İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi, 21(42), Güz 2022, 183-193.

İSTANBUL TİCARET ÜNİVERSİTESİ
FEN BİLİMLERİ DERGİSİ

İstanbul Commerce University Journal of Science

http://dergipark.org.tr/ticaretfbd

Araştırma Makalesi / Research Article

KARDİYOVASKÜLER HASTALIK TAHMİNİNDE MAKİNE ÖĞRENMESİ SINIFLANDIRMA ALGORİTMALARININ KARŞILAŞTIRILMASI*

COMPARISON OF THE MACHINE LEARNING CLASSIFICATION ALGORITHMS IN THE CARDIOVASCULAR DISEASE PREDICTION

Gamze KABA[1] **Seda BAĞDATLI KALKAN[2]**

https://doi.org/10.55071/ticaretfbd.1145660

Sorumlu Yazar / Corresponding Author
gamze.arkaba@gmail.com

Geliş Tarihi / Received
19.07.2022

Kabul Tarihi / Accepted
01.09.2022

Öz

Makine öğrenmesi teknikleri, günümüzde birçok alanda kullanılmakta olup veri yığınlarını sınıflandırmaya ve tahmine dayalı analizler ile veriden faydalı bilgiler çıkarmamıza olanak sağlamaktadır. Gelişen teknoloji ile sağlık alanında kayıt altına alınan veri sayısında ciddi artışlar yaşanmaktadır. Sağlık sektöründe oluşan veri yığınlarının makine öğrenmesi yöntemleri ile analiz edilerek yorumlanması, birçok hastalığın erken teşhisinde önem arz etmektedir. Bu çalışmada Kardiyovasküler Hastalığın erken teşhisine katkı sağlamak için makine öğrenmesi algoritmaları ile çalışmada kullanılan veriler üzerinde en başarılı sınıflandırma tahminini yapan algoritmaya ulaşmak hedeflenmiştir. Naive Bayes, Lojistik Regresyon, Rastgele Orman, K-En Yakın Komşu ve Destek Vektör Makineleri olmak üzere beş farklı makine öğrenmesi yöntemi kullanılarak performansları karşılaştırılmıştır. En başarılı performansı veren yöntem tespit edilmiştir. Olası bir kalp hastalığı tahmini üzerine yapılacak olan çalışmalar için makine öğrenmesi algoritmalarından analize uygun yöntem seçiminde fikir vermek amaçlanmıştır. Aynı zamanda, sağlık alanında yapılacak olan benzer çalışmaların güncel tutulması hastalığın erken teşhisine ve tedavisine katkı sağlayabilmektedir.

Anahtar Kelimeler: Kardiyovasküler hastalık, makine öğrenmesi, sınıflandırma algoritmaları.

Abstract

Machine learning techniques are used in many fields nowadays and allow us to classify data piles and extract useful information from data with predictive analysis. With developing technology, there is a significant increase in the number of recorded data in the field of health. With the machine learning methods, analysing and interpreting the data stacks in the field of health is important for the early diagnosis of many diseases. In this study, it is aimed to reach the algorithm that makes the most successful classification prediction on the data used in the study with machine learning algorithms in order to contribute to the early diagnosis of Cardiovascular Disease. The performances were compared by using five different machine learning methods; Naive Bayes, Logistic Regression, Random Forest, K-Nearest Neighbor and Support Vector Machines. The method that gives the most successful estimation performance has been determined. It is aimed to give an idea in the selection of the appropriate method from machine learning algorithms for the studies to be made on the prediction of a possible heart disease. At the same time, keeping up-to-date similar studies in the field of health can contribute to the early diagnosis and treatment of the disease.

Keywords: Cardiovascular disease, classification algorithms, machine learning.

Bu yayın Gamze KABA isimli öğrencinin İstanbul Ticaret Üniversitesi Fen Bilimleri Enstitüsü, İstatistik Programındaki Lisansüstü tezinden üretilmiştir.

[1]İstanbul Ticaret Üniversitesi, Fen Bilimleri Enstitüsü, İstatistik Anabilim Dalı, İstanbul, Türkiye.
gamze.arkaba@gmail.com, Orcid.org/ 0000-0001-5375-0161.

[2]İstanbul Ticaret Üniversitesi, İnsan ve Toplum Bilimleri Fakültesi, İstatistik Bölümü, İstanbul, Türkiye.
sbagdatli@ticaret.edu.tr, Orcid.org/ 0000-0003-3002-2983.

1. GİRİŞ

Günümüzde insanları birçok alanda etkisi altına alan ve hızla gelişen teknoloji ile yaşanan veri birikimi makineler sayesinde depolanmaktadır. Kayıt altına alınan bu veri yığınları çeşitli makine öğrenmesi algoritmaları kullanılarak analiz edilebilir. Verilerin sınıflandırılması ve analiz edilmesi hem insanlara kolaylık sağlamaktadır hem de ileriye dönük güçlü tahminlerde bulunulmasında yardımcı olmaktadır. Makine öğrenmesi algoritmalarının kullanımı, sağlık alanında özellikle kanser ya da kalp krizi gibi bazı ölümcül hastalıkların tahmin çalışmalarında erken teşhis ve iyileşme sürecine erken geçişin sağlanması adına oldukça önemlidir. Ölümcül hastalıklar, birçok farklı faktöre bağlı olarak gelişerek insanların hayatını tehdit altına almaktadır. Makine öğrenmesi algoritmaları hastalıklara etki eden risk faktörlerini değerlendirerek, bireylerin taşıdığı risk faktörlerini tespit eder ve hastalığın erken teşhisini tahmin etmemizi sağlamaktadır.

Uzun yıllardır küresel çapta ölümcül hastalıkların birinci sırasında yer alan hastalık türü olan Kardiyovasküler Hastalıklardan (KVH) her yıl milyonlarca insan hayatını kaybetmektedir. Dünya Sağlık Örgütü (WHO, World Health Organization) verilerine göre hayatını kaybeden insanların çoğunluğunun ölüm nedenleri kalp krizi veya inme gibi ani gelişen hastalıklar ile gerçekleşmiştir (WHO, 2021). Küresel olarak tüm insanları tehdit eden bu hastalığın tahmininde makine öğrenmesi algoritmaları kullanmak, sağlık alanında tedavi çalışmalarının hızlanmasına ve gelişmesine büyük katkı sağlayacaktır.

Bu çalışmada Kardiyovasküler Hastalığın tanımı, önemi ve risk faktörleri ele alınmıştır. Makine öğrenmesi kavramının tarihsel gelişimi, literatürdeki yeri, türleri, aşamaları ve model başarı ölçütleri sırasıyla detaylı olarak açıklanmıştır. Çalışmanın devamında ele alınan verilerin analizinde kullanılacak makine öğrenmesi sınıflandırma algoritmaları teorik ve uygulamalı olarak açıklanmaya çalışılmıştır. Uygulanan yöntemlere ait bulgular değerlendirilerek, başarı performansları karşılaştırılmıştır. Sonuçlar doğrultusunda, bu alanda yapılacak olan çalışmalarda kullanılan yöntemlerin fikir oluşturması ve hastalık tanısı için bir tahmin üretme modeli kurulması amaçlanmıştır.

2. LİTERATÜR İNCELEMESİ

Her yıl küresel olan büyük can kayıplarına sebep olan Kardiyovasküler Hastalıklar üzerine yapılan çalışmaların artması ve güncel kalması sağlık alanında büyük önem arz etmektedir. Hastalıkların teşhisi çalışmalarında makine öğrenmesi sınıflandırma yöntemleri sıkça kullanılmaktadır.

Literatürde, kalp hastalıklarını etkileyen faktörler üzerine yapılan çalışmalar ve kalp hastalığı tahmininde makine öğrenmesi algoritmalarının karşılaştırılmasına dayanan bazı güncel çalışmalar incelenmiştir.

Erkuş (2015) tarafından yapılan "Veri Madenciliği Yöntemleri İle Kardiyovasküler Hastalık Tahmini Yapılması" adlı tez çalışmasında kalp ve damar hastalıklarının erken teşhisine katkı sağlamak adına veri madenciliği yöntemleri kullanarak performans karşılaştırmaları yapmıştır. Çalışmasında kullandığı veri setindeki 604 hastanın 297'si KVH tanısı konulmuş, 307'si KVH tanısı konulmamış hastalardan oluşmaktadır. İstatistiksel gruplama tekniğini kullanarak 3 ayrı veri kümesi oluşturmuştur ve her birinin üzerinde ayrı ayrı değerlendirmeler yapmıştır. Model performansını arttırmak adına nitelik seçme yöntemlerinden yararlanmıştır. Kullandığı sınıflandırıcı veri madenciliği yöntemlerinden Hidden Naive Bayes (HNB) algoritması %84,8 başarı oranı ile en iyi performansı verdiği sonucuna ulaşılmıştır.
Çilhoroz ve Çilhoroz (2021) tarafından yapılan "Kardiyovasküler Hastalıklara Bağlı Ölümleri Etkileyen Faktörlerin Belirlemesi: OECD Ülkeleri Üzerinde Bir Araştırma" adlı çalışmalarında,

Ekonomik Kalkınma ve İşbirliği Örgütü (OECD) veri tabanı üzerinden alınan veriler ile kardiyovasküler hastalıklara bağlı ölümleri etkileyen faktörler incelenmek istenmiştir. Çalışmada En Küçük Kareler (EKK) yöntemi kullanılmıştır. Çalışma sonucunda sigara içme ve alkol kullanımı zararlı alışkanlıklarının KVH'a bağlı ölümler üzerinde pozitif etkileri olduğu tespit edilmiştir.

Kim ve ark. (2021) tarafından hazırlanan "Makine Öğrenimi Tabanlı Kardiyovasküler Hastalık Tahmini Modeli: Kore Ulusal Sağlık Sigortası Hizmeti Verileri Üzerine Bir Kohort Çalışması" adlı çalışmada National Health Insurance Service (Ulusal Sağlık Sigortası Hizmeti, NHIS) sağlık taraması veri setine en iyi tahmin modeli kuran makine öğrenmesi algoritmasını bulmak amacıyla Lojistik Regresyon, K-En Yakın Komşu, Karar Ağaçları, Rastgele Orman, Ekstra Ağaçlar, XGBoosting, Gradyan Arttırma, AdaBoost, Destek Vektör Makineleri ve Çok Katmanlı Algılayıcılar yöntemlerini uygulamışlardır. Çıkan sonuçlara göre tüm tahmin modelleri sonuçları performans ölçütlerine göre karşılaştırmışlardır ve XGBoosting, Gradient Boosting ve Rastgele Orman algoritmaları ile en iyi tahmin modeli kurulduğu sonucuna varılmıştır.

3. MAKİNE ÖĞRENMESİ SINIFLANDIRMA YÖNTEMLERİ

Sınıflandırma yöntemleri denetimli makine öğrenmesi yöntemlerinden olup, veri seti içerisinde kategorik yapıda bağımlı değişkenin bulunduğu problemlerde kullanılabilmektedir. Sınıflandırma problemlerinde kullanılabilecek birçok sınıflandırma algoritmasının bulunması ve problem üzerinde farklı tekniklerin test edilerek karşılaştırılması ile doğruluk oranı en yüksek sonuca ulaşılabilir. Bu amaç doğrultusunda çalışma kapsamında sınıflandırma algoritmalarından Lojistik Regresyon, Naive Bayes, Rastgele Orman, K-En Yakın Komşu ve Destek Vektör Makineleri olmak üzere 5 farklı yöntem uygulanmış ve performans karşılaştırılmaları yapılmıştır. Çalışmada kullanılan algoritmalar kısaca açıklanmıştır.

3.1. Lojistik Regresyon Analizi

Lojistik regresyon analizinde amaç, her bir girdi değişkeninin çıktı değişkenindeki hangi gruba atanacağını belirlemek için doğrusal bir regresyon modeli oluşturmaktır. Algoritma sınıflandırma işlemini gerçekleştirirken bağımlı ve bağımsız değişkenler arasındaki ilişkiyi inceler. Lojistik regresyon analizinin uygulanabilmesi için bağımlı değişkenin kesikli ve kategorik yapıda olması gerekmektedir.

Lojistik regresyon analizinde kullanılan yöntemler bağımlı değişkenin yapısına göre farklılık gösterir. Bu yöntemler; bağımlı değişken iki kategorili yapıdan oluşuyorsa İkili (Binary) Lojistik Regresyon Analizi, ikiden fazla sıralı kategoriden oluşuyorsa Sıralı (Ordinal) Lojistik Regresyon Analizi, ikiden fazla sırasız kategoriden oluşuyorsa Çok Değişkenli (Multinominal) Lojistik Regresyon Analizi olarak gruplandırılmaktadır (Alpar, 2020).

Lojistik dağılım fonksiyonunda bağımlı değişken 0 ile 1 arasında değer alırken, bağımsız değişkenler '-∞ ile +∞' arasında değerler alabilir. Lojistik regresyon analizinde sınıflama yaparken kullanılacak olan logit model fonksiyonu başlangıçta doğrusal bir yapıda değildir. Denklemi doğrusal bir yapıya getirmek için gerekli logit dönüşümler uygulanarak doğrusal model olan Logit Modele şu şekilde ulaşılmaktadır (Güriş, 2019).

$$L_i = \beta_0 + \beta_1 X_i \quad (1)$$

3.2. Naive Bayes Algoritması

Naive Bayes algoritması, denetimli makine öğrenmesi yöntemlerinden olup bir olayın gerçekleşme olasılığı ve belli bir sınıfta bulunması durumunu koşullu olasılık yöntemi ile hesaplar. Naive bayes algoritması, bağımlı değişken kategorisi fazla olan büyük veri setlerinde daha başarılı sonuçlar verebilmektedir.

Naive bayes sınıflandırıcı algoritması adını İngiliz matematikçi olan Thomas Bayes tarafından almıştır. Bu yöntem, basit bir işleyişe sahip olduğu için ve de düşük hata oranları verdiği için günümüzde sınıflama problemlerinde fazlasıyla tercih edilmektedir. Sınıflanmamış verilerde olasılıksal tahminlerde bulunarak, verileri bulunan sınıflar arasında ait olma olasılığı en yüksek olan sınıfa atama işlemini gerçekleştirmektedir (Doğan, 2015).

Naive Bayes teoremi kısaca özetlenmiştir (Han ve ark., 2011):

n boyutlu bağımsız değişkenler kümesi $X = (x_1, x_2, \dots, x_n)$ ve m sınıflı bağımlı değişkenler $C = (C_1, C_2, \dots, C_m)$ olsun. Sınıfları henüz belirlenmemiş bağımsız değişkenler üzerinde yapılan olasılık tahmin değerleri sırasıyla $A_1, A_2, \dots, A_n$. Böylece sınıfları bilinmeyen bir X veri seti verildiğinde X' in X koşuluna bağlı olarak en yüksek sonsal olasılığa sahip C sınıflarından birisine ait olduğunu tahmin eder. $P(C_i|X) > P(C_j|X)$ ise $1 \leq j \leq m, j \neq i$ *olduğunda* $P(C_i|X)$ maksime edilmiş olur ve maksimum sonsal hipotez olarak adlandırılarak denklemi şu şekilde gösterilmektedir (Han ve ark., 2011).

$$P(C_i|X) = \frac{P(Ci)P(X|Ci)}{P(X)} \quad (2)$$

$P\ (X)$ değerleri tüm sınıflar için sabit olduğundan sadece $P\ (X|C_i)\ P\ (C_i)$ maksimize edilmeye ihtiyaç duyar. Sınıfın önceki olasılıkları bilinmediği zamanlarda, genel olarak eşit derecede olası oldukları, yani $P\ (C_1) = P\ (C_2) = \cdots = P\ (C_m)$ olduğu varsayılır ve $P\ (X|C_i)$ maksimize edilir. Nitelik sayısı fazla olduğunda $P\ (X|C_i)$ hesabı zor olacağı için sınıf koşullu bağımsızlık varsayımı yapılır. Bu varsayımda sınıf etiketleri verilerek, girdi değişkenleri birbirinden koşullu olarak bağımsız varsayılır ve şu şekilde hesaplanmaktadır (Han ve ark., 2011).

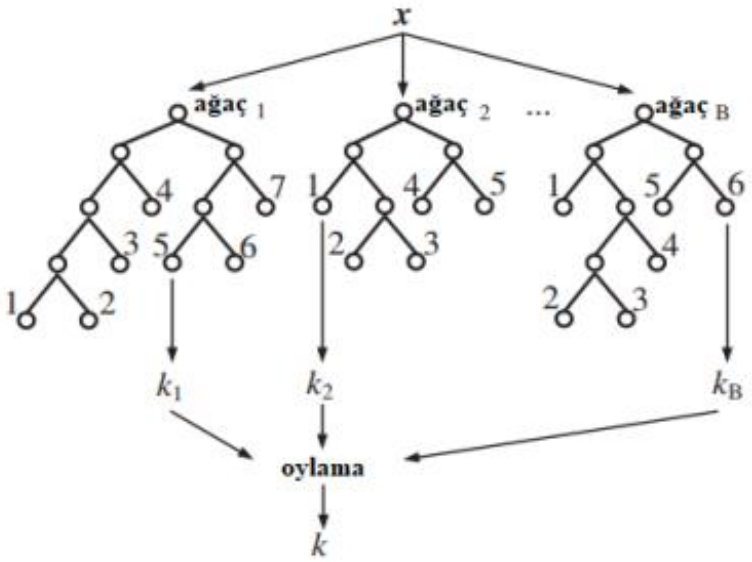
$$P(X|Ci) = \prod_{k=1}^{n} P(x_k|Ci) \quad (3)$$

İşlemlerin devamında bağımsız değişkenlerin kategorik değişken ya da sürekli değişken olma durumuna bakıldıktan sonra gerekli hesaplamalar yapılmaktadır (Han ve ark., 2011).

3.3. Rastgele Orman (Random Forest) Algoritması

Rastgele Orman (Random Forest) algoritması, birden fazla karar ağacı oluşturarak bir orman meydana getirir ve her ağacın sonuçlarını oylama (voting) yoluyla eleyerek ya da ortalamalarını alarak çözümler (Breiman, 2001). Bu yöntemde dallanan her bir ağaç klasik karar ağaçlarına göre daha az doğru olma eğilimindedir. Bu sebepten dolayı, dallanan her bir ağaç tahmininin toplu tahmini elde edilerek daha güçlü tahmin sonuçlarına ulaşmamız sağlanır. Tek bir ağacın tahmini, eğitim setinde küçük bir değişiklik yapılmış olsa dahi çok hassas yapıda olduğu için durumdan fazlasıyla etkilenebilmektedir. Rastgele ormandaki her ağaç tek başına zayıf bir öğrenciyken, rastgele orman güçlü bir öğrenci olabilmektedir (Lewis, 2017).

Rastgele orman algoritması ağaç yapısı Şekil 1'de verilmiştir.



Şekil 1. Rastgele orman algoritması ağaç yapısı (Englund & Verikas, 2012; Cihan, 2018).

Rastgele orman algoritmasında öncelikle veri setinden örneklemin 2/3'ü eğitim, 1/3'ü test verisi olarak ayırmak için bootstrap yöntemi ile n tane örnekleme çekilir. Düğümler arasında rastgele olarak m tane tahmine dayalı değişken seçilir ve aralarından en iyi dallanmayı verecek olan belirlenir. En iyi dallanmayı verecek olan değişken belirlenirken Gini indeksi ile hesaplama yapılarak, ulaşılan değere göre veri seti her düğümde iki alt dala ayrılmaktadır. Tüm işlemler mümkün olan en geniş dallanma sayesinde yaprak düğüm elde edilinceye kadar tekrarlanarak, bütün ağaçların ayrı ayrı tahminleri birleştirilmektedir. Sınıflama ağaçları için en çok oyu alan tahmin değeri kabul edilirken, regresyon ağaçları için oylamaların ortalaması alınarak tahmin değeri elde edilmektedir (Akman ve ark., 2011).

3.4. K-En Yakın Komşu Algoritması

K-en Yakın Komşu (KNN) algoritması birbirine yakın olan gözlemlerin benzer olacağı düşüncesine dayanan parametrik olmayan bir sınıflandırma algoritmasıdır. Sınıfı henüz belirlenmemiş bir gözlem değerinin diğer her bir gözleme olan uzaklığına bakılarak k ağırlık merkezli olan bir alan belirlenmektedir. Bu alan içerisinde sınıf kategorilerinden hangisi daha fazla gözlemlenebiliyorsa, gözlem değerinin o sınıfa ait olduğu tahmin edilmektedir.

KNN algoritmasında iki gözlem değeri arasındaki uzaklık ölçütünü belirlemek için genellikle Öklid yöntemini kullanmaktadır ve 4. eşitlikteki formülasyon ile hesaplanmaktadır (Altuncu, 2021).

$$\mathrm{d}(X_i, X_k) = \sqrt{\sum_{j=1}^{D}(X_i^{(j)} - X_k^{j})^2} \quad (4)$$

Öncelikle komşu sayısı olan k ve uzaklık ölçüsü belirlenmektedir. Daha sonra sınıfı belirlenmek istenen verinin k-en yakın komşuları bulunarak, sınıf etiketine oy çokluğu ile karar verilmektedir (Altuncu, 2021).

K-en yakın komşu algoritması, veri seti içerisinde aykırı gözlem ve birbiri ile uyuşmayan nitelikler var olduğunda doğruluktan uzaklaşarak düşük başarı oranı vermektedir. Yöntemin bu açıdan düşük performans sergilemesinin sebebi, her özniteliğe göre eşit ağırlık atayan mesafeye dayalı yaklaşım tekniği kullanmasıdır. KNN yöntemi, bu sorunu çözmek için algoritmaya öznitelik ağırlıklandırılması ve aykırı gözlem budanması sağlanarak geliştirilmiştir (Han ve ark., 2011).

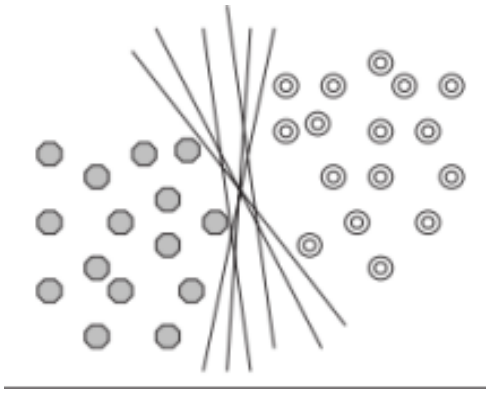
3.5. Destek Vektör Makineleri (Support Vector Machines) Algoritması

Destek Vektör Makineleri (SVM), regresyon ve sınıflandırma problemlerinde sıkça kullanılan denetimli makine öğrenmesi algoritmasıdır. SVM algoritmasının ilk kullanım alanı sınıflandırma problemleri olmuştur ve bu yöntemin amacı doğrusal bir hiper düzlem oluşturarak sınıfların ayrımını analize en uygun şekilde sağlamaktır. Bazı veri setlerinde sınıfların ayrımı hiper düzlem tekniğiyle doğrusal olarak ayrılmayabilmektedir. Veri setine farklı yaklaşım teknikleriyle bu problem çözümlenebilmektedir.

Destek vektör makineleri yönteminde, doğrusal olarak ayrım yapılabilen veri setlerinde her bir X_i girdi değişkeni Y_i sınıf etiketine sahip olur ve bu sınıflar -1 ve +1 olarak etiketlenir. Eğitim verilerinin ayrımını, bir hiper düzlem oluşturarak en optimum şekilde ayırmak hedeflenmektedir. Değişkenler arasındaki sınırı maksimum noktaya ulaştıran düzlemlere destek vektörler denir. Doğrusal olarak ayrılabilen verilerde karar fonksiyonu 5. eşitlikteki fonksiyon ile hesaplanır (Kavzoğlu & Çölkesen, 2010).

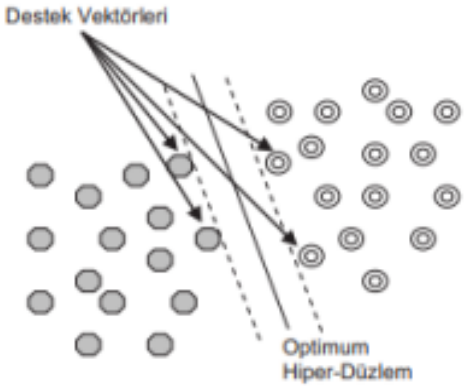
$$f_{(x)} = sign(\sum_{i=1}^{k} \lambda_i \, y_i (x.x_i) \; + b) \tag{5}$$

İki sınıflı bir problem için hiper düzlem örneği Şekil 2'de verilmiştir.



Şekil 2. İki sınıflı bir problem için hiper düzlemler (Kavzoğlu & Çölkesen, 2010).

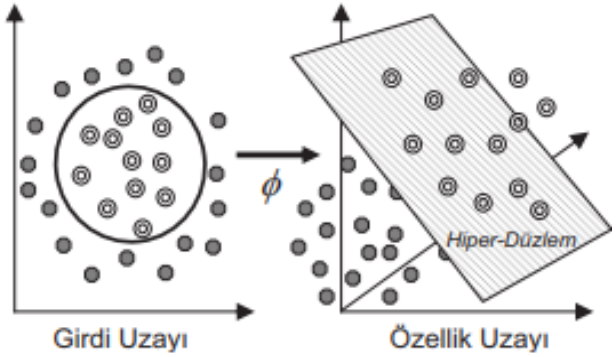
Şekil 2'de görüldüğü gibi iki sınıflı eğitim verilerini en uygun şekilde ayırmak için hiper düzlemler kullanılmaktadır. Böylece iki sınıf birbirlerinden olabildiğince uzakta kalabilmektedir. Hiper düzleme en yakın verilerin arasındaki uzaklık marjin olarak adlandırılır ve marjin mesafesi optimum olana kadar çalışma devam eder. Hiper düzlem ile ayrılan iki ayrı sınıflı etiketli veriler Şekil 3'te olduğu gibi destek vektörler ile sınırlandırılmaktadır.



Şekil 3. Optimum hiper düzlem ve destek vektörleri (Kavzoğlu & Çölkesen, 2010).

Bazı problemlerde veri seti içerisindeki verilerin dağılımından ya da verilerin yüksek ilişkili olması gibi durumlardan kaynaklı verilerin doğrusal olarak ayrılması mümkün olmamaktadır. Bu problemler genellikle veride boyut arttırma işlemi ile çözülebilmektedir.

Şekil 4'te girdi uzayında veride doğrusal ayrılma sağlanamadığı için veri özellik uzayı olarak adlandırılan daha yüksek boyutlu bir uzaya taşınmıştır. Böylece hiper düzlem ile veri sınıfları arasındaki optimum ayrım sağlanmış olur ve bu işlemi gerçekleştirmek için Kernel fonksiyonlarından yararlanılmaktadır (Kavzoğlu & Çölkesen, 2010).



Şekil 4. Kernel fonksiyonu ile verinin daha yüksek bir boyuta dönüştürülmesi (Kavzoğlu & Çölkesen, 2010).

Destek vektör makineleri çalışmalarında, veri boyutunun çok fazla olması eğitim süresini ve maliyeti arttırabilmektedir. Bu bağlamda, destek vektör makinelerinin kullanımının küçük boyutlu verilerde tercih edilmesi daha iyi olabilmektedir (Ergün & İlhan, 2021).

3.6. Model Performansı Karşılaştırma Ölçütleri

Çalışma kapsamında kullanılan veri seti, farklı makine öğrenmesi algoritmaları ile test edildikten sonra kurulan modellerin başarı performansları değerlendirilirken bazı istatistiksel ölçütlere bakılmaktadır. Model performans karşılaştırma aşamasında, tahmin değerleri ile gerçek değerler olan test değerleri karşılaştırılmalıdır. Sınıflandırma problemlerinde Karışıklık Matrisi (Confusion Matrix) yani hata matrisine bakılması gerekmektedir. Örnek bir karışıklık matrisine Tablo 1'de yer verilmiştir.

Tablo 1. Örnek karışıklık matrisi

Karışıklık Matrisi	**Hasta (Gerçek Sınıf)**	**Hasta Değil (Gerçek Sınıf)**
Hasta (Tahmini Sınıf)	DP	YP
Hasta Değil (Tahmini Sınıf)	YN	DN

Tablo 1'i incelediğimizde DP hasta olarak tahmin ettiğimiz ve gerçekte de hasta olarak sınıflandırılan bireyleri, YP hasta olarak tahmin ettiğimiz fakat gerçekte hasta olmayan bireyleri, YN hasta değil olarak tahmin ettiğimiz fakat gerçekte hasta olarak sınıflandırılan bireyleri, DN hasta değil olarak tahmin ettiğimiz ve gerçekte de hasta değil olarak sınıflandırılan bireyleri temsil etmektedir.

Buradan yola çıkarak model performans ölçütleri aşağıda açıklanmaktadır.

Doğruluk ölçütü, modelde doğru sınıflandırılan veri sayısının toplam veri sayısına oranı ile bulunur.

$$\text{Doğruluk} = \frac{DP+DN}{DP+DN+YP+YN} \quad (6)$$

Kesinlik ölçütü, modelde doğru olarak sınıflandırılan verilerin oranının bulunmasını sağlar.

$$\text{Kesinlik} = \frac{DP}{DP+YP} \quad (7)$$

Duyarlılık ölçütü, pozitif değerlerin doğru sınıflandırılan değerlerinin bulunması ile belirlenir.

$$\text{Duyarlılık} = \frac{DP}{DP+YN} \quad (8)$$

Duyarlılık ve kesinlik ölçütlerinin tutarsız olduğu durumlarda iki ölçütün harmonik ortalaması alındığında F1 skor ölçütü elde edilir. Bu sayede dengesiz veri setlerinde F1 skorlama ile daha güçlü bir ölçüt değerine ulaşılmaktadır ve şu şekilde hesaplanmaktadır (Uğuz, 2021).

$$\text{F1 Skor} = 2\frac{Duyarlılık \times Kesinlik}{Duyarlılık+Kesinlik} \quad (9)$$

4. UYGULAMA

Bu çalışma kapsamında kullanılan veri seti, kaggle.com üzerinden alınan Kalp Yetmezliği Tahmin Veri Seti (2021), "UCI Machine Learning Repository" veri tabanına kaynağına aittir. Bu veri seti, 5 farklı veri kümelerinden 11 ortak özellik altında birleştirilerek oluşturulmuş hazır veridir. Veri kaynakları; Cleveland Klinik Şirketi (Ph. D. Robert Detrano, Amerika), Macaristan Kardiyoloji Enstitüsü (Andras Janosi, Budapeşte), Zürih Üniversite Hastanesi (William Steinburnn, İsviçre) ve Basel Üniversite Hastanesi (Matthias Pfisterer, İsviçre) tarafından oluşturulmuştur. Veri seti 918 hastaya ait bilgileri içermektedir. Veri setindeki bağımsız değişkenlerden 6'sı kategorik, 5'i nümerik yapıdan oluşmaktadır. Bağımlı değişken olan Kalp Hastalığı ise (0,1) sınıflı olan kategorik yapılı bir değişken türüdür. Bağımlı değişken kategorisindeki 0 kalp hastası olmama durumunu gösterirken, 1 kalp hastası olma durumunu belirtmektedir. Veri seti 28-77 yaş aralığındaki bireylerden oluşmaktadır.

Bu çalışmanın uygulama bölümünde, günümüz programlama dillerinde kullanımı popüler olan Python 3.7 programlama dili kullanılmış ve Jupyter Notebook 6.1.4 üzerinden analiz edilmiştir. Algoritmalar Sckit-Learn (Sklearn) Kütüphanesi kullanılarak çalıştırılmış ve modeller kurulmuştur.

Veri seti analize başlamadan önce veri ön işleme aşaması kuralları takibinde çalıştırılmaya hazır hale getirilmiştir. Veri seti Holdout yöntemi kullanılarak %70'e %30 olacak şekilde eğitim verisi ve test verisi olarak ikiye ayrılmıştır. Eğitim veri setine K-Katlı Çapraz Doğrulama yöntemi ile 10 katlı çapraz doğrulama uygulanmıştır. Sınıflandırma yöntemlerinden Destek Vektör Makineleri algoritmasının uygulanma sürecinde doğrusal çekirdek (linear kernel) fonksiyonu kullanılmıştır. Algoritmalar çalıştırılarak elde edilen sonuçlar Tablo 2'de karşılaştırılmış ve gösterilmiştir.

Tablo 2. Sınıflama Algoritmalarının Karşılaştırılması

Modeller	Kesinlik	Duyarlılık	F1 Skoru	Doğruluk
Rastgele Orman	0.8750	0.9089	0.8908	0.8738
Lojistik Regresyon	0.8765	0.8923	0.8836	0.8676
Naive Bayes	0.8816	0.8815	0.8811	0.8660
K-En Yakın Komşu	0.7089	0.7358	0.7216	0.6793
SVM	0.7321	0.7481	0.6935	0.6618

Tablo 2'de makine öğrenmesi Rastgele Orman, Lojistik Regresyon, Naive Bayes, K-En Yakın Komşu ve Destek Vektör Makineleri sınıflandırma algoritmalarının tahmin performans sonuçları verilmiştir.

Genel olarak duyarlılık ve kesinlik oranları birbirine yakın değerler olduğu için herhangi bir tutarsızlık gözlenmemiştir. Elde edilen performans sonuçlarına göre Rastgele Orman algoritması ile bireylerin hastalık durumu %87,38 doğruluk ve %87,50 kesinlik oranı ile doğru tahmin edilmiştir ve böylece algoritmalar arasında en başarılı sınıflandırmayı Rastgele Orman algoritması yapmıştır.

Lojistik Regresyon analizi performansı %86,76 doğruluk oranı ile Rastgele Orman algoritmasından sonra ikinci başarılı sonucu veren algoritma olmuştur. Naive Bayes algoritması bireylerin hastalık durumunu %86,60 oranı ile başarılı olarak tahmin etmiştir. K-En Yakın Komşu algoritması sonucuna baktığımızda başarı oranında ciddi bir düşüş yaşanarak %67,93 olarak hesaplanmıştır. Son olarak Destek Vektör Makineleri diğer algoritmalara göre en düşük performans oranı olan %66,18 olarak hesaplanmıştır.

5. SONUÇLAR

Günümüzde gelişen teknoloji ve artan nüfus ile sağlık alanında kayıt altına alınan veriler çok büyük veri yığınları oluşturmaktadır. Kayıt altına alınan bu verilerin güvenilir bir şekilde depolanması ve bazı istatistiksel yöntemler ile analiz edilmesi sağlık alanında atılacak adımlarda ve izlenecek süreçlerde ciddi faydalar sağlayacağı için son derece önemlidir. Bu alanda yapılacak olan çalışmaların güncel tutulması hastalar için ileriye dönük tahminlerde bulunulmasına ve erken tedavi sürecine girilmesinde fayda sağlayabilmektedir.

Bu çalışma kapsamında kullanılan veriler Kaggle platformu üzerinden alınan UCI tabanlı gerçek verilerden oluşan Kalp Hastalığı veri setidir. Veri seti 28-77 yaş aralığında olan 725 erkek, 193 kadın bireylerden oluşan 918 gözlem değerindeki hastalardan oluşmaktadır. Çalışmada ele alınan kalp hastalığı tahmini bir sınıflandırma problemi olarak değerlendirilerek bu doğrultuda algoritmalar kullanılmıştır.

Çalışmada denetimli makine öğrenmesi sınıflandırma algoritmalarından olan Lojistik Regresyon, Naive Bayes, Rastgele Orman, K-En Yakın Komşu ve Destek Vektör Makineleri kullanılmıştır. Algoritmalar uygulanmadan önce veri seti ön işleme aşaması kuralları doğrultusunda Holdout yöntemi ile %70'e %30 olacak şekilde eğitim ve test verisi olarak ayrılmıştır. Ayrıca, eğitim veri setine K-Katlı Çapraz Doğrulama yöntemi ile 10 katlı çapraz doğrulama uygulanmıştır. Eğitim veri seti üzerinden algoritmalar çalıştırılarak kurulan modeller test verisi ile sınanmıştır.

Tüm algoritma sonuçları model başarı ölçütlerine göre değerlendirilerek, en başarılı modele ulaşmak hedeflenmiştir. Böylece en yüksek başarı performansını %87,38 doğruluk oranı ile Rastgele Orman algoritması verirken, en düşük başarı performansını %66,18 doğruluk oranı ile

Destek Vektör Makineleri vermiştir. Rastgele Orman algoritması kullanılarak %87,38 tahmin oranı ile güçlü bir performans yakalanmıştır. Bu doğrultuda, sağlık alanında yapılacak olan hastalık teşhisi çalışmalarında Rastgele Orman algoritması tercih edildiğinde başarılı bir tahmin modeli oluşturulacaktır.

Algoritmaların performanslarına bakıldığında, çalışmada kullanılan veri seti doğrultusunda kalp hastalıklarının tahmininde makine öğrenmesi sınıflandırma algoritmalarının başarılı sonuçlar verebileceği öngörülmüştür. Bu çalışmada, makine öğrenmesi yöntemleri kullanılarak yapılacak olan benzer çalışmalarda fikir oluşturması ve hastalık tanısı için başarılı bir model kurulması amaçlanmıştır. Kullanılan makine öğrenmesi algoritmalarından olan Rastgele Orman algoritması %87,38 tahmin oranı ile oldukça güçlü başarı oranı yakalamıştır ve hastalık tahmini çalışmalarında tercih edilmesi gereken bir algoritma olarak belirlenmiştir. Kardiyovasküler Hastalıklar sebebiyle yaşanan ölüm oranlarında her geçen gün ciddi artışlar yaşanmaktadır ve bu alanda yapılacak olan çalışmalarda veri seti boyutu arttırılarak analizlere devam edildiğinde yüksek başarı performansları yakalanabilir.

Yazarların Katkısı
Yazarların makaleye katkıları eşit orandadır.

Çatışma Beyanı
Yazarlar arasında herhangi bir çıkar çatışması bulunmamaktadır.

Araştırma ve Yayın Etiği Beyanı
Yapılan çalışmada araştırma ve yayın etiğine uyulmuştur.

KAYNAKÇA

Akman, M., Genç, Y. & Ankaralı, H. (2011). Random forests yöntemi ve sağlık alanında bir uygulama. *Türkiye Klinikleri Journal of Biostatistics.* 3(1), 36-48.

Alpar, R. (2020). Uygulamalı çok değişkenli istatistiksel yöntemler. *Detay Yayıncılık,* Ankara.

Altuncu, M. A. (2021). *Makine öğrenmesi ve derin öğrenme yöntemleri kullanılarak saldırı tespit ve önleme sistemi geliştirilmesi* [Doktora Tezi]. Kocaeli Üniversitesi Fen Bilimleri Enstitüsü. Kocaeli.

Breiman, L. (2001). Random forests. *Machine Learning*. 45(1), 5-32.

Cihan, Ş. (2018). *Koroner arter hastalığı riskinin makine öğrenmesi ile analiz edilmesi* [Yüksek Lisans Tezi]. Kırıkkale Üniversitesi Fen Bilimleri Enstitüsü. Kırıkkale.

Çilhoroz, İ. A., & Çilhoroz, Y. (2021). Kardiyovasküler hastalıklara bağlı ölümleri etkileyen faktörlerin belirlemesi: OECD ülkeleri üzerinde bir araştırma. *Acıbadem Üniversitesi Sağlık Bilimleri Dergisi.* 12(2), 340-345.

Doğan, A. (2015). *Bireysel araç kredilerinin yasal takibe girme durumları hakkında tahmin modellerinin oluşturulması* [Yüksek Lisans Tezi]. Bahçeşehir Üniversitesi Fen Bilimleri Enstitüsü. İstanbul.

Englund, C., & Verikas, A. (2012). A novel approach to estimate proximity in a random forest: An exploratory study. *Expert Systems with Applications.* 39 (17), 13046-13050.

Ergün, Ö. N., & İlhan, H. O. (2021). Early stage diabetes prediction using machine learning methods. *Avrupa Bilim ve Teknoloji Dergisi.* (29), 52-57.

Erkuş, S. (2015). *Veri madenciliği yöntemleri ile kardiyovasküler hastalık tahmini yapılması* [Yüksek Lisans Tezi]. Bahçeşehir Üniversitesi Fen Bilimleri Enstitüsü. İstanbul.

Güriş, S., &Astar, M. (2019). SPSS ile istatistik. *Der Yayınları,* İstanbul.

Han, J., Kamber, M., & Pei, J. (2011). Data mining concepts and techniques. *Morgan Kaufmann Publication*, USA.

Kavzoğlu, T., & Çölkesen, İ. (2010). Destek vektör makineleri ile uydu görüntülerinin sınıflandırılmasında kernel fonksiyonlarının etkilerinin incelenmesi. *Harita Dergisi.* 144(7), 73-82.

Kim, J. O., Jeong, Y. S., Kim, J. H., Lee, J. W., Park, D., & Kim, H. S. (2021). machine learning-based cardiovascular disease prediction model: A cohort study on the Korean national health ınsurance service health screening database. *Diagnostics.* 11(6), 943.

Lewis, N. D. C. (2017). Machine learning made easy with R: An ıntuitive step by step blueprint for beginners. *CreateSpace Independent Publishing Platform*, USA.

Uğuz, S. (2021). Makine öğrenmesi teorik yönleri ve python uygulamaları ile bir yapay zeka ekolü. *Nobel Akademik Yayıncılık*, Ankara.

Who Health Organization (WHO) (2021). Cardiovascular Diseases. https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1 adresinden 11 Haziran 2021 tarihinde alınmıştır.