliğinden Organize Olan Harita (SOM) modelleri kullanılmıştır. Kullanılan modeller ile elde edilen başarı oranları karşılaştırılmış ve %99,88 doğruluk oranı olan PRNN ile en iyi başarı performansı elde edilmiştir. Farzaliyev ve arkadaşları (Farzaliyev et al., 2023), çocuklarda anemi hastalığının teşhisinde Topluluk Öğrenme Yöntemlerinin (Esemble Learning) kullanılmasını amaçlamışlardır. Çalışmada birçok MÖ yöntemi kullanılmıştır; bu yöntemler arasında DT, SVM, RF, LR, KNN yer almaktadır. Ayrıca, Topluluk öğrenme yöntemleri olarak Torbalama (Bagging), Artırma (Boosting) ve İstifleme (Staking) yöntemleri de kullanılmıştır. Çalışma sonucunda, MÖ yöntemlerinden %98 doğruluk oranı ile DT en yüksek başarı performansını elde ederken, Topluluk öğrenme yöntemlerinden %91 doğruluk oranı ile Boosting en yüksek başarı performansını elde etmiştir.

Literatür taramalarına göre, talasemi tahmini için yapılan çalışmalarda kullanılan MÖ algoritmaların yanı sıra, MLP algoritması ve MCH, MCHC ve RDW gibi farklı parametreler kullanılmıştır (Ferih et al., 2023). Ayrıca, Uyarlanabilir Güçlendirme (ADA Boosting), Xgboost, DT, MLP ve Gradient Boosting algoritmaları ile gerçek veri seti üzerinden talasemi tahmini için çalışmalar yapılmıştır (Devanath et al., 2022).

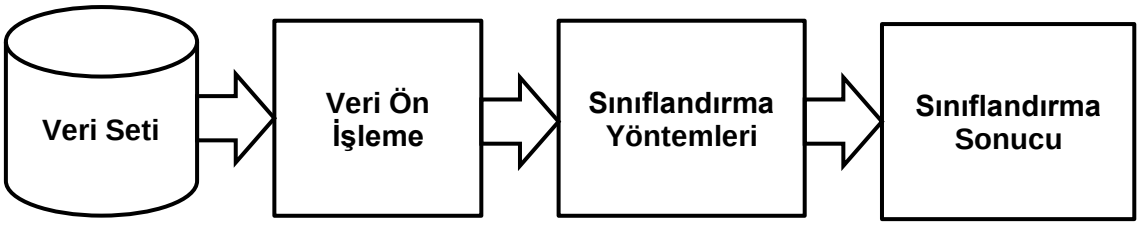
MÖ algoritmalarının kullanımı, hastalıkların erken teşhis ve tedavi süreçlerinin iyileştirilmesi açısından büyük öneme sahiptir. Özellikle büyük ve gerçek veri setleri ile çalışılması durumunda daha doğru ve hızlı teşhisler konulması mümkün olmaktadır. Bu çalışmada, talasemi hastalığının

tahmin edilmesine yönelik farklı MÖ algoritmaları gerçek veri seti üzerinde kullanılacak ve performans karşılaştırmaları yapılacaktır. Çalışma ile talasemi hastalığı tahmini için yüksek performanslı ve daha doğru analiz edilebilecek bir model elde edilmesi planlanmaktadır. MÖ yöntemlerinden NB, KNN, SVM, LR, RF ve DT algoritmaları kullanılarak talasemi hastalığı için tahminleme yapılacaktır.

2. Materyal ve Metot

Bu çalışmada, veri seti Erzurum Atatürk Üniversitesi Araştırma Hastanesine gelen hasta bilgilerinden oluşmaktadır. Bu çalışma için etik kurul izni Atatürk Üniversitesi Girişimsel Olmayan Klinik Araştırmalar Etik Kurulu'nun 07/06/2024 tarihli ve B.30.2.ATA.0.01.00/378 numaralı kararı ile alınmıştır. Talasemi hastalığının tahmini için NB, KNN, SVM, LR, RF ve DT gibi sınıflandırma yöntemleri kullanılmıştır. Bu yöntemler, farklı parametrelerle karşılaştırılarak sağlıklı ve talasemi hastası olan kişileri başarılı bir şekilde sınıflandıracak en yüksek doğruluk değerinin elde edildiği yöntem belirlenmeye çalışılmıştır. Çalışma, Google Colaboratory (Colab, 2024) bulut ortamında ve Python programlama dili ile gerçekleştirilmiştir.

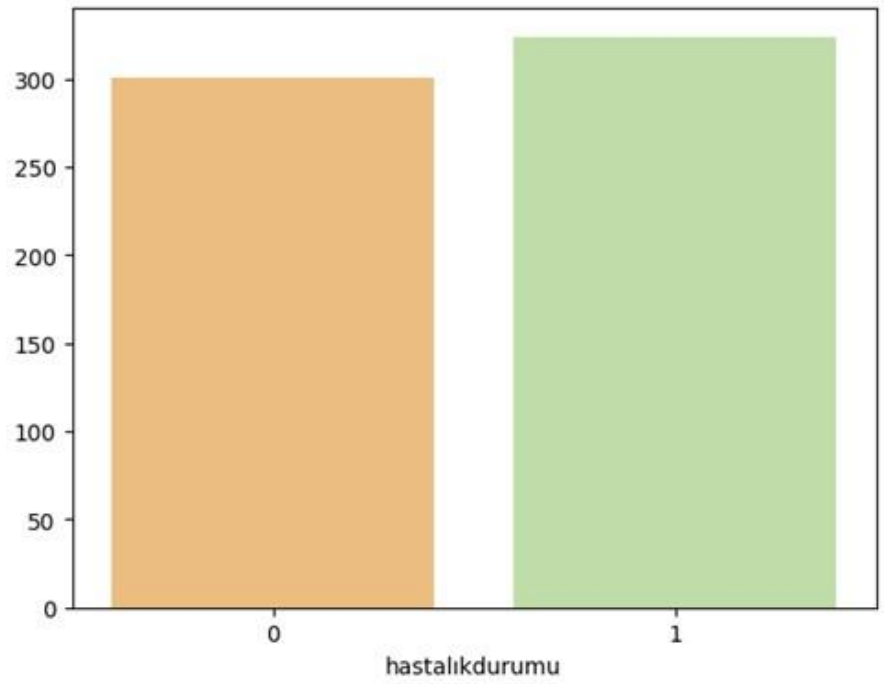
Talasemi hastalığının tahmini için sınıflandırma yöntemleri ile gerçekleştirilen çalışmanın akış şeması Şekil 1'de gösterilmiştir. Bu şemadaki her aşama, aşağıdaki alt bölümlerde detaylandırılmıştır.



Şekil 1. Çalışma akış şeması.

2.1. Veri Seti

Bu çalışmada, talasemi hastalığı tahmini için kullanılan veri seti, Erzurum Atatürk Üniversitesi Araştırma Hastanesi'ne gelen hastalardan oluşmaktadır. Veri seti, Şekil 2'de görüldüğü gibi, hastaneye gelen 625 hasta içerisinden 324 kişinin talasemi hastası, 301 kişinin ise diğer hastalardan oluştuğu verilerini içermektedir.



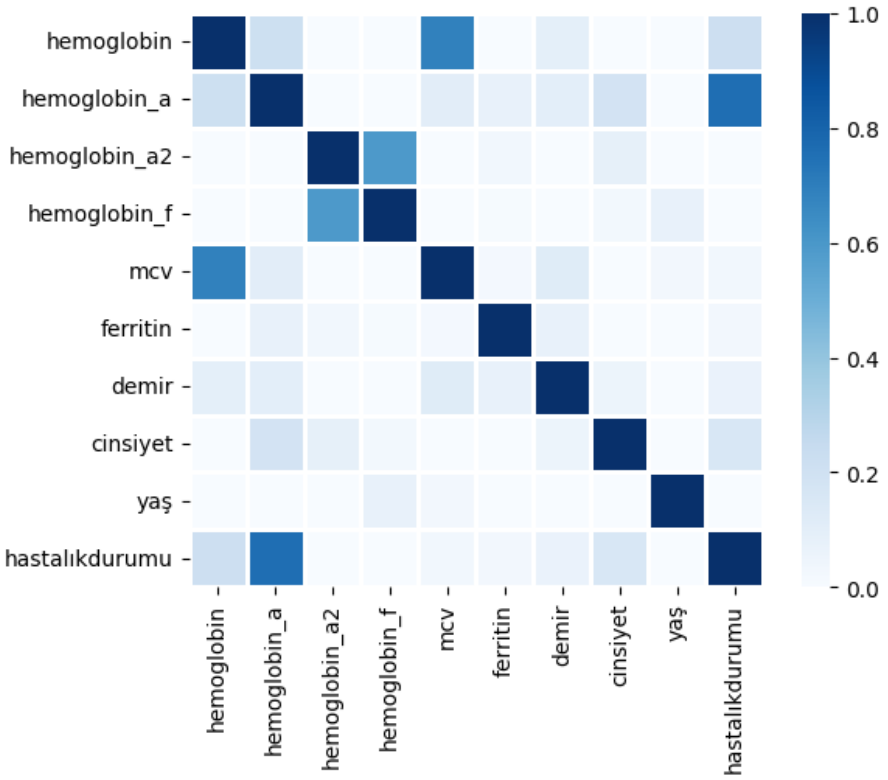
Şekil 2. Talasemi hastalığı olan (1) ve olmayan hasta (0) dağılımı.

Veri seti, 9 bağımsız değişken (hemoglobin, hemoglobin a, hemoglobin a2, hemoglobin f, mcv, ferritin, demir, cinsiyet, yaş) ve 1 bağımlı değişkenden (hastalık durumu) oluşmaktadır. Tablo 1'de özellik, açıklama ve birer örnek ile ayrıntılı bir şekilde gösterilmiştir.

Tablo 1. Talasemi hastalığı tahmini için veri seti özellikleri.

Öznitellik	Açıklama	Örnek
hemoglobin	kandaki hemoglobin değeri	56
hemoglobin a	kandaki hemoglobin a değeri	94,26
hemoglobin a2	kandaki hemoglobin a2 değeri	7,2
hemoglobin f	kandaki hemoglobin f değeri	18,49
mcv	ortalama hücre hacmi	69,7
ferritin	kandaki ferritin değeri	8,05
demir	kandaki demir değeri	13
cinsiyet	kadın, erkek	(0,1)
yaş	yaş	26
hastalık durumu	sağlıklı, hasta	(0,1)

Özniteliklerin birbirleriyle ilişkileri, Şekil 3'te korelasyon ısı haritası ile gösterilmektedir. Isı haritasında, 1.0'a olan yakınlık öznitelikler arasındaki pozitif ilişkiyi, 0.0'a yakınlık ise öznitelikler arasındaki negatif ilişkiyi göstermektedir. Isı haritasına bakıldığında, öznitelikler arasında pozitif yönde yüksek ilişkiye sahip hemoglobin a2, hemoglobin f ve hemoglobin, mcv özniteliklerinin olduğu gözlemlenmektedir. Bu ilişkiye göre, hemoglobin a2 özniteliğinde artan değer ile birlikte hemoglobin f öznitelikteki değerin de arttığı söylenebilir.

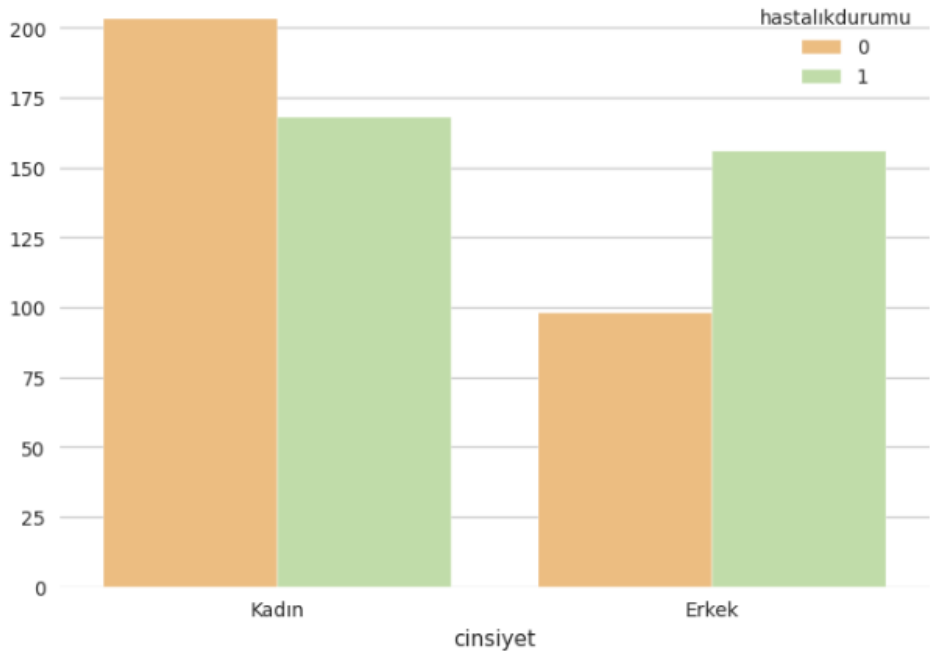


Şekil 3. Değişkenler arasındaki ısı haritası.

2.2. Veri Ön İşleme

Yapay Zeka uygulamalarının doğru ve yüksek performanslı çalışmalarına engel olan birçok etken bulunmaktadır. Bu etkenler arasında eksik gözlem, sınıflar arası dengesizlik ve aykırı gözlemler yer almaktadır (Karadağ, 2021). Veri ön işleme, kullanılan MÖ yöntemleri için önemli bir adımdır. Bu çalışmada kullanılan veri setinde sınıflandırma yapabilmek için ön işleme yapmak ve veri setini hazır hale getirmek gerekmektedir. MÖ yöntemlerini uygulamadan önce veri setinde eksik verilerin var olup olmadığı kontrol edilmiş ve eksik veri bulunmadığı görülmüştür. Veri setinde aykırı değerler tespit edilmiş ve bu aykırı değerler çeyrekler açıklığı yöntemi ile düzeltilmiştir. Bu işlemler sonrasında veri setine normalizasyon uygulanmıştır. Normalizasyonun uygulanma sebebi, veri setinde farklı ölçeklerde olan değerlerin belirli bir ölçeğe çekilerek sınıflandırma performanslarını karşılaştırmaktır. Ayrıca, KNN ve SVM gibi yöntemlerde veriler arasındaki uzaklıklar kullanıldığı için normalizasyon yapılmıştır. Bu çalışmada, normalizasyon için 'MinMaxScaler' yöntemi uygulanmıştır. Bu yöntemle veri setindeki tüm özellikler aynı ölçek aralığına (0 ve 1 arasına) getirilmiş ve bu sayede ölçek farklılıklarının MÖ yöntemleri üzerinde olumsuz etki oluşturmasının önüne geçilmiştir. Veri seti, %70 eğitim, %30 test verisi olarak ayrılmıştır. Eğitim ve test verisi olarak ayrılan veri setinde özniteliklerin dağılımına bağlı olarak sapmalar olabilmektedir. Bu sapmaların önüne geçebilmek için kullanılan modelde k katlı çapraz doğrulama (k fold cross validation) yöntemi

uygulanmıştır. Veri setinde yer alan hastaların cinsiyete göre dağılımını gösteren sütun grafiği Şekil 4'te gösterilmektedir. Şekil 4'e göre, talasemi tanısı olan kadın bireylerin, erkek bireylere göre daha fazla olduğu söylenebilir.



Şekil 4. Cinsiyete göre talasemi hastalığı olan (1) ve olmayan (0) hasta dağılımı.

2.3. Sınıflandırma Yöntemleri

Veri seti ön işlemeden sonra MÖ yöntemlerinden NB, KNN, SVM, LR, RF ve DT kullanılmıştır.

NB, Bayes teoremine dayanan bir denetimli öğrenme yöntemidir. Bu yöntem, bir sınıfta belirli bir özelliğin varlığının diğer özelliklerle bağımsız olduğunu varsayar. Eğitim süreci oldukça basit ve hızlıdır; bu, bağımsızlık varsayımına dayanmaktadır. NB sınıflandırıcıları, özellikler arasında güçlü bağımsızlık varsayımlarına sahiptir ve basit olasılıklı sınıflandırıcılar kategorisine girer (Alzubi et al., 2018). Ayrık ve sürekli verileri işleyebilir ve ikili ile çoklu sınıflandırma için kullanılabilir. Modeller uygun şekilde eğitildiğinde ve ayarlandığında, basit olmalarına rağmen iyi performans gösterebilirler (Ibrahim & Abdulazeez, 2021). KNN, bir denetimli öğrenme algoritmasıdır ve eski, basit sınıflandırma yöntemlerinden biridir. Bu algoritma, seçilen bir gözlemin kendisine en yakın olan gözlem arasındaki yakınlığı kullanarak sınıflandırma yapar. KNN, veri noktaları arasındaki Öklid mesafesini kullanır ve veriler arasında komşular elde eder. k değeri kullanıcı tarafından belirlenen bir sabittir ve bu değerin seçimi önemlidir; dikkatli seçilmesi gerekmektedir (Gonaygunta, 2023). Çalışmada k değeri veri boyutu ve yapısına bağlı olarak belirlenmiştir. SVM, en yaygın olarak

kullanılan sınıflandırma yöntemlerinden biridir. Sınıflandırma ve regresyon analizi için kullanılan denetimli öğrenme algoritmasıdır. SVM, doğrusal veya doğrusal olmayan sınıflandırmaları verimli bir şekilde gerçekleştirebilir. Bu, sınıflar arasında en iyi boşluğu bularak ve sınırlar çizerek yapmaktadır. Kenar boşlukları, sınıflar arasındaki aralığı maksimize ederek sınıflandırma hatalarını en aza indirir (Mahesh, 2020). LR, denetimli bir MÖ algoritmasıdır ve bir kategorik bağımlı değişkenin olasılığını tahmin etmek için ikili sınıf etiketlerinde kullanılır (Gonaygunta, 2023). Bu bağımlı değişken, 1 (evet/doğru) veya 0 (hayır/yanlış) olan verileri içeren ikili bir değişkendir. RF, denetimli bir MÖ algoritmasıdır ve sınıflandırma ile regresyon analizinde kullanılabilir. Birden fazla DT'nin tahminlerini bir araya getirir. Büyük verileri kümeler ve küçük ağaçlar oluşturur; küçük ağaç sayısı arttıkça kesinlik artar (Vatansever et al., 2021). DT, en çok kullanılan denetimli öğrenme algoritmalarından biridir. DT, düğüm, dal ve yapraktan oluşur. Her düğüm, özellik üzerindeki yapılan testi, her dal, yapılan test sonucunu ve her yaprak ise düğümde bulunan sınıf etiketini temsil eder (Nematzadeh et al., 2015). DT, regresyon ve sınıflandırma analizinde kullanılabilir.

2.4. Sınıflandırma Algoritmalarının Performansını Test Etme Ölçütleri

Çalışmada kullanılan sınıflandırma sonuçlarının ve başarı performansların belirlenmesinde çeşitli yaklaşımlar uygulanmıştır. Sınıflandırmada kullanılan algoritmaların performans değerlendirmesinde karmaşıklık matrisleri (confusion matrix) oluşturulmuş ve bu matrisler kullanılarak kesinlik (precision), duyarlılık (recall), f1 skoru (f1 score), doğruluk (accuracy), işlem karakteristik eğrisi (ROC-AUC curve), log loss (logaritmik kayıp) gibi metrikler aracılığıyla algoritma performanslarının başarı oranları karşılaştırılmıştır.

İkili sınıflandırmada, problem için elde edilen sonuçlar Pozitif (Positive-P) ve Negatif (Negative-N) olarak etiketlenmektedir. 2x2 karmaşıklık matrisi, olası dört sonuç için her biri tarafından tahmin edilen örneklerin sayısını göstermektedir. Bu tahminler; doğru pozitif (True Positive-TP), yanlış pozitif (False Positive-FP), doğru negatif (True Negative-TN) ve yanlış negatif (False Negatif-FN) kullanılarak değerlendirilir (Ozcift & Gulten, 2011). Kullanılan sınıflandırıcıların performans ölçütlerini türetmek için Tablo 2'de gösterilen karmaşıklık matris tablosu kullanılmaktadır.

Tablo 2. Karmaşıklık matrisi.

	Gerçek Pozitif (1)	Gerçek Negatif (0)
Tahmin Pozitif (1)	True Positive (TP)	False Positive (FP)
Tahmin Negatif (0)	False Negative (FN)	True Negative (TN)

TP: Talasemi hastası olarak etiketlenmiş örneğin talasemi hastası olması

FP: Talasemi hastası olmayan bir örneğin talasemi hastası olarak etiketlenmesi

TN: Talasemi hastası olmayan bir örneğin hasta olmaması

FN: Talasemi hastası olan bir örneğin talasemi hastası olmadığının etiketlenmesi olarak açıklanabilir.

2.4.1. Kesinlik

Kesinlik metriği, oluşturulan modelde pozitif olarak sınıflandırılan örneklerin ne kadarının gerçekten pozitif olduğunu göstermektedir. İyi bir sınıflandırıcıda kesinlik değerinin 1 olması beklenir. Payda değeri (TP + FP) ne kadar büyürse, kesinlik değeri o kadar azalır. Kesinlik değeri aşağıdaki formül (1) ile hesaplanmaktadır.

$$Kesinlik = TP/(TP + FP) \quad (1)$$

2.4.2. Duyarlılık

Duyarlılık metriği; oluşturulan modelde pozitif olarak sınıflandırılan örneklerin ne kadarının gerçekten pozitif olduğunu göstermektedir. İyi bir sınıflandırıcıda duyarlılık değerinin 1 olması beklenir. Payda değeri (TP + FN) ne kadar büyürse, duyarlılık değeri o kadar azalır. Duyarlılık değeri aşağıdaki formül (2) ile hesaplanmaktadır.

$$Duyarlılık = TP/(TP + FN) \quad (2)$$

2.4.3. F1 Skoru

F1 Skoru, kesinlik ve duyarlılık metrikleri arasındaki dengeyi gösterir ve bu metriklerin harmonik ortalamasını kullanarak hesaplanır (Sevli, 2022). Genellikle doğruluk değerinden daha kapsamlı bir metrik olduğu kabul edilir. Duyarlılık ve kesinlik değeri ne kadar büyük olursa, f1 skor değeri de o kadar büyük olur. F1 skor değeri aşağıdaki formül (3) ile hesaplanmaktadır.

$$F1\ Skoru = 2\ x\ (Duyarlılık\ x\ Kesinlik)/(Duyarlılık + \ Kesinlik) \quad (3)$$

2.4.4. Doğruluk

Doğruluk, oluşturulan modelin genel başarısını göstermektedir. Veri setinin dengeli dağılması durumunda doğruluk metriği etkili bir performans göstergesi olarak kabul edilebilir. Doğruluk değeri aşağıdaki formül (4) ile hesaplanmaktadır.

$$Doğruluk = (TP + TN) \,/\, (TP + FP + TN + FN) \quad (4)$$

2.4.5. ROC-AUC Eğrisi

ROC bir olasılık eğrisi olarak adlandırılabilir. Duyarlılık ve FP ile oranından elde edilir. Eksende (0,0) ve (1,1) noktaları arasında yer almaktadır. AUC, ROC eğrisi altında kalan alanı temsil eder ve 0 ile 1 aralığında değer almaktadır. AUC değeri ne kadar 1'e yakın olursa model performans başarısı o kadar yüksektir denilebilir (Eröz, 2010). Model yetersiz ise AUC değeri 0 veya 0'a yakın bir değer alır.

2.4.6. Log Loss

Log Loss değeri, çalışmada yapılan tahmindeki olasılık değerlerine dayanmakta olup sınıflandırma problemleri için önemli bir ölçüttür. Bu ölçüt her bir örnek için tahmin edilen olasılık dağılımı ile gerçek sınıf etiketlerini kullanmaktadır. Olasılıkların, gerçek sınıf etiketlerine yakınlığını hesaplamaktadır. Log loss değerleri ne kadar düşük olursa çalışmada uygulanan model başarısı o kadar yüksek olmaktadır. Log Loss değeri aşağıdaki formül (5) ile hesaplanmaktadır. Log loss değeri her bir iterasyon için hesaplanır ve daha sonra hesaplanan bu değerlerin ortalaması alınmaktadır.

$$LogLoss = -\frac{1}{n}\sum_{i=0}^{n}[y_i \log(p) + (1 - y_i) \log(1 - p)] \quad (5)$$

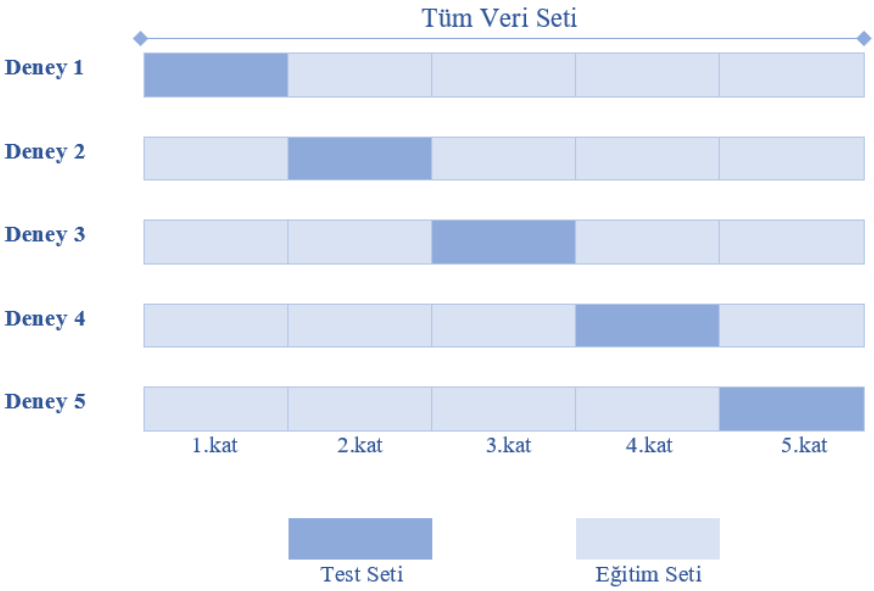
n: Veri setinde bulunan örnek sayısı

y: Gerçek sınıf etiketi (0,1)

p: Tahmin edilen olasılık

2.5. K Katlı Çapraz Doğrulama

K katlı çapraz doğrulama, oluşturulan modelin performansını değerlendirmede önemli bir adımdır. k değerinin seçimi, performans başarısı açısından oldukça önemlidir. Genel olarak, literatürde k değeri sıklıkla 5 veya 10 olarak belirtilmektedir (Nti et al., 2021). Ancak, bu değerler veri setine bağlı olarak her çalışmada farklılık gösterebilir. Şekil 5'te 5 adet k değeri için eğitim ve test verilerine göre çapraz doğrulama adımları gösterilmiştir.



Şekil 5. K katlı çapraz doğrulama.

3. Bulgular ve Tartışma

Veri seti, Erzurum Atatürk Üniversitesi Araştırma Hastanesi'ne gelen hastalardan oluşmaktadır. Veri setinde, 324 talasemi hastası ve 301 sağlıklı birey olmak üzere toplam 625 hasta bulunmaktadır. Veri seti, gerekli veri ön işleme adımlarından geçerek sınıflandırma için hazır hale getirilmiştir. Sınıflandırmaya başlamadan önce, veri seti %70 eğitim ve %30 test verisi olarak ikiye ayrılmıştır. Bu çalışmada, sınıflandırma yöntemleri kullanılarak talasemi hastalığı tahmini yapmak için farklı modeller oluşturulmuştur. MÖ yöntemlerinden NB, KNN, SVM, LR, RF ve DT kullanılmıştır. Dengeli dağılmayan veri setlerinde, oluşturulan modellerin performansını değerlendirmek için doğruluk ölçütü tek başına yeterli olmayabilir. Bu nedenle, kullanılan sınıflandırma yöntemlerinin performansı, kesinlik, duyarlılık, f1 skoru, doğruluk, işlem karakteristik eğrisi, log loss gibi istatistiksel performans metrikleri dikkate alınarak değerlendirilmiştir.

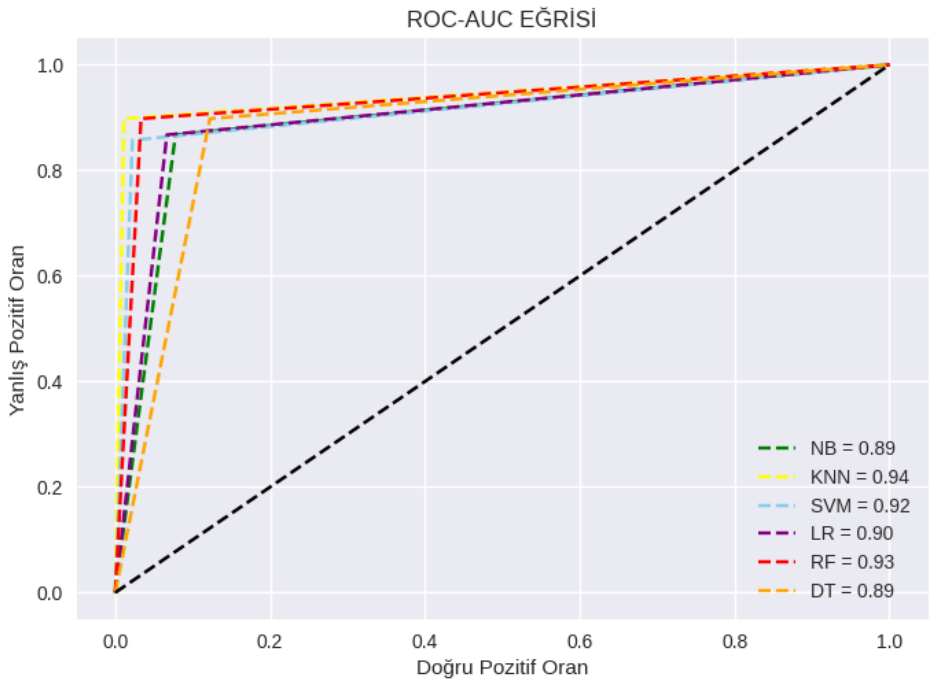
Tablo 3. Çalışmada kullanılan sınıflandırma yöntemlerinin performans yüzdeleri.

Yöntem	Keskinlik (%)	Duyarlılık (%)	F1 – Skor (%)	AUC (%)	Doğruluk (%)	K Katlı Çapraz Doğrulama (%)
NB	89.55	89.36	89.37	89.55	89.36	90.40
KNN	94.58	94.15	94.15	**94.25**	**94.14**	87.52
SVM	92.22	91.49	91.48	92.12	91.48	91.84
LR	90.15	89.89	89.90	90.14	89.89	89.12
RF	93.35	93.09	93.09	93.06	93.08	**93.92**
DT	87.32	87.23	87.24	89.11	87.23	92.48

Tablo 3'te görüldüğü üzere, sınıflandırma yöntemlerinin keskinlik, duyarlılık, f1 skor, doğruluk gibi başarı sonuçları verilmiştir. Çalışmada her bir yöntem için belirli parametre değerleri, başarı

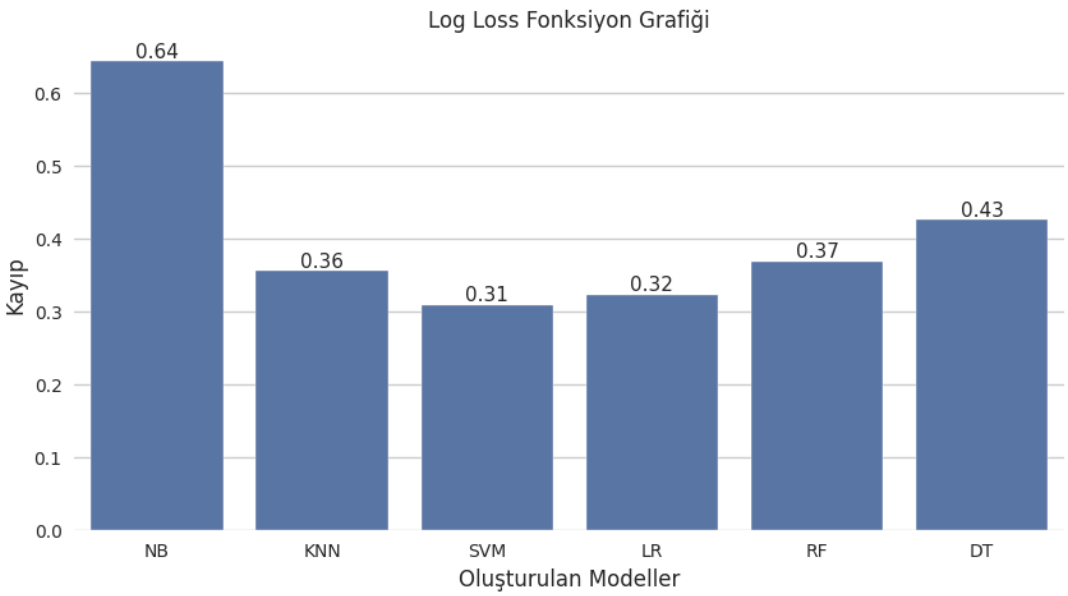
sonuçlarına göre belirlenmiştir. Örneğin, KNN için 'n_neighbors' parametesi 2 ile 17 arasında tüm değerler denenmiş ve en optimum değer 11 olarak belirlenmiştir. Ayrıca, uzaklık ölçütü olarak 'minkowski' ölçüm yöntemi kullanılmıştır. SVM için 'linear' çekirdek fonksiyonu kullanılmıştır. LR için 'solver' parametresi 'lbfgs' ve 'random_state' parametresi 42 olarak belirlenmiştir. 'random_state' parametresi, sınıflandırma aşamasında veri setinde farklı yerlerden ayrım yaparak ezberlemenin önüne geçilmesini sağlar. RF için karar ağacı sayısı 20 olarak seçilmiş ve karar ağaçlarındaki elde edilen değerlerin ortalaması ile sonuç üretilmiştir. DT için ağaç derinliği, yani 'max_depth' parametresi 5 olarak belirlenmiştir. Ağaç derinliği, kök düğüm ile oluşan yaprak arasındaki yolu belirtir ve ağacın maksimum derinliğini ifade eder. Sınıflandırmada kullanılan parametreler için farklı değerler verilmiş ve sonuçlar değerlendirilmiş, en uygun olan parametreler tercih edilmiştir.

Çalışmada kullanılan 6 sınıflandırma yönteminin doğruluk değerlerine bakıldığında, tüm yöntemlerin genel olarak başarılı olduğu söylenebilir. Deneysel sonuçlar karşılaştırıldığında, performans değerlendirmeleri şu şekilde açıklanabilir: Hiçbir yöntem uygulanmadan elde edilen sonuçlara göre %94.14 doğruluk değeri ile KNN, %93.08 doğruluk değeri ile RF, %91.48 doğruluk değeri ile SVM, %89.89 doğruluk değeri ile LR, %89.36 doğruluk değeri ile NB ve %87.23 doğruluk değeri ile DT başarılı performanslar göstermiştir. MÖ sınıflandırma yöntemlerine k katlı çapraz doğrulama yöntemi uygulanmıştır. Bu yöntemde k değeri 5 olarak alınmıştır. K katlı çapraz doğrulama uygulandıktan sonra sırasıyla %93.92 doğruluk değeri ile RF, %92.48 doğruluk değeri ile DT, %91.84 doğruluk değeri ile SVM, %90.40 doğruluk değeri ile NB, %89.12 doğruluk değeri ile LR ve %87.52 doğruluk değeri ile KNN elde edilmiştir. KNN sınıflandırma algoritması, k katlı çapraz doğrulama yöntemi kullanılmadan önce %94.14 doğruluk değeri ile en başarılı performansı göstermiş olup, bu yöntemde %94.58 keskinlik, %94.15 duyarlılık ve %94.15 f1-skoru değerleri elde edilmiştir. En düşük başarı performansı ise %87.23 doğruluk değeri ile DT sınıflandırma yöntemi olmuştur. K katlı çapraz doğrulama yöntemi kullanıldıktan sonra %93.92 doğruluk değeri ile RF sınıflandırma algoritması en başarılı performansı göstermiş olup, bu yöntemde %93.35 keskinlik, %93.09 duyarlılık ve %93.09 f1-skoru değerleri elde edilmiştir. En düşük başarı performansı ise %87.52 doğruluk değeri ile KNN sınıflandırma yöntemi olmuştur.

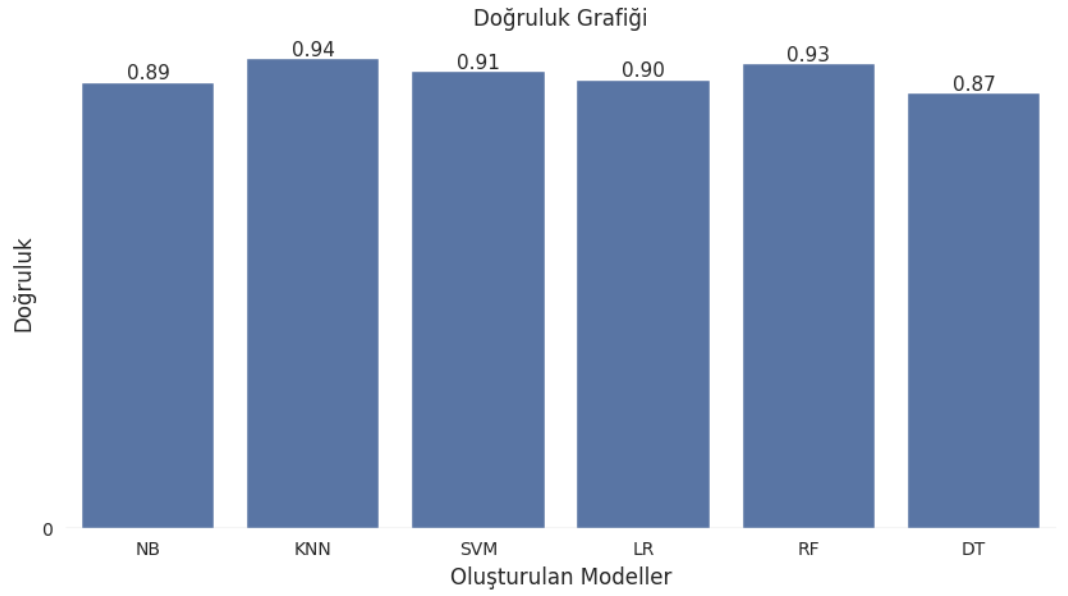


Şekil 6. Deneysel sonuçların ROC-AUC eğrisi.

Çalışmada sonucunda, Şekil 6'da görüldüğü gibi, NB, KNN, SVM, LR, RF ve DT sınıflandırma yöntemlerinin sırasıyla, k katlı çapraz doğrulama yöntemi kullanılmadan elde edilen ROC eğrisi ve AUC değerleri gösterilmiştir. Grafikte (0,0) noktasından (1,1) noktasına giden siyah eğri, ROC eğrisinin anlamlandırılmasında bir referans noktası olarak kullanılmaktadır. Bu eğrinin altında kalan alan, grafiğin %50'sini oluşturmaktadır ve başarılı bir sınıflandırmada bu eğri üzerinde grafik oluşması beklenmektedir. Çalışma sonucunda %94 ile en yüksek AUC değerine sahip olan KNN yöntemi, %89 ile en düşük AUC değerine sahip olan NB ve DT sınıflandırma yöntemleri olarak belirlenmiştir. Sırası ile %93 AUC değeri ile RF, %92 AUC değeri ile SVM, %90 AUC değeri ile LR, %89 AUC değeri ile NB ve %89 AUC değeri ile DT sınıflandırma yöntemleri elde edilmiştir. Bu değerler, oluşturulan yöntemlerin yüksek doğrulukla sınıflandırma yapabildiğini göstermektedir.



Şekil 7. Log loss fonksiyon grafiği.



Şekil 8. Doğruluk değeri grafiği.

Çalışmada, her iterasyon için kayıp değerler incelenmiş ve bu kayıp değerlerin ortalamaları hesaplanmıştır. Kayıp değerlerin ortalamasını elde etmek amacıyla log loss fonksiyonu kullanılmıştır. Her sınıflandırma algoritması için ayrı ayrı log loss fonksiyonu çalıştırılmış ve Şekil 7'de gösterildiği gibi, bulunan kayıp değer ortalamaları barplot grafiği ile sunulmuştur. Log loss değeri ne kadar düşükse, modelin başarısı o kadar yüksek olmaktadır. Log loss fonksiyonu ile elde edilen en düşük ortalama kayıp değeri, SVM yöntemi ile 0.31 olarak bulunmuş; Şekil 8'de görüldüğü üzere, SVM yönteminde doğruluk değeri 0.91 olarak elde edilmiştir. Log loss fonksiyonu sonucunda elde edilen ortalama kayıp değerler sırasıyla; 0.32 LR, 0.36 KNN, 0.37 RF, 0.43 DT ve 0.64 NB olarak

belirlenmiştir. Her bir iterasyon için doğruluk değerleri ise başarı sırasına göre; 0.94 SVM, 0.93 RF, 0.91 SVM, 0.90 LR, 0.89 NB ve 0.87 DT olarak elde edilmiştir.

Tablo 4. Literatürde yer alan çalışmalar ile gerçekleştirilen çalışmanın karşılaştırılması.

Çalışma Adı	Uygulanan Modeller	Başarılı Model	Doğruluk
Anemi hastalığının yapay sinir ağları yöntemleri kullanılarak sınıflandırılması (Yağmur et al., 2023)	LVQ, CLNN, PRNN, SOM	PRNN	%99.88
An Application of Machine Learning to Thalassemia Diagnosis (Liu, 2024)	PCA-LR, PLS	PLS	%92.50
Discrimination of β-thalassemia and iron deficiency anemia through extreme learning machine and regularized extreme learning machine based decision support system (Çil et al., 2020)	LR, KNN, SVM, ELM, RELM	SVM	%94.37
Applications of Artificial Intelligence in Thalassemia: A Comprehensive Review (Ferih et al., 2023)	KNN, NB, DT, MLP	MLP	%92
Thalassemia Prediction using Machine Learning Approaches (Devanath et al., 2022)	KNN, LR, SVM, NB, RF, ADABOOST, Xgboos, DT, MLP, Gradient Boosting	ADABOOST	%100
The TVGH-NYCU Thal-Classifier: Development of Machine-Learning Classifier for Differentiating Thalassemia and Non-Thalassemia Patients (Fu et al., 2021)	SVM	SVM	%76
Gerçekleştirilen Çalışma	NB, KNN, SVM, LR, RF, DT	KNN	%94.14

Tablo 4'te görüldüğü üzere, literatür taraması gerçekleştirilmiştir. Bu tarama sonucunda, çalışmaya benzer diğer araştırmalarda kullanılan başarılı modeller ve bu modellerin doğruluk oranları listelenmiştir. Diğer çalışmalarda, bu çalışmada kullanılan modellerden farklı modeller uygulanmıştır; bunlar arasında LVQ, CLNN, PRNN, SOM, PCA-LR, PLS, ELM, RELM, MLP, ADABOOST, Xgboos, Gradient Boosting gibi yöntemler yer almaktadır. Tabloda yer alan çalışmaların doğruluk sonuçlarına göre, ADABOOST yönteminin %100 başarı performansı ile en başarılı yöntem olduğu söylenebilir. %76 başarı performansı ile SVM, diğer çalışmalardaki oranlara göre daha düşük bir başarı göstermiştir. Gerçekleştirilen çalışma, SVM yöntemi ile %94.14 doğruluk elde ederek oldukça yüksek bir başarı göstermiştir. Veri setinde veri azlığı ve parametre eksikliği modellerimizi daha yüksek doğruluk değerine ulaşılamamasının sebebi olarak gösterilebilir.

5. Sonuçlar ve Öneriler

Kalıtsal hastalıkların her geçen gün artış gösterdiği ve dünya genelinde yaygın bir gen hastalığı olan talasemi, erken teşhis açısından büyük önem taşımaktadır. Teknolojinin ilerlemesi ile birlikte tıp alanında MÖ yöntemlerinin kullanımı yaygınlaştırılmıştır. MÖ sınıflandırma yöntemleri, hastalıkların erken teşhis ve tedavi süreçlerinin iyileştirilme açısından büyük potansiyele sahiptir. Bu

çalışmada, MÖ kullanılarak talasemi hastalığı tahmin edilmesi ve bu tahminle birlikte talasemi ile ilişkili risklerin en aza indirilip, var olan uygun tedavi yöntemleriyle daha kaliteli bir yaşam sürülmesine katkı sağlanması amaçlanmıştır.

Çalışmada, MÖ yöntemleri kullanılarak 625 hastanın özellikleri analiz edilmiştir. Bu analiz için 6 farklı sınıflandırma yöntemi uygulanmıştır; NB, KNN, SVM, LR, RF, DT. Çalışmada, genel performans açısından en başarılı ve en başarısız yöntemler, kullanılan MÖ yöntemleri dikkate alınarak yapılan analizler sonucunda belirlenmiştir. Çalışma ile kişilerin talasemi hastası olup olmadıklarını tahmin edebilecek modeller oluşturulması amaçlanmıştır.

İlk olarak veri seti analizi yapılmış ve hemoglobin, hemoglobin a, hemoglobin a2, hemoglobin f, mcv bağımsız değişkenlerin, bireylerin talasemi hastası olmasında en etkili öznitelikler olduğu görülmüştür. Ancak öznitelikler tek başına talasemi hastalığını belirlemek için yeterli olmamaktadır; bu nedenle diğer performans metrikleri de incelenmiştir. Veri ön işleme adımları kapsamında eksik veri kontrolü ve aykırı değerlerin ele alınması gerçekleştirilmiştir. Veri seti, %70 eğitim, %30 test verisi olarak ikiye ayrılmıştır. Veri ön işleme adımları sonrasında veri seti modelleme için hazır hale getirilmiştir.

NB, KNN, SVM, LR, RF, DT algoritmaları için modeller oluşturulmuştur. İlk olarak, hiçbir yöntem uygulanmadan her bir yöntemin için performans sonuçları alınmıştır. Daha sonra, eğitim ve test verisi olarak ayrılan veri setinde sapmaların önüne geçmek amacıyla k katlı çapraz doğrulama yöntemi uygulanmış ve her model için ayrı performans sonuçları elde edilmiştir. Oluşturulan modellerin performans ölçütleri olarak kesinlik, duyarlılık, f1 skoru, doğruluk, işlem karakteristik eğrisi ve log loss gibi istatistiksel performans metrikleri dikkate alınarak değerlendirilmiştir.

Çalışmada, hiçbir yöntem uygulanmadan kurulan modeller arasında KNN yöntemi en yüksek doğruluk değeri olan %94.14 ve AUC değeri olan %94 ile en başarılı sonuçları vermiştir. DT yöntemi ise %87.23 doğruluk ve %89 AUC değerleri ile en düşük performansı göstermiştir. K katlı çapraz doğrulama yöntemi kullanıldıktan sonra RF yöntemi en yüksek doğruluk değeri olan %93.92'yi, KNN yöntemi ise en düşük doğruluk değeri olan %87.52'yi elde etmiştir. K katlı çapraz doğrulama uygulandıktan sonra KNN yönteminin performansında küçük bir azalma görülmüştür.

Log loss fonksiyonu ile her bir iterasyon için kayıp değerler hesaplanmış ve ortalamaları Şekil 7'de gösterilmiştir. Kayıp değerleri ne kadar düşükse, model performansı o kadar yüksek olmaktadır. Log loss fonksiyonu sonucunda SVM yöntemi 0.31 kayıp değeri ile en düşük, NB yöntemi ise 0.64 kayıp değeri ile en yüksek sonuçları vermiştir. Şekil 8'de, her bir iterasyonda modellerden elde edilen başarı değerlerinin ortalamaları verilmiştir. SVM yöntemi 0.94 doğruluk değeri ile en yüksek, DT yöntemi ise 0.87 doğruluk değeri ile en düşük doğruluk değerini elde etmiştir.

Çalışmada değerlendirilen performans metriklerine göre, %94.14 başarı oranı ile KNN yönteminin talasemi hastalığının teşhisinde başarılı bir şekilde kullanılabileceği sonucuna

ulaşılmıştır. Ayrıca, kullanılan diğer yöntemlerin başarı performanslarının birbirine oldukça yakın olduğu görülmüştür. Literatürdeki farklı veri setleri ve parametrelerle yapılan MÖ sınıflandırma yöntemleri, talasemi tahmininde başarılı performans sonuçları elde etmiştir. Çalışmanın özgünlüğü, Erzurum Atatürk Üniversitesi Araştırma Hastanesine gelen hastalardan oluşan veri seti ve kullanılan farklı parametrelerle sağlanmıştır. Bu çalışma ile bireylerin talasemi hastası olup olmadıklarının doğru bir şekilde erken teşhis edilmesi sağlanarak doktorların hastalık için erken ve doğru karar vermesine olanak tanınacaktır. Gelecekte yapılacak çalışmalar, daha büyük veri setleri üzerinde ve farklı parametreler eklenerek, derin öğrenme yöntemleri ile hibrit modeller oluşturmayı ve bu modellerle etkili tahminleme ve analizler yapmayı hedeflemektedir.

Yazarların Katkısı

Tüm yazarlar çalışmaya eşit katkıda bulunmuştur.

Çıkar Çatışması Beyanı

Yazarlar arasında herhangi bir çıkar çatışması bulunmamaktadır.

Araştırma ve Yayın Etiği Beyanı

Yapılan çalışmada araştırma ve yayın etiğine uyulmuştur.

Kaynaklar

Akgül, İ., Kaya, V., Karavaş, E., Aydın, S., & Baran, A. (2024). A Novel Artificial Intelligence-Based Hybrid System to Improve Breast Cancer DetectionUsing DCE-MRI. *BULLETIN OF THE POLISH ACADEMY OF SCIENCES. TECHNICAL SCIENCES, 72*(3).

Alzubi, J., Nayyar, A., & Kumar, A. (2018). Machine learning from theory to algorithms: an overview. Journal of physics: conference series,

Colab. (2024). *Google Colaboratory*. Retrieved 2024 from https://colab.research.google.com/

Çil, B., Ayyıldız, H., & Tuncer, T. (2020). Discrimination of β-thalassemia and iron deficiency anemia through extreme learning machine and regularized extreme learning machine based decision support system. *Medical hypotheses*, *138*, 109611.

Devanath, A., Akter, S., Karmaker, P., & Sattar, A. (2022). Thalassemia Prediction using Machine Learning Approaches. 2022 6th International Conference on Computing Methodologies and Communication (ICCMC),

Eröz, B. (2010). *Veri yapısına bağlı olarak Roc eğrisi altında kalan alana ilişkin istatistiksel yöntemlerin karşılaştırılması* Sağlık Bilimleri Enstitüsü].

Farzaliyev, E., Saihood, Q., & Sonuç, E. (2023). Çocuklarda Anemi Hastalığının Teşhisinde Topluluk Öğrenme Yöntemlerinin Kullanılması. 1 st International Conference on Recent Academic Studies,

Ferih, K., Elsayed, B., Elshoeibi, A. M., Elsabagh, A. A., Elhadary, M., Soliman, A., Abdalgayoom, M., & Yassin, M. (2023). Applications of artificial intelligence in thalassemia: A comprehensive review. *Diagnostics*, *13*(9), 1551.

Fu, Y., Liu, H., Lee, L., Chen, Y., Chien, S., Lin, J., Chen, W., Cheng, M., Lin, P., & Lai, J. (2021). The TVGH-NYCU Thal-Classifier: Development of a Machine-Learning Classifier for Differentiating Thalassemia and Non-Thalassemia Patients. Diagnostics (Basel) 2021; 11 (9): 1725. *DOI: https://doi. org/10.3390/diagnostics11091725. PMID: https://www. ncbi. nlm. nih. gov/pubmed/34574066*.

Gao, J., & Liu, W. (2022). Advances in screening of thalassaemia. *Clinica Chimica Acta*, *534*, 176-184.

Gonaygunta, H. (2023). Machine learning algorithms for detection of cyber threats using logistic regression. *Department of Information Technology, University of the Cumberlands*.

Ibrahim, I., & Abdulazeez, A. (2021). The role of machine learning algorithms for diagnosing diseases. *Journal of Applied Science and Technology Trends*, *2*(01), 10-19.

Karaaziz, M., & Okyayuz, Ü. H. (2020). Bir talasemi hastasının hastalık ile uyumluluğunun incelenmesi: olgu sunumu. *Cukurova Medical Journal*, *45*(1), 362-369.

Karadağ, K. (2021). Kan vermeye elverişli donörlerin makine öğrenme yöntemleri ile tespiti. *Adıyaman Üniversitesi Mühendislik Bilimleri Dergisi*, *8*(15), 508-514.

Laengsri, V., Shoombuatong, W., Adirojananon, W., Nantasenamat, C., Prachayasittikul, V., & Nuchnoi, P. (2019). ThalPred: a web-based prediction tool for discriminating thalassemia trait and iron deficiency anemia. *BMC medical informatics and decision making*, *19*, 1-14.

Liu, S. (2024). An Application of Machine Learning to Thalassemia Diagnosis. *Journal of Computer and Communications*, *12*(2), 211-230.

Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, *9*(1), 381-386.

Masala, G. L., Golosio, B., Cutzu, R., & Pola, R. (2013). A two-layered classifier based on the radial basis function for the screening of thalassaemia. *Computers in biology and medicine*, *43*(11), 1724-1731.

Mohammed, M. Q., & Al-Tuwaijari, J. M. (2021). A Survey on various Machine Learning Approaches for thalassemia detection and classification. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, *12*(13), 7866-7871.

Nematzadeh, Z., Ibrahim, R., & Selamat, A. (2015). Comparative studies on breast cancer classifications with k-fold cross validations using machine learning techniques. 2015 10th Asian control conference (ASCC),

Nti, I. K., Nyarko-Boateng, O., & Aning, J. (2021). Performance of machine learning algorithms with different K values in K-fold cross-validation. *International Journal of Information Technology and Computer Science*, *13*(6), 61-71.

Ozcift, A., & Gulten, A. (2011). Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms. *Computer methods and programs in biomedicine*, *104*(3), 443-451.

Phirom, K., Charoenkwan, P., Shoombuatong, W., Charoenkwan, P., Sirichotiyakul, S., & Tongsong, T. (2022). DeepThal: A Deep Learning-Based Framework for the Large-Scale Prediction of the α+-Thalassemia Trait Using Red Blood Cell Parameters. *Journal of Clinical Medicine*, *11*(21), 6305.

Sevli, O. (2022). Farklı sınıflandırıcılar ve yeniden örnekleme teknikleri kullanılarak kalp hastalığı teşhisine yönelik karşılaştırmalı bir çalışma. *Journal of Intelligent Systems: Theory and Applications*, *5*(2), 92-105.

Vatansever, B., Aydın, H., & Çetinkaya, A. (2021). Heart Disease Prediction with Machine Learning Algorithm Using Feature Selection by Genetic Algorithm. *Bilim, Teknoloji ve Mühendislik Araştırmaları Dergisi*, *2*(2), 67-80.

Yağmur, N., Temurtaş, H., & İdris, D. (2023). Anemi Hastalığının Yapay Sinir Ağları Yöntemleri Kullanılarak Sınıflandırılması. *Journal of Scientific Reports-B*(008), 20-34.