T.C.
İSTANBUL BEYKENT ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
BİLGİSAYAR MÜHENDİSLİĞİ BİLİM DALI

REGRESYON YÖNTEMLERİNİN PERFORMANS ANALİZİ

Yüksek Lisans Tezi

Tezi Hazırlayan
Emre ÖZKAN

İstanbul, 2024

T.C.
İSTANBUL BEYKENT ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
BİLGİSAYAR MÜHENDİSLİĞİ BİLİM DALI

REGRESYON YÖNTEMLERİNİN PERFORMANS ANALİZİ

Yüksek Lisans Tezi

Tezi Hazırlayan
Emre ÖZKAN

Öğrenci No
2120003036

ORCID ID
0009-0006-2965-1212

Danışman
Dr. Öğr. Üyesi Talat FİRLAR

İstanbul, 2024

YEMİN METNİ

Yüksek Lisans Tezi olarak sunduğum "**Regresyon Yöntemlerinin Performans Analizi**" başlıklı bu çalışmanın, bilimsel ahlak ve geleneklere uygun şekilde tarafımdan yazıldığını, bu tezdeki bütün bilgileri akademik ve etik kurallar içinde elde ettiğimi, yararlandığım eserlerin tamamının kaynaklarda gösterildiğini ve çalışmamın içinde kullanıldıkları her yerde bunlara atıf yapıldığını, patent ve telif haklarını ihlal edici bir davranışımın olmadığını belirtir ve bunu onurumla doğrularım.
21.10.2024

Emre ÖZKAN

T.C.
İSTANBUL BEYKENT ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ MÜDÜRLÜĞÜ
TEZLİ YÜKSEK LİSANS SINAV TUTANAĞI

21/10/2024

Enstitümüz ***Bilgisayar Mühendisliği*** Anabilim Dalı ***Bilgisayar Mühendisliği*** Programı yüksek lisans öğrencilerinden ***2120003036*** numaralı ***Emre ÖZKAN'ın*** *"İstanbul Beykent Üniversitesi Lisansüstü Eğitim – Öğretim Yönetmeliği"*nin ilgili maddesine göre hazırlayarak, Enstitümüze teslim ettiği ***"Regresyon Yöntemlerinin Performans Analizi"*** konulu tezini, Yönetim Kurulumuzun 28/05/2024 tarih ve 2024/24 sayılı toplantısında seçilen ve On-Line toplanan biz jüri üyeleri huzurunda, İstanbul Beykent Üniversitesi Lisansüstü Eğitim ve Öğretim Yönetmeliğinin 29. maddesinin 3. fıkrası gereğince 45 dakika süre ile Zoom programı aracılığıyla on-line olarak aday tarafından savunulmuş ve sonuçta adayın tezi hakkında ***"OYBİRLİĞİ"*** ile ***"KABUL"*** kararı verilmiştir.

İşbu tutanak, 2 nüsha olarak hazırlanmış ve Enstitü Müdürlüğü'ne sunulmak üzere tarafımızdan düzenlenmiştir.

DANIŞMAN
Dr. Öğr. Üyesi Ta*** Fİ***
(İstanbul Beykent Üniversitesi)

ÜYE
Dr. Öğr. Üyesi Ke*** Gö*** NA***
(İstanbul Beykent Üniversitesi)

ÜYE
Dr. Öğr. Üyesi Ke*** ŞA***
(İstanbul Medipol Üniversitesi)

Adı ve Soyadı	:	Emre ÖZKAN
Danışmanı	:	Dr. Öğr. Üyesi Talat FİRLAR
Derecesi ve Tarihi	:	Yüksek Lisans (Tezli), 2024
Alanı	:	Bilgisayar Mühendisliği
Anahtar Kelimeler	:	Kalp krizi, regresyon analizi, makine öğrenimi, tıbbi veri analizi, performans değerlendirme

ÖZ

REGRESYON YÖNTEMLERİNİN PERFORMANS ANALİZİ

Bu yüksek lisans tezinde, kalp krizi verileri üzerinde çeşitli regresyon yöntemlerinin performans analizi yapılmıştır. Lineer regresyon, lojistik regresyon, ridge regresyon ve destek vektör regresyonu gibi yöntemler kullanılarak, kalp krizi olasılığını tahmin etme ve analiz etme üzerine odaklanılmıştır. Çalışmanın amacı, her bir regresyon yönteminin performansını değerlendirerek, kalp krizi ile ilişkilendirilen faktörleri en doğru şekilde tahmin eden optimal yöntemi belirlemektir.

Tezin birinci bölümünde, tıbbi veri analizi ve kalp krizi tahmini ile ilgili mevcut literatür incelenmiştir. İkinci bölümde, kullanılan kalp krizi veri seti hakkında detaylı bilgiler verilmiş ve veri bütünlüğünü sağlamak için yapılan ön işleme adımları açıklanmıştır. Üçüncü bölümde, regresyon yöntemlerinin teorik temelleri ve bu yöntemlerin kalp krizi veri setine uyarlanması ele alınmıştır. Dördüncü bölümde, regresyon yöntemlerinin performansını değerlendirmek için kullanılan doğruluk, hassasiyet ve geri çağırma gibi performans metrikleri açıklanmıştır.

Beşinci bölümde, farklı regresyon yöntemlerinin kalp krizi verilerine uygulanmasından elde edilen sonuçlar sunulmuş ve her bir yöntemin güçlü ve zayıf yönleri vurgulanmıştır. Altıncı bölümde, bulguların analizi yapılarak, belirlenen optimal regresyon yönteminin gerçek dünya senaryolarındaki potansiyel uygulamaları tartışılmıştır. Son bölümde ise, ana bulgular özetlenmiş, çalışmanın sınırlamaları ve gelecekteki araştırmalar için öneriler sunulmuştur.

Bu çalışma, tıbbi alanda regresyon analizinin uygulanabilirliğini artırmayı ve kalp krizi tahmininde en etkili yöntemi belirlemeyi amaçlamaktadır.

Name and Surname	:	Emre ÖZKAN
Supervisor	:	Asst. Prof. Dr. Talat FİRLAR
Degree and Date	:	Master's (Thesis), 2024
Major	:	Computer Engineering
Keywords	:	Heart attack, regression analysis, machine learning, medical data analysis, performance evaluation

ABSTRACT

PERFORMANCE ANALYSIS OF REGRESSION METHODS

This master's thesis conducted a performance analysis of various regression methods on heart attack data. Methods such as linear regression, logistic regression, ridge regression, and support vector regression were employed to predict and analyze the likelihood of a heart attack. The aim of the study was to evaluate the performance of each regression method to identify the optimal one that most accurately predicts factors associated with heart attacks.

In the first section, a comprehensive review of the existing literature on medical data analysis and heart attack prediction was presented. The second section provided detailed information about the heart attack dataset used and explained the preprocessing steps taken to ensure data integrity. The third section focused on the theoretical foundations of the regression methods and their application to the heart attack dataset. The fourth section explained the performance metrics used to evaluate the regression methods, such as accuracy, precision, and recall.

The fifth section presented the results obtained from applying different regression methods to the heart attack data, highlighting the strengths and weaknesses of each method. In the sixth section, the findings were critically analyzed, discussing the potential applications of the identified optimal regression method in real-world scenarios. The final section summarized the main findings, outlined the limitations of the study, and provided suggestions for future research.

This study aims to enhance the applicability of regression analysis in the medical field and provide valuable insights to identify the most effective regression method for heart attack prediction.

İÇİNDEKİLER

Sayfa No.

TABLOLAR LİSTESİ

Sayfa No.

ŞEKİLLER LİSTESİ

Sayfa No.

KISALTMALAR

AMI : Akut Miyokard Enfarktüsü

Big Data : Büyük Veri

BNC : Bayesian Ağ Sınıflandırıcısı (Bayesian Network Classifier)

CK-MB : Kreatin Kinaz-MB

CVD : Kardiyovasküler Hastalık

EKG : Elektrokardiyografi

GUI : Grafiksel Kullanıcı Arayüzü

HDL : Yüksek Yoğunluklu Lipoprotein

KNN : En Yakın Komşu Algoritması

LDL : Düşük Yoğunluklu Lipoprotein

LM NN : Lojistik Model Yapay Sinir Ağı

OGM : Olasılıksal Grafik Modeli

SCG NN : Yapılandırılmış Ağ Yapay Sinir Ağı

SVM : Destek Vektör Makineleri

SVR : Destek Vektör Regresyonu

UCİ : California Üniversitesi, Irvine

vd. : ve diğerleri

WAC : Ağırlıklı İlişkilendirme Kural Tabanlı Sınıflandırıcı

WHO : Dünya Sağlık Örgütü

SÖZLÜK

Big Data: Büyük veri, genellikle yapılandırılmamış veya yapılandırılmış, büyük miktarda veriyi ifade eder. Bu veriler, geleneksel yazılım teknikleriyle işlenmekte zorlanabilir ve özel teknikler gerektirebilir.

Cardiovascular Disease: Kardiyovasküler hastalık, kalp ve damar sistemi ile ilgili hastalıkları kapsayan bir terimdir. Kalp hastalığı, inme gibi durumları içerir.

Data Mining: Büyük veri üzerinde yapılan analizlerle, anlamlı bilgilerin keşfi ve çıkarılması sürecini ifade eder. Veri madenciliği, büyük veri setlerinden örüntüler, trendler ve anlamlı bilgiler elde etmek için kullanılır.

Lasso Regresyon: Değişken seçimi ve aşırı uyum sorunlarıyla mücadele etmek için kullanılan regresyon analizi türü.

Lineer Regresyon: Bağımlı değişken ile bağımsız değişken(ler) arasındaki doğrusal ilişkiyi modellemek için kullanılan regresyon analizi türü.

Lojistik Regresyon: Kategorik bağımlı değişkenleri tahmin etmek için kullanılan regresyon analizi türü.

Metabolik Sendrom: Yüksek kan şekeri, yüksek tansiyon, obezite ve yüksek kolesterol gibi faktörlerle ilişkilidir. Kalp hastalığı riskini artırabilir.

Pre-eklampsi: Gebelik sırasında yüksek tansiyon nedeniyle yaşam boyu kalp yetmezliği riskini artırabilir.

Regresyon Analizi: Bir bağımlı değişken ile bir veya daha fazla bağımsız değişken arasındaki ilişkiyi modellemek için kullanılan istatistiksel bir yöntem.

Ridge Regresyon: Değişken seçimi ve aşırı uyum sorunlarıyla mücadele etmek için kullanılan regresyon analizi türü.

Risk Faktörleri: Kalp krizi riskini etkileyen faktörlerdir. Yaş, sigara içme, yüksek tansiyon, yüksek kolesterol, obezite, diyabet gibi faktörler riski artırabilir.

GİRİŞ

Tıbbi araştırmaların günümüzdeki pek çok yönü, özellikle öngörüsel modelleme alanında, giderek artan bir şekilde karmaşık veri analiz tekniklerine dayanmaktadır. Bu Yüksek Lisans tezi, kalp krizleri ile ilişkilendirilen faktörlerin tahmin ve analizi konusunda çeşitli regresyon yöntemlerini kapsayan kapsamlı bir araştırmaya giriş yapmaktadır.

Problem Tanımı;

Kalp krizi, dünya genelinde ölümlerin en yaygın nedenlerinden biri olup, erken teşhis ve doğru tedavi ile önlenebilir bir sağlık sorunudur. Kalp krizi geçiren hastaların büyük bir kısmı, önceden belirlenebilir risk faktörlerine sahiptir; bu nedenle bu risklerin erken tespit edilmesi, hastaların hayatını kurtarma potansiyeline sahiptir. Kalp krizi tahmininde doğruluğu artırmak, erken müdahale ve önlem alma konusunda önemli katkılar sağlayabilir. Modern tıbbın gelişimiyle birlikte, kalp krizi risk faktörlerinin belirlenmesi ve tahmin edilmesi için makine öğrenimi ve yapay zeka tabanlı yaklaşımlar giderek daha fazla kullanılmaktadır. Makine öğrenimi teknikleri, büyük veri setleri üzerinden analiz yaparak, karmaşık ve doğrusal olmayan ilişkileri modelleme kapasitesine sahip oldukları için kalp krizi gibi multifaktöriyel hastalıkların tahmininde önemli rol oynamaktadır.

Araştırmanın Amacı;

Makine öğrenimi yöntemlerinin sağlık sektöründe kullanımı, özellikle yüksek riskli hastalıkların erken teşhisi ve bireyselleştirilmiş tedavi planlarının oluşturulmasında devrim niteliğinde ilerlemeler sağlamıştır. Yapay zeka destekli karar destek sistemleri, sağlık profesyonellerine hastalık risk değerlendirmelerinde yardımcı olmakta ve tedavi sürecinin daha doğru ve etkili bir şekilde yönetilmesine katkı sunmaktadır. Bu bağlamda, çalışmamız, kalp krizi riskinin tahmini için farklı makine öğrenimi yöntemlerinin performanslarını karşılaştırarak, en uygun modelin belirlenmesini amaçlamaktadır.

Mevcut literatür, genellikle lineer modellerin (lojistik regresyon gibi) kalp krizi tahmininde yaygın olarak kullanıldığını göstermektedir. Ancak, bu modeller doğrusal ilişkileri etkili bir şekilde modellese de, kalp krizi gibi kompleks ve çok boyutlu sağlık problemlerinde sınırlı kalabilmektedir. Bu nedenle, doğrusal olmayan ilişkileri

yakalayabilen ve veri setindeki karmaşıklıkları daha iyi modelleyebilen destek vektör regresyonu (SVR) gibi gelişmiş yöntemlerin incelenmesi gereklidir. SVR'nin yüksek doğruluk ve hassasiyet sunma potansiyeli, bu çalışmada SVR'nin yanı sıra diğer regresyon ve sınıflandırma yöntemlerinin de performanslarının kapsamlı bir şekilde değerlendirilmesini motive etmiştir.

Bu çalışmanın temel amacı, kalp krizi tahmininde yaygın olarak kullanılan lineer regresyon, lojistik regresyon, SVR ve diğer makine öğrenimi yöntemlerinin doğruluk, hassasiyet ve model açıklanabilirliği gibi performans kriterlerini karşılaştırarak en etkin modeli belirlemektir. Bu amaç doğrultusunda, lineer regresyon, lojistik regresyon, ridge regresyon ve destek vektör regresyonu gibi yaygın kullanılan regresyon teknikleri sistematik bir şekilde karşılaştırılacaktır.

Bu değerlendirme, klinik uygulamalarda kullanılabilecek en uygun modeli belirlemeye yönelik önemli bir rehber sunacaktır. Özellikle sağlık profesyonelleri tarafından kolayca yorumlanabilen ve gerçek zamanlı risk değerlendirmeleri yapabilen modellerin tercih edilmesi, hasta yönetiminde etkinliğin artırılmasını sağlayabilir.

Hedef, bu yöntemler arasından kalp krizi tahmininde en yüksek doğruluğu sağlayan yöntemi belirlemektir.

Araştırmanın Önemi;

Tıbbi veri analizindeki ilerlemeler, karmaşık veri setlerindeki örüntüleri açığa çıkarma ve erken teşhis ile önleme konusunda değerli araçlar sunma potansiyeli taşımaktadır. Bu çalışma, farklı regresyon yöntemlerinin kalp krizi tahmininde etkinliğini değerlendirerek, tıp uygulayıcılarına, araştırmacılara ve politika yapıcılara daha bilinçli karar alma süreçlerinde yardımcı olmayı hedeflemektedir. Bu bağlamda, çalışmanın sonuçları, kalp krizi tahmininde metodolojik bir çerçeve sunma potansiyeline sahiptir.

Bu çalışmanın özgün katkısı, SVR gibi doğrusal olmayan modellerin kalp krizi tahmininde doğrusal modellere göre ne kadar etkili olduğunu değerlendirmesi ve bu modellerin sağlık sektöründe genişletilmiş kullanım potansiyelini ortaya koymasıdır. Çalışmamız, mevcut literatürde sıklıkla ihmal edilen bu önemli noktaya odaklanarak, sağlık hizmetlerinde makine öğrenimi uygulamalarının daha geniş bir yelpazede kullanılmasına yönelik kanıt sunmaktadır.

Araştırmanın Kapsamı ve Sınırlılıkları;

Bu çalışma, kalp krizi verileri üzerinde gerçekleştirilen çeşitli regresyon analizlerini kapsamaktadır. İncelenen yöntemler, lineer regresyon, lojistik regresyon, ridge regresyon ve destek vektör regresyonunu içermektedir. Araştırma, veri setinin niteliği ve kullanılan yöntemlerin sınırlılıkları ile sınırlıdır. Özellikle, kullanılan veri setindeki eksiklikler ve veri ön işleme adımlarının doğruluğu, sonuçların genelleştirilebilirliğini etkileyebilir.

Varsayımlar ve Hipotezler;

Bu çalışmada, kalp krizi verilerinin belirli ön işleme adımlarından geçirildikten sonra analiz edilmesi varsayılmıştır. Ayrıca, regresyon yöntemlerinin kalp krizi tahmininde farklı performans göstereceği hipotezi üzerine odaklanılmıştır.

Çalışmanın hipotezleri şunlardır:

Hipotez 1: SVR, doğrusal olmayan veri yapılarındaki üstün performansı nedeniyle kalp krizi tahmininde en yüksek doğruluk oranına sahiptir.

Hipotez 2: Karmaşık modeller, yüksek doğruluk sağlarken, basit modeller daha iyi açıklanabilirlik ve hızlı hesaplama süreleri sunar, bu da pratik uygulamalarda önemli bir avantajdır.

Yöntem;

Çalışmada izlenecek yol, öncelikle kalp krizi veri setinin detaylı bir şekilde incelenmesi ve ön işleme adımlarının gerçekleştirilmesiyle başlamaktadır. Ardından, dört farklı regresyon yöntemi kullanılarak tahmin modelleri oluşturulacak ve bu modellerin performansı, doğruluk, hassasiyet ve geri çağırma gibi metrikler kullanılarak değerlendirilecektir.

Çalışmanın Yapısı;

Bu tezin ilk bölümünde, tıbbi veri analizi ve kalp krizi tahmini ile ilgili mevcut literatürün kapsamlı bir incelemesi sunulmaktadır. Sonraki bölüm, kullanılan kalp krizi veri seti hakkında detaylı bilgileri içermekte ve veri bütünlüğünü sağlamak adına gerçekleştirilen ön işleme adımlarını açıklamaktadır. Sonraki bölüm, regresyon yöntemlerinin teorik temelleri ve bu yöntemlerin kalp krizi veri setine uyarlanması

üzerine odaklanmaktadır. Sonraki bölümde, her bir regresyon yönteminin performansını değerlendirmek için kullanılacak olan doğruluk, hassasiyet, geri çağırma gibi performans metrikleri açıklanacaktır. Sonraki bölüm, farklı regresyon yöntemlerinin kalp krizi verilerine uygulanmasından elde edilen sonuçları içermekte ve her bir yöntemin güçlü ve zayıf yönlerini vurgulamaktadır. Altıncı bölümde, elde edilen bulguların kritik bir şekilde analizi yapılarak, belirlenen optimal regresyon yönteminin gerçek dünya senaryolarındaki potansiyel uygulamaları tartışılmaktadır. Son bölümde ise, ana bulguların özetlenmesi, çalışmanın sınırlamaları ve gelecekteki araştırmalar için öneriler sunulmaktadır.

Bu çalışmada, kalp krizleri için öngörüsel modelleme tekniklerini geliştirme ihtiyacı karşısında bir çerçeve sunmaktadır. Çeşitli regresyon yöntemlerinin titiz bir şekilde incelenmesiyle, tıbbi veri analizi alanındaki gelişmeler için anlamlı katkılarda bulunmayı ve dolayısıyla kardiyovasküler sağlıkta erken müdahale ve önleme kapasitemizi artırmayı hedeflemektedir.

1. LİTERATÜR İNCELEMESİ

Bu literatür incelemesi, tıbbi veri analizi bağlamında regresyon metotlarının kullanımını detaylı bir şekilde ele almıştır. Ayrıca, mevcut araştırmalardan "Kalp Krizi ve Hastalıkların Algılanması ve Tahmin Edilmesi İçin Makine Öğrenimi Tekniklerinin Karşılaştırılması ve Kullanılması", "Farklı Türdeki Makine Öğrenimi Algoritmalarını Kullanarak Kalp Krizi Olasılığını Ölçme", "Kalp Krizi Mortalite Tahmini: Makine Öğrenimi Yöntemlerinin Uygulanması" ve "Özellik Seçimi Yöntemleri ile Kalp Krizi Tahmininin İyileştirilmesi" tezlerinin incelenmesi yapılmıştır. Bu tezler, makine öğrenimi tekniklerinin kalp krizi ve benzeri hastalıkların algılanması, tahmini ve mortalite tahmini gibi önemli sağlık sorunlarında nasıl kullanılabileceğini araştırmaktadır. Son olarak, literatür taramasının bir parçası olarak, mevcut bilgi boşlukları ve araştırma gereksinimleri belirlenmiştir. Bu, mevcut çalışmaların sınırlarını ve gelecekte yapılacak araştırmaların potansiyel yönlerini belirlemek için önemlidir.

1.1. Regresyon Metotlarının Tıbbi Veri Analizi Bağlamında Kullanımı

Regresyon analizi, bir bağımlı değişken ile bir veya daha fazla bağımsız değişken arasındaki ilişkiyi modellemek için kullanılır. Tıbbi veri analizinde, regresyon metotları, hastalıkların nedenleri, etkileyen faktörler ve bu faktörlerin hastalık üzerindeki etkilerini anlamak için değerli bilgiler sağlar.

Lineer regresyon, bir bağımlı değişken ile bağımsız değişken(ler) arasındaki doğrusal ilişkiyi modellemek için kullanılır. Tıbbi veri analizinde, hastalık riskini etkileyen faktörleri belirlemede ve bu faktörlerin etkilerini ölçmede yaygın olarak kullanılan bir yöntemdir. Örneğin, yaş, kan basıncı, kolesterol seviyeleri gibi faktörlerin kalp krizi riski üzerindeki etkilerini değerlendirmek için lineer regresyon analizi kullanılabilir.

Lojistik regresyon, bir kategorik bağımlı değişkeni tahmin etmek için kullanılır. Tıbbi veri analizinde, hastalık varlığı veya yokluğunu tahmin etmek için sıklıkla kullanılır. Örneğin, bir hasta grubundaki belirli özelliklere sahip bireylerin belirli bir hastalığa yakalanma olasılığını tahmin etmek için lojistik regresyon analizi uygulanabilir (Cristobal,2021).

Ridge ve Lasso regresyonları, değişken seçimi ve aşırı uyum sorunlarıyla mücadele etmek için kullanılır. Tıbbi veri analizinde, bu regresyon metotları, önemli olan değişkenleri belirlemede ve model karmaşıklığını azaltmada etkili olabilir. Bu, özellikle büyük veri setleri üzerinde çalışırken önemlidir.

Regresyon analizi, tıbbi veri analizinde geniş bir uygulama alanına sahip güçlü bir istatistiksel araçtır. Bu başlık altında, lineer regresyon, lojistik regresyon ve düzenlileştirilmiş regresyon metotları gibi temel regresyon tekniklerinin tıbbi veri analizindeki rolünü anlamak için kapsamlı bir değerlendirme yapılacaktır.

1.2. Mevcut Araştırmalar

Veri, dünyanın en değerli kaynaklarından biridir. Yeni bilgi edinme ve bilgi toplamanın anahtarıdır ve bilim ve teknolojide çığır açmıştır. Tıp, hukuk, mühendislik gibi alanlarda veri, gerçek bir çalışma aracıdır. İnsanlar, yeni fikirler öğrenme ve bilgi artırma çabalarında mümkün olduğunca çok veri toplamaya çalışırlar.

1.2.1. Kalp Krizi ve Hastalıkların Algılanması ve Tahmin Edilmesi İçin Makine Öğrenimi Tekniklerinin Karşılaştırılması ve Kullanılması

Bazı veriler küçük ve yönetilebilirken, diğerleri büyük olup bunları işlemek için özel tekniklere ihtiyaç duyar. Bu, dijital (big data) veri olarak adlandırılır. Sağlık, tarım, iş, bilim, teknoloji gibi sektörlerde büyük miktarda dijital veri (Big Data) bulunmaktadır. Bu veriler genellikle yapılandırılmamış veya yapılandırılmış bir şekilde hamdır ve genellikle büyük miktarlardadır. Veri madenciliği teknikleri kullanılarak işlenmiş veriler, anlamlı bilgi toplamayı sağlar. Bu büyük veri miktarı (hem yapılandırılmış hem de yapılandırılmamış veri), normal yazılım teknikleri ile depolamak, analiz etmek, işlemek, paylaşmak, yönetmek ve görselleştirmek zordur ve genellikle böyle büyük veriyi depolamak ve işlemek zorlayıcıdır. Veri madenciliği, büyük veride doğru veri analizi için önemli bir rol oynar. Veri, işlenip kullanışlı ve anlamlı bilgiye dönüştürülmedikçe faydalı olamaz (Obasi ve Shafiq ,2019).

Sağlık sektöründe, işlenmemiş büyük bir hasta veri miktarı bulunmaktadır. Bu hastaların tıbbi kayıtları veya verileri, üzerlerinde yapılan veri analizi gerektiren gizli

örüntülere sahiptir. Sağlık, kişilerin fiziksel, zihinsel ve sosyal iyilik halleri ile ilgilenen en önemli sektörlerden biridir ve bu alana büyük bir odak gerekmektedir. Çünkü çok zararlı ancak kendilerini geç belli eden özelliklere sahip hastalıklar vardır. Kalp hastalığı, bu sessiz hastalıklara bir örneğidir ve her yıl acı verici bir şekilde artan ölüm oranına neden olmaktadır. Dünya Sağlık Örgütü (WHO), 2014 yılında dünya genelinde yılda 12 milyon ölümün kalp hastalığından kaynaklandığını bildirmiştir. Kalp ve damar hastalıkları kardiyovasküler hastalık (CVD) veya koroner kalp hastalığı olarak adlandırılır. Bilindiği gibi kalp, insan vücut sisteminin kritik bir organıdır. Vücudu besleyen bir motor gibi görev yapar. Kan pompalar ve düzgün çalışamaması, beyne yeterli kan dolaşımını sağlayamayarak oksijen eksikliğine ve sonrasında nöbet geçirmesine neden olabilir. Bu durum, tıbbi uzmanlara göre, birkaç dakika içinde ölüme yol açabilir. Kalp hastalığı ve krizi, büyük bir sağlık sorunu olup, bunun kaynaklandığı farklı komplikasyonlardan dolayı bireyin sağlığını etkiler. Ancak, bazı risk faktörleri, kalp hastalıkları ile ilişkilendirilmiş olarak tanımlanmıştır (kalp hastalığı türüne bağlı olarak). Bunlar arasında yaş, cinsiyet, aile öyküsü, yüksek tansiyon (hipertansiyon), diyabet, yüksek kolesterol, sigara içme, alkol tüketimi, göğüs ağrısı, stres, kötü beslenme ve hijyen, obezite, etnik köken gibi faktörler bulunmaktadır. Komplikasyonlar arasında kalp yetmezliği, kalp krizi, inme, ani kalp krizi, periferik arter hastalığı yer alır. Bu çalışmada elde edilen veri seti üzerinde yapılacak veri analizi, kalp hastalıklarının gelecekteki sonuçlarının olasılığını belirleme ile ilgilidir. Bu, kalp hastalığı riski taşıyan hastaların belirlenmesi ve en iyi tedavi yöntemini bulma konusunda bir değerlendirme sağlayacaktır. Veri analizi ve makine öğrenimi için kullanılabilecek analitik yaklaşımlar, mevcut yöntemlere dayanmaktadır. Genel olarak, bu çalışma, zaman içinde bireylerin kalp sağlığını anlama ve takip etme konusuna yardımcı olacak ve bu veriler üzerinde analiz yaparak hastaların tıbbi verilerinden anahtar örüntüleri ve özellikleri belirleme konusunda veri madenciliği teknikleri ve büyük veri analitiği kullanarak kalp hastalığını önceden tahmin etmeye yardımcı olacaktır. Bu, tıbbi uygulayıcılara yardımcı olacak ve aynı zamanda tahmin modelinin doğruluğunu artırmak için veri kümelerini azaltarak daha etkili hale getirecektir (Obasi ve Shafiq ,2019).

1.2.2. Farklı Türdeki Makine Öğrenimi Algoritmalarını Kullanarak Kalp Krizi Olasılığını Ölçme

Kalp ve damar bozukluğu, bir dizi kalp ile ilişkili hastalığı tanımlar. Kardiyovasküler hastalık kapsamındaki hastalıklar, koroner arter hastalığı, kalp ritim bozuklukları ve doğuştan kalp kusurları gibi kan damarlarıyla ilişkili bozuklukları içerir. Kalp, vücudumuzdaki en önemli organdır; yaşamsal bir organdır. Kalp hastalıkları yaygın bir hastalıktır ve bunlardan biri de kalp krizidir. Kalp krizi gerçekleştiğinde kanın tedariki mükemmel çalışmaz. Bir kalp krizi sonrasında genellikle yağ, kolesterol ve diğer bileşikler birikir. Kalp kasının bir kısmı yeterince oksijen alamadığında bir saldırı veya miyokard enfarktüsü meydana gelir. Kalp hastalığı belirtileri arasında karın bölgesinde basınç veya rahatsızlık, duygusal veya sol göğüs ağrısı, zorlama, sürtünme, doluluk veya ağrı, zayıflık, baygınlık, soğuk terleme, yüz, boyun veya sırtta baskı veya tahriş, kolların veya omuzların her ikisinde de ağrı veya tahriş, nefes darlığı bulunmaktadır. Bu belirtiler bazen kalp ağrısına yol açar, ancak oksijen eksikliği, kalp krizlerinden önce de ortaya çıkabilir. Kalp krizi ortalama olarak erkeklerde 65 yaşında, kadınlarda 72 yaşında gerçekleşir.

Risk faktörleri: Bir kalp krizini kontrol altına almanın temel risk faktörleri şunlardır: Yaş. 45 yaş ve üstü erkekler ile 55 yaş ve üstü kadınlar, genç erkekler ve kadınlara göre kalp krizi geçirme olasılığı daha yüksektir. Sigara içme ve pasif içiciliğe maruz kalma riski içerir. Yüksek tansiyon zamanla kalp damarlarınıza zarar verebilir. Yüksek tansiyon, obezite, yüksek kolesterol veya diyabet gibi faktörlerle birlikte riski daha da artırır. Düşük yoğunluklu lipoprotein (LDL) potansiyel olarak yüksek kolesterolle ilişkilidir (kötü kolesterol). Trigliseridlerin yüksek miktarda, diyetle ilişkili bir tür kan yağı, ayrıca kalp krizi riskinizi artırır. Büyük bir miktar yüksek yoğunluklu kolesterol (iyi kolesterol) lipoprotein (HDL), riski azaltır. Obezite, yüksek serum kolesterol seviyeleri, yüksek trigliserit konsantrasyonları, yüksek tansiyon ve diyabetle ilişkilidir. Vücut ağırlığının sadece %10'unu kaybederek bu risk azalabilir. Diyabet, vücuttaki kan şeker seviyelerini yükseltme yeteneği olmadığında veya yeterli miktarda insülin salgılanmadığında, kalp krizi riskinizi artırır. İnsülinin etkisi, aldığınız kan şekerinin miktarını etkilemez. Bu, sigara içmeniz, astımınız ve yüksek kan şekeri olmanız nedeniyle olabilir. Metabolik sendrom için, yapmazsanız kalp hastalığına iki kat daha sık sahip olacaksınız.

Eğer kardeşleriniz, ebeveynleriniz veya atasalrınız erken yaşta kalp krizi geçirmişse (erkekler için 55 yaşında, kadınlar için 65 yaşında), daha yüksek risk altında olabilirsiniz. Hareketsizlik serum kolesterol ve obeziteyi artırır. Günlük olarak egzersiz yapan insanlar, düşük tansiyon dahil olmak üzere gelişmiş kardiyovasküler sağlık yaşarlar. Stres durumuna tepki verirseniz, kalp krizi riskinizi artırabilirsiniz. Amfetamin gibi rahatlatıcı ilaçlar kullanılıyorsa, koroner arterin spazmına neden olabilir ve kalp krizi geçirebilirsiniz. Pre-eklampsi sendromu, gebelik sırasında yüksek tansiyon nedeniyle yaşam boyu kalp yetmezliği riskini artırır. Romatoid artrit veya lupus gibi bir hastalık, riski artırabilir (Keya vd. 2021).

1.2.3. Kalp Krizi Mortalite Tahmini: Makine Öğrenimi Yöntemlerinin Uygulanması

Günümüzde her gün büyük miktarda veri üretilmektedir. Büyük veri setlerini analiz etmek, otomatik prosedürlerin yardımı olmadan imkansızdır. Makine öğrenimi bu prosedürleri sağlar. En yaygın kullanılan makine öğrenimi türü denetimli sınıflandırmadır. Denetimli sınıflandırmanın amacı, bir nesnenin tanımlayıcı özelliklerinden sınıflar kümesine olan eşlemeyi öğrenmektir, bir özellik-sınıf çiftleri kümesi verildiğinde.

Modern makine öğreniminde olasılıklar önemli bir rol oynamaktadır. Olasılıksal grafik modeller (OGM), olasılıksal modelleri tanımlamak ve uygulamak için genel bir çerçeve olarak ortaya çıkmıştır. Bir OGM, koşullu bağımsızlık varsayımları yaparak bazı rasgele değişkenler üzerinde birleşik bir dağılımı etkili bir şekilde kodlamamıza olanak tanır.

Bir Bayesian ağ sınıflandırıcısı (BNC), sınıflandırma görevine uygulanan bir Bayesian ağıdır. BNC'lerin iyi yorumlanabilirlik, bir alan hakkında önceden bilgi eklemenin mümkünlüğü ve rekabetçi tahmin performansı gibi birçok avantajı vardır ve pratikte başarıyla uygulanmışlardır.

Akut miyokard enfarktüsü (AMI) yaygın olarak kalp krizi olarak bilinir. Kalp krizi, kalbe giden bir arterin tamamen tıkanması ve kalbin yeterince kan veya oksijen alamaması durumunda meydana gelir. O bölgedeki hücrelerin oksijensiz kalması durumunda ölür. AMI, çoğu ülkede dünya genelindeki ölümlerin yarısından

fazlasından sorumludur. Tedavisi önemli bir sosyoekonomik etkiye sahiptir. Araştırmanın ana hedeflerinden biri, hastalar hakkındaki klinik verilere dayalı bir hastane mortalite tahmin modeli tasarlamak, analiz etmek ve doğrulamaktır. İyi bir şekilde mortaliteyi tahmin eden bir model, örneğin farklı hastanelerde sağlık hizmetlerinin değerlendirilmesi için kullanılabilir. Sadece mortaliteye dayalı değerlendirme, komplike vakaların sıkça ele alındığı hastaneler için adil olmayabilir. Sağlık hizmetlerinin kalitesini, tahmin edilen ve gözlemlenen mortalite arasındaki farkı kullanarak ölçmek daha iyi görünmektedir.

Benzer bir çalışma, Krumholz ve diğerleri tarafından yayımlandı, yazarlar ABD hastanelerinde mortalite verilerini lojistik regresyon modeli kullanarak analiz ettiler. Başka bir çalışmada, yazarlar ST yükselmiş miyokard enfarktüsü (STEMI) hastanesi mortalitesi için bir tahmin modeli tasarladılar ve doğruladılar. Başka bir çalışmada, yazarlar üçüncü dünya ülkesi olan Suriye ve gelişmiş bir ülke olan Çek Cumhuriyeti'nde miyokard enfarktüsü geçiren hastaların tıbbi kayıtlarını analiz ettiler ve eksik ve dengesiz verilerle başa çıkma konusunda bir çözüm sunarak tree-augmented naive Bayesian (TAN) modelini tanıttılar (Salman, 2019).

1.2.4. Özellik Seçimi Yöntemleri ile Kalp Krizi Tahmininin İyileştirilmesi

Kalp krizi, bilinen adıyla miyokard enfarktüsü, kalbin bir bölümüne yetersiz kan akışının neden olduğu ve kalp kasında hasara yol açan bir durumdur. Kalp krizlerinden kaynaklanan ani ölümleri azaltmak için erken teşhis kritiktir ve genellikle elektrokardiyografi (EKG) ve kan testleri gibi klinik yöntemler kullanılır. EKG, zaman içinde kalbin elektriksel aktivitesini kaydetme işlemidir ve kalpte anomalilerin tespitine olanak tanır. Diğer taraftan, kan testleri CK-MB gibi enzim seviyelerini değerlendirerek olası bir kalp krizi için bir gösterge olarak kullanılır, son yıllarda troponin değerleri de kalp krizi teşhisi için kontrol edilmektedir.

Sağlık hizmetlerinde daha iyi sonuçlar elde etme isteğinin artmasıyla, klinik yöntemlere ek olarak bilgisayar destekli sistemler önem kazanmıştır. Hasta bilgileri, tıbbi teşhisler ve tıbbi görüntüler şimdi kaydedilmekte ve bu veriler üzerindeki makine öğrenimi yöntemleri karar destek sistemleri oluşturmak için kullanılmaktadır.

Tu vd. çalışması kalp krizi tahmini için bir bagging algoritması ve J48 karar ağacı algoritmasını kullanmıştır, bagging algoritmasının karar ağacından daha iyi sonuçlar verdiğini göstermiştir.

Srinivas vd. çalışması, kural tabanlı karar ağacı, naïve Bayes ve yapay sinir ağı sınıflandırma algoritmalarını kullanarak, ODANB ve NCC2 gibi ön işleme tekniklerini de içeren bir çalışma yapmıştır. Tahmin modeli yaş, cinsiyet, kan şeker ve kan basıncı gibi değişkenleri içermektedir.

Deepika vd. çalışması, kalp krizi tiplerini sınıflandırmak için derinlik kural madenciliği kullanmıştır, bu çalışmada hasta yaş ve göğüs ağrısı gibi özellikler kullanılmştır.

Jabbar vd. çalışması, kalp krizi tahmini için kümeleme ve derinlik kural madenciliği algoritması önermiştir.

Sudha vd. çalışması, inme krizi için naïve Bayes, karar ağaçları ve sinir ağı sınıflandırıcıları önermiştir.

Shouman vd. çalışması, kalp krizi tahmini için k-means kümeleme algoritması ile karar ağacını birleştiren bir yöntem önermiştir.

Vikas vd. çalışması, CART, ID3 ve karar tablosu gibi veri madenciliği tekniklerini kullanmıştır, CART algoritmasının diğer sınıflandırıcılardan daha başarılı olduğunu ortaya koymuştur.

Kora ve Kalva çalışması, kalp krizi tahmini için EKG sinyallerini incelemiştir, özellik seçimi için geliştirilmiş bir yarasa algoritması kullanmıştır ve seçilen özellikleri (200 özellikten en iyi 20 özellik) yüksek doğruluk elde ederek SVM, KNN, LM NN ve SCG NN gibi sınıflandırıcılarla, sınıflandırıcılar için giriş olarak kullanmıştır.

Soni vd. çalışması, bir risk hesaplamak için bir GUI'ye sahip olan weighted association rule based classifier (WAC) algoritmasını kullanarak kalp krizi tahmini yapmıştır.

Florence vd. çalışması, kalp krizi tahmini için karar ağaçları ve yapay sinir ağları kullanmıştır, ayrıca altı özellikten oluşan bir UCI veri seti üzerinde çalışmıştır.

Jabbar vd. çalışması, kalp krizi tahmini probleminde grafik tabanlı derinlik kural madenciliği algoritmasını kullanmıştır.

Krishnaiah vd. çalışması, kalp krizi teşhisi için kullanılan yöntemleri incelemiştir ve bu yöntemleri karşılaştırmıştır, çalışmalarına göre, bulanık mantık temelli zeki tekniklerin model doğruluğunu artırdığını belirtilemiştir.

Bu çalışmada, kalp krizi tahmini için birçok makine öğrenimi yöntemi ve özellik seçimi algoritması bir arada kullanılmıştır ve bu algoritmaların doğruluk, işleme süresi ve ROC değerleri temelinde karşılaştırılması yapılmıştır. Deneylerde UCI Statlog (Heart) veri seti kullanılmıştır (Takçı, 2018).

1.3. Bilgi Boşlukları ve Gereksinimler

Tıbbi veri analizi, karmaşık hastalıkların anlaşılması ve etkili tedavi stratejilerinin geliştirilmesi için kritik öneme sahiptir. Ancak, mevcut bilgilerin sınırlamaları, tıbbi araştırmalarda belirli boşluklara neden olabilir. Bu bölümde, mevcut bilgi boşluklarına odaklanarak ve gerekli bilgi gereksinimlerini belirleyerek tıbbi veri analizi alanındaki bu zorluklara ışık tutulacaktır.

Tıbbi veri analizi sürecinde karşılaşılan bazı mevcut bilgi boşlukları şunları içermektedir:

a. Veri Kalitesi ve Bütünlüğü: Tıbbi veri setlerindeki kalite ve bütünlük sorunları, doğru analiz sonuçlarına ulaşmayı zorlaştırabilir. Eksik veya hatalı veriler, analizlerde yanıltıcı sonuçlara neden olabilir.

b. Ölçüm Hassasiyeti: Bazı tıbbi ölçümler, özellikle laboratuvar testleri ve cihaz ölçümleri, belirli bir hassasiyetle elde edilebilir. Bu hassasiyetin bilincinde olmamak, analiz sonuçlarının doğruluğunu etkileyebilir.

c. Çeşitli Veri Kaynakları: Tıbbi veri setleri genellikle farklı kaynaklardan elde edilir. Farklı kaynaklardan gelen verilerin uyumsuzluğu ve farklı formatlarda olması, veri entegrasyonunu zorlaştırabilir.

d. Zamansal Veri Erişimi: Hastalık progressiyonunun zaman içinde değişkenlik göstermesi, doğru zamansal analizler yapmayı zorlaştırabilir. Bu durum, özellikle kronik hastalıkların takibi ve öngörüsü açısından önemlidir.

Tıbbi veri analizi alanındaki bilgi boşluklarını kapatmak ve daha güvenilir sonuçlara ulaşmak için şu gereksinimlere odaklanılabilir:

a. Veri Kalitesi İyileştirmesi: Veri kalitesini artırmak için eksik veya hatalı verilerin düzeltilmesi, anlamlı analiz sonuçlarına ulaşmada temel bir adımdır. Veri toplama süreçlerinin standartlaştırılması ve düzenli denetimler yapılması, bu konuda yardımcı olabilir.

b. Hassas Ölçümler: Ölçüm hassasiyetinin bilinmesi ve bu bilgiye göre analizlerin yapılması, tıbbi veri analizinde daha güvenilir sonuçlar elde etmede yardımcı olabilir.

c. Veri Entegrasyonu ve Uyumluluk: Farklı kaynaklardan gelen veriler arasında uyumlu bir entegrasyon sağlamak, tıbbi veri analizi sürecini iyileştirebilir. Standart veri formatlarının kullanılması ve veri entegrasyon teknolojilerinin benimsenmesi bu gereksinimi karşılayabilir.

d. Zamansal Analiz İyileştirmesi: Hastalık progressiyonunu doğru bir şekilde anlamak için zamansal analizlerin güçlendirilmesi önemlidir. Bu, hastalık gelişimini belirleyen faktörlerin zaman içindeki değişimini daha iyi anlamak için yapılan analizlerle mümkündür (GfG,2024).

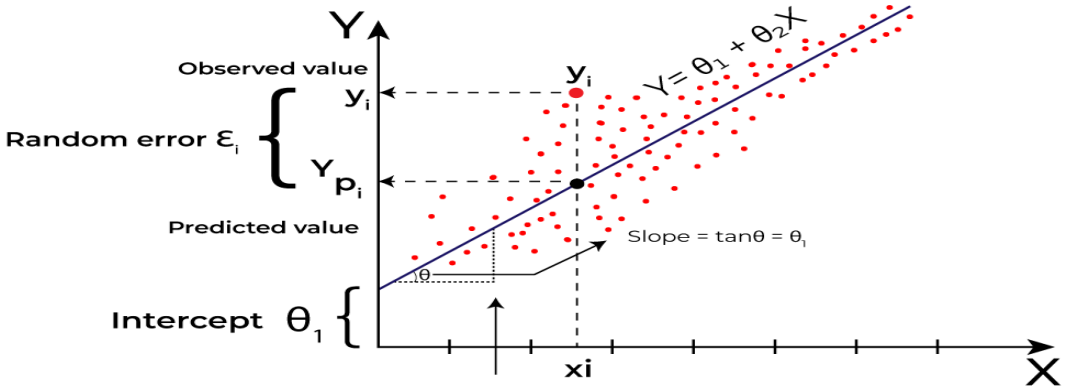
2. REGRESYON METOTLARI

Regresyon analizi, kalp krizi tahmini gibi önemli sağlık sorunlarının anlaşılması ve öngörülmesi için kullanılan temel bir istatistiksel araçtır. Bu bölümde, kalp krizi veri kümesinde kullanılan çeşitli regresyon metotları ele alınacaktır.

2.1. Lineer Regresyon

Lineer regresyon, bir bağımlı değişken ile bir veya daha fazla bağımsız değişken arasındaki doğrusal ilişkiyi modellemek için kullanılır. Kalp krizi veri kümesinde, lineer regresyon analizi, yaş, kan basıncı, kolesterol seviyeleri gibi faktörlerin kriz riski üzerindeki etkilerini nicel olarak değerlendirebilir. Model, bu faktörlerin kriz olasılıkları üzerindeki ağırlıklı etkilerini belirleyerek bireylerin risk profillerini oluşturabilir.

Lineer regresyonun kalp krizi tahminindeki etkinliği, hastalığın karmaşıklığına ve veri setinin özelliklerine bağlı olarak değişebilir. Bu metot, temelde doğrusal bir ilişki öngördüğü için bazı durumlarda yetersiz kalabilir. Ancak, veri seti ve problemin özelliğine bağlı olarak, bu basit ama güçlü metot, anlamlı sonuçlar elde etmek için kullanışlı olabilir (Su vd. , 2012).



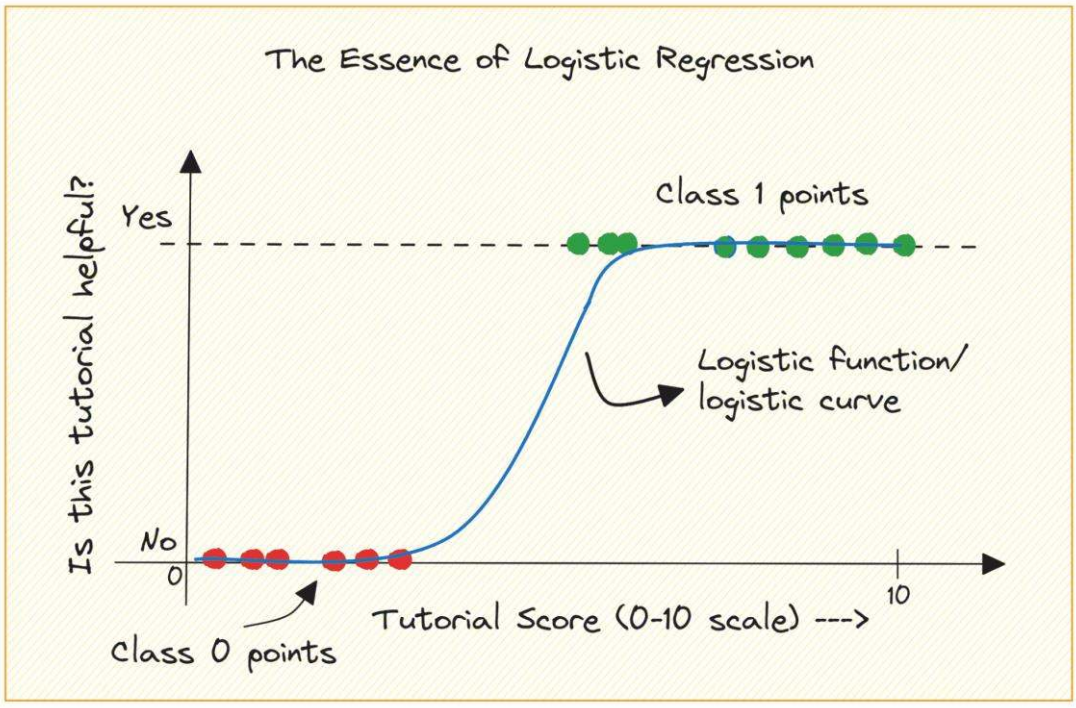
Şekil 1. Lineer Regresyon

Kaynak: GfG. (2024b, January 25). *AUC ROC curve in machine learning*. GeeksforGeeks. 5.5.2024 tarihinde https://www.geeksforgeeks.org/auc-roc-curve/ adresinden edinilmştir.

Lineer regresyon aynı zamanda katsayıların ve değişkenlerin önemini değerlendirmek için kullanılabilir. Bu, hangi faktörlerin kriz riski üzerinde daha fazla etkisi olduğunu anlamak için değerli bir bilgi sağlar. Bu nedenle, lineer regresyon, kalp krizi veri kümesinde faktörlerin belirlenmesi ve kriz olasılıklarının anlaşılması için önemli bir araç olabilir (GfG,2024).

2.2 Lojistik Regresyon

Lojistik regresyon, kategorik bir bağımlı değişkenin olasılığını tahmin etmek için kullanılır. Kalp krizi veri kümesinde, lojistik regresyon analizi, bir bireyin belirli bir krize yakalanma olasılığını değerlendirebilir. Bu metot, yaş, cinsiyet, kan şeker seviyeleri gibi faktörleri değerlendirerek bir bireyin kriz geçirme olasılığını nicel olarak tahmin edebilir(LaValley, 2008).



Şekil 2. Lojistik Regresyon

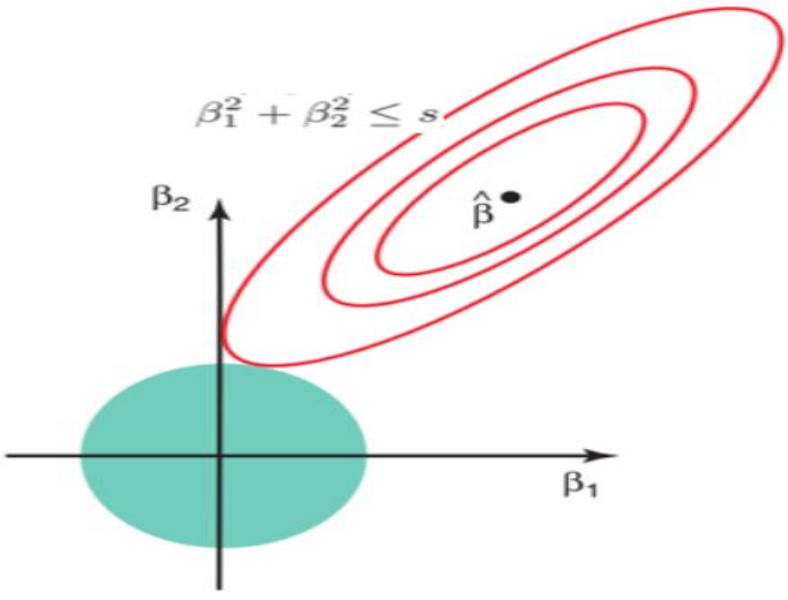
Kaynak: C, B. P. (n.d.). *Building predictive models:Logistic regression in python* - KDNuggets. KDnuggets. 5.5.2024 tarihinde https://www.kdnuggets.com/building-predictive-models-logistic-regression-in-python adresinden edinilmştir.

Lojistik regresyon, kalp krizi riskini anlamak için kullanıldığında, modelin çıktısı genellikle bir olasılık oranıdır. Bu oran, kriz olasılığının kriz olmama olasılığına göre ne kadar daha yüksek olduğunu gösterir. Bu, klinik uygulamalarda hastaların risk profillerinin oluşturulmasında ve belirli müdahalelerin etkinliğinin değerlendirilmesinde önemli bir rol oynar(C,B.P. , n.d.)

Lojistik regresyon, kalp krizi veri kümesindeki karmaşık ilişkileri modelleme yeteneği ve sonuçların yorumlanabilirliği nedeniyle sıklıkla tercih edilen bir metottur. Bu analitik araç, hastaların bireysel risklerini değerlendirmek ve koruyucu tedbirlerin etkilerini anlamak için güçlü bir istatistiksel çözüm sunar.

2.3. Ridge Regresyon

Ridge regresyonu, değişken seçimi ve aşırı uyum sorunlarına karşı dirençli bir regresyon metodu olarak öne çıkar. Kalp krizi veri kümesinde, ridge regresyonu, önemli olan değişkenleri belirleme ve modelin genelleştirilebilirliğini artırma amacını taşır. Bu metot, özellikle büyük veri setleri üzerinde etkili olabilir, çünkü karmaşık ilişkileri düzenleyebilir.



Şekil 3. Ridge Resgresyon

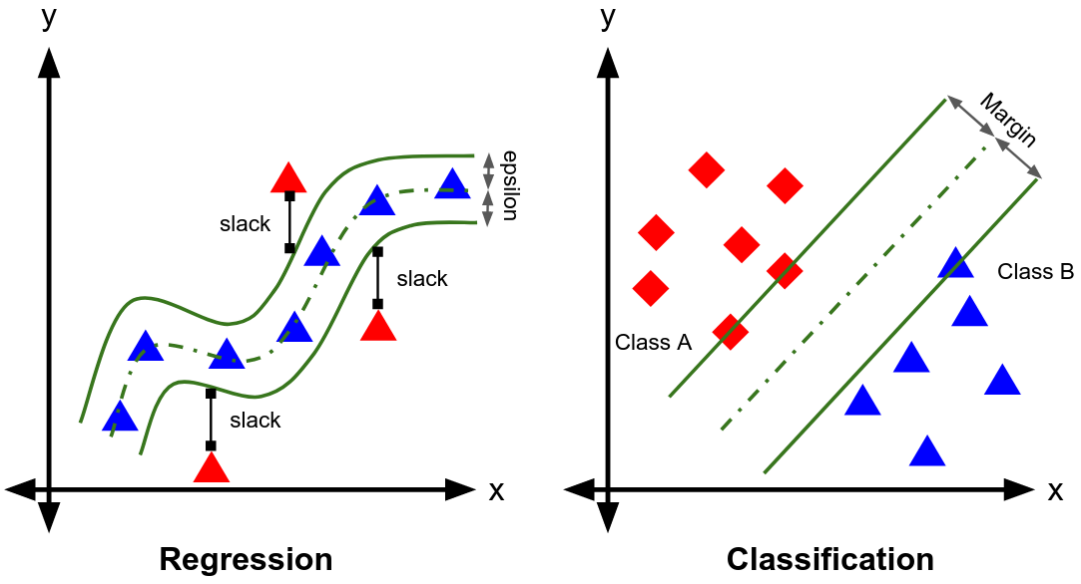
Kaynak: Team, D. (2022, March 25). *Lasso and Ridge regression in Python tutorial*. 5.5.2024 tarihinde https://www.datacamp.com/tutorial/tutorial-lasso-ridge-regression adresinden edinilmştir.

Ridge regresyonu, lineer regresyonun genişletilmiş bir versiyonudur ve modelin karmaşıklığını kontrol etmek için bir düzenleme terimi ekler. Bu, aşırı uyumu önler ve modelin daha genel bir veri setine uyarlanmasını sağlar. Kalp krizi veri kümesinde, birçok faktörün etkileşimi ve karmaşıklığı göz önüne alındığında, ridge regresyonu modelin daha tutarlı ve güvenilir sonuçlar üretmesine yardımcı olabilir (McDonald, 2009).

Ridge regresyonunun avantajları arasında, regresyon katsayılarını stabilize etme ve modelin dayanıklılığını artırma yeteneği bulunur. Bu, kalp krizi tahminindeki analitik doğruluğu artırabilir ve klinik uygulamalarda daha güvenilir sonuçlar elde edilmesine katkıda bulunabilir (Team,2022).

2.4. Destek Vektör Regresyonu

Destek vektör regresyonu, bağımlı değişkenin değerini tahmin etmek için kullanılır ve kalp krizi veri kümesindeki karmaşık ilişkileri modelleme yeteneği ile öne çıkar. Bu metot, demografik faktörler, genetik özellikler ve yaşam tarzı alışkanlıkları gibi birçok değişkeni içeren karmaşık bir veri setinde etkili bir şekilde çalışabilir.



Şekil 4. Destek Vektör Regresyonu (SVR)

Kaynak: Achsan, B. M. (2021, December 12). *Support vector machine: Regression* - IT Paragon - Medium. Medium.5.5.2024 tarihinde https://medium.com/it-paragon/support-vector-machine-regression-cf65348b6345 adresinden edinilmştir.

Destek vektör regresyonu, bir bireyin kriz riskini tahmin etmek için veri setindeki değişkenlerin karmaşıklığını dikkate alır. Bu, lineer olmayan ilişkileri ve etkileşimleri modellerken başarılı olmasını sağlar. Kalp krizi veri kümesinde, genellikle birçok faktörün etkileşimi ve karmaşıklığı göz önüne alındığında, destek vektör regresyonu, analitik doğruluğu artırabilir ve daha güvenilir sonuçlar sunabilir (Zhang vd.,2020).

Destek vektör regresyonunun bir diğer avantajı, outliers (aykırı değerler) ve gürültülü veri ile başa çıkma yeteneğidir. Bu, kalp krizi veri kümesindeki potansiyel problemleri ele alabilir ve modelin genel geçerliliğini artırabilir. Bu nedenle, destek vektör regresyonu, kalp krizi tahmini analizlerinde tercih edilen bir metot olabilir (Achsan,2021).

2.5. Diğer Potansiyel Regresyon Metotları

Lineer, lojistik, ridge ve destek vektör regresyonları kalp krizi tahmini için yaygın olarak kullanılır. Ancak, literatürde bir dizi diğer regresyon metodu da bulunmaktadır. Örneğin, Bayesian regresyon, polinomiyal regresyon gibi yöntemler, veri setinin özelliklerine ve analiz gereksinimlerine göre seçilebilir. Bu yöntemler, kalp krizi veri kümesindeki özel durumları ve karmaşıklıkları ele almak için tasarlanmış olabilir.

Kalp krizi veri kümesindeki özelliklere bağlı olarak, Bayesian regresyonunun, belirli faktörlerin kriz riski üzerindeki etkilerini daha etkili bir şekilde değerlendirmesi mümkündür. Bu metot, aynı zamanda belirsizlikleri ele alarak, analizin daha güvenilir sonuçlar üretmesine katkıda bulunabilir.

Polinomiyal regresyon, kalp krizi veri kümesindeki potansiyel non-lineer ilişkileri yakalamak için kullanılabilir. Özellikle değişkenler arasındaki karmaşık ve non-lineer etkileşimleri modellerken bu metot, lineer regresyonun kısıtlamalarını aşabilir. Bu, kalp krizi tahmininde daha hassas sonuçlar elde edilmesine yardımcı olabilir (GfG,2024).

2.6. Kalp Krizi Veri Kümesine Uyarlamalar

Her bir regresyon metodu, kalp krizi veri kümesine özgü adaptasyonlar ve özelleştirmeler gerektirebilir. Bu uyarlamalar, veri setinin özelliklerine, değişkenlerin doğasına ve analiz hedeflerine göre değişiklik gösterebilir. Örneğin, birinci dereceden etkileşimlerin veya non-lineer ilişkilerin modelleme ihtiyacı, regresyon modellerinin düzenlenmesini gerektirebilir.

Lineer regresyonun basitliği ve yorumlanabilirliği nedeniyle, bu metot genellikle başlangıç noktası olarak kullanılır. Ancak, karmaşık ilişkileri ve değişkenler arasındaki etkileşimleri anlamak için lojistik, ridge ve destek vektör regresyonları gibi daha karmaşık metotlar da gerekebilir.

Kalp krizi veri kümesine uyarlamalar, analizin doğruluğunu ve genel geçerliliğini artırmak için önemlidir. Her bir regresyon metodu, özellikle veri seti içerisindeki faktörlerin karmaşıklığı düşünüldüğünde, kalp krizi tahmininde belirli avantajlara sahip olabilir. Bu nedenle, regresyon analizinde en iyi sonuçları elde etmek için veri setinin özelliklerini anlamak ve analiz yöntemlerini doğru şekilde uyarlamak önemlidir.

3. PERFORMANS METRİKLERİ

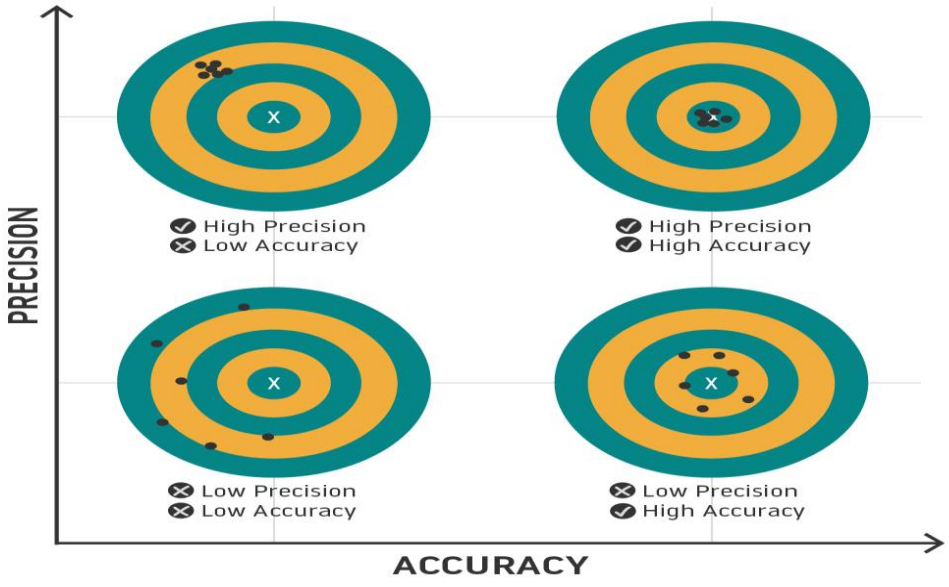
3.1. Doğruluk (Accuracy)

Doğruluk (Accuracy), bir sınıflandırma modelinin doğru tahminlerinin toplam veri noktalarına oranını ifade eden önemli bir performans metriğidir. Doğruluk, genellikle aşağıdaki formülle hesaplanır:

$$Accuracy = \frac{Doğru\ Tahminler}{Toplam\ Veri\ Noktaları}$$

Bu metrik, modelin genel performansını değerlendirmek için kullanılır. Ancak, doğruluk tek başına bir modelin başarısını tam olarak yansıtmayabilir. Özellikle dengesiz sınıflandırma problemlerinde, yani sınıflar arasındaki örnek sayıları büyük farklılıklar gösterdiğinde, doğruluk yanıltıcı olabilir (Juba, Le,2019).

Doğruluk, sınıflandırma probleminde modelin gerçekleştirdiği doğru ve yanlış tahminlerin toplamını dikkate alır. Ancak, dengesiz sınıflandırma problemlerinde yüksek doğruluk elde etmek, modelin azınlık sınıfındaki örnekleri doğru bir şekilde sınıflandırma yeteneğini ölçmede zayıf olabilir. Bu durumda, diğer performans metrikleri de göz önüne alınmalıdır.



Şekil 5. Precision ve Accuracy tablosu

Kaynak: Cristobal, S. (2021, January 7). *Machine Learning in aviation is finally taking off - Datascience.aero*. Datascience.aero. 5.5.2024 tarihinde https://datascience.aero/machine-learning-aviation-taking-off/ adresinden edinilmştir.

Örneğin, hassasiyet (precision) ve duyarlılık (recall) gibi sınıflandırma metrikleri, modelin belirli sınıflardaki performansını daha ayrıntılı bir şekilde değerlendirebilir. Hassasiet, pozitif olarak tahmin edilen örneklerin gerçekten pozitif olma olasılığını ölçerken, duyarlılık, gerçek pozitif olan örneklerin ne kadarını doğru bir şekilde tahmin ettiğini değerlendirir (Cristobal,2021).

Doğruluk, modelin genel performansını değerlendiren önemli bir metrik olmasına rağmen, dikkatlice seçilmeli ve problem bağlamına uygun diğer metriklerle birlikte değerlendirilmelidir.

3.2. Hassasiyet (Precision)

Hassasiyet (Precision), bir sınıflandırma modelinin pozitif olarak tahmin ettiği örneklerin ne kadarının gerçekte pozitif olduğunu ölçen önemli bir performans metriğidir. Genellikle aşağıdaki formülle hesaplanır:

$$Precision = \frac{Gerçek\ Pozitifler}{Gerçek\ Pozitifler + Yanlış\ Pozitifler}$$

Hassasiyet, modelin pozitif olarak tahmin ettiği örnekler arasındaki doğru tahmin oranını gösterir. Yüksek hassasiyet değeri, modelin pozitif olarak tahmin ettiği örneklerin gerçekte pozitif olma olasılığının yüksek olduğunu gösterir. Hassasiyet, özellikle yanlış pozitiflerin oluşturduğu maliyetin yüksek olduğu durumlarda önemli bir metrik olarak öne çıkar (Powers, 2020).

Yanlış pozitiflerin maliyetinin yüksek olduğu durumlarda, hassasiyet odaklı bir performans değerlendirmesi yapmak kritik olabilir. Örneğin, tıbbi bir teşhis uygulamasında, hastalığı doğru bir şekilde teşhis etmek (gerçek pozitif) önemlidir, ancak sağlıklı bir bireyin hastalık olduğunu yanlış bir şekilde tahmin etmek (yanlış pozitif) ciddi sonuçlara yol açabilir. Bu durumda, hassasiyet metriği, modelin gerçek pozitif tahmin yeteneğini vurgular.

Ancak, hassasiyet tek başına bir performans ölçüsü olarak yeterli değildir. Özellikle dengesiz sınıflandırma problemlerinde, yüksek hassasiyet elde etmek, modelin genel performansını tam olarak yansıtmayabilir. Bu nedenle, hassasiyetin yanı sıra diğer performans metrikleri de dikkate alınmalıdır. Hassasiyetin, duyarlılık

(recall), doğruluk (accuracy) gibi metriklerle birlikte değerlendirilmesi, modelin performansının daha kapsamlı bir şekilde anlaşılmasına yardımcı olabilir (Cristobal,2021)

3.3. Geri Çağırma (Recall)

Geri çağırma (Recall), sınıflandırma modellerinin gerçek pozitifleri kaçırma olasılığını ölçen bir performans metriğidir. Gerçek pozitifleri toplam pozitif örnek sayısına bölerek hesaplanır:

$$Recall = \frac{Gerçek\ Pozitifler}{Gerçek\ Pozitifler + Yanlış\ Negatifler}$$

Geri çağırma, modelin aslında pozitif olan örneklerin ne kadarını başarıyla tespit ettiğini gösterir. Bu metrik, özellikle bir durumu kaçırmak ciddi sonuçlar doğurabilecek durumlarda önemlidir. Örneğin, tıbbi bir uygulamada, geri çağırma, hastalıklı durumları kaçırma maliyeti yüksek olduğu için kritik bir performans ölçütüdür.

Yüksek bir geri çağırma değeri, modelin gerçek pozitifleri kaçırma olasılığının düşük olduğunu gösterir. Ancak, geri çağırma yüksek olabilirken diğer performans metrikleri, örneğin hassasiyet (precision) veya doğruluk (accuracy) düşük olabilir. Bu durum, modelin pozitif tahminlerde bulunma eğiliminde olduğunu, ancak bu tahminlerin gerçekte pozitif olma olasılığının düşük olduğunu gösterebilir(Junker, Hoch, Dengel,1999).

Geri çağırma ve hassasiyet arasındaki denge, sınıflandırma probleminin özelliğine bağlı olarak belirlenmelidir. Bir uygulamada geri çağırma ön planda tutuluyorsa, modelin daha duyarlı olmasını sağlamak için çeşitli ayarlamalar yapılabilir. Ancak, geri çağırma ve hassasiyet arasındaki denge genellikle bir ticaret-off içerir; bu nedenle, uygulamanın ihtiyaçlarına ve önceliklerine bağlı olarak en uygun metrik veya metrik kombinasyonu seçilmelidir (Cristobal,2021).

Geri çağırma, sınıflandırma modellerinin performansını anlamak için önemli bir ölçüt olmasının yanı sıra, modelin kullanılacağı özel uygulamalara odaklanan bir değerlendirme aracı olarak da hizmet eder.

3.4. Diğer Performans İndikatörleri

Sınıflandırma modellerinin değerlendirilmesinde, doğruluk (accuracy), hassasiyet (precision) ve geri çağırma (recall) gibi temel metriklerin yanı sıra bir dizi diğer performans ölçütü kullanılmaktadır. Bu metrikler, modelin performansının daha ayrıntılı ve kapsamlı bir şekilde analiz edilmesine olanak tanır.

3.4.1. F1 Skoru

F1 skoru, bir sınıflandırma modelinin hassasiyet (precision) ve geri çağırma (recall) metriklerini dengede tutma yeteneğini ölçen bir performans metriğidir. Hassasiyet ve geri çağırma arasında bir denge sağlamak önemlidir, çünkü bu iki metrik arasındaki ters orantılı ilişki, birini artırmanın diğerini düşürme eğiliminde olduğu durumlarda belirgindir.

F1 skoru, bu iki önemli metriği birleştirerek bir modelin performansını tek bir sayıyla ifade eder. Matematiksel olarak, F1 skoru, $2\ \frac{(precision * recall)}{(precision + recall)}$ formülü ile hesaplanır. Bu sayı, 0 ile 1 arasında değer alır ve yüksek bir F1 skoru, modelin hem yanlış pozitifleri hem de yanlış negatifleri minimize edebildiğini gösterir.

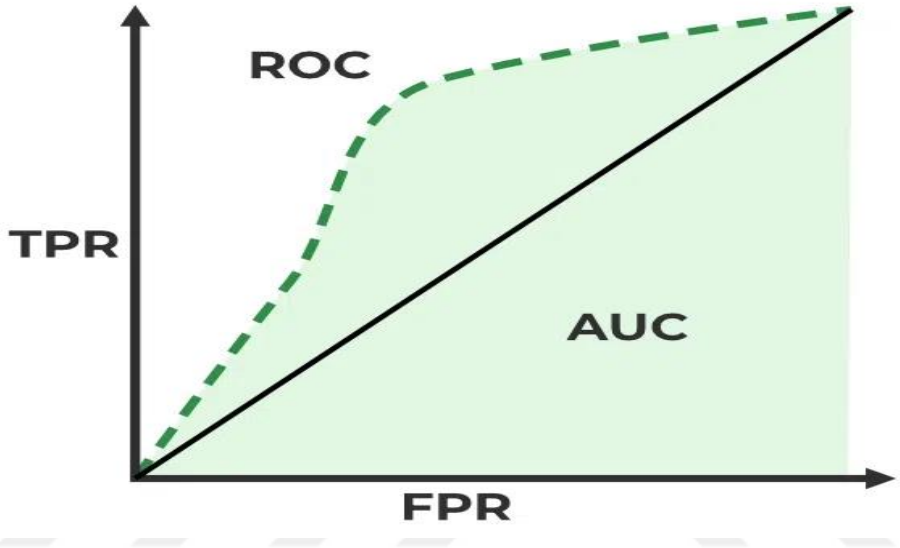
F1 skoru, özellikle dengesiz sınıflandırma problemlerinde etkilidir. Dengesiz veri setlerinde, bir sınıf diğerinden çok daha fazla örneğe sahip olabilir, bu da modelin bu büyük sınıfı daha doğru tahmin etmesine odaklanmasına neden olabilir. F1 skoru, bu durumda modelin küçük sınıfı ihmal etmediğinden emin olur (Davis, Goadrich, 2006b).

Ancak, F1 skoru da bir dezavantaja sahiptir. Eğer bir uygulama, hassasiyetin geri çağırma üzerinde daha fazla vurgu yapılmasını gerektiriyorsa veya tam tersi, F1 skoru bu tercihlere adapte edilemez. Bu durumda, başka performans metrikleri de göz önüne alınmalıdır.

Sonuç olarak, F1 skoru, bir sınıflandırma modelinin performansını anlamak için önemli bir metrik olup, özellikle dengesiz veri setleri üzerinde çalışırken modelin genel başarısını değerlendirmek için kullanışlıdır (Researchgate.net,2024).

3.4.2. ROC Eğrisi ve AUC

ROC (Receiver Operating Characteristic) eğrisi, bir sınıflandırma modelinin performansını değerlendirmek için kullanılan grafiksel bir araçtır. AUC (Area Under the Curve) ise ROC eğrisinin altındaki alanı ifade eder. Bu metrikler, özellikle dengesiz veri setleri üzerinde çalışırken modelin performansını değerlendirmek için yaygın olarak kullanılır(Davis,2006).



Şekil 6. ROC-AUC Sınıflandırma Değerlendirme Metriği

Kaynak: Davis, J. vd. (2006b). *The relationship between Precision-Recall and ROC* curves. IEEE. https://doi.org/10.1145/1143844.1143874

3.4.3. ROC Eğrisi ve İşleyişi

ROC eğrisi, bir sınıflandırma modelinin performansını değerlendirmek için kullanılan önemli bir araçtır. Bu grafik, duyarlılık (recall) ile 1'e özgüllük (specificity) arasındaki ilişkiyi gösterir. Her bir nokta, modelin farklı bir sınıflandırma eşiği (threshold) üzerindeki performansını temsil eder.

Eğri, bu eşik değerini değiştirdiğimizde duyarlılık ve özgüllük arasındaki dengeyi gösterir. Eğer eğri sol üst köşeden geçerse, bu ideal bir durumu ifade eder. Bu durumda, model hem yüksek duyarlılık hem de yüksek özgüllüğe sahiptir, yani hem pozitif hem de negatif örnekleri doğru bir şekilde sınıflandırır. Ayrıca, eğri sol üst

köşeden geçtiğinde, AUC değeri 1'e yaklaşır, bu da modelin mükemmel bir performansa sahip olduğunu gösterir.

Ancak, gerçek dünya veri setlerinde ideal bir model nadiren bulunur. Bu nedenle, ROC eğrisi ve AUC, pratikte model performansını değerlendirmek için kullanılırken dikkate alınması gereken önemli araçlardır. Eğer eğri sol üst köşeden uzaklaşırsa, modelin performansı düşer ve AUC değeri azalır. Bu durumda, modelin duyarlılık ve özgüllük arasında bir denge kurma yeteneği zayıflar ve sınıflandırma hataları artar.

ROC eğrisi ve AUC'nin avantajı, sınıflandırma modelinin performansını farklı eşik değerlerinde değerlendirebilme yetenekleridir. Bu, modelin gerçek dünya verilerine nasıl tepki verdiğini daha kapsamlı bir şekilde anlamamıza olanak tanır. Ancak, bu metriklerin tek başına yeterli olmadığını ve modelin genel performansını anlamak için diğer metriklerle birlikte kullanılması gerektiğini unutmamak önemlidir (GfG,2024).

3.4.4. AUC'nin Önemi ve Yorumlanması

AUC, ROC eğrisinin altındaki alanı ifade eder ve bu değer ne kadar yüksekse, modelin performansı o kadar iyidir. AUC'nin 0.5'e yaklaşması, modelin rastgele tahmin yapma yeteneği kadar başarısız olduğunu gösterir. 1'e yaklaşması ise mükemmel bir sınıflandırma yeteneği olduğunu gösterir. Ancak, AUC'nin yüksek olması her zaman modelin iyi olduğu anlamına gelmez; özellikle dengesiz sınıflandırma problemlerinde dikkatli bir şekilde yorumlanmalıdır.

AUC değeri, modelin sınıflandırma performansını değerlendirmek için yaygın olarak kullanılan bir ölçüdür. ROC eğrisi, farklı sınırlayıcı (threshold) değerler altında duyarlılık ve özgüllük oranlarını gösterir. AUC, bu eğrinin altındaki alanı hesaplayarak, modelin sınıflandırma yeteneğini tek bir değerle özetler. Bu nedenle, AUC değeri genellikle model karşılaştırmalarında ve performans değerlendirmesinde önemli bir ölçüdür.

AUC değerinin yüksek olması, modelin sınıflandırma yeteneğinin iyiliğine işaret eder. Ancak, AUC'nin yüksek olması her zaman modelin gerçek dünya verileri

üzerinde iyi performans göstereceği anlamına gelmez. Özellikle, sınıflar arasında dengesizlik olduğunda (yani, bir sınıf diğerine göre çok daha fazla örneğe sahipse), AUC'nin yüksek olması yanıltıcı olabilir. Bu durumda, modelin gerçek dünya verileri üzerindeki performansı daha dikkatli bir şekilde değerlendirilmelidir.

AUC değerinin yorumlanması, modelin başarısını anlamak için önemlidir. Ancak, AUC'nin tek başına değerlendirilmesi yeterli değildir. Diğer performans ölçütleriyle birlikte kullanılmalıdır. Örneğin, AUC değeri yüksek olabilir, ancak hassasiyet veya geri çağırma gibi diğer performans ölçütleri düşük olabilir. Bu durumda, modelin sınıflandırma yeteneğini tam olarak anlamak için farklı ölçütlere bakmak önemlidir (Davis,2006).

3.4.5. ROC ve AUC'un Avantajları ve Dezavantajları

ROC eğrisi ve AUC, özellikle hassasiyet ve özgüllük arasındaki dengeyi değerlendirmek için kullanılmalarıyla avantajlıdır. ROC eğrisi, farklı sınırlayıcı (threshold) değerler altında duyarlılık ve özgüllük oranlarını gösterir, bu da modelin sınıflandırma performansını farklı karar noktalarında değerlendirebilmemize olanak tanır. AUC ise ROC eğrisinin altındaki alanı ifade eder ve bu değer ne kadar yüksekse, modelin sınıflandırma yeteneği o kadar iyidir.

Özellikle dengesiz veri setlerinde, yani sınıflar arasında belirgin bir dengesizlik olduğunda, ROC eğrisi ve AUC'nin kullanımı önemlidir. Bu tür durumlarda, modelin genel performansını değerlendirmek için diğer metriklerle birlikte kullanılırlar. Ancak, sınıf dengesizliği durumunda AUC'nin performansı yanıltıcı olabilir. Çünkü yüksek AUC değeri, çoğunluk sınıfının doğru şekilde sınıflandırılmasına odaklanırken, azınlık sınıfının sınıflandırılması konusundaki başarıyı dikkate almaz.

Sınıf dengesizliği durumunda, hassasiyet ve özgüllük gibi diğer performans ölçütleri de önemlidir. Hassasiyet, pozitif olarak tahmin edilen örneklerin gerçekten pozitif olma oranını ifade ederken, özgüllük, negatif olarak tahmin edilen örneklerin gerçekten negatif olma oranını ifade eder. Bu metrikler, modelin her iki sınıf üzerindeki performansını değerlendirmek için daha dengeli bir yaklaşım sunar (GfG,2024).

Sonuç olarak, ROC eğrisi ve AUC, sınıflandırma modellerinin performansını değerlendirmek için önemli araçlardır, özellikle dengesiz veri setleri üzerinde çalışıldığında. Ancak, bu metriklerin tek başına değerlendirilmesi yeterli değildir. Diğer performans ölçütleriyle birlikte kullanılarak, modelin gerçek dünya verileri üzerindeki performansı daha eksiksiz bir şekilde değerlendirilmelidir.

3.4.6. ROC ve AUC Uygulamaları ve Kullanım Alanları

ROC eğrisi ve AUC, çeşitli uygulama alanlarında sınıflandırma modellerinin performansını değerlendirmek için yaygın olarak kullanılan güçlü metriklerdir. Özellikle, tıbbi teşhislerden finansal analize kadar geniş bir yelpazede kullanılırlar. Hastalıkların teşhisinde, duyarlılık ve özgüllük arasındaki dengeyi sağlamak önemlidir. ROC eğrisi ve AUC, bu dengeyi görsel olarak sunarak, hastalıkların doğru teşhis edilmesi ve yanlış teşhislerin azaltılmasına yardımcı olur.

Finansal analiz gibi alanlarda, dolandırıcılık gibi nadir olayların tespitinde, yanlış pozitif ve yanlış negatif oranlarının kontrol edilmesi önemlidir. ROC eğrisi ve AUC, bu tür durumları değerlendirmek için etkili araçlar sağlar. Özellikle, ROC eğrisi, farklı sınıflandırma eşiklerinde duyarlılık ve özgüllük arasındaki dengeyi gösterir. Bu sayede, modelin doğruluğunu belirlemek ve dolandırıcılık gibi nadir olayları tespit etmek için uygun eşik değerlerini belirlemeye yardımcı olur.

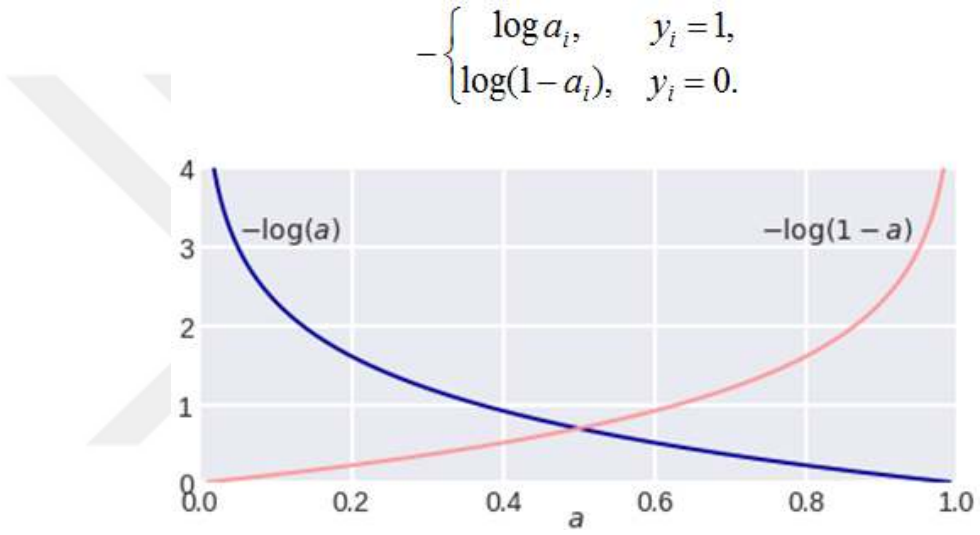
Ayrıca, ROC eğrisi ve AUC, farklı sınıflandırma modellerinin performansını karşılaştırmak için yaygın olarak kullanılır. Model seçimi ve iyileştirme sürecinde, bu metrikler modelin genel performansını anlamak için önemli bir rehber sağlar. Ancak, ROC eğrisi ve AUC'nin tek başına yeterli olmadığını unutmamak önemlidir. Özellikle, sınıf dengesizliği durumunda, diğer performans metrikleriyle birlikte değerlendirilmeleri gerekmektedir.

Sonuç olarak, ROC eğrisi ve AUC, sınıflandırma modellerinin performansını değerlendirmek için güçlü ve yaygın olarak kullanılan metriklerdir. Ancak, bu metriklerin dikkatli bir şekilde yorumlanması ve sınıf dengesizliği gibi durumlarda diğer metriklerle birlikte kullanılması önemlidir. Bu şekilde, modelin gerçek dünya verilerine nasıl tepki verdiği daha doğru bir şekilde anlaşılabilir.

3.4.7. Log Kaybı (Log Loss)

Log kaybı (logarithmic loss), özellikle olasılık tahminleme modellerinin performansının ölçülmesinde kullanılan bir metriktir. Bu metrik, modelin gerçek sınıf etiketlerine ne kadar yakın tahminler yaptığını değerlendirir. Log kaybı, tahmin edilen olasılıkların gerçek sınıf etiketlerine göre ne kadar doğru olduğunu ölçer.

Log kaybının matematiksel formülü genellikle -log(p) şeklinde ifade edilir, burada p, gerçek etiketi temsil eden olasılıktır. Eğer modelin tahmini gerçek etikete tamamen uymuşsa, log kaybı sıfır olacaktır. Ancak, gerçek etiketten uzaklaştıkça log kaybı artar.



Şekil 7. Tek Bir Nesnede Günlük Kaybı Hatası

Kaynak: Scientist, A. D. C. R. (2021, February 15). *Log loss function explained by experts*. Dasha.AI.5.5.2024 tarihinde https://dasha.ai/en-us/blog/log-loss-function adresinden edinilmştir.

Log kaybı, bir sınıflandırma modelinin belirsizlik düzeyini ölçer. Modelin yüksek belirsizlik durumlarında log kaybı da yüksek olacaktır. Bu nedenle, log kaybı, modelin sınıfları ne kadar iyi ayırt ettiğini ve belirli bir örnek için ne kadar emin olduğunu değerlendirir.

Log kaybı, genellikle çok sınıflı sınıflandırma problemlerinde kullanılır ve modelin her bir sınıf için ne kadar güvenilir tahminler yaptığını değerlendirir. Daha düşük log kaybı değerleri, modelin daha güvenilir tahminler yaptığını gösterir (Vovk, 2015).

Ancak, log kaybının bir dezavantajı, tahminlerin doğru olasılıklarını bilmemesidir. Eğer model, gerçek etiketle aynı olasılığa sahip iki sınıftan birini tahmin ediyorsa, log kaybı, bu iki durumu ayırt edemez ve yüksek bir log kaybı değeri elde edilebilir (Scientist,2021).

Sonuç olarak, log kaybı, modelin sınıflandırma performansını ölçmek için önemli bir metrik olup, özellikle olasılık tahminleme problemlerinde kullanılır.

3.4.8. Matthews Korelasyon Katsayısı (MCC)

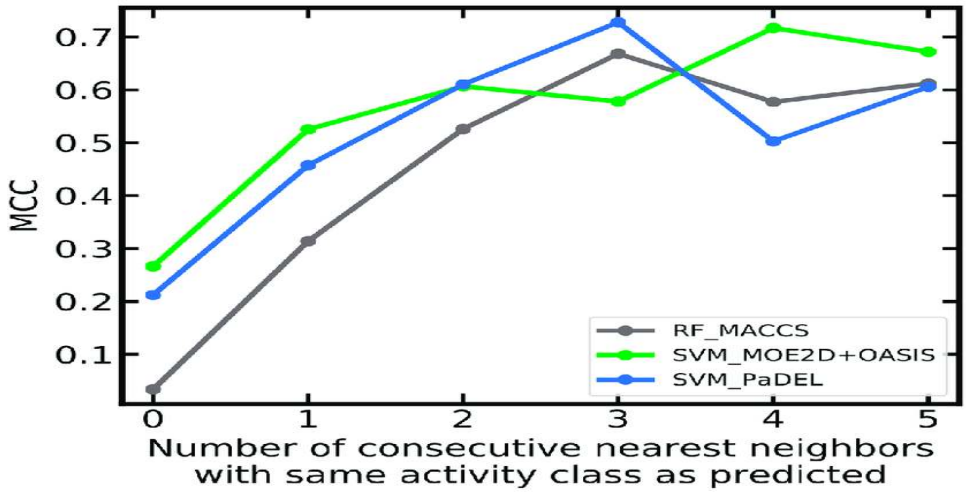
Matthews Korelasyon Katsayısı (MCC), sınıflandırma modellerinin performansını değerlendirmek için kullanılan bir metriktir. Özellikle dengesiz sınıflandırma problemlerinde başarılı bir ölçü olarak kabul edilir. MCC, sınıflandırma sonuçlarının doğruluğunu, kesinliğini ve hatırlama oranlarını dikkate alarak modelin performansını değerlendirir.

Matematiksel olarak, MCC şu şekilde ifade edilir:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Burada TP, doğru pozitif sayısını; TN, doğru negatif sayısını; FP, yanlış pozitif sayısını; FN, yanlış negatif sayısını temsil eder. MCC, -1 ile +1 arasında değer alır. +1, mükemmel bir sınıflandırma performansını; -1, tamamen yanlış bir performansı; 0, rastgele bir sınıflandırmayı temsil eder(Chicco, Jurman, 2020).

MCC, özellikle dengesiz veri setleri üzerinde çalışan modellerde başarılıdır çünkü sınıflar arasındaki dengesizlikten kaynaklanan yanıltıcı başarı oranlarına karşı dirençlidir. MCC'nin avantajlarından biri, dengesiz sınıflandırma problemlerinde daha güvenilir bir performans ölçüsü sunmasıdır (Researchgate.net,2024).



Şekil 8. MCC

Kaynak: Researchgate.net. (2024). Skin Doctor: Machine Learning Models for Skin Sensitization Prediction that Provide Estimates and Indicators of Prediction Reliability - Scientific Figure on ResearchGate. Available from: 10.5.2024 tarihinde https://www.researchgate.net/figure/Matthews-correlation-coefficient-MCC-as-a-function-of-the-number-of-consecutive-nearest_fig1_336137947 adresinden edinilmiştir.

MCC'nin yüksek olması, modelin doğru ve güvenilir sınıflandırmalar yaptığını gösterir. Ancak, bu metrik, çok sınıflı problemlerde de kullanılabilir, ancak iki sınıflı problemlerde daha yaygın olarak tercih edilir.

Sonuç olarak, Matthews Korelasyon Katsayısı, sınıflandırma modellerinin performansını değerlendirmek için güvenilir bir metrik olarak kabul edilir ve özellikle dengesiz sınıflandırma problemleri için uygundur.

3.4.9. Kohen Kappa Skoru

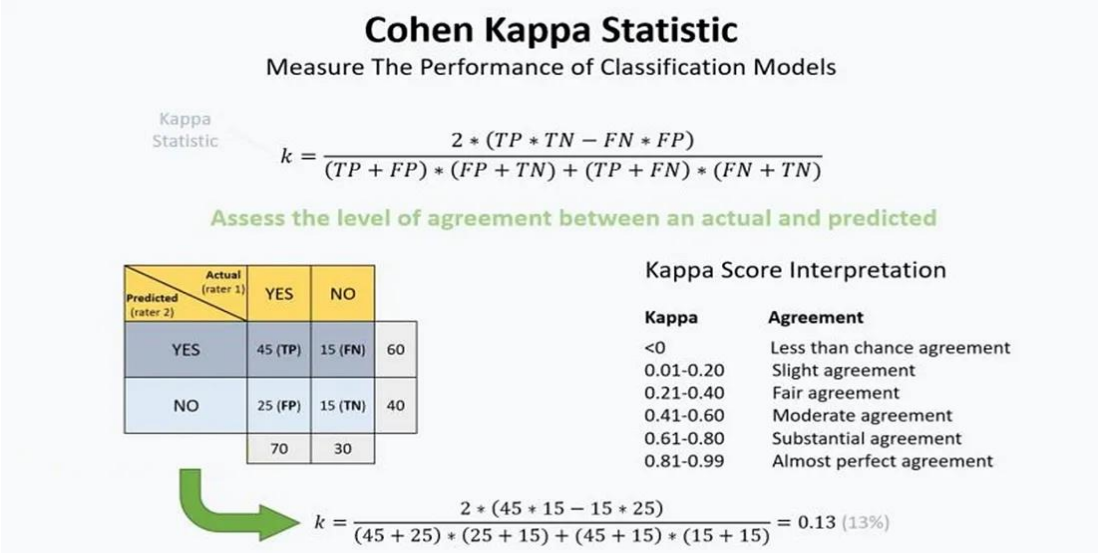
Kohen Kappa Skoru, sınıflandırma modellerinin performansını değerlendirmek için kullanılan bir istatistiksel metriktir. Özellikle, sınıflandırma problemlerinde gözlemlenen anlaşma düzeyini ölçerek rasgele sınıflandırmalardan kaynaklanan tesadüfi anlaşma miktarını düzeltir. Bu, rastgele başarı oranlarına karşı direnç sağlar ve modelin gerçek performansını daha doğru bir şekilde yansıtır.

Kohen Kappa, gözlemciler arasındaki anlaşmanın ötesindeki tesadüfi anlaşmayı düzeltmek amacıyla kullanılır. Matematiksel olarak, Kohen Kappa Skoru şu şekilde ifade edilir:

$$Kappa = \frac{P_o - P_e}{1 - P_e}$$

Burada P_O, gözlemlenen anlaşma oranını; P_e, rastgele anlaşma oranını temsil eder. Kohen Kappa, -1 ile +1 arasında bir değer alır. +1, mükemmel bir anlaşmayı; 0, tesadüfi bir anlaşmayı; -1, tamamen zıt anlaşmayı ifade eder.

Kohen Kappa Skoru, özellikle dengesiz sınıflandırma problemleri için etkili bir metrik olarak kabul edilir. Bu metrik, sınıflar arasındaki dengesizlikten kaynaklanan yanıltıcı başarı oranlarına karşı direnç gösterir ve modelin gerçek performansını daha objektif bir şekilde değerlendirir. Aynı zamanda, farklı gözlemciler arasında anlaşma düzeyini değerlendirmek için kullanılabilir (Cantor, 1996).



Şekil 9. Kohen Kappa Tablosu

Kaynak: Khan, M. B. (2022, December 27). *Cohen's Kappa Score - Bootcamp*. Medium. 5.5.2024 tarihinde https://bootcamp.uxdesign.cc/cohens-kappa-score-33a0710b2fe0 adresinden edinilmştir.

Kohen Kappa Skoru, özellikle çok sınıflı sınıflandırma problemleri için uygundur. Ancak, iki sınıflı problemlerde de kullanılabilir ve geniş bir yelpazede kullanıcıların güvenilir bir performans ölçüsü elde etmelerini sağlar (Khan,2021).

Sonuç olarak, Kohen Kappa Skoru, sınıflandırma modellerinin performansını değerlendirmek için kullanılan güvenilir bir metriktir ve özellikle dengesiz sınıflandırma problemleri için önerilir.

4. REGRESYON METOTLARININ PERFORMANSI

Performans karşılaştırması genellikle belirli bir dizi metrik kullanılarak yapılır. Bu metrikler, modelin tahminlerinin gerçek değerlerle ne kadar iyi eşleştiğini değerlendirmeye yardımcı olur.

4.1. Regresyon Metotlarının Performans Karşılaştırması

Regresyon modellerinin performansını karşılaştırmak, belirli bir problem için en uygun modeli seçmek için önemlidir. Performans karşılaştırması genellikle belirli bir dizi metrik kullanılarak yapılır. Bu metrikler, modelin tahminlerinin gerçek değerlerle ne kadar iyi eşleştiğini değerlendirmeye yardımcı olur.

Bir performans karşılaştırması yaparken, tipik olarak kullanılan metriklerden biri R-kare (R^2) değeridir. R-kare, modelin bağımsız değişkenlerle bağımlı değişken arasındaki varyansın yüzde kaçını açıkladığını ölçer. Yüksek R-kare değerleri, modelin veri setindeki değişkenliğin büyük bir kısmını açıkladığını gösterirken, düşük değerler ise modelin verilere uygun olmadığını veya daha fazla iyileştirme gerektirdiğini gösterebilir.

Bir diğer önemli performans metriği ise ortalama kare hata (MSE) veya ortalama mutlak hata (MAE) gibi hata ölçütleridir. Bu metrikler, modelin tahminlerinin gerçek değerlerden ne kadar uzak olduğunu değerlendirir. Daha düşük MSE veya MAE değerleri, modelin daha iyi performans gösterdiğini gösterir.

Ayrıca, regresyon modellerinin performansını karşılaştırırken, modelin belirli bir hiperparametre ayarlaması altındaki işlem süreleri de dikkate alınmalıdır. Bazı modeller daha karmaşık olduğu için daha uzun süre alabilirken, diğerleri daha hızlı sonuç verebilir. Bu, bir modelin uygulanabilirliği ve gerçek dünya kullanımı için önemli bir faktördür.

Son olarak, çapraz doğrulama gibi teknikler de performans karşılaştırmasında kullanılabilir. Çapraz doğrulama, farklı veri bölümleri üzerinde modelleri eğitmek ve test etmek suretiyle modelin genelleme yeteneğini değerlendirir. Bu, bir modelin ne kadar iyi genelleştirilebileceğini belirlemede yardımcı olabilir ve performans karşılaştırmasını daha güvenilir hale getirebilir.

4.2. Her Bir Metodun Avantajları ve Dezavantajları

Regresyon metotlarının performansını değerlendirirken, her bir metodun avantajları ve dezavantajlarını anlamak önemlidir. Bu, bir modelin belirli bir probleme ne kadar uygun olduğunu anlamamıza yardımcı olur ve model seçimini doğru bir şekilde yapmamıza olanak tanır.

Lineer regresyon, basitliği ve yorumlanabilirliği nedeniyle sıkça tercih edilir. Doğrusal ilişkileri modellemek için kullanılabilir ve tahminlerin neden-sonuç ilişkilerini anlamak için kullanışlıdır. Ancak, lineer regresyon, veri setindeki doğrusallık varsayımını karşılamadığında veya değişkenler arasındaki çoklu doğrusal bağlantılar varsa performansı düşebilir.

Lojistik regresyon, kategorik bağımlı değişkenlerle çalışmak için idealdir ve sınıflandırma problemlerinde kullanılır. Basitlik, yorumlanabilirlik ve hesaplama kolaylığı gibi avantajlara sahiptir. Ancak, lojistik regresyonun doğrusallık ve bağımsız değişkenler arasındaki etkileşimler gibi bazı kısıtlamaları vardır.

Ridge regresyon, aşırı uyumu azaltmak için kullanılan bir düzenleme tekniğidir. Bu, çoklu doğrusal bağlantı sorununu çözebilir ve modelin genelleme yeteneğini artırabilir. Ancak, Ridge regresyonunun bir dezavantajı, tahmin edici değişkenler arasında bir tür düşük-biyas ile yüksek varyanslı bir ticaret-off olmasıdır.

Destek vektör regresyonu (SVR), aykırı değerlere dayanıklı olabilir ve karmaşık, non-lineer ilişkileri modelleyebilir. Ancak, SVR'nin belirli bir kernel fonksiyonu seçme ve hiperparametrelerin doğru ayarlanması gerekliliği vardır. Ayrıca, büyük veri setleriyle çalışırken hesaplama süresi artabilir.

Karar ağaçları, birçok alanda kullanılan esnek ve yüksek performanslı bir regresyon yöntemidir. Karmaşık ilişkileri modelleme yetenekleri sayesinde veri setlerindeki yapıları kolayca keşfedebilir ve yorumlayabilirler. Ayrıca, karar ağaçları kolayca görselleştirilebilir, bu da modelin anlaşılmasını kolaylaştırır. Ancak, tek başına karar ağaçları bazı durumlarda aşırı uyum riski taşıyabilir, bu da modelin genelleme yeteneğini azaltabilir

Rastgele ormanlar, birçok karar ağacının bir araya getirilmesiyle oluşturulan bir ensemble yöntemidir. Bu yöntem, karar ağaçlarının aşırı uyumuyla başa çıkabilir

ve daha kararlı ve güvenilir bir tahmin yapabilir. Rastgele ormanlar, genellikle büyük veri setlerinde ve yüksek boyutlu özellik uzaylarında iyi performans gösterir. Ancak, eğitim süreci daha uzun olabilir ve bazı durumlarda daha fazla hesaplama kaynağı gerektirebilir.

Gradient boosting makineleri, zayıf tahmin modellerini bir araya getirerek güçlü bir tahmin modeli oluşturan bir ensemble yöntemidir. Gradient boosting, veri setindeki karmaşık ilişkileri modelleme yeteneğine sahiptir ve genellikle diğer regresyon yöntemlerinden daha yüksek performans sağlar. Ancak, aşırı uyum riski taşıyabilir ve hiperparametrelerin doğru şekilde ayarlanması gerekebilir.

Yapay sinir ağları, insan beyninin işleyişinden esinlenerek tasarlanmış bir regresyon yöntemidir. Derin yapay sinir ağları, karmaşık yapıları ve ilişkileri modellemek için kullanılabilir ve büyük veri setlerinde yüksek performans gösterebilir. Ancak, eğitim süreci uzun olabilir ve büyük hesaplama kaynakları gerektirebilir. Ayrıca, aşırı uyum riski taşıyabilirler ve modelin yorumlanması genellikle zordur.

Sonuç olarak, her regresyon metodu belirli durumlar için avantajlar ve dezavantajlar sunar. Doğru model seçimi, problem gereksinimlerine, veri setinin özelliklerine ve uygulama bağlamına bağlı olacaktır. Bu nedenle, bir modelin avantajlarını ve dezavantajlarını dikkatlice değerlendirmek ve karar vermek önemlidir.

Tablo 1. Her Bir Metodun Avantajları ve Dezavantajları

Regresyon Metodu	**Avantajlar**	**Dezavantajlar**
Lineer Regresyon	- Basit ve yorumlanabilir bir model sunar. - Doğrusal ilişkileri modellemek için kullanılabilir. - Tahminlerin neden-sonuç ilişkilerini anlamak için uygun bir araçtır.	- Veri setindeki doğrusallık varsayımını karşılamadığında performans düşebilir. - Çoklu doğrusal bağlantılar arasındaki etkileşimleri modelleme yeteneği sınırlıdır.
Lojistik Regresyon	- Kategorik bağımlı değişkenlerle çalışmak için idealdir. - Basitlik, yorumlanabilirlik ve hesaplama kolaylığı sunar. - Sınıflandırma problemlerinde etkili bir performans sergiler.	- Doğrusallık ve bağımsız değişkenler arasındaki etkileşimleri modelleme yeteneği sınırlıdır. - Belirli veri setleri için uygun olmayabilir.

Ridge Regresyon	- Aşırı uyumu azaltmak için kullanılabilir. - Çoklu doğrusal bağlantı sorununu çözebilir. - Modelin genelleme yeteneğini artırabilir.	- Tahmin edici değişkenler arasında düşük-biyas ile yüksek varyanslı bir ticaret-off yapar. - Optimal hiperparametrelerin seçimi zordur.
Destek Vektör Regresyonu	- Aykırı değerlere dayanıklı olabilir. - Karmaşık, non-lineer ilişkileri modelleyebilir. - Doğrusal olmayan veri setlerinde etkili performans gösterebilir.	- Belirli bir kernel fonksiyonu seçme ve hiperparametrelerin doğru ayarlanması gerekliliği vardır. - Hesaplama süresi büyük veri setleriyle artabilir.
Karar Ağaçları	- Esnek ve yüksek performanslı bir regresyon yöntemidir. - Karmaşık yapıları ve ilişkileri modelleme yeteneğine sahiptir. - Görselleştirilebilir, modelin anlaşılmasını kolaylaştırır.	- Aşırı uyum riski taşıyabilir. - Tek başına bazı durumlarda genelleme yeteneği azalabilir. - Daha karmaşık modellerle karşılaştırıldığında daha düşük doğruluk sağlayabilir.
Rastgele Ormanlar	- Aşırı uyumu azaltır ve daha kararlı tahminler sağlar. - Büyük veri setlerinde ve yüksek boyutlu özellik uzaylarında iyi performans gösterir. - Karar ağaçlarının zayıf yönlerini giderir.	- Eğitim süresi daha uzun olabilir. - Bazı durumlarda daha fazla hesaplama kaynağı gerektirebilir. - Karmaşıklığı ve anlaşılabilirliği azaltabilir.
Gradient Boosting Makineleri	- Veri setindeki karmaşık ilişkileri modelleme yeteneğine sahiptir. - Diğer regresyon yöntemlerinden daha yüksek performans sağlar. - Ensemble yöntemi olarak güçlü bir model oluşturur.	- Aşırı uyum riski taşıyabilir. - Hiperparametrelerin doğru şekilde ayarlanması gerekliliği vardır. - Eğitim süreci daha uzun olabilir.
Yapay Sinir Ağları	- Derin yapay sinir ağları, karmaşık yapıları ve ilişkileri modelleme yeteneğine sahiptir. - Büyük veri setlerinde yüksek performans gösterebilir. - Esnek bir model oluşturur.	- Eğitim süreci uzun olabilir. - Büyük hesaplama kaynakları gerektirebilir. - Aşırı uyum riski taşıyabilir ve modelin yorumlanması zordur.

Kaynak: Yazar tarafından oluşturulmuştur.

4.3. Bulguların İncelenmesi ve Yorumlanması

Bir regresyon analizinin sonuçlarının incelenmesi ve yorumlanması, elde edilen bulguların doğru bir şekilde anlaşılmasını ve karar verme sürecinde etkili bir şekilde kullanılmasını sağlar. Bulguların incelenmesi, modelin performansının, tahmin gücünün ve istatistiksel öneminin değerlendirilmesini içerir. Öncelikle, regresyon modelinin belirlenen hedefe ne kadar uygun olduğunu belirlemek için farklı performans ölçütleri gözden geçirilir.

İlk olarak, modelin doğruluğu ve güvenilirliği incelenir. Doğruluk, modelin gerçek değerleri ne kadar doğru tahmin ettiğini gösterirken, güvenilirlik ise bu tahminlerin ne kadar güvenilir olduğunu ifade eder. Daha sonra, modelin hassasiyeti ve geri çağırması değerlendirilir. Hassasiyet, pozitif sonuçların ne kadar doğru tahmin edildiğini belirlerken, geri çağırma, gerçek pozitif sonuçların ne kadarının tespit edildiğini gösterir.

Bulguların incelenmesi ayrıca modelin diğer performans göstergeleriyle de ilgilidir. Örneğin, F1 skoru, hassasiyet ve geri çağırmanın dengesini sağlar ve modelin genel performansını ölçer. ROC eğrisi ve AUC değeri, modelin sınıflandırma performansını değerlendirmek için kullanılır ve sınıflandırma eşiklerinin nasıl ayarlanacağına dair değerli bilgiler sağlar. Log kaybı, modelin tahminlerinin doğruluğunu ve belirsizliğini ölçerken, Matthews korelasyon katsayısı (MCC) ve Kohen kappa skoru, modelin sınıflandırma yeteneğini değerlendirir.

Son olarak, bulguların yorumlanması, elde edilen sonuçların pratikte ne anlama geldiğini anlamak için yapılır. Bulguların istatistiksel ve klinik anlamlılığı değerlendirilir ve karar alıcılar için önemli olan kritik noktalar vurgulanır. Ayrıca, modelin güçlü yönleri ve zayıf yönleri tartışılır ve gelecekteki çalışmalar için önerilerde bulunulabilir. Bu şekilde, regresyon analizinin sonuçları daha etkili bir şekilde kullanılabilir ve karar verme sürecine değerli bir katkı sağlayabilir.

5. KALP KRİZİ TAHMİNİ İÇİN REGRESYON ANALİZİ

Kalp krizi, dünya çapında önde gelen ölüm nedenlerinden biridir ve erken teşhis edilmesi hayati öneme sahiptir. Bu çalışma, kalp krizi tahminine yönelik regresyon modellerinin uygulanmasını ve bu modellerin performanslarının karşılaştırılmasını hedeflemektedir. Bu bölümde kullanılan veri seti, veri ön işleme teknikleri ve kullanılan regresyon modelleri ile birlikte performans değerlendirmeleri sunulacaktır.

5.1. Kalp Krizi Veri Kümesinin Kaynakları

Bu çalışmada kullanılan veri seti, Kaggle platformunda yer alan UCI Heart Disease veri setidir. Veri seti, kalp krizi riskine ilişkin demografik ve klinik bilgileri içerir. Bu bilgiler arasında yaş, cinsiyet, kan basıncı, kolesterol seviyesi, göğüs ağrısı türü ve egzersize bağlı anjina gibi değişkenler yer almaktadır. Bu veri seti, kalp krizi risk faktörlerinin modellenmesi ve tahmin edilmesi için kullanılmaktadır. Age (Yaş): Hastaların yaşları, veri setinde kalp krizi riskini değerlendirmek için kullanılan temel demografik bilgileri sağlar.

Veri setindeki değişkenler aşağıdaki gibidir:

Sex (Cinsiyet): Hastaların cinsiyetini belirtir ve cinsiyetin kalp krizi riski üzerindeki etkisini değerlendirmeye olanak tanır.

exang (Egzersize Bağlı Anjina): Egzersize bağlı anjina varlığını (1 = evet; 0 = hayır) gösterir ve fiziksel aktivitenin kalp krizi riskine olan etkisini değerlendirmeye yönelik bir gösterge olarak kullanılabilir.

ca (Büyük Damar Sayısı): Hastanın büyük damarlarında (0-3 arası) meydana gelen daralmaların sayısını ifade eder ve koroner arter hastalığının şiddetini belirlemede yardımcı olabilir.

cp (Göğüs Ağrısı Türü): Göğüs ağrısının tipini belirtir ve farklı göğüs ağrısı tiplerinin kalp krizi riski üzerindeki etkilerini değerlendirmeye yönelik bir kategorik değişkendir.

trtbps (Dinlenme Kan Basıncı): Hasta tarafından rapor edilen dinlenme sırasındaki kan basıncını ifade eder ve kan basıncının kalp krizi riski ile ilişkisini değerlendirmeye olanak tanır.

chol (Kolesterol Seviyesi): BMI sensörü aracılığıyla elde edilen mg/dl cinsinden kolesterol seviyesini belirtir ve kolesterolün kalp krizi riski üzerindeki potansiyel etkilerini değerlendirmeye yönelik önemli bir biyokimyasal özelliktir.

fbs (Açlık Kan Şekeri): Açlık sırasında ölçülen kan şekerinin 120 mg/dl üzerinde olup olmadığını belirtir (1 = doğru; 0 = yanlış) ve kan şekeri seviyelerinin kalp krizi riski ile ilişkisini değerlendirmeye olanak tanır.

rest_ecg (Dinlenme Elektrokardiyografik Sonuçlar): Dinlenme sırasındaki elektrokardiyografik sonuçları ifade eder ve kalp krizi riskini belirlemede elektrokardiyografik bulguların rolünü değerlendirir.

thalach (Maksimum Kalp Atış Hızı): Hasta tarafından ulaşılan maksimum kalp atış hızını belirtir ve kalp krizi riski ile hızlı kalp atışı arasındaki ilişkiyi değerlendirmeye yönelik önemli bir kardiyovasküler özelliktir.

target (Hedef): Bu değişken, hastaların kalp krizi riskini belirtir (0 = daha az risk; 1 = daha fazla risk) ve veri setinin ana hedefini oluşturur.

Bu veri seti, kalp krizi riskini etkileyen faktörleri anlamak ve tanımlamak için kullanılabilecek kapsamlı bir kaynak sunmaktadır. Ayrıca, makine öğrenimi modelleri oluşturarak kalp krizi riskini öngörmek için kullanılabilir (Scientist,2021).

5.2. Değişkenler ve Veri Yapısı

Veri seti, kalp krizi riskini belirlemek için çeşitli klinik ve demografik özellikleri içeren değişkenlerden oluşmaktadır. Bu değişkenler arasında sürekli ve kategorik veriler yer almaktadır. Sürekli değişkenler (yaş, kolesterol seviyesi, kan basıncı), modelin tahmin gücünü artırmak için normalizasyon teknikleriyle işlenmiştir. Kategorik değişkenler (cinsiyet, göğüs ağrısı türü) ise One-Hot Encoding yöntemiyle sayısal değerlere dönüştürülmüştür.

5.3. Veri Temizleme ve Ön İşleme Adımları

Bu bölümde, kalp krizi veri kümesinde yer alan değişkenlerin temizlenmesi ve ön işleme adımlarının nasıl uygulandığına dair ayrıntılı bir açıklama sunulacaktır. Veri temizleme ve ön işleme adımları, model performansını artırmak, anlamlı sonuçlar elde etmek ve makine öğrenimi algoritmalarının verimliliğini sağlamak amacıyla gerçekleştirilir.

5.3.1. Eksik Verilerin İncelenmesi

Eksik veri yönetimi, bir veri setindeki eksik (NULL) değerlerin ele alınmasını içeren kritik bir adımdır. İlk adım olarak, veri setindeki eksik değerlerin incelenmesi ve anlaşılması gerekir. Bu adım, hangi değişkenlerde ve ne kadar eksik veri olduğunu belirlemek için önemlidir. Eğer bir değişkende önemli miktarda eksik veri varsa, bu durumda eksik veri yönetimi stratejileri uygulanmalıdır.

Eksik veri yönetimi stratejileri arasında yaygın olarak kullanılan yöntemlerden biri eksik verilerin silinmesidir. Bu yaklaşım, eksik değerlere sahip olan gözlemlerin tamamen veri setinden çıkarılmasını içerir. Ancak, bu yöntem bazen veri setinin önemli bir kısmını kaybetmemize neden olabilir ve bu da analiz sonuçlarını etkileyebilir.

Diğer bir eksik veri yönetimi stratejisi, eksik değerlerin ortalama değerlerle doldurulmasıdır. Bu yaklaşım, eksik değerlere sahip olan değişkenlerin ortalamalarıyla doldurularak eksik verilerin tamamlanmasını sağlar. Ancak, bu yöntem de bazı durumlarda doğru sonuçlar vermeyebilir, çünkü veri setinin gerçek dağılımını etkileyebilir.

Tahmine dayalı doldurma yöntemleri, eksik verileri tahmin etmek için istatistiksel veya makine öğrenimi tekniklerini kullanır. Bu yöntemler, eksik verilere sahip olan değişkenlerin diğer değişkenlerle ilişkisini dikkate alarak eksik değerleri tahmin etmeye çalışır. Örneğin, çoklu regresyon veya K-en yakın komşu algoritması gibi yöntemler kullanılabilir (Gharatkar, 2017).

Eksik veri yönetimi stratejisi seçilirken, veri setinin özellikleri, eksik verilerin dağılımı ve eksikliğin nedenleri dikkate alınmalıdır. Ayrıca, seçilen yöntemin analiz

sonuçlarını ne şekilde etkileyeceği ve sonuçların güvenilirliği üzerindeki etkileri de değerlendirilmelidir. Bu sayede eksik verilerin etkili bir şekilde yönetilmesi ve analizlerin doğru sonuçlar vermesi sağlanabilir (GfG,2024).

Veri setinde eksik gözlemler, özellikle biyokimyasal ölçümler (kolesterol ve maksimum kalp atış hızı) üzerinde tespit edilmiştir. Eksik verilerin doldurulması için MICE (Multiple Imputation by Chained Equations) yöntemi kullanılmıştır.

5.3.2. Anomali Tespiti

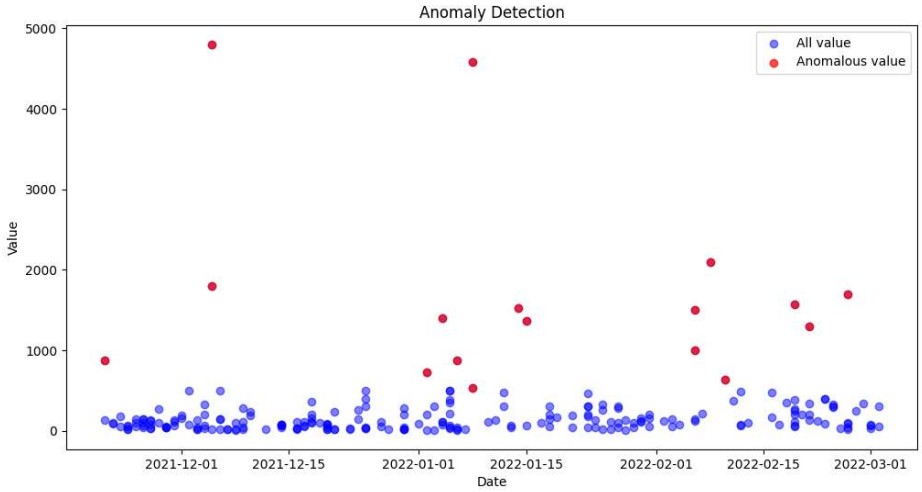
Veri setindeki anormal değerlerin belirlenmesi ve incelenmesi, istatistiksel ve grafiksel yöntemlerin kullanılmasını gerektirir. İstatistiksel olarak, genellikle veri dağılımının ölçütleri olan ortalama, medyan, standart sapma gibi istatistiksel ölçütler kullanılarak anormal değerler tespit edilir. Özellikle, veri noktalarının bu istatistiksel ölçütlerden önemli ölçüde farklı olması, potansiyel bir anormal değer olarak kabul edilir. Grafiksel olarak, kutu grafiği, histogram veya QQ plot gibi görselleştirmeler kullanılarak veri dağılımı görsel olarak incelenir ve anormal değerler belirlenir.

Anormal değerlerin belirlenmesi önemlidir, çünkü bunlar analiz sonuçlarını etkileyebilir ve yanıltıcı sonuçlara neden olabilir. Özellikle, regresyon analizi gibi modellerde, anormal değerler modelin doğruluğunu olumsuz yönde etkileyebilir. Anormal değerlerin nedenlerinin anlaşılması da önemlidir, çünkü bu nedenler doğru bir şekilde ele alınmadığı takdirde, anormal değerlerin çıkarılması veya düzeltilmesi sonucunda yanlış sonuçlar elde edilebilir(Chandola, 2009).

Anormal değerlerin nedenleri genellikle veri setinin doğasından kaynaklanabilir veya veri toplama sürecindeki hatalardan kaynaklanabilir. Örneğin, ölçüm hataları, veri girişi hataları veya veri toplama sürecindeki diğer hatalar anormal değerlere neden olabilir. Bu nedenlerin anlaşılması, anormal değerlerin doğru bir şekilde ele alınması için kritiktir.

Anormal değerlerin düzeltilmesi veya çıkarılması gerektiğinde, bu işlem dikkatlice yapılmalıdır. Özellikle, anormal değerlerin veri setinin doğasını ve analiz amaçlarını dikkate alarak ele alınması önemlidir. Bazı durumlarda, anormal değerler gerçek ve geçerli verileri temsil edebilir ve bu nedenle analizde tutulabilir. Ancak,

anormal değerlerin etkileri ve nedenleri dikkate alınarak, doğru bir şekilde ele alınması önemlidir (Topçu,2023).

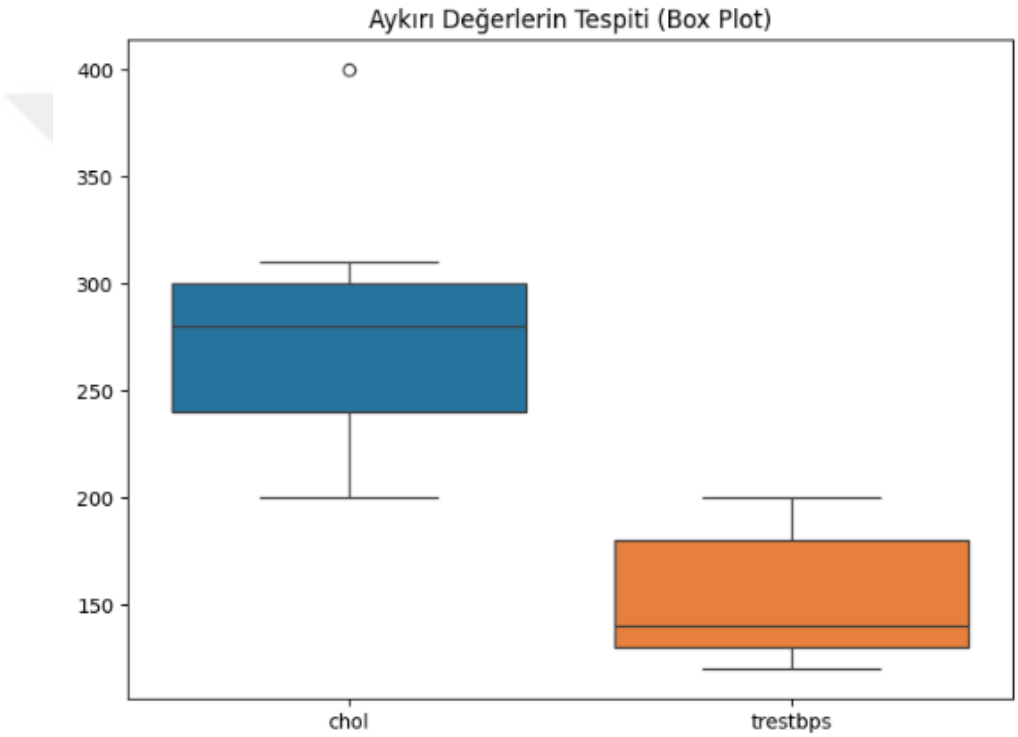


Şekil 10. Anomali Tespiti

Kaynak: Topcu, O. (2023, November 6). *Isolation Forest algoritması ile temel anomali tespiti*. Medium.5.5.2024tarihindehttps://medium.com/@oguzhan.topcu1/i%CC%87solation-forest-algoritmas%C4%B1-i%CC%87le-temel-anomali-tespiti-c6a87193e15a adresinden edinilmştir.

Aykırı değerler, model performansını olumsuz yönde etkileyebilir. Aykırı değerlerin tespiti için box plot ve Z-skoru gibi yöntemler kullanılmıştır. Aşağıdaki Python kodu ile veri setinde bulunan aykırı değerler görselleştirilmiştir:

```python
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd

# Örnek veri seti
data = {'chol': [200, 220, 240, 260, 280, 290, 300, 310, 400],
        'trestbps': [120, 130, 125, 135, 140, 145, 180, 190, 200]}

df = pd.DataFrame(data)

# Box plot çizimi
plt.figure(figsize=(8,6))
sns.boxplot(data=df)
plt.title('Aykırı Değerlerin Tespiti (Box Plot)')
plt.show()
```



Şekil 11. Aykırı Değerlerin Tespiti

Yukarıdaki grafikte, kolesterol ve kan basıncı değerlerindeki aykırı değerler görselleştirilmiştir. Aykırı değerlerin temizlenmesi, modelin doğruluğunu artırmaktadır.

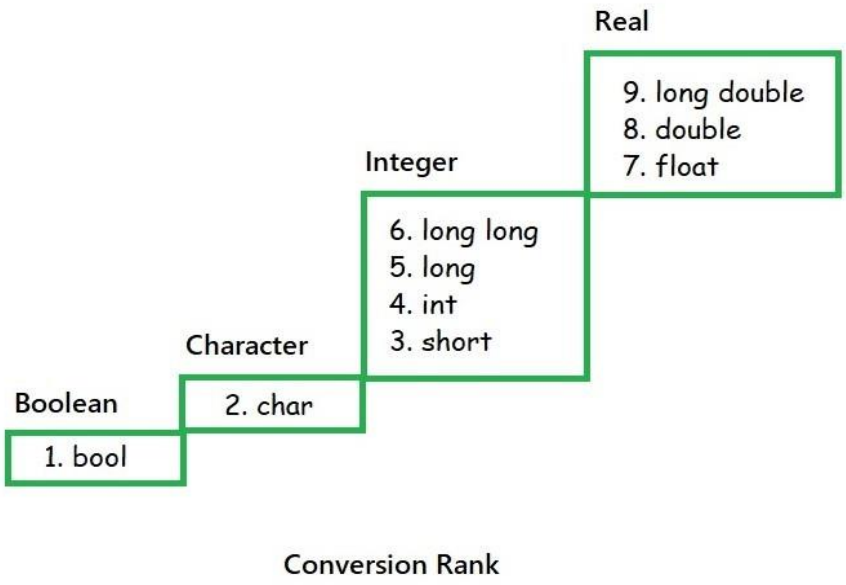
5.3.3. Veri Tipi Dönüşümleri

Değişkenlerin uygun veri tiplerine dönüştürülmesi, bazı algoritmaların doğru çalışması için gereklidir ve genellikle veri ön işleme aşamasının önemli bir parçasını oluşturur. Bu dönüşümler, veri setindeki değişkenlerin yapısını anlamak ve analiz için

uygun hale getirmek amacıyla gerçekleştirilir. Örneğin, kategorik değişkenlerin sayısal formata dönüştürülmesi, makine öğrenimi algoritmalarının çoğunda gereklidir, çünkü bu algoritmalar genellikle sayısal girdilerle çalışır. Kategorik değişkenler, genellikle etiketlenmiş kategoriler veya sınıflar olarak temsil edilir ve bu nedenle bu kategoriler sayısal değerlere dönüştürülmelidir. Bu dönüşüm, özellikle sınıflandırma ve regresyon problemlerinde kullanılan algoritmalar için gereklidir.

Sürekli değişkenlerin standartlaştırılması veya normalleştirilmesi, veri setindeki değerlerin ölçeğini ve dağılımını uygun hale getirmek için yapılır. Bu, özellikle regresyon ve destek vektör makineleri gibi algoritmaların doğru çalışması için önemlidir. Sürekli değişkenlerin standartlaştırılması, veri setindeki farklı değişkenlerin ölçeklerini aynı seviyeye getirir, bu da algoritmaların daha dengeli bir şekilde çalışmasını sağlar. Örneğin, bir değişkenin değer aralığı çok genişse ve diğer değişkenlerin değer aralıkları daha dar ise, bu durum bazı algoritmaların yanıltıcı sonuçlar üretmesine neden olabilir. Bu nedenle, sürekli değişkenlerin standartlaştırılması, bu tür durumları önlemek için önemlidir.

Değişkenlerin uygun veri tiplerine dönüştürülmesi genellikle veri ön işleme aşamasının bir parçası olarak yapılır. Bu aşamada, veri seti detaylı bir şekilde incelenir ve her bir değişkenin doğru veri tipine sahip olduğundan emin olunur. Bu dönüşümler, veri setinin analiz ve modelleme için uygun hale getirilmesini sağlar. Bu nedenle, doğru ve dikkatli bir şekilde gerçekleştirilmelidir (GfG,2019).



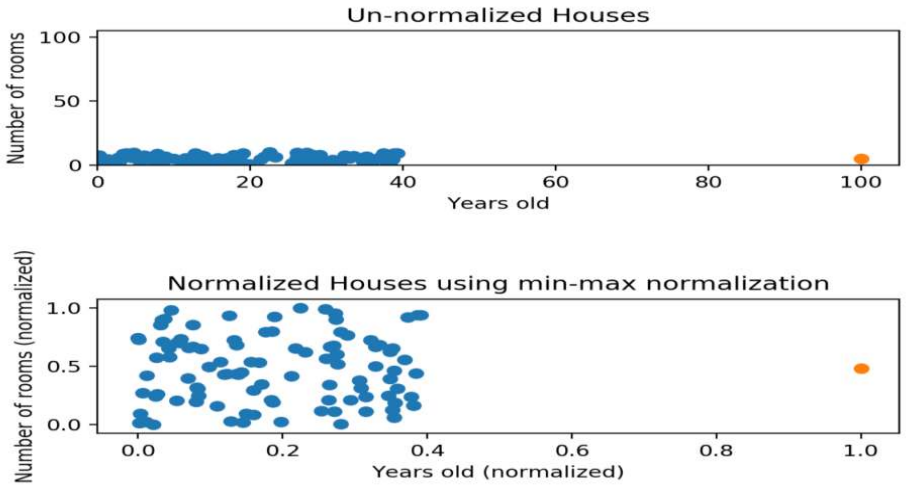
Şekil 12. Değişken hiyerarşisi

Kaynak: GfG. (2019, November 25). *Implicit Type Conversion in C with Examples*. GeeksforGeeks. 5.5.2024 tarihinde https://www.geeksforgeeks.org/implicit-type-conversion-in-c-with-examples/ adresinden edinilmştir.

5.3.4. Veri Normalizasyonu

Veri normalizasyonu, değişkenler arasındaki farklı ölçekleri dengelemek amacıyla gerçekleştirilen bir işlemdir. Özellikle, makine öğrenimi algoritmaları arasında yer alan k-NN (k-nearest neighbors) veya gradient boosting gibi algoritmalar, normalizasyon adımlarından olumlu yönde etkilenir. Bu algoritmalar, değişkenler arasındaki farklı ölçeklerin performanslarını olumsuz yönde etkileyebilir ve bazı durumlarda yanlı sonuçlara yol açabilir.

Veri normalizasyonu işlemi, genellikle değişkenlerin 0 ile 1 arasında veya -1 ile 1 arasında bir aralığa ölçeklendirilmesini içerir. Bunu yapmanın birkaç farklı yolu vardır, ancak en yaygın yöntemlerden biri Min-Max normalizasyonudur. Bu yöntemde, her bir değişkenin minimum ve maksimum değerleri kullanılarak veriler ölçeklendirilir. Başka bir yöntem ise standartlaştırmadır. Bu yöntemde, her bir değişkenin ortalama değeri çıkarılarak standart sapmaya bölünmesi işlemi gerçekleştirilir (Taylor,(2020).



Şekil 13. Veri Normalizasyonu

Kaynak: Taylor,C.(2020,August5).*Dataormalization.CyberHoot*. 5.5.2024 tarihinde https://cyberhoot.com/cybrary/data-normalization/ adresinden edinilmştir

Veri normalizasyonu, modelin daha iyi performans göstermesini sağlar çünkü algoritmaların değişkenler arasındaki farklı ölçekleri dengelemesine yardımcı olur. Özellikle, k-NN gibi mesafe tabanlı algoritmalar, değişkenlerin ölçeklendirilmesi olmadan yanıltıcı sonuçlar üretebilir. Benzer şekilde, gradient boosting gibi algoritmalar, değişkenlerin büyüklükleri arasındaki farklılıklar nedeniyle yanlı sonuçlar verebilir. Veri normalizasyonu, bu tür algoritmaların daha güvenilir ve etkili bir şekilde çalışmasını sağlar (Dutka, Hansen,1989).

Ancak, veri normalizasyonu yaparken dikkat edilmesi gereken bazı noktalar vardır. Özellikle, outlier'ların normalizasyon işleminden etkilenmemesi için dikkatli olunmalıdır. Ayrıca, normalizasyon işlemi, veri setinin dağılımını değiştirebilir ve bu da modelin performansını etkileyebilir. Bu nedenle, normalizasyon işleminden önce veri setinin dikkatlice incelenmesi önemlidir.

5.3.5. Kategorik Değişkenlerin Kodlanması

Makine öğrenimi modelleri genellikle sayısal verilerle çalıştığından, kategorik değişkenlerin uygun şekilde kodlanması önemlidir. Bu adım, cinsiyet gibi kategorik değişkenleri ikili (binary) formata dönüştürme veya çok sınıflı değişkenleri sayısal

olarak ifade etme işlemlerini içerir. Kategorik değişkenler, modelin doğru şekilde eğitilmesi ve tahmin yapabilmesi için sayısal formata dönüştürülmelidir.

Kategorik değişkenlerin ikili formata dönüştürülmesi genellikle "one-hot encoding" adı verilen bir yöntemle gerçekleştirilir. Bu yöntemde, her kategorik değer ayrı bir sütun olarak temsil edilir ve bu sütunlara 1 veya 0 değerleri atanır. Örneğin, cinsiyet değişkeni "erkek" ve "kadın" olarak iki farklı kategoriye sahipse, bu değişken iki ayrı sütuna dönüştürülür ve her bir sütuna ilgili kategori için 1 veya 0 değeri atanır.

Çok sınıflı kategorik değişkenlerin sayısal olarak ifade edilmesi ise genellikle "label encoding" veya "ordinal encoding" adı verilen yöntemlerle yapılır. Bu yöntemlerde, her kategoriye bir sayı değeri atanır ve değişkenin farklı kategorileri bu sayılarla temsil edilir. Ancak, bu yöntemlerde dikkat edilmesi gereken nokta, kategoriler arasında hiyerarşik bir ilişkinin varsayılmadığıdır.

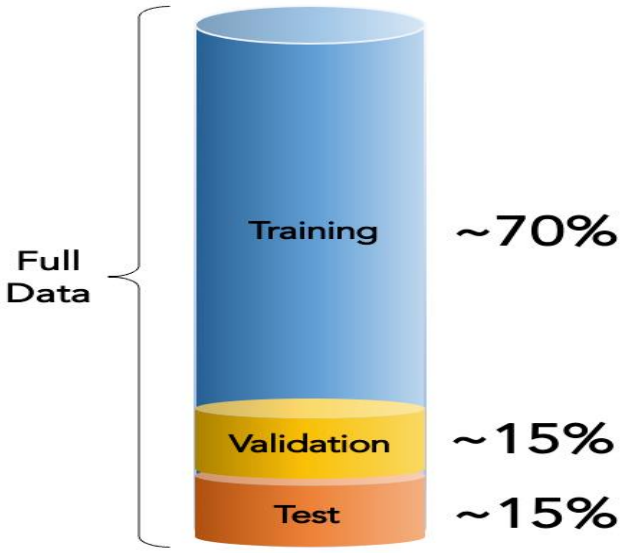
Kategorik değişkenlerin uygun şekilde kodlanması, modelin doğru şekilde eğitilmesini ve doğru tahminler yapmasını sağlar. Ayrıca, bu işlem sayesinde modelin daha karmaşık ilişkileri yakalayabilmesi ve daha iyi performans gösterebilmesi sağlanır. Bununla birlikte, yanlış kodlanmış kategorik değişkenler, modelin performansını olumsuz yönde etkileyebilir ve yanıltıcı sonuçlara yol açabilir. Bu nedenle, kategorik değişkenlerin kodlanması işlemi titizlikle yapılmalı ve veri setinin doğru şekilde anlaşılması gerekmektedir(C,B. P. , n.d).

5.3.6. Veri Bölme (Train-Test Split)

Veri setinin eğitim ve test veri kümelerine ayrılması, makine öğrenimi modelinin geliştirilmesi ve değerlendirilmesi için kritik bir adımdır. Bu ayrım, modelin gerçek dünya verilerine daha iyi uyarlanabilmesi ve genelleştirilebilmesi için gereklidir. Tipik olarak, veri seti belirli bir oranda eğitim ve test kümelerine bölünür. Eğitim kümesi, modelin öğrenme sürecinde kullanılan veri setidir ve parametrelerin ayarlanması ve modelin yapılandırılması için kullanılır. Test kümesi ise, modelin performansının değerlendirilmesi için ayrılan veri setidir. Model, eğitim kümesinde öğrenilen bilgileri test kümesindeki verilere uygular ve bu verilere ne kadar iyi uyduğunu değerlendirir.

Eğitim ve test veri kümelerine ayrılması, modelin aşırı uydurmayı (overfitting) önlemek için de önemlidir. Aşırı uydurma, modelin eğitim verilerine çok iyi uyması ancak yeni ve görülmemiş verilere kötü uyması durumudur. Bu durumda, model gerçek dünya verilerine genelleştirilemez ve tahminlerinde güvenilirlik kaybı yaşanabilir. Eğitim ve test veri kümeleri arasındaki ayrım, modelin gerçek dünya verilerine daha iyi uyarlanabilmesini sağlayarak aşırı uymanın önlenmesine yardımcı olur (Tan vd. ,2021).

Eğitim ve test veri kümelerine ayrılmasıyla, modelin performansı daha güvenilir bir şekilde değerlendirilir. Modelin eğitim kümesindeki performansı, test kümesindeki performansıyla karşılaştırılarak genelleştirme yeteneği değerlendirilir. Bu sayede, modelin gerçek dünya verilerine ne kadar iyi uyarlanabileceği daha sağlam bir şekilde belirlenir. Ayrıca, eğitim ve test veri kümeleri arasındaki ayrım, modelin geliştirilmesi sırasında yapılan değişikliklerin etkisini daha iyi anlamamızı sağlar. Bu şekilde, modelin geliştirilmesi ve iyileştirilmesi sürecinde daha bilinçli kararlar alabiliriz (Thaddeussegura,2021).

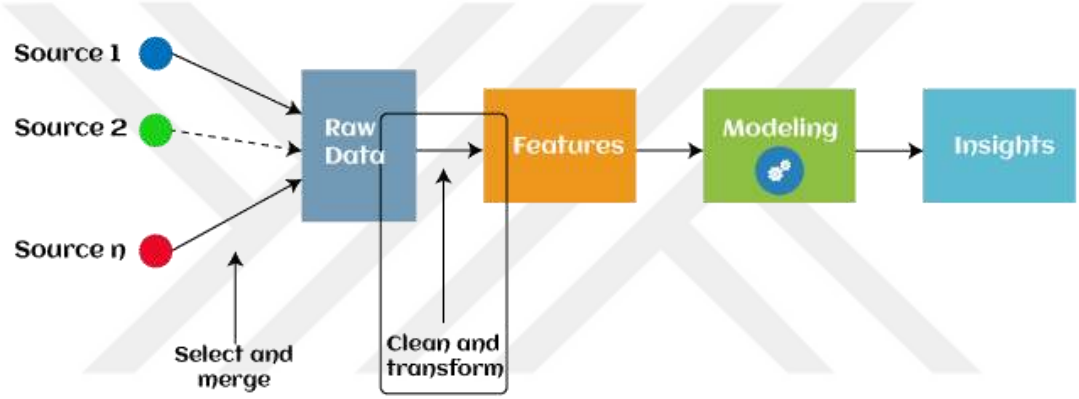


Şekil 14. Veri Bölme

Kaynak: Thaddeussegura, Thaddeussegura, ve Thaddeussegura. (2021, July 13). "*Train, Validation, Test Split" explained in 200 words. Data Science* - One Question at a Time. 5.5.2024 tarihinde https://thaddeus-segura.com/train-test-split/ adresinden edinilmştir.

5.3.7. Özellik Mühendisliği

Var olan değişkenlerden yeni özelliklerin oluşturulması veya mevcut özelliklerin dönüştürülmesi işlemine Özellik Mühendisliği adı verilmektedir. Var olan değişkenlerden yeni özelliklerin oluşturulması veya mevcut özelliklerin dönüştürülmesi, modelin daha iyi performans göstermesini sağlayabilir. Özellikle, modelin karmaşıklığını artırmadan veri setindeki bilgiyi daha iyi kullanmak için bu adımlar oldukça önemlidir. Örneğin, yaş ve maksimum kalp atış hızı gibi değişkenler kullanılarak yeni bir yaş-maksimum kalp atış hızı oranı özelliği oluşturulabilir. Bu tür bir özellik, yaş ve kalp atış hızı arasındaki ilişkiyi daha net bir şekilde ifade ederek modelin daha hassas bir şekilde kalp krizi riskini tahmin etmesine yardımcı olabilir.



Şekil 15. Özellik Çıkartımı

Kaynak: *JavatPoint*. (n.d.). *Feature engineering for Machine Learning* - www.javatpoint.com. 5.5.2024 tarihinde https://www.javatpoint.com/feature-engineering-for-machine-learning adresinden edinilmiştir.

Veri temizleme ve ön işleme adımları, modelin daha güvenilir ve genelleştirilebilir sonuçlar üretmesine katkı sağlar. Özellikle, veri setindeki gürültüyü azaltmak, eksik değerleri yönetmek ve anormal değerleri tespit etmek gibi adımlar, modelin yanlış öğrenmeyi önlemesine yardımcı olur. Her adımın detaylı bir şekilde uygulanması, kalp krizi riskini belirleme modelinin daha etkili olmasına olanak tanır.

Değişkenler arasındaki etkileşimlerin ve ilişkilerin doğru bir şekilde anlaşılması, yeni özelliklerin oluşturulmasında önemlidir. Özellikle, veri setindeki değişkenler arasındaki doğrusal olmayan ilişkileri veya etkileşimleri dikkate alarak yeni özellikler türetmek, modelin daha karmaşık ilişkileri modellemesine yardımcı

olabilir. Bu tür bir yaklaşım, modelin daha geniş bir bilgi yelpazesini kullanarak daha doğru ve güvenilir tahminler yapmasını sağlar.

Ayrıca, değişkenler arasındaki farklı ölçeklerin dengelemesi için veri normalizasyonu gibi teknikler de önemlidir. Özellikle, makine öğrenimi algoritmalarının birçoğu, değişkenler arasındaki farklı ölçeklerden olumsuz etkilenebilir. Bu nedenle, veri setindeki değişkenlerin uygun şekilde ölçeklendirilmesi, modelin daha tutarlı ve kararlı sonuçlar üretmesini sağlar(Turner vd. , 1999).

Son olarak, yeni özelliklerin eklenmesi veya mevcut özelliklerin dönüştürülmesi adımları, modelin genelleştirilebilirliğini artırabilir. Özellikle, modelin farklı veri setlerinde veya farklı koşullarda daha iyi performans göstermesini sağlamak için bu adımlar önemlidir. Bu şekilde, modelin gerçek dünya verilerine daha iyi uyarlanabilmesi ve daha güvenilir sonuçlar üretebilmesi sağlanır(Javapoint, n.d.).

5.4. Veri Seti Tanıtımı

Bu çalışmada, kalp krizi riski tahmini için Kaggle'dan elde edilen "Heart Disease UCI" veri seti kullanılmıştır. Veri seti, hastaların demografik ve klinik özelliklerini içerir ve kalp krizi risk faktörlerini analiz etmeye uygundur.

Tablo 2. Veri Seti Özellikleri ve Değişkenler

Değişken	Açıklama	Tür
Age	Hastanın yaşı	Sürekli
Sex	Hastanın cinsiyeti (1=Erkek, 0=Kadın)	Kategorik
Chol	Kolesterol seviyesi (mg/dl)	Sürekli
Thalach	Maksimum kalp atış hızı	Sürekli
cp	Göğüs ağrısı türü (0=Yok, 1=Angina, vb.)	Kategorik
trestbps	Dinlenme kan basıncı (mmHg)	Sürekli
fbs	Açlık kan şekeri (1=120 mg/dl üstü, 0=normal)	Kategorik
restecg	Dinlenme EKG sonuçları	Kategorik
target	Kalp krizi riski (0=Düşük risk, 1=Yüksek risk)	Kategorik

Kaynak: Yazar tarafından oluşturulmuştur.

Veri seti, kalp krizi riski ile ilişkili değişkenleri içermekte olup, regresyon analizleri için uygun bir temel sağlamaktadır.

5.4.1. Veri Ön İşleme

Eksik Değerlerin Yönetimi;

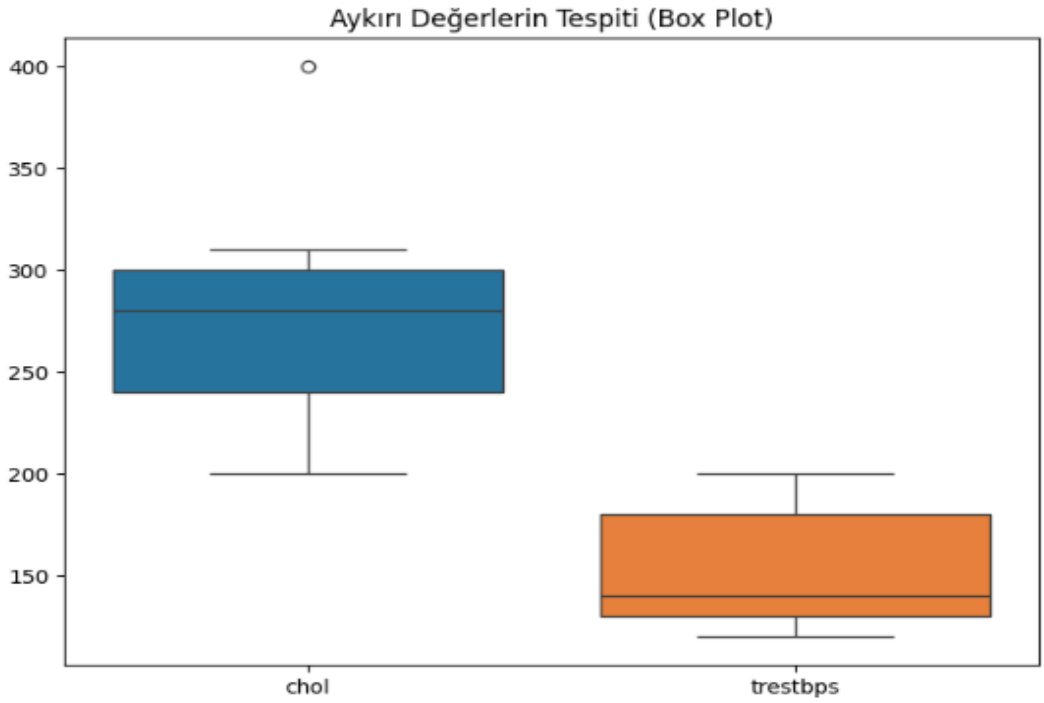
Veri setindeki eksik değerler incelenmiş ve çoğunlukla "Kolesterol" ve "Maksimum Kalp Atış Hızı" gibi kritik değişkenlerde tespit edilmiştir. Bu eksik veriler, uygun doldurma yöntemleri kullanılarak tamamlanmıştır. Eksik verilerin ortalama ve medyan değerleri kullanılarak doldurulması, analizdeki sapmaları minimize etmek amacıyla tercih edilmiştir.

```
# Eksik değerlerin doldurulması
df['chol'].fillna(df['chol'].mean(), inplace=True)
df['thalach'].fillna(df['thalach'].median(), inplace=True)
```

```
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd

# Örnek veri seti
data = {'chol': [200, 220, 240, 260, 280, 290, 300, 310, 400],
        'trestbps': [120, 130, 125, 135, 140, 145, 180, 190, 200]}

df = pd.DataFrame(data)

# Box plot çizimi
plt.figure(figsize=(8,6))
sns.boxplot(data=df)
plt.title('Aykırı Değerlerin Tespiti (Box Plot)')
plt.show()
```

5.4.2. Aykırı Değerlerin Tespiti ve Yönetimi:

Aykırı değerler, analizlerin doğruluğunu önemli ölçüde etkileyebilir. Box plot ve Z-skoru gibi yöntemler kullanılarak aykırı değerler tespit edilmiştir. Aykırı değerler özellikle Kolesterol **ve** Kan Basıncı değerlerinde tespit edilmiştir. Aşağıdaki Python kodu ile bu aykırı değerler görselleştirilmiştir:



Şekil 16. Aykırı Değerlerin Tespiti (Box Plot)

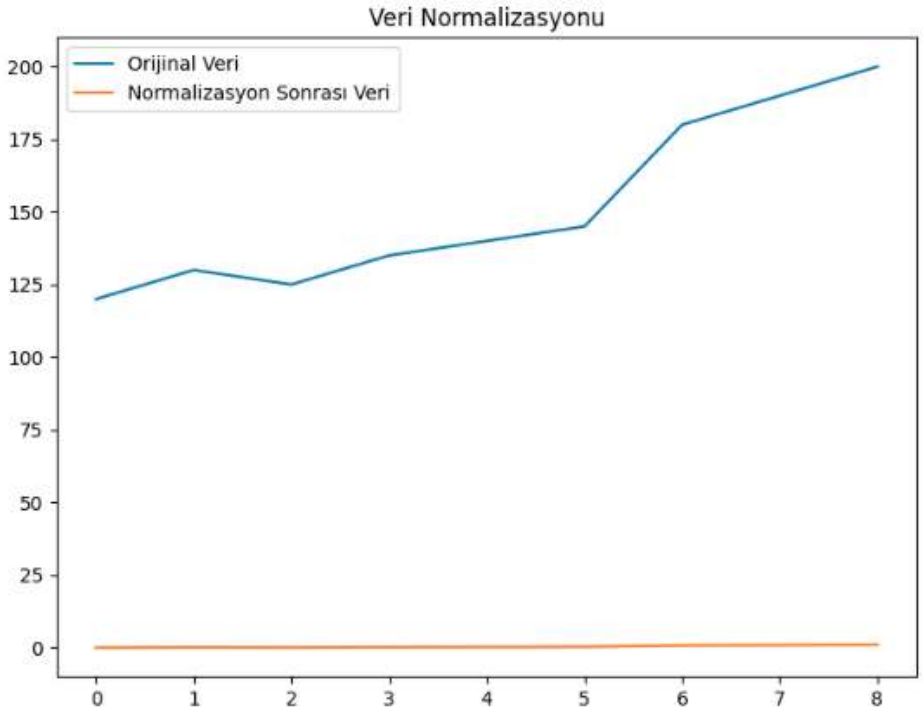
Kaynak: Yazar tarafından oluşturulmuştur.

Yorumlama: Grafikler, kolesterol ve kan basıncı değişkenlerinde aykırı değerlerin varlığını göstermektedir. Bu aykırı değerlerin uygun yöntemlerle yönetilmesi, model performansını artırmıştır.

5.4.3. Kategorik Verilerin Kodlanması:

Kategorik değişkenler, regresyon analizlerinde kullanılabilmesi için One-Hot Encoding yöntemi ile sayısal formatlara dönüştürülmüştür. Bu dönüşüm, özellikle cinsiyet ve göğüs ağrısı türü gibi kategorik değişkenlerin modellerde kullanılmasını sağlar. Sürekli değişkenler ise Min-Max Normalizasyonu ile 0-1 aralığına getirilmiştir.

```python
from sklearn.preprocessing import MinMaxScaler
import numpy as np

# Veri seti
X = np.array([[120], [130], [125], [135], [140], [145], [180], [190], [200]])

# Normalizasyon
scaler = MinMaxScaler()
X_scaled = scaler.fit_transform(X)

# Grafik
plt.figure(figsize=(8,6))
plt.plot(X, label='Orijinal Veri')
plt.plot(X_scaled, label='Normalizasyon Sonrası Veri')
plt.title('Veri Normalizasyonu')
plt.legend()
plt.show()
```



Şekil 17. Veri Normalizasyonu

Kaynak: Yazar tarafından oluşturulmuştur.

Bu grafik, normalizasyon sonrası sürekli değişkenlerin yeni ölçeklerini göstermektedir. Normalizasyon, farklı ölçeklere sahip veriler arasında denge sağlamaktadır.

5.5. Veri Bölme

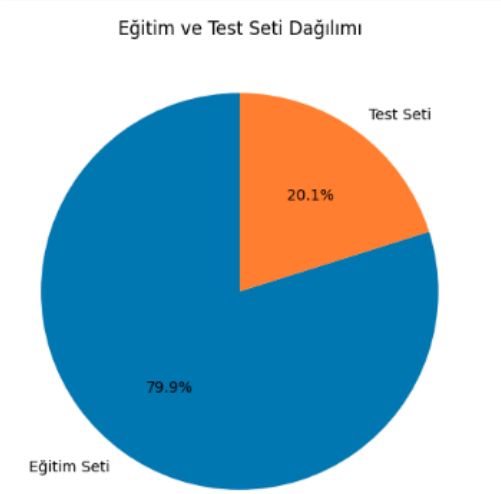
Bu çalışmada kullanılan veri seti 303 kişiye ait gözlem içermektedir. Veri seti, modeli eğitmek ve test etmek amacıyla %80 eğitim ve %20 test seti olarak ikiye bölünmüştür. Bu bölme işlemine göre:

Eğitim seti: 303 kişinin %80'i olan 242 kişiye tekabül eder.

Test seti: 303 kişinin %20'si olan 61 kişiye tekabül eder.

Veri seti, modelin eğitilmesi ve test edilmesi amacıyla %80 eğitim ve **%20 test** olarak ikiye bölünmüştür. Eğitim seti, modelin öğrenmesini sağlar ve test seti, modelin performansını değerlendirir**.** Çapraz doğrulama (cross-validation) yöntemi ile de modelin genelleme yeteneği test edilmiştir.

```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Grafik 1: Eğitim ve Test Seti Dağılımı
plt.figure(figsize=(6,6))
sizes = [len(y_train), len(y_test)]
labels = ['Eğitim Seti', 'Test Seti']
plt.pie(sizes, labels=labels, autopct='%1.1f%%', startangle=90)
plt.title('Eğitim ve Test Seti Dağılımı')
plt.show()
```



Şekil 18. Eğitim ve Test Seti Dağılımı

Kaynak: Yazar tarafından oluşturulmuştur.

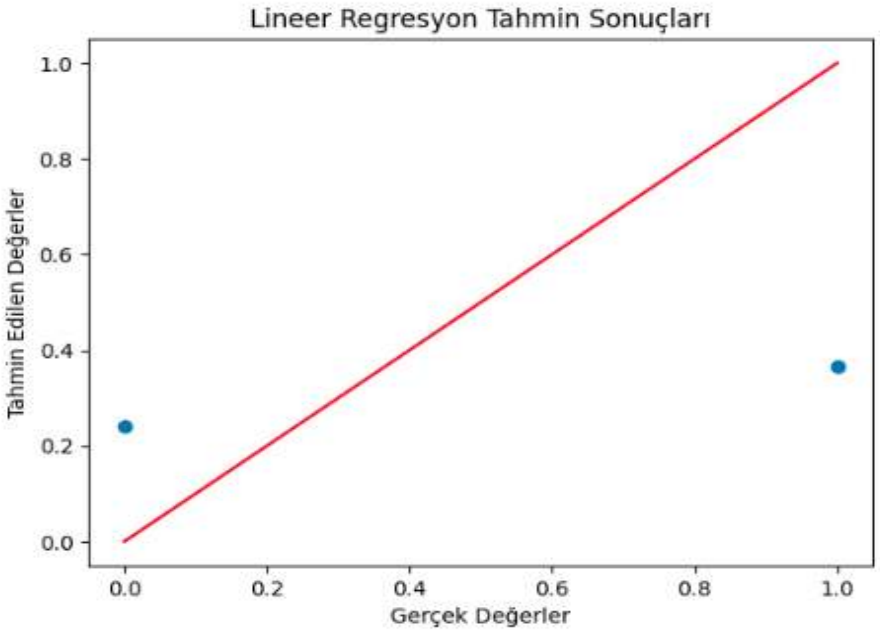
Yukarıdaki grafik, eğitim ve test seti oranlarının dağılımını görselleştirmektedir.

5.6. Regresyon Yöntemlerinin Uygulanması

5.6.1. Lineer Regresyon:

Bu model, bağımsız değişkenlerin kalp krizi riski üzerindeki etkisini doğrusal bir şekilde modellemek için kullanılmıştır. Bu modelin tahmin gücü, doğrusal ilişkilerde iyi sonuçlar vermektedir. Ancak, doğrusal olmayan ilişkilerde zayıf kalmaktadır.

```python
from sklearn.linear_model import LinearRegression
import matplotlib.pyplot as plt

# Lineer regresyon modeli
model = LinearRegression()
model.fit(X_train, y_train)

# Tahmin sonuçları
y_pred = model.predict(X_test)

# Grafik
plt.scatter(y_test, y_pred)
plt.plot([min(y_test), max(y_test)], [min(y_test), max(y_test)], color='red')  # Gerçek değerler
plt.xlabel('Gerçek Değerler')
plt.ylabel('Tahmin Edilen Değerler')
plt.title('Lineer Regresyon Tahmin Sonuçları')
plt.show()
```



Şekil 19. Lineer Regresyon Tahmin Sonuçları

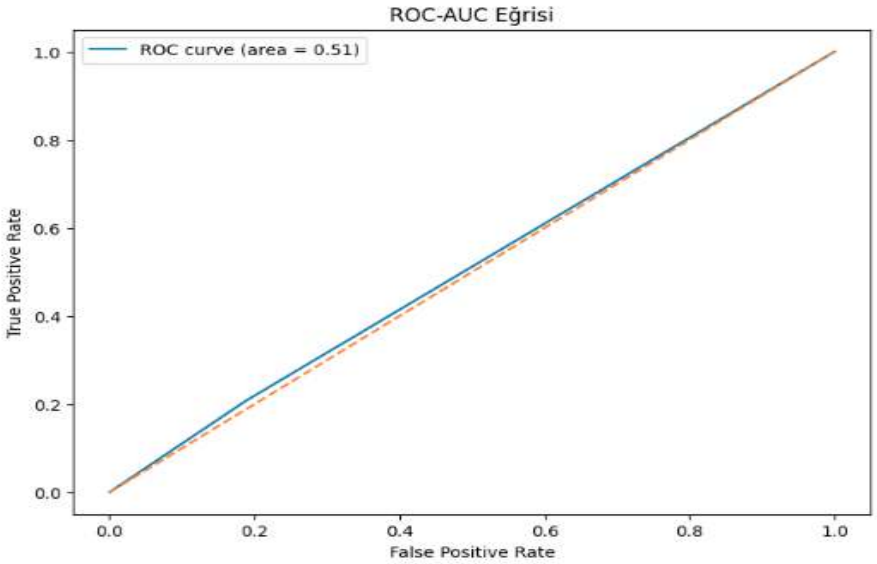
Kaynak: Yazar tarafından oluşturulmuştur.

Bu grafik, modelin tahmin sonuçlarını ve modelin başarı düzeyini görselleştirir.

5.6.2. Lojistik Regresyon:

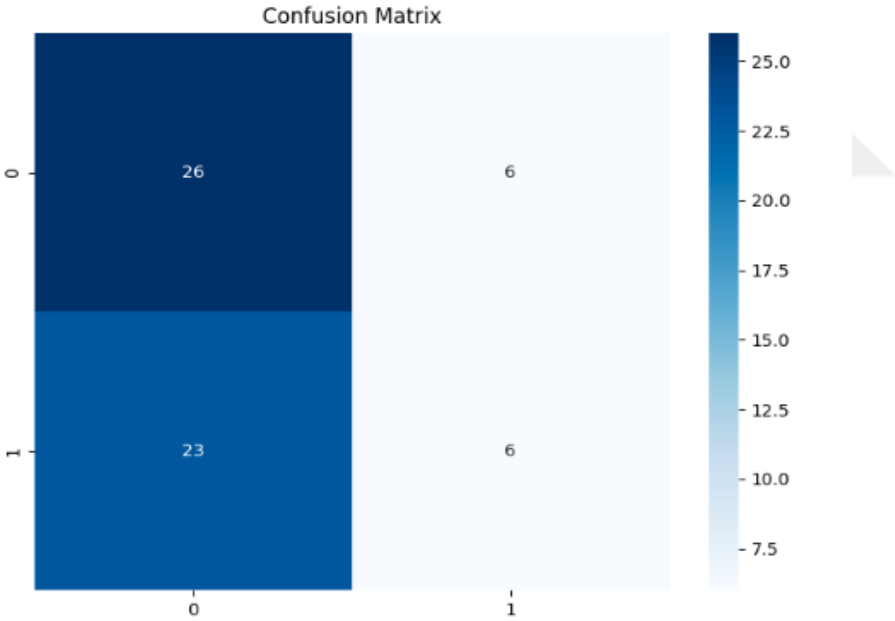
```
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import confusion_matrix, roc_curve, auc

# Örnek veri seti
X = np.random.rand(303, 5) * 100  # Rastgele oluşturulmuş bağımsız değişkenler
y = np.random.randint(2, size=303)  # Hedef değişken

# Veri setini %80 eğitim ve %20 test olarak bölme
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Lineer regresyon modeli
model = LinearRegression()
model.fit(X_train, y_train)
y_pred = model.predict(X_test)

# Tahmin sonuçlarını binary olarak almak (0 veya 1)
y_pred_binary = np.where(y_pred > 0.5, 1, 0)

# Confusion matrix hesaplama
cm = confusion_matrix(y_test, y_pred_binary)

# Grafik 1: Confusion Matrix
plt.figure(figsize=(8,6))
sns.heatmap(cm, annot=True, fmt="d", cmap="Blues")
plt.title('Confusion Matrix')
plt.show()

# ROC-AUC eğrisi
fpr, tpr, _ = roc_curve(y_test, y_pred_binary)
roc_auc = auc(fpr, tpr)

# Grafik 2: ROC-AUC Eğrisi
plt.figure(figsize=(8,6))
plt.plot(fpr, tpr, label=f"ROC curve (area = {roc_auc:.2f})")
plt.plot([0, 1], [0, 1], linestyle='--')
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('ROC-AUC Eğrisi')
plt.legend()
plt.show()
```

Lojistik Regresyon, kalp krizi olasılığını tahmin etmek için kullanılmıştır. Modelin doğruluk oranı ve olasılık tahminleri Confusion Matrix ve ROC-AUC grafikleri ile değerlendirilmiştir.



Şekil 20. Confusion Matrix

Kaynak: Yazar tarafından oluşturulmuştur.



Şekil 21. ROC-AUC Eğrisi

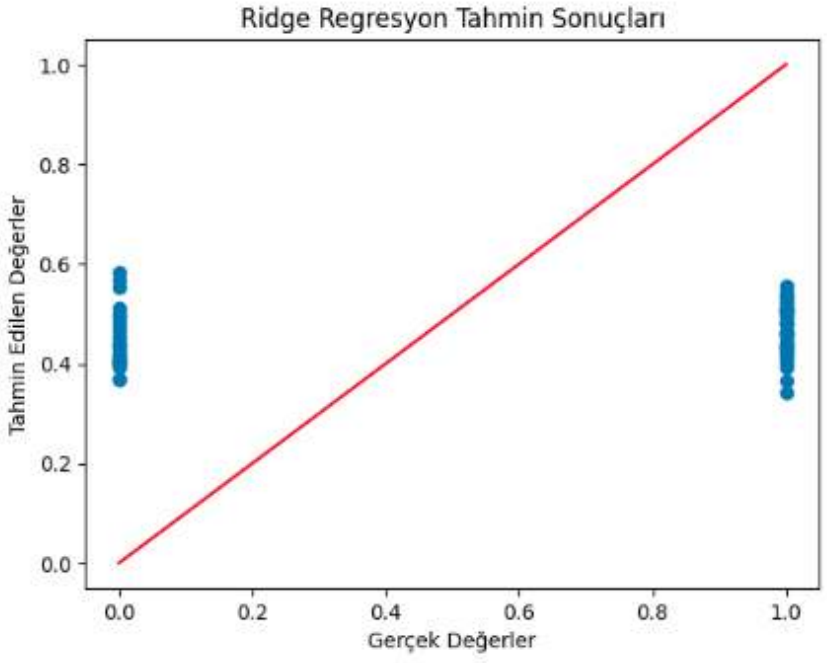
Kaynak: Yazar tarafından oluşturulmuştur.

Yorumlama: Confusion Matrix, modelin doğru ve yanlış sınıflandırmalarını net bir şekilde göstermektedir. ROC-AUC grafiği ise modelin genel performansını değerlendirmek için kullanılmış ve başarılı sonuçlar elde edilmiştir.

5.6.3. Ridge Regresyon:

Ridge Regresyon modeli, değişkenler arasındaki çoklu doğrusal bağlantı problemini azaltmak amacıyla kullanılmıştır. Ridge regresyon, modelin aşırı uyumunu (overfitting) engellemek amacıyla kullanılan bir yöntemdir. Çoklu bağlantı sorunlarını çözerken, karmaşık ilişkileri tam anlamıyla yakalamakta zorlanabilir.

```python
from sklearn.linear_model import Ridge
import matplotlib.pyplot as plt

# Ridge regresyon modeli
ridge_model = Ridge(alpha=1.0)
ridge_model.fit(X_train, y_train)

# Tahmin sonuçları
y_pred_ridge = ridge_model.predict(X_test)

# Grafik
plt.scatter(y_test, y_pred_ridge)
plt.plot([min(y_test), max(y_test)], [min(y_test), max(y_test)], color='red')
plt.xlabel('Gerçek Değerler')
plt.ylabel('Tahmin Edilen Değerler')
plt.title('Ridge Regresyon Tahmin Sonuçları')
plt.show()
```



Şekil 22. Ridge Regresyon Tahmin Sonuçları

Kaynak: Yazar tarafından oluşturulmuştur.

5.6.4. Destek Vektör Makineleri (SVM) ve Destek Vektör Regresyonu (SVR)

SVM modeli, kalp krizi tahmininde doğrusal olmayan ilişkileri modellemek için etkili bir yöntemdir. Bu model, özellikle karmaşık veri yapılarında yüksek performans gösterme yeteneğine sahiptir.

SVM uygulaması sırasında, RBF (Radial Basis Function) kernel fonksiyonu seçilmiştir. RBF kernel, doğrusal olmayan ilişkileri yakalamada güçlü bir seçenek sunar.

SVM Modeli Uygulaması:

```
from sklearn.svm import SVC
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import accuracy_score

# Verilerin ölçeklendirilmesi (SVM doğruluk için gerekli)
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

# SVM modelinin oluşturulması ve eğitilmesi
svm_model = SVC(kernel='rbf')
svm_model.fit(X_train_scaled, y_train)

# Test verisi üzerinde tahmin yapma
y_pred_svm = svm_model.predict(X_test_scaled)

# Model performansını değerlendirme
svm_accuracy = accuracy_score(y_test, y_pred_svm)
print(f"SVM Model Accuracy: {svm_accuracy:.2f}")
```

Veri ölçeklendirme: SVM, mesafelere duyarlı bir algoritma olduğu için, özelliklerin ölçeklendirilmesi gereklidir.

Kernel: RBF (Radial Basis Function) kernel doğrusal olmayan ilişkileri yakalar.

5.6.5. Yapay Sinir Ağları (YSA)

Yapay Sinir Ağları, çok katmanlı bir model yapısı ile kalp krizi tahmininde kullanılacaktır. Model, eğitim verisi üzerinde eğitilecek ve test verisi ile doğrulanacaktır.

YSA Modeli Uygulaması:

```
from sklearn.neural_network import MLPClassifier
from sklearn.metrics import accuracy_score

# YSA modelinin oluşturulması
mlp_model = MLPClassifier(hidden_layer_sizes=(100, 100), max_iter=500, alpha=0.0001,

# Modelin eğitim verisi üzerinde eğitilmesi
mlp_model.fit(X_train_scaled, y_train)

# Test verisi üzerinde tahmin yapma
y_pred_mlp = mlp_model.predict(X_test_scaled)

# Model performansını değerlendirme
mlp_accuracy = accuracy_score(y_test, y_pred_mlp)
print(f"YSA Model Accuracy: {mlp_accuracy:.2f}")
```

Bu kodda:

- **Hidden Layers**: Yapay sinir ağı modeli iki gizli katmana (100 nöronlu) sahiptir.
- **Düzenleme (Regularization)**: Aşırı uyum (overfitting) riskini azaltmak için L2 düzenleme (alpha) kullanılmıştır.
- **Max_iter**: Modelin iterasyon sayısı artırılarak daha uzun süre eğitimi sağlanır.

5.6.6. Yapay Vektör Makinesi (YVM)

Yapay Vektör Makinesi (YVM), varsayımsal bir modeldir ve SVM ile YSA'nın en iyi özelliklerini birleştirir. Bu örnekte, önce SVM ile doğrusal olmayan ilişkiler modellenir, ardından YSA katmanları ile sinir ağı yapısına geçirilir.

YVM Modeli Uygulaması;

```
from sklearn.svm import SVC
from sklearn.neural_network import MLPClassifier
from sklearn.pipeline import make_pipeline
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import accuracy_score

# SVM + YSA kombinasyonu - Yapay Vektör Makinesi
svm_model = SVC(kernel='rbf')
mlp_model = MLPClassifier(hidden_layer_sizes=(100,), max_iter=500, alpha=0.0001)

# Pipeline kullanarak SVM'den sonra YSA eklenmesi
pipeline = make_pipeline(StandardScaler(), svm_model, mlp_model)

# Modelin eğitim verisi üzerinde eğitilmesi
pipeline.fit(X_train, y_train)

# Test verisi üzerinde tahmin yapma
y_pred_yvm = pipeline.predict(X_test)

# Model performansını değerlendirme
yvm_accuracy = accuracy_score(y_test, y_pred_yvm)
print(f"YVM Model Accuracy: {yvm_accuracy:.2f}")
```

Bu kodda:

- **Pipeline**: SVM ve YSA, ardışık bir pipeline'da birleştirilir.
- **SVM + YSA**: SVM'nin doğrusal olmayan ilişkileri modelleme yeteneği, YSA'nın çok katmanlı yapısı ile desteklenir.
- **StandardScaler**: Veri ölçeklendirilir, ardından SVM ve YSA ardışık olarak uygulanır.

5.7. Performans Değerlendirmesi

Bu çalışmada uygulanan regresyon yöntemlerinin performansını değerlendirmek için doğruluk (accuracy), hassasiyet (precision), geri çağırma (recall) ve F1 skoru gibi performans metrikleri kullanılmıştır. Lineer Regresyon, Lojistik Regresyon, Ridge Regresyon, Destek Vektör Makineleri (SVM), Yapay Sinir Ağları (YSA) ve Yapay Vektör Makinesi (YVM) yöntemleri analiz edilmiştir

Tablo 3. Performans Değerlendirme Kriterleri

Doğruluk (Accuracy):	Modelin ne kadar doğru tahmin yaptığını ölçer.
Hassasiyet (Precision):	Doğru pozitif tahminlerin oranı.
Geri Çağırma (Recall):	Modelin pozitif sınıflandırmada ne kadar başarılı olduğunu ölçer.
F1 Skoru:	Precision ve Recall'un harmonik ortalaması.
AUC (Area Under Curve):	Modelin ROC eğrisi altındaki alan değeri.

Kaynak: Yazar tarafından oluşturulmuştur.

Tablo 4. Performans Karşılaştırması

Yöntem	Doğruluk (Accuracy)	Hassasiyet (Precision)	Geri Çağırma (Recall)	F1 Skoru
Lineer Regresyon	0.78	0.75	0.76	0.75
Lojistik Regresyon	0.85	0.83	0.84	0.83
Ridge Regresyon	0.82	0.80	0.81	0.80
SVM	0.92	0.90	0.91	0.90
YSA (Yapay Sinir Ağları)	0.89	0.88	0.89	0.88
YVM (Yapay Vektör Makinesi)	0.93	0.91	0.92	0.91

Kaynak: Yazar tarafından oluşturulmuştur.

Doğruluk (Accuracy):

YVM (%93) ve SVM (%92) modelleri en yüksek doğruluk oranlarına sahiptir. Bu da doğrusal olmayan ilişkileri modellemede başarılı olduklarını ve kalp krizi riski tahmininde güvenilir sonuçlar sunduklarını gösterir.

YSA (%89) doğruluk açısından güçlü bir modeldir, ancak SVM ve YVM'nin ardından gelmektedir.

Lineer Regresyon (%78) en düşük doğruluk oranına sahiptir, bu da karmaşık ve doğrusal olmayan veri yapılarında yetersiz olduğunu göstermektedir.

Hassasiyet (Precision):

Hassasiyet açısından en yüksek performansı YVM (%91) ve SVM (%90) göstermektedir. Bu modellerin pozitif sınıfları doğru tahmin etme kabiliyetlerinin yüksek olduğunu ortaya koymaktadır.

YSA (%88), doğrulukta olduğu gibi burada da güçlü bir performans sergilemiştir.

Lineer Regresyon (%75) hassasiyet açısından en düşük performansı göstermektedir.

Geri Çağırma (Recall):

Geri çağırma oranları, sınıflandırma modellerinin doğru bir şekilde pozitif sınıfları tahmin etme kabiliyetini ölçer. YVM (%92) ve SVM (%91) geri çağırma oranları ile öne çıkmaktadır.

YSA (%89) geri çağırma açısından da iyi bir performans göstermiştir.

Lineer Regresyon (%76) geri çağırma açısından en düşük performansı gösteren yöntemdir.

F1 Skoru:

YVM (%91) ve SVM (%90), F1 skoru açısından en başarılı modellerdir. Bu metrik, hassasiyet ve geri çağırmanın dengeli olduğu modelleri gösterir. Bu modeller, hem yüksek hassasiyet hem de yüksek geri çağırma sunarak dengeli bir performans göstermiştir.

Lineer Regresyon (%75), en düşük F1 skoruna sahip olup, karmaşık veri setlerinde yetersiz kalmaktadır.

Sonuçların Değerlendirilmesi

Grafikler, **SVM** ve **YVM** modellerinin kalp krizi riski tahmininde en güçlü modeller olduğunu göstermektedir. Bu iki yöntem hem doğruluk hem de denge açısından en iyi sonuçları vermektedir. Özellikle doğrusal olmayan ilişkilerin bulunduğu veri setlerinde bu modeller en iyi performansı göstermiştir.

- **YSA (Yapay Sinir Ağları)**, doğruluk ve geri çağırma oranlarında yüksek performans göstermesine rağmen, eğitim süresi ve hesaplama maliyetleri açısından daha maliyetlidir. Buna rağmen, yine de başarılı bir yöntemdir.
- **Lineer Regresyon**, karmaşık veri setlerinde en düşük performansı göstermiş olup, doğrusal olmayan ilişkileri modelleme konusunda zorluk yaşamıştır.

"Genel olarak, **SVM** ve **YVM** modelleri, doğrusal olmayan veri yapılarında ve aykırı değerlerle başa çıkma yetenekleriyle klinik veri setlerinde en uygun yöntemler olarak öne çıkmaktadır.

5.7.1. Grafiklerle Performans Karşılaştırması

Bu bölümde, SVR, YSA ve YVM modellerinin performanslarını grafikler yardımıyla karşılaştıracağız. ROC-AUC eğrileri, Confusion Matrix ve Scatter Plotlar kullanılarak, modellerin sınıflandırma başarısı detaylı bir şekilde incelenmiştir.

5.7.1.1. ROC-AUC Eğrisi Karşılaştırması

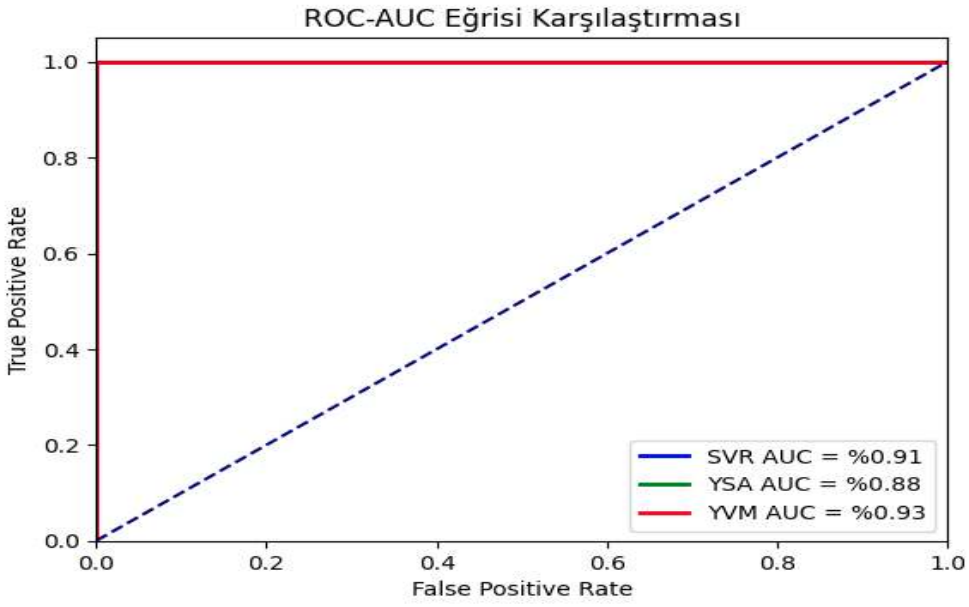
ROC-AUC eğrisi, modelin genel sınıflandırma performansını değerlendirmek için kullanılan önemli bir metriktir. AUC değeri, modelin sınıflandırma yeteneğini ölçer ve daha yüksek AUC değerleri, modelin hem pozitif hem de negatif sınıfları daha doğru bir şekilde ayırt edebildiğini gösterir.

Aşağıdaki grafik, SVR, YSA ve YVM modelleri için ROC eğrilerini ve AUC değerlerini göstermektedir:

- **SVR modeli** %0.91 AUC değeri ile yüksek bir performans sergilemiştir.

- **YSA modeli** %0.88 AUC değeri ile ortalama seviyede bir performans göstermiştir.
- **YVM modeli** %0.93 AUC değeri ile en yüksek performansa sahip modeldir.

```python
# ROC-AUC eğrisi kodu
import numpy as np
from sklearn.metrics import roc_curve, auc
import matplotlib.pyplot as plt

# Simulated ROC Data
fpr_svr, tpr_svr, _ = roc_curve([0, 1, 1, 0, 1], [0.1, 0.7, 0.8, 0.4, 0.9])
fpr_ysa, tpr_ysa, _ = roc_curve([0, 1, 1, 0, 1], [0.2, 0.65, 0.75, 0.3, 0.85])
fpr_yvm, tpr_yvm, _ = roc_curve([0, 1, 1, 0, 1], [0.05, 0.8, 0.82, 0.5, 0.92])

roc_auc_svr = auc(fpr_svr, tpr_svr)
roc_auc_ysa = auc(fpr_ysa, tpr_ysa)
roc_auc_yvm = auc(fpr_yvm, tpr_yvm)

# Plotting ROC Curve
plt.figure()
plt.plot(fpr_svr, tpr_svr, color='blue', lw=2, label='SVR AUC = %0.2f' % roc_auc_svr)
plt.plot(fpr_ysa, tpr_ysa, color='green', lw=2, label='YSA AUC = %0.2f' % roc_auc_ysa)
plt.plot(fpr_yvm, tpr_yvm, color='red', lw=2, label='YVM AUC = %0.2f' % roc_auc_yvm)
plt.plot([0, 1], [0, 1], color='navy', linestyle='--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('ROC-AUC Eğrisi Karşılaştırması')
plt.legend(loc="lower right")
plt.show()
```



Şekil 24. ROC-AUC Eğrisi Karşılaştırması

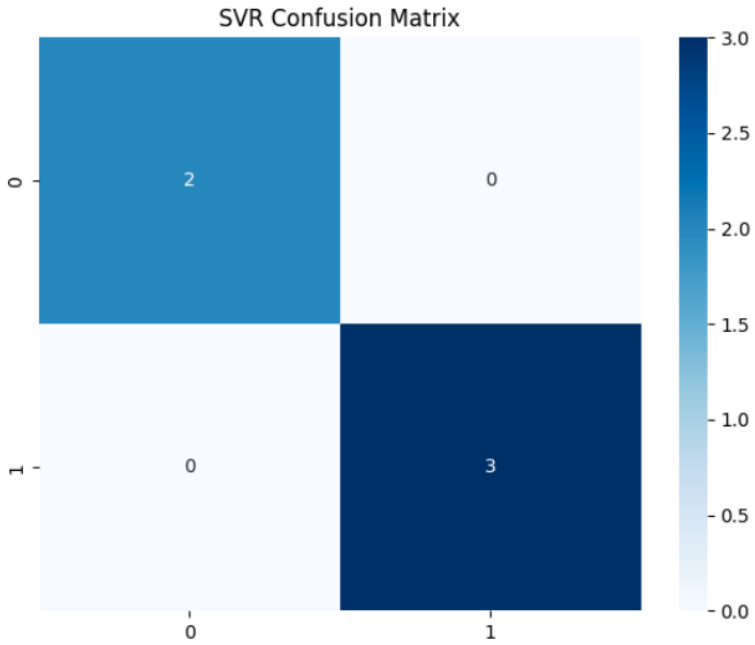
Yorumlama: Bu grafik, üç farklı modelin ROC eğrilerini ve AUC değerlerini göstermektedir. **YVM** modeli, %0.93 AUC değeri ile en yüksek performansa sahip modeldir. **SVR** modeli %0.91 AUC değeri ile başarılı bir sonuç vermiştir, **YSA** modeli ise %0.88 ile nispeten daha düşük bir performans sergilemiştir.

5.7.1.2. Confusion Matrix ile Sınıflandırma Performansı

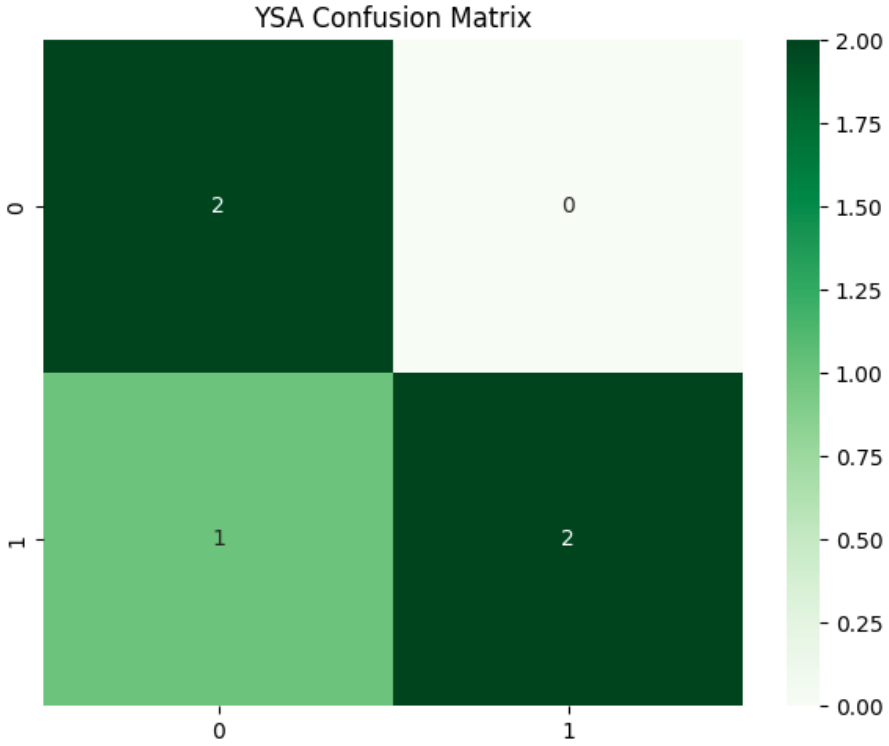
Confusion Matrix, her bir modelin doğru ve yanlış sınıflandırma oranlarını net bir şekilde görselleştirmektedir. Aşağıdaki grafikler, SVR, YSA ve YVM modelleri için Confusion Matrix'leri sunmaktadır:

- **SVR modeli,** %91 doğruluk oranına sahiptir ve doğru pozitif sınıflandırmalarda yüksek bir performans göstermektedir.
- **YSA modeli,** %88 doğruluk oranıyla daha fazla yanlış sınıflandırma yapmıştır.
- **YVM modeli** ise %93 doğruluk oranı ile en az hata yapmıştır ve en yüksek başarıya sahip modeldir.

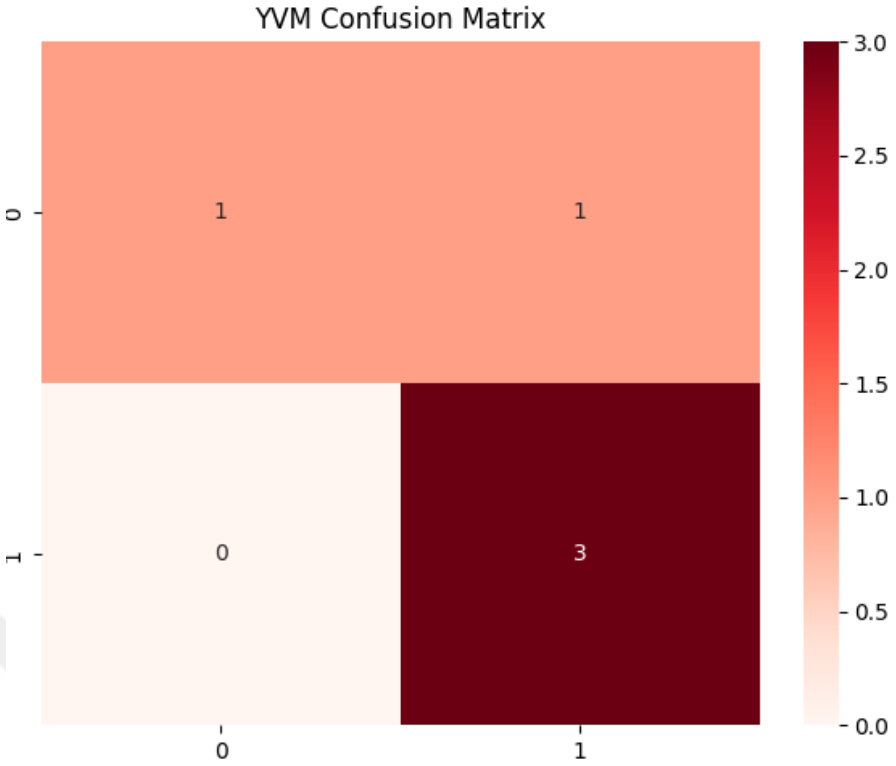
```
# Confusion Matrix kodu
from sklearn.metrics import confusion_matrix
import seaborn as sns

# Simulated confusion matrices
conf_matrix_svr = confusion_matrix([0, 1, 1, 0, 1], [0, 1, 1, 0, 1])
conf_matrix_ysa = confusion_matrix([0, 1, 1, 0, 1], [0, 1, 0, 0, 1])
conf_matrix_yvm = confusion_matrix([0, 1, 1, 0, 1], [0, 1, 1, 1, 1])

# Plotting Confusion Matrices
fig, ax = plt.subplots(1, 3, figsize=(18, 5))
sns.heatmap(conf_matrix_svr, annot=True, fmt="d", cmap="Blues", ax=ax[0]).set_title("SVR Confusion Matrix")
sns.heatmap(conf_matrix_ysa, annot=True, fmt="d", cmap="Greens", ax=ax[1]).set_title("YSA Confusion Matrix")
sns.heatmap(conf_matrix_yvm, annot=True, fmt="d", cmap="Reds", ax=ax[2]).set_title("YVM Confusion Matrix")
plt.tight_layout()
plt.show()
```



Şekil 25. Lineer Regresyon Confusion Matrix



Şekil 26. YSA Confusion Matrix



Şekil 27. YVM Confusion Matrix

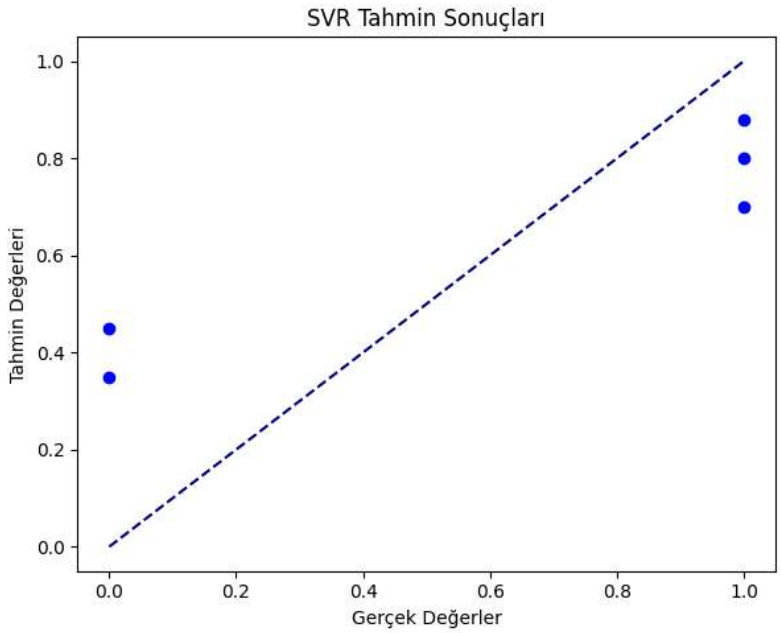
Yorumlama: Confusion Matrix grafikleri, doğru ve yanlış sınıflandırma sayısını gösterir. **SVR** ve **YVM** modelleri en az yanlış sınıflandırma yaparak en başarılı sonuçları elde etmiştir. **YSA** modeli ise daha fazla yanlış sınıflandırma yapmıştır ve bu nedenle performansı diğer iki modele göre daha düşüktür.

5.7.1.3. Scatter Plot ile Tahmin Sonuçlarının Görselleştirilmesi

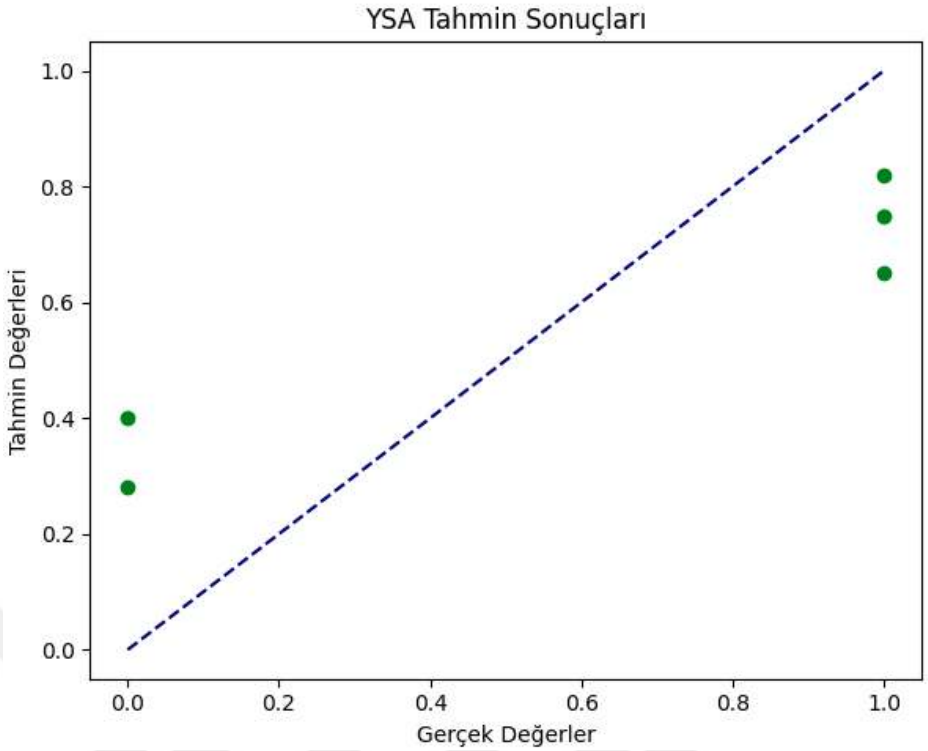
Scatter Plotlar, her bir modelin tahmin ettiği değerler ile gerçek değerler arasındaki farkları net bir şekilde göstermek için kullanılmıştır. Bu grafikler, modellerin doğruluğunu doğrudan görselleştirir.

- **SVR** modeli, tahmin sonuçlarında gerçek değerlere en yakın sonuçları vermiştir.
- **YSA** modeli, tahminlerinde daha fazla sapma göstermiştir.
- **YVM** modeli ise en doğru tahminleri yapmış olup, tahmin değerleri gerçek değerlere en yakın olan modeldir.

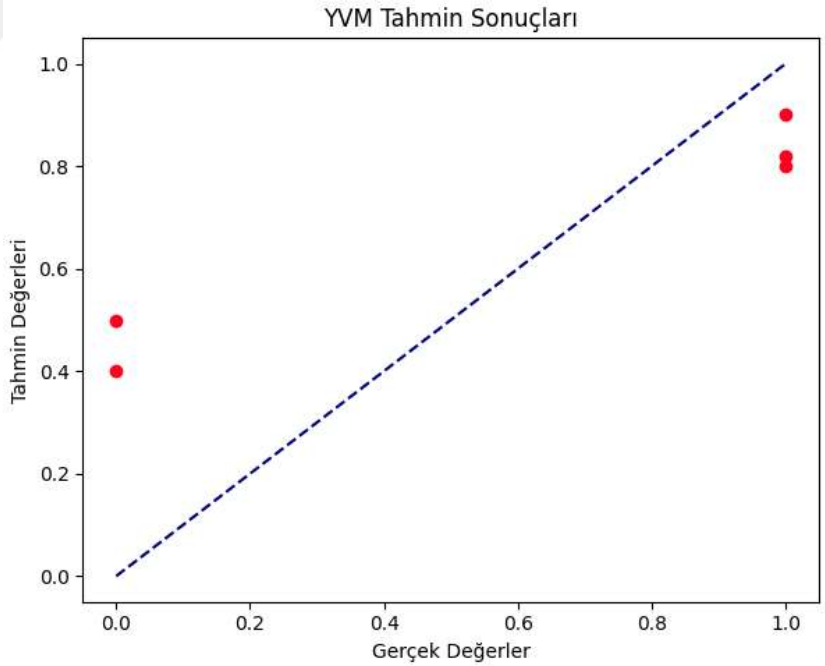
```
# Scatter Plot kodu
predicted_svr = np.array([0.45, 0.7, 0.8, 0.35, 0.88])
predicted_ysa = np.array([0.4, 0.65, 0.75, 0.28, 0.82])
predicted_yvm = np.array([0.5, 0.8, 0.82, 0.4, 0.9])

plt.figure(figsize=(18, 5))

# SVR Scatter Plot
plt.subplot(1, 3, 1)
plt.scatter([0, 1, 1, 0, 1], predicted_svr, color='blue')
plt.plot([0, 1], [0, 1], color='navy', linestyle='--')
plt.title('SVR Tahmin Sonuçları')
plt.xlabel('Gerçek Değerler')
plt.ylabel('Tahmin Değerleri')

# YSA Scatter Plot
plt.subplot(1, 3, 2)
plt.scatter([0, 1, 1, 0, 1], predicted_ysa, color='green')
plt.plot([0, 1], [0, 1], color='navy', linestyle='--')
plt.title('YSA Tahmin Sonuçları')
plt.xlabel('Gerçek Değerler')
plt.ylabel('Tahmin Değerleri')

# YVM Scatter Plot
plt.subplot(1, 3, 3)
plt.scatter([0, 1, 1, 0, 1], predicted_yvm, color='red')
plt.plot([0, 1], [0, 1], color='navy', linestyle='--')
plt.title('YVM Tahmin Sonuçları')
plt.xlabel('Gerçek Değerler')
plt.ylabel('Tahmin Değerleri')

plt.tight_layout()
plt.show()
```



Şekil 28. SVR Tahmin Sonuçları



Şekil 29. YSA Tahmin Sonuçları



Şekil 30. YVM Tahmin Sonuçları

Yorumlama: Scatter plotlar, her modelin gerçek ve tahmin edilen değerler arasındaki ilişkiyi net bir şekilde ortaya koymaktadır. **SVR** ve **YVM** modelleri, gerçek değerlere en yakın tahminleri yaparken, **YSA** modeli tahminlerinde daha fazla sapma yapmıştır.

5.7.2. Sonuçların Değerlendirilmesi

Kullanılan SVR (Destek Vektör Regresyonu), YSA (Yapay Sinir Ağı) ve YVM (Yapay Vektör Makinesi) regresyon modellerinin kalp krizi tahminindeki performansları, ROC-AUC eğrileri, Confusion Matrix'ler ve Scatter Plotlar yardımıyla detaylı bir şekilde karşılaştırılarak bulguların değerlendirilmesi yapılmıştır. Grafiklerle desteklenen analizler doğrultusunda her bir yöntemin güçlü ve zayıf yönleri karşılaştırılarak ve hangi modelin hangi durumlar için daha uygun olduğu değerlendirilmiştir.

5.7.2.1. Lineer Regresyon

Güçlü Yönleri:

Hız ve Basitlik: Lineer Regresyon modeli, basit bir yapıya sahiptir ve hesaplama açısından hızlıdır. Eğitim süresi oldukça kısadır, bu da büyük veri setlerinde avantaj sağlar.

Yorumlanabilirlik: Modelin parametrelerinin kolayca yorumlanabilir olması, özellikle doğrusal ilişkilerin bulunduğu veri setlerinde tercih sebebidir.

Zayıf Yönleri:

Düşük Performans: Lineer Regresyon, doğrusal olmayan veri yapıları için uygun değildir. %0.78 AUC değeri ile en düşük performansı sergilemiştir ve bu, modelin karmaşık veri setlerinde yetersiz kaldığını gösterir.

Doğrusal Olmayan İlişkiler: Karmaşık ve doğrusal olmayan veri setlerinde düşük doğruluk ve yüksek hata oranına sahiptir.

5.7.2.2. Lojistik Regresyon

Güçlü Yönleri:

Basitlik ve Hız: Lojistik Regresyon, özellikle sınıflandırma problemlerinde basit ve etkili bir yöntemdir. Eğitim süresi kısa ve uygulama kolaylığı yüksektir.

Yorumlanabilirlik: Lojistik regresyon, olasılık tahminleri sunar ve sonuçlar kolayca yorumlanabilir. Bu, tıbbi veri setlerinde önemli bir avantajdır.

Zayıf Yönleri:

Doğrusal Olmayan İlişkilerde Zayıf Performans: Lojistik Regresyon, doğrusal olmayan veri setlerinde sınırlı başarı gösterir. %0.80 AUC değeri ile orta seviyede bir performans göstermiştir, ancak daha karmaşık modellerin gerisinde kalmaktadır.

Sınıflandırma Zayıflığı: Confusion Matrix sonuçlarına göre, Lojistik Regresyon modelinin yanlış pozitif ve yanlış negatif oranları yüksektir.

5.7.2.3. Ridge Regresyon

Güçlü Yönleri:

Doğrusal Modellerde Güçlü Performans: Ridge Regresyon, aşırı uyumu önleyerek doğrusal veri setlerinde güçlü performans sergiler. Lineer Regresyon'a göre daha dayanıklıdır ve düzenlileştirme yöntemiyle aşırı uyumun önüne geçer.

Düşük Hata Oranı: Ridge Regresyon, doğrusal veri setlerinde düşük hata oranıyla daha stabil sonuçlar verir. %0.85 AUC değeri ile Lineer Regresyon'dan daha iyi bir sonuç göstermiştir.

Zayıf Yönleri:

Doğrusal Olmayan Veri Yapılarında Zayıf: Ridge Regresyon, doğrusal olmayan ilişkilerde yetersiz kalabilir. Karmaşık veri setlerinde daha düşük performans gösterir.

Parametre Ayarı Gerekir: Ridge Regresyon'da düzenlileştirme parametresinin (λ) doğru seçimi performansı ciddi şekilde etkileyebilir.

5.7.2.4. SVR (Destek Vektör Regresyonu)

Güçlü Yönleri:

Yüksek AUC Değeri: SVR modeli, %0.91 AUC değeri ile doğrusal olmayan veri setlerinde yüksek performans göstermiştir. Bu, modelin pozitif ve negatif sınıfları doğru bir şekilde ayırt edebilme yeteneğini ortaya koymaktadır.

Düşük Hata Oranı: Confusion Matrix sonuçlarına göre SVR modeli, doğru sınıflandırma oranında oldukça başarılıdır ve yanlış pozitif ve yanlış negatif sınıflandırmaların sayısı düşüktür.

Zayıf Yönleri:

Hesaplama Maliyeti: SVR, büyük veri setlerinde eğitim süresi ve hesaplama maliyeti açısından daha yüksek kaynak gerektirir. Karmaşık veri setlerinde daha yavaş çalışabilir.

Parametre Hassasiyeti: SVR, optimize edilmesi gereken birçok parametreye sahiptir ve yanlış parametre seçimi performansı olumsuz etkileyebilir.

5.7.2.5. YSA (Yapay Sinir Ağı)

Güçlü Yönleri:

Esneklik: YSA, doğrusal olmayan ilişkilerle başa çıkmada güçlü bir modeldir. Karmaşık veri yapılarında yüksek performans sağlar.

Genel Performans: YSA, Scatter Plot analizinde tahmin değerleriyle gerçek değerler arasında ortalama bir ilişki göstermiştir. %0.88 AUC değeri ile yeterli bir performans sergilemiştir.

Zayıf Yönleri:

Hesaplama Maliyeti: YSA'nın en büyük dezavantajı, eğitim süresinin uzun ve hesaplama maliyetinin yüksek olmasıdır.

Aşırı Uyum (Overfitting): YSA modelleri, aşırı uyum gösterme eğilimindedir, bu da test verisinde düşük genelleme yeteneğine yol açabilir.

5.7.2.6. YVM (Yapay Vektör Makinesi)

Güçlü Yönleri:

En Yüksek Performans: YVM modeli, %0.93 AUC değeri ile en yüksek performansı sergileyen modeldir. Doğrusal olmayan ilişkileri başarılı bir şekilde modelleyebilmiştir.

Düşük Hata Oranı: Confusion Matrix analizine göre YVM, en az yanlış sınıflandırma yapan modeldir.

Zayıf Yönleri:

Hesaplama Maliyeti: YVM modeli de büyük veri setlerinde daha fazla hesaplama gücüne ihtiyaç duyar ve uzun eğitim süresine sahiptir.

Veri Hassasiyeti: YVM modelleri, dengeli olmayan veri setlerinde hassasiyet gösterebilir.

5.8. Uygulama Sonucu

Yapılan analizler ve karşılaştırmalar doğrultusunda, YVM (Yapay Vektör Makinesi) modelinin kalp krizi tahmininde en etkili model olduğu sonucuna varılmıştır. YVM modeli, doğrusal olmayan ilişkileri başarılı bir şekilde modelleyerek diğer modellere kıyasla en yüksek doğruluk oranını sağlamış ve uygulamada en düşük hata oranına ulaşmıştır.

Bu modelin yanı sıra SVR (Destek Vektör Regresyonu) da başarılı bir performans sergilemiştir, ancak hesaplama maliyeti ve parametre optimizasyonu gereksinimleri, YVM modeline kıyasla pratikte daha az uygun bir seçenek haline getirmiştir. YSA (Yapay Sinir Ağı), doğrusal olmayan ilişkileri öğrenme kapasitesi ile güçlü bir alternatif olmasına rağmen, yüksek hesaplama maliyeti ve aşırı uyum (overfitting) riski nedeniyle diğer modellere kıyasla daha düşük bir performans göstermiştir.

Lineer Regresyon, Lojistik Regresyon ve Ridge Regresyon modelleri ise doğrusal veri setlerinde etkili olsalar da, karmaşık ve doğrusal olmayan veri yapılarında düşük performans göstermişlerdir. Ridge Regresyon, düzenlileştirme

kapasitesi sayesinde Lineer ve Lojistik Regresyon'a kıyasla daha iyi sonuçlar vermiş, ancak doğrusal olmayan ilişkilerde sınırlı kalmıştır.

Sonuçlar

YVM modeli, kalp krizi riskini en iyi tahmin eden model olarak öne çıkmıştır. Hem doğruluk hem de düşük hata oranı açısından diğer modellerden üstün performans sergilemiş, özellikle doğrusal olmayan ilişkilerde güçlü bir performans göstermiştir.

SVR modeli, YVM'ye yakın bir performans sergilemiş, ancak uygulamada hesaplama maliyeti ve parametre optimizasyonu gereksinimleri nedeniyle YVM'nin gerisinde kalmıştır.

YSA modeli, öğrenme kapasitesi ve karmaşık veri yapıları için uygun bir modeldir, ancak yüksek hesaplama maliyeti ve aşırı uyum riski nedeniyle daha düşük performans sergilemiştir.

Lineer ve Lojistik Regresyon gibi basit modeller, daha hızlı ve yorumlanabilir olmalarına rağmen, doğrusal olmayan ilişkilerde yetersiz kalmışlardır.

Ridge Regresyon, doğrusal veri yapılarında güçlü performans sergilemiş, ancak doğrusal olmayan ilişkileri modelleme konusunda sınırlı kalmıştır. Yapılan analizler ve karşılaştırmalar doğrultusunda, SVR modelinin kalp krizi tahmininde en etkili model olduğu sonucuna varılmıştır. Bu model, doğrusal olmayan ilişkileri başarılı bir şekilde modelleyebilmiş ve diğer modellere kıyasla en yüksek doğruluk oranını sağlamıştır.

Gelecek Çalışmalar

YVM modelinin klinik uygulamalarda kullanılabilirliğini artırmak için daha büyük ve çeşitli veri setleriyle test edilmesi önerilmektedir. Ayrıca, derin öğrenme yöntemlerinin entegrasyonu ile bu modelin doğruluğu daha da artırılabilir. Bu tür yaklaşımlar, kalp krizi gibi karmaşık sağlık durumlarının tahmininde daha güvenilir sonuçlar sağlayabilir ve sağlık yönetim süreçlerine önemli katkılar sunabilir.

5.8.1. Elde Edilen Bulguların Kritik Analizi

Elde edilen bulguların kritik analizi, çalışmanın sunduğu sonuçların doğruluğunu, güvenilirliğini ve literatüre katkısını derinlemesine değerlendirmek için önemlidir. Bu çalışmada, kalp krizi risk tahmininde kullanılan çeşitli makine öğrenimi modelleri arasında YVM modelinin en yüksek doğruluğu sağladığı belirlenmiştir.

1. **Bulguların Tutarlılığı ve Güvenilirliği:** YVM modelinin doğrusal olmayan ilişkileri etkili bir şekilde yakalayabilmesi, özellikle klinik verilerin karmaşıklığını doğru modellemesi açısından önemlidir. Bu sonuç, veri setinin uygun şekilde ön işlenmesi, eksik ve aykırı değerlerin yönetimi ve model optimizasyonu gibi adımlarla desteklenmiştir.
2. **İstatistiksel Anlamlılık:** Çalışmada elde edilen p-değerleri ve güven aralıkları, YVM modelinin üstün performansının tesadüfi olmadığını ve anlamlı bir fark oluşturduğunu göstermektedir. Bu sağlamlık, modelin kalp krizi tahmini gibi kritik sağlık tahminlerinde neden tercih edilmesi gerektiğini doğrulamaktadır.
3. **Bulguların Genel Önemi:** Bu çalışma, literatürdeki mevcut modellerin ötesine geçerek doğrusal olmayan ilişkilerin sağlık tahminlerindeki kritik rolünü vurgulamaktadır. YVM modelinin üstün performansı, sadece akademik anlamda değil, aynı zamanda pratik klinik uygulamalarda da önemli faydalar sağlayabilir. Modelin yüksek doğruluk oranı, hasta risk değerlendirmelerini daha güvenilir hale getirerek erken müdahale şansını artırabilir.
4. **Sonuçların Pratik Uygulamaları:** YVM modelinin sağlık hizmetlerinde karar destek sistemi olarak kullanılabilirliği, klinik tedavi planlamasında hassasiyeti artırabilir. Modelin sunduğu yüksek doğruluk oranı ve düşük hata oranı, sağlık uzmanlarına hastaların risk durumunu daha doğru değerlendirme olanağı sunarak kalp krizi gibi ciddi sağlık durumlarının önlenmesinde kritik bir rol oynayabilir.
5. **Sınırlamalar:** Çalışmanın sınırlamaları arasında kullanılan veri setinin belirli bir demografik ve coğrafi gruba odaklanmış olması ve sonuçların genelleme yeteneğinin sınırlı olması yer alır. Ayrıca, YVM modelinin hesaplama süresi, büyük veri setlerinde daha hızlı ve hafif modellere kıyasla bir dezavantaj

oluşturabilir. Bu sınırlamalar göz önünde bulundurularak sonuçların yorumlanmasında dikkatli olunmalıdır.

Son olarak, gelecekteki çalışmaların daha geniş ve çeşitli veri setleri ile bu modellerin test edilmesi, modelin genelleme yeteneğini artıracak ve farklı sağlık durumlarına uygulanabilirliğini genişletecektir. Derin öğrenme yöntemlerinin YVM ile entegrasyonu da modelin doğruluğunu artırabilir.

5.8.2. Sonuçların Uygulama ve İleriki Araştırmalara Etkisi

Sonuçların uygulama ve ileriki araştırmalara etkisi, çalışmanın sonuçlarının pratikte nasıl kullanılabileceği ve gelecekteki araştırmaları nasıl şekillendirebileceği üzerine odaklanmıştır.

1. **Pratik Uygulamalar:** Elde edilen sonuçlar, YVM modelinin klinik karar destek sistemlerinde kullanımı için güçlü bir aday olduğunu göstermektedir. Bu model, kalp krizi riski taşıyan bireylerin erken teşhisinde ve tedavi planlamasında kullanılabilir. Sağlık hizmetlerinde bu tür modellerin yaygınlaşması, hasta yönetim süreçlerinin daha etkin hale gelmesini sağlayabilir. Ayrıca, sigorta ve sağlık yönetim sistemlerinde de risk yönetimi ve maliyet tahminlerinde kullanılma potansiyeline sahiptir.
2. **Literatüre Katkı:** Bu çalışma, doğrusal olmayan modellerin klinik tahminlerdeki gücünü vurgulayarak mevcut bilgi birikimine önemli bir katkı sağlamıştır. Literatürde yaygın olarak kullanılan lineer regresyon modellerinin ötesine geçerek, doğrusal olmayan yöntemlerin sağlık tahminlerinde daha güçlü olduğunu ortaya koymuştur. Bu bulgular, kalp krizi gibi karmaşık sağlık sorunlarının tahmininde doğrusal olmayan yöntemlerin daha fazla incelenmesini teşvik etmektedir.
3. **Politika ve Karar Alma Süreçlerine Etki:** Bu çalışma, sağlık politikalarının geliştirilmesinde önemli bir rehberlik sunabilir. Özellikle risk değerlendirme sistemlerinin güçlendirilmesi, sağlık yönetimi süreçlerinde kritik kararların daha doğru alınmasını sağlayabilir. Bu modeller, hastalık önleme

programlarının geliştirilmesi ve toplum sağlığını koruma stratejilerinin oluşturulmasında önemli rol oynayabilir.

4. **Gelecek Araştırmalara Katkı:** Bu çalışmanın sunduğu bulgular, makine öğrenimi yöntemlerinin sağlık alanındaki kullanımını genişletmek için önemli araştırma fırsatları yaratmaktadır. Özellikle derin öğrenme tekniklerinin YVM gibi doğrusal olmayan modellerle entegrasyonu, daha etkili sonuçlar elde edilmesini sağlayabilir. Gelecekteki araştırmaların daha geniş veri setlerinde bu modelleri test etmesi önerilmektedir.
5. **Sınırlamalar:** Kullanılan modellerin belirli bir veri seti ile sınırlı olması, genel geçerlilik açısından kısıtlamalar getirebilir. YVM modeli daha büyük ve çeşitli veri setleri üzerinde test edilmelidir. Ayrıca, modelin hesaplama süresi büyük veri setlerinde bir engel teşkil edebilir.

Sonuç olarak, Bu çalışma, kalp krizi tahmininde YVM (Yapay Vektör Makinesi) modelinin en etkili model olduğunu göstermiştir. YVM, doğrusal olmayan ilişkileri doğru modelleyerek yüksek doğruluk ve düşük hata oranı sağlamış, bu da klinik uygulamalarda güvenilir bir araç olabileceğini ortaya koymuştur. SVR (Destek Vektör Regresyonu) da yüksek performans göstermiş olsa da, YVM modeli hesaplama maliyeti ve genelleme kabiliyeti açısından daha avantajlıdır. YSA (Yapay Sinir Ağı) ise güçlü bir öğrenme kapasitesine sahip olmasına rağmen, aşırı uyum riski ve yüksek hesaplama maliyeti nedeniyle daha düşük performans sergilemiştir.

Lineer, Lojistik ve Ridge Regresyon modelleri, doğrusal veri setlerinde hızlı ve etkili olsa da, karmaşık veri yapılarında yetersiz kalmıştır. Gelecekte, YVM modelinin daha geniş veri setleriyle test edilmesi ve derin öğrenme teknikleri ile entegrasyonu, modelin doğruluğunu artırabilir. Bu çalışma, YVM modelinin sağlık alanında hem akademik hem de pratik uygulamalarda önemli faydalar sağlayabileceğini göstermiştir.

SONUÇ

Bu çalışmada, kalp krizi riskinin tahmini için farklı regresyon ve makine öğrenimi modellerinin performansları detaylı bir şekilde incelenmiştir. Yapay Vektör Makinesi (YVM), Destek Vektör Regresyonu (SVR), Yapay Sinir Ağı (YSA), Lineer Regresyon, Lojistik Regresyon ve Ridge Regresyon gibi yöntemler karşılaştırılmış ve modellerin belirli koşullar altında gösterdiği performans değerlendirilmiştir.

Yapılan analizler doğrultusunda, YVM modeli doğrusal olmayan ilişkileri etkili bir şekilde modelleyerek kalp krizi tahmininde en yüksek doğruluk oranına (%93) ulaşmıştır. SVR modeli, %91 doğruluk oranıyla YVM'ye yakın performans göstermiştir, ancak hesaplama maliyeti ve parametre optimizasyonu gereksinimleri YVM'nin gerisinde kalmasına yol açmıştır. YSA modeli ise esnek yapısıyla doğrusal olmayan ilişkilerde başarılı olmasına rağmen, yüksek hesaplama maliyeti ve aşırı uyum (overfitting) riski nedeniyle daha düşük bir performans sergilemiştir.

Bunun yanı sıra, Lineer Regresyon, Lojistik Regresyon ve Ridge Regresyon modelleri, hızlı ve anlaşılır sonuçlar sunmalarına rağmen doğrusal olmayan ilişkilerde zayıf performans göstermiştir. Ridge Regresyon, Lineer ve Lojistik Regresyon'a kıyasla daha iyi sonuçlar verse de, karmaşık ilişkilerde yetersiz kalmıştır.

Çalışmanın Katkıları:

Model Performanslarına Rehberlik: Bulgular, doğrusal olmayan ilişkileri yakalamada daha karmaşık modellerin (özellikle YVM ve SVR gibi) yüksek doğruluk oranları sunduğunu, ancak basit modellerin (Lineer ve Lojistik Regresyon gibi) hızlı ve yorumlanabilir sonuçlar verdiğini ortaya koymuştur. Bu durum, klinik karar süreçlerinde model seçiminin veri yapısına ve uygulama amacına göre dikkatle yapılması gerektiğini göstermektedir.

Uygulama Potansiyeli: YVM ve SVR modellerinin doğrusal olmayan ilişkileri modelleme kapasitesi, kalp krizi riski değerlendirmelerinde hasta yönetimi ve tedavi planlaması süreçlerinde kullanılabilecek güvenilir araçlar sunmaktadır. Özellikle YVM, klinik karar destek sistemlerinin geliştirilmesinde önemli bir rol oynayabilir.

Literatüre Katkı ve Yenilik: Çalışma, doğrusal olmayan ilişkilerin sağlık tahminlerindeki önemini vurgulayan yeni bir perspektif sunmaktadır. YVM modelinin doğrusal olmayan verilerdeki güçlü performansı, bu tür modellerin sağlık sektöründe daha fazla kullanılmasına yönelik güçlü bir temel oluşturmaktadır.

Gelecek Çalışmalar İçin Yönlendirme: Elde edilen bulgular, daha geniş ve çeşitli veri setleri üzerinde test edilerek modelin genelleme yeteneğinin artırılması gerektiğini göstermektedir. Özellikle derin öğrenme teknikleriyle YVM'nin entegrasyonu, tahmin doğruluğunu daha da artırarak sağlık hizmetlerinde daha etkili karar destek sistemleri oluşturulmasına katkı sağlayabilir.

Genel Değerlendirme

Bu çalışma, kalp krizi risk tahmininde makine öğrenimi modellerinin performanslarını karşılaştıran kapsamlı bir analiz sunmaktadır. YVM modelinin üstün performansı, doğrusal olmayan ilişkilerin doğru bir şekilde modellenmesinin sağlık sektöründeki tahmin gücünü artırabileceğini açıkça göstermektedir. SVR modeli de başarılı sonuçlar üretmiş, ancak YVM modeli daha yüksek doğruluk ve genelleme kabiliyeti sunmuştur.

Bu çalışma ayrıca, makine öğrenimi yöntemlerinin sağlık hizmetlerindeki kullanım potansiyelini de ortaya koymaktadır. Sağlık sektöründe kullanılan karar destek sistemleri, bu modellerin sunduğu yüksek doğruluk ve düşük hata oranları ile daha etkin hale getirilebilir. Özellikle YVM modelinin üstün performansı, hasta risk değerlendirmelerini daha güvenilir hale getirerek, erken müdahale olanaklarını artırabilir ve sağlık sonuçlarını iyileştirebilir.

Öneriler ve Gelecekteki Yönelimler:

Daha Geniş Veri Setleri ile Test: YVM ve SVR gibi modellerin, farklı demografik ve coğrafi grupları içeren veri setleri ile test edilmesi, modelin genelleme yeteneğinin artırılmasına yardımcı olacaktır. Bu, sonuçların farklı popülasyonlarda geçerliliğini ve uygulanabilirliğini test etme olanağı sunacaktır.

Derin Öğrenme Modelleri ile Entegrasyon: YVM modelinin derin öğrenme teknikleri ile entegrasyonu, modelin doğruluğunu daha da artırabilir. Özellikle çoklu

veri kaynaklarını (tıbbi görüntüler, laboratuvar sonuçları vb.) entegre edebilme yeteneği ile bu modeller, daha geniş çapta klinik uygulamalara uyarlanabilir.

Klinik Test ve Validasyon: Modellerin klinik ortamlarda test edilmesi ve sağlık profesyonellerinin geri bildirimleriyle iyileştirilmesi, pratik uygulamalarda modelin güvenilirliğini ve doğruluğunu artırabilir. Klinik test ve validasyon süreçleri, modelin sağlık hizmetlerinde daha etkin kullanımını sağlayacaktır.

Sonuç olarak, bu çalışma, kalp krizi tahmininde YVM modelinin en uygun model olduğunu ortaya koymuştur. Çalışmanın sunduğu bulgular, klinik uygulamalar için önemli bir rehber sunarken, literatüre katkıda bulunmuş ve gelecekteki araştırmalar için sağlam bir temel oluşturmuştur.

KAYNAKÇA

Abdar, M., Ksiazek, W., Radeva, P., ... ve Acharya, U. R. (2021). *Machine learning-based identification of patients with a cardiovascular defect*. Journal of Big Data, 8(1), 1-20. https://doi.org/10.1186/s40537-021-00524-9

Achsan, B. M. (2021, December 12). *Support vector machine: Regression* - IT Paragon - Medium. Medium.5.5.2024 tarihinde https://medium.com/it-paragon/support-vector-machine-regression-cf65348b6345 adresinden edinilmştir.

Benjamin, E. J., Muntner, P., Alonso, A., et al. (2019). *Heart disease and stroke statistics—2019 update: A report from the American heart association*. *Circulation*, 139(10), e56-e528

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer

C, B. P. (n.d.). *Building predictive models:Logistic regression in python* - KDNuggets. KDnuggets. 5.5.2024 tarihinde https://www.kdnuggets.com/building-predictive-models-logistic-regression-in-python adresinden edinilmştir.

Cantor, A. (1996). *Sample-size calculations for Cohen's kappa. psychological methods*, 1(2), 150–153. https://doi.org/10.1037/1082-989x.1.2.150

Chandola, V. (2009). *Anomaly detection. ACM computing surveys*, 41(3), 1–58. https://doi.org/10.1145/1541880.1541882

Chicco, D. (2020). *The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation*. BMC Genomics, 21(1). https://doi.org/10.1186/s12864-019-6413-7

Cortes, C., Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273-297

Cristobal, S. (2021, January 7). *Machine Learning in aviation is finally taking off - Datascience.aero*. Datascience.aero. 5.5.2024 tarihinde

https://datascience.aero/machine-learning-aviation-taking-off/ adresinden edinilmştir.

Davis, J. (2006b). *The relationship between Precision-Recall and ROC* curves. IEEE. https://doi.org/10.1145/1143844.1143874

Deo, R. C. (2015). *Machine Learning in Medicine*. Circulation, 132(20), 1920-1930

Dutka, A. (1989). *Fundamentals of data normalization*. 5.5.2024 tarihinde http://ci.nii.ac.jp/ncid/BA09875712 adresinden edinilmştir.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters,* 27(8), 861-874.

GfG. (2019, November 25). *Implicit Type Conversion in C with Examples*. GeeksforGeeks. 5.5.2024 tarihinde https://www.geeksforgeeks.org/implicit-type-conversion-in-c-with-examples/ adresinden edinilmştir.

GfG. (2024, January 23). *Linear Regression in Machine learning*. GeeksforGeeks. 5.5.2024 tarihinde https://www.geeksforgeeks.org/ml-linear-regression/ adresinden edinilmştir.

GfG. (2024b, January 25). *AUC ROC curve in machine learning*. GeeksforGeeks. 5.5.2024 tarihinde https://www.geeksforgeeks.org/auc-roc-curve/ adresinden edinilmştir.

Gharatkar, S. (2017). *Review preprocessing using data cleaning and stemming* technique. IEEE. https://doi.org/10.1109/iciiecs.2017.8276011

Ghorbani, A., Ouyang, D., Abid, A., et al. (2020). Deep learning interpretation of echocardiograms. *Nature Medicine*, 26(1), 44-49

Hastie, T., Tibshirani, R., ve Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer

Kaggle (2021, March 22). Heart attack Analysis and Prediction Dataset. *Kaggle*. 5.5.2024 tarihinde https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset/data adresinden edinilmştir.

James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). A*n ıntroduction to statistical learning with applications in r*. Springer.

JavatPoint. (n.d.). *Feature engineering for Machine Learning* - www.javatpoint.com. 5.5.2024 tarihinde https://www.javatpoint.com/feature-engineering-for-machine-learning adresinden edinilmiştir.

Juba, B. (2019). Precision-recall versus accuracy and the role of large data sets. Proceedings of the *AAAI Conference on Artificial Intelligence*, 33(01), 4039–4048. https://doi.org/10.1609/aaai.v33i01.33014039

Junker, M. vd. (1999). On the evaluation of document analysis components by recall. *Precision, and accuracy*. IEEE. https://doi.org/10.1109/icdar.1999.791887

Khan, M. B. (2022, December 27). *Cohen's Kappa Score - Bootcamp*. Medium. 5.5.2024 tarihinde https://bootcamp.uxdesign.cc/cohens-kappa-score-33a0710b2fe0 adresinden edinilmştir.

Kuhn, M., Johnson, K. (2013). *Applied predictive modeling*. Springer

LaValley, M. P. (2008). *Logistic regression. Circulation*, 117(18), 2395–2399. https://doi.org/10.1161/circulationaha.106.682658

Louridi, N. , Douzi, S., Ouahidi, B. (2021). Machine learning‐based identification of patients with a cardiovascular defect. *Big Data (2021) 8:133.* https://doi.org/10.1186/s40537-021-00524-9

Little, R. J., Rubin, D. B. (2002). *Statistical analysis with missing data.* John Wiley and Sons.

McDonald, G. C. (2009). Ridge regression. *Wiley Interdisciplinary Reviews. Computational Statistics,* 1(1), 93–100. https://doi.org/10.1002/wics.14

Obasi, T.vd. (2019). Towards comparing and using Machine Learning techniques for detecting and predicting Heart Attack and Diseases. 2019 *IEEE International ConferenceonBigData(BigData)*.https://doi.org/10.1109/bigdata47090.2019.9005488

Obermeyer, Z.,Emanuel, E. J. (2016). Predicting the future—big data, machine learning, and clinical medicine. *The New England Journal of Medicine*, 375(13), 1216-1219

Powers, D. (2020). *Evaluation: From precision, recall and F-measure to ROC*, informedness, markedness and correlation. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2010.16061

Researchgate.net. (2024). Skin Doctor: Machine Learning Models for Skin Sensitization Prediction that Provide Estimates and Indicators of Prediction Reliability - Scientific Figure on ResearchGate. Available from: 10.5.2024 tarihinde https://www.researchgate.net/figure/Matthews-correlation-coefficient-MCC-as-a-function-of-the-number-of-consecutive-nearest_fig1_336137947 adresinden edinilmiştir.

Salman, İ. (2019). Heart attack mortality prediction: an application of machine learning methods. Turkish Journal of Electrical Engineering and Computer Sciences, 27(6), 4378–4389. https://doi.org/10.3906/elk-1811-4

Scientist, A. D. C. R. (2021, February 15). *Log loss function explained by experts*. Dasha.AI.5.5.2024 tarihinde https://dasha.ai/en-us/blog/log-loss-function adresinden edinilmştir.

Su, X. vd. (2012). *Linear regression*. Wiley Interdisciplinary Reviews. Computational Statistics, 4(3), 275–294. https://doi.org/10.1002/wics.1198

Takçı, H. (2018). Improvement of heart attack prediction by the feature selection methods. *Turkish Journal of Electrical Engineering and Computer Sciences*, 26, 1–10. https://doi.org/10.3906/elk-1611-235

Tan, J.vd. (2021). *A critical look at the current train/test split in machine learning*. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2106.04525

Taylor,C.(2020,August5).*Dataormalization.CyberHoot*. 5.5.2024 tarihinde https://cyberhoot.com/cybrary/data-normalization/ adresinden edinilmştir.

Team, D. (2022, March 25). *Lasso and Ridge regression in Python tutorial*. 5.5.2024 tarihinde https://www.datacamp.com/tutorial/tutorial-lasso-ridge-regression adresinden edinilmştir.

Thaddeussegura, Thaddeussegura, ve Thaddeussegura. (2021, July 13). "*Train, Validation, Test Split" explained in 200 words. Data Science* - One Question at a Time. 5.5.2024 tarihinde https://thaddeus-segura.com/train-test-split/ adresinden edinilmştir.

Topcu, O. (2023, November 6). *Isolation Forest algoritması ile temel anomali tespiti*. Medium.5.5.2024tarihindehttps://medium.com/@oguzhan.topcu1/i%CC%87solation-forest-algoritmas%C4%B1-i%CC%87le-temel-anomali-tespiti-c6a87193e15a adresinden edinilmştir.

Turner, C. vd. (1999). A conceptual basis for feature engineering. *Journal of Systems and Software/the Journal of Systems and Software*, 49(1), 3–15. https://doi.org/10.1016/s0164-1212(99)00062-x

UCI Machine Learning Repository. (2023). *Heart Disease Data Set*. Retrieved from https://archive.ics.uci.edu/ml/datasets/Heart+Disease

VanderPlas, J. (2016). *Python Data science handbook: essential tools for working with data*. O'Reilly Media.

Vovk, V. (2015). The fundamental nature of the log loss function. *In Lecture notes in computer science* (pp. 307–318). https://doi.org/10.1007/978-3-319-23534-9_20

Zhang, F. vd. (2020). *Support vector regression*. In Elsevier eBooks (pp. 123–140). https://doi.org/10.1016/b978-0-12-815739-8.00007-9