A picture containing text, font, logo, graphics

Description automatically generated A picture containing horse, stallion

Description automatically generated

**Kalp yetmezlİĞİ TEŞHİSİ**

**Kategorik Veri Çözümlemesi**

Berke Sevim - 2018285047

Sevimonline@hotmail.com

1. VERİ SETİ TANITIMI

Bu Projede Python programlama dili ve NumPy, pandas, Matplotlib, Seaborn, SciPy, Scikit-learn gibi çeşitli Python kütüphaneleri kullanılmıştır.

Kardiyovasküler hastalıklar (KVH'ler), her yıl tahmini 17,9 milyon can alarak dünya çapında 1 numaralı ölüm nedenidir ve bu, dünya çapındaki tüm ölümlerin %31'ini oluşturmaktadır. 5KVH ölümlerinden dördü kalp krizi ve felçten kaynaklanır ve bu ölümlerin üçte biri 70 yaşın altındaki kişilerde erken gerçekleşir.

Kalp yetmezliği, KVH'lerin neden olduğu yaygın bir olaydır ve bu veri seti, olası bir kalp hastalığını tahmin etmek için kullanılabilecek 11 özellik içerir. Veri setinde kayıp gözlem bulunmamaktadır.

Kardiyovasküler hastalığı olan veya yüksek kardiyovasküler risk altında olan kişiler (hipertansiyon, diyabet, hiperlipidemi veya önceden belirlenmiş hastalık gibi bir veya daha fazla risk faktörünün varlığından dolayı) erken tespit ve yönetime ihtiyaç duyar ve bu durumda bir makine öğrenimi modeli çok yardımcı olabilir.

1. KEŞİFSEL VERİ ANALİZİ

Keşifsel veri analizi, veri setindeki desenleri, ilişkileri ve önemli noktaları keşfetmek için kullanılan bir veri analizi yöntemidir. Ayrıca veri setinin yapısını anlamak ve içindeki bilgileri ortaya çıkarmak için de kullanılır.

Keşifsel veri analizi, genellikle görselleştirme teknikleri, özet istatistikleri ve veri madenciliği yöntemlerini kullanır.

1. Veriye Genel Bakış: İlk adım, veri setini anlamak ve keşfetmeye başlamaktır. Bu adımda, veri setinin boyutlarına, değişkenlerine ve genel yapısına göz atılır.

2. Veri Görselleştirme: Görselleştirme teknikleri kullanılarak veri seti grafikler, tablolar veya grafikler aracılığıyla görselleştirilir. Bu, veri setindeki desenleri, dağılımları, eğilimleri ve ilişkileri daha iyi anlamak için önemli bir adımdır.

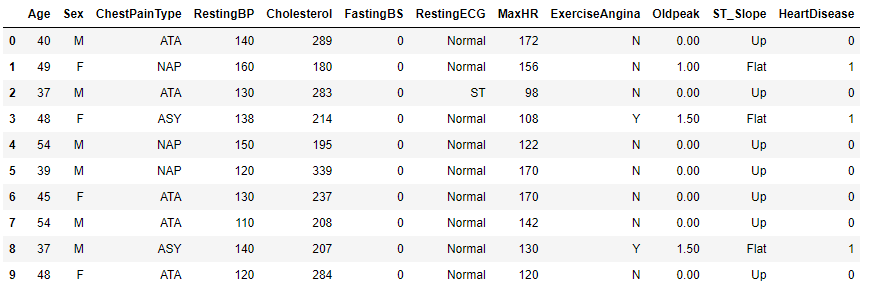
3. Özet (Tanımlayıcı) İstatistikleri: Veri setinin özetini sağlayan istatistiksel hesaplamalar yapılır. Bu hesaplamalar, merkezi eğilim ölçütleri (ortalama, medyan), değişkenlik ölçütleri (standart sapma, varyans), dağılım ölçütleri (minimum, maksimum) gibi istatistiksel özetler içerir.

4. Veri Keşfi: Veri setindeki desenleri ve ilişkileri keşfetmek için veri madenciliği yöntemleri uygulanır. Örnek olarak, kümeleme analizi, birlikte görülme analizi veya ilişki kuralı analizi gibi yöntemler kullanılabilir.

5. Anlam ve Yorumlama: Keşfedilen desenler ve ilişkiler anlamlandırılır ve yorumlanır. Bu adımda, elde edilen sonuçlar ve bulgular, ilgili konu veya probleme uygulanarak anlamlı hale getirilir.

Keşifsel veri analizi, veri setinin daha derinlemesine anlaşılması, veri hikayesinin oluşturulması ve ileri analizlerin yönlendirilmesi için önemli bir adımdır. Veri bilimciler, analistler ve araştırmacılar, keşifsel veri analizi yöntemlerini kullanarak yeni bilgilerin ortaya çıkmasına yardımcı olurlar.

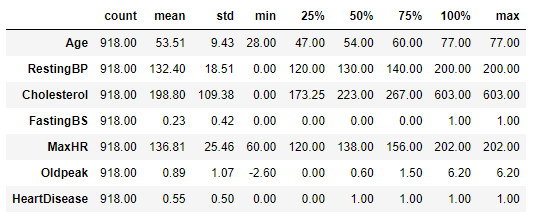
* 1. Veriye Genel Bakış



Şekil 1.1 : Verinin ilk 10 gözlemi

Şekil 1.1 de verinin genel görünümü gözükmektedir. Veride 918 gözlem, 12 değişken vardır. Veride kayıp gözlem yoktur. Değişkenlerin tanımı aşağıdaki şekilde açıklanmaktadır.

* Age: hastanın yaşı [yıl]
* Sex: hastanın cinsiyeti [M: Erkek, F: Kadın]
* ChestPainType: göğüs ağrısı türü [TA: Tipik Angina, ATA: Atipik Angina, NAP: Anjinal Olmayan Ağrı, ASY: Asemptomatik]
* RestingBP: dinlenme kan basıncı [mm Hg]
* Cholesterol: serum kolesterolü [mm/dl]
* FastingBS: açlık kan şekeri [1: Eğer FastingBS > 120 mg/dl ise, 0: Aksi takdirde]
* RestingECG: istirahat elektrokardiyogram sonuçları [Normal: Normal, ST: ST-T dalgası anormalliği (T dalgası inversiyonları ve/veya ST yükselmesi veya > 0,05 mV'lik depresyon), LVH: Estes kriterlerine göre olası veya kesin sol ventrikül hipertrofisi gösteriyor]
* MaxHR: ulaşılan maksimum kalp atış hızı [60 ile 202 arasındaki sayısal değer]
* Exercise Angina: egzersize bağlı anjina [Y: Evet, N: Hayır]
* Oldpeak: oldpeak = ST [Düşüşte ölçülen sayısal değer]
* ST\_Slope: zirve egzersizinin eğimi ST segmenti [Yukarı: yukarı eğimli, Düz: düz, Aşağı: aşağı eğimli]
* HeartDisease: çıktı sınıfı [1: kalp hastalığı, 0: Normal]
  1. Tanımlayıcı İstatistikler



Şekil 1.2 : Verinin tanımlayıcı istatistikleri

Veriyi tanıtmak, değişkenler hakkında ön bilgi edinmek ve kayıp gözlemlerin olup olmadığını tespit etmek amacıyla tanımlayıcı istatistikler elde edilmiştir.

Elde edilen istatistiklere göre:

* Veride ortalama yaş 53-54 kırılımındadır. Minimum yaş 28, maximum yaş 77’dir.
* Dinlenme kan basıncı (RestingBP) ortalama 132.40’tır. Minimum dinlenme kan basınca 0, maximum 200’dür
* Kolesterol (Cholesterol) ortalama 198.80 mm/dl ‘dir. . Minimum Kolesterol 0, maximum kolesterol 603 mm/dl ‘dir.
* Açlık kan şekeri (FastingBS) ortalama 0 ’dir. Minimum açlık kan şekeri 0,

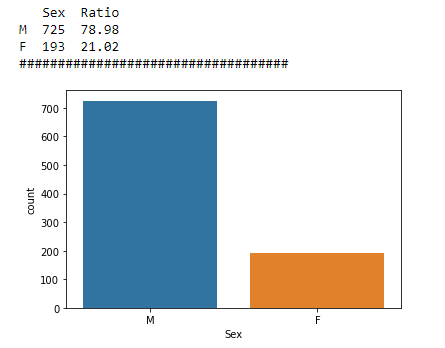
maximum 1’dir. Kategorik bir değişken olduğu için 0 ya da 1 değerlerini almaktadır. Bu değişken numerik veri tipiyle ifade edilmiştir.

(Açlık kan şekerinin 1 olması FastingBS > 120 mg/dl olması anlamına gelmektedir)

* Hastalarda ulaşılan maximum kalp hızı (MaxHR) ortalama 136.81’dir. Minimum kalp hızı 60, Maximum 202’dir.
* Ortalama Oldpeak 0.89’dur. Minimum Oldpeak -2.60, maximum 6.20’dir.
* HeartDisease bağımlı değişkendir. Çıktı sınıfı [1: kalp hastalığı, 0: Normal] olarak ele alınmaktadır.
  1. Veri Görselleştirme – Veri Keşfi

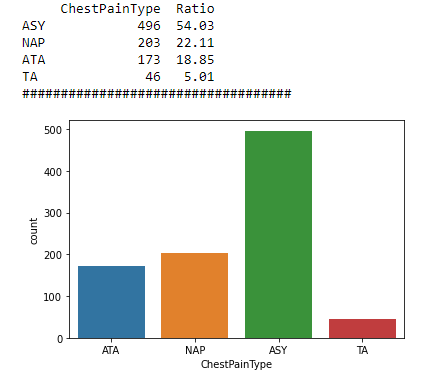
Veri görselleştirme, veri analizinde kullanılan önemli bir araçtır. Verilerin grafikler, tablolar veya grafiksel temsiller aracılığıyla görsel olarak ifade edilmesini sağlar. Bu, verilerin daha anlaşılır, kapsamlı ve etkili bir şekilde analiz edilmesine yardımcı olur. Kalp yetmezliği verisinin özelliklerinin birbiri ile olan ilişkilerini daha iyi anlamak için veri görselleştirme tekniğine başvurulmuştur.

* **Kategorik Değişkenlerin Dağılımı**



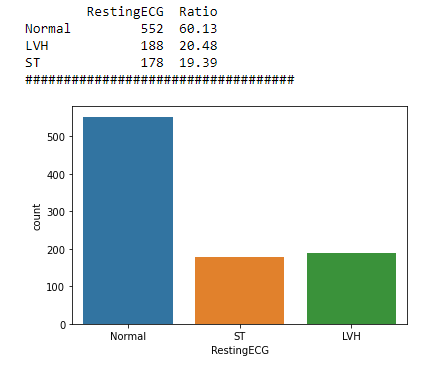
Şekil 1.3 : Cinsiyet Değişkeninin Dağılımı

Şekil 1.3’te %78.98 lik erkek dağılımı, %21.02 kadın dağılımı vardır. 725 erkek, 193 kadın bulunmaktadır.



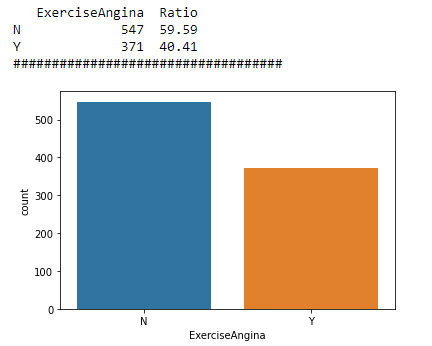
Şekil 1.4 : Göğüs Ağrısı Türü (ChestPainType) Değişkeninin Dağılımı

Şekil 1.4’de %54.03’lük ASY (Asemptomatik), %22.11’lik NAP (Anjinal Olmayan Ağrı), %18.85’lik ATA (Atipik Angina), %5.01’lik TA (Tipik Angina) dağılım mevcuttur.

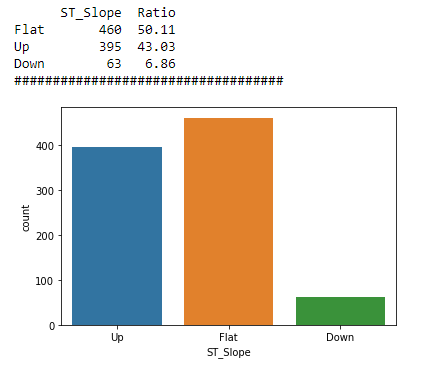


Şekil 1.5 : İstiharat Elektrokardiyogram Sonuçları (RestingECG) Değişkeninin Dağılımı

Şekil 1.5’de %60.13’lük Normal, %20.48’lik LVH (Estes kriterlerine göre olası veya kesin sol ventrikül hipertrofisi gösteriyor), %19.39’luk ST (ST-T dalgası anormalliği (T dalgası inversiyonları ve/veya ST yükselmesi veya > 0,05 mV'lik depresyon)) dağılım mevuttur.

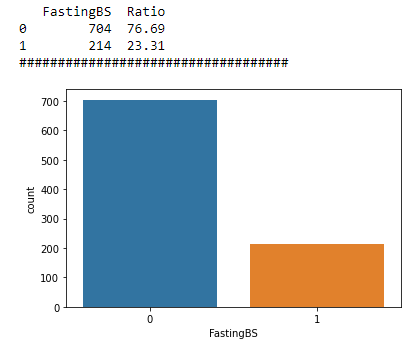


Şekil 1.6 : Egzersize bağlı anjina (ExerciseAngina [Y: Evet, N: Hayır] ) Değişkeninin Dağılımı



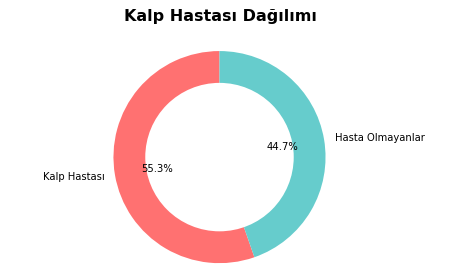
* Şekil 1.6 : Zirve Egzersizinin Eğimi (ST\_Slope) Değişkeninin Dağılımı

Şekil 1.6’da %50.11’lik Flat (Düz), %43,03’lük Up (Yukarı Eğimli), %6.86’lık Down (Aşağı Eğimli) bir dağılım mevcuttur.

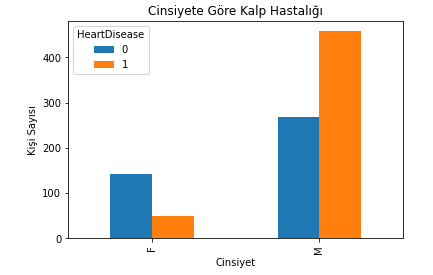


* Şekil 1.7 : Açlık Kan Şekeri (FastingBS) Değişkeninin Dağılımı

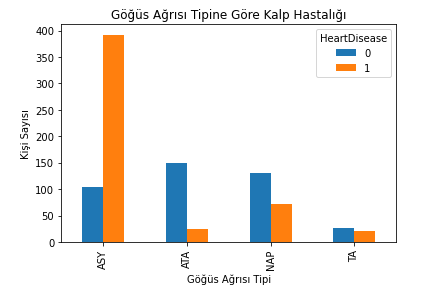
Şekil 1.7’de %76.69’luk 0, %23.31’lik 1 dağılım mevcuttur. [1: Eğer FastingBS > 120 mg/dl ise, 0: Aksi takdirde]



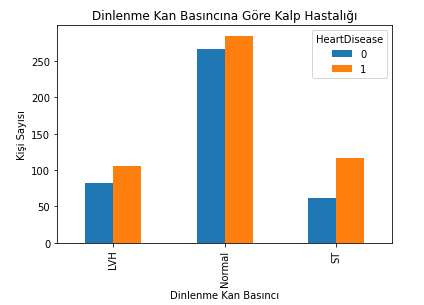
Şekil 1.8 : Kalp Hastası (HeartDisease) Değişkeninin Dağılımı



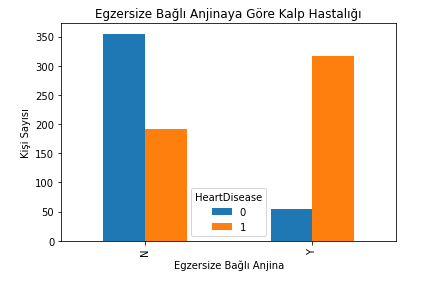
Şekil 1.9 : Cinsiyete Göre Kalp Hastalığı Dağılımı (F:Kadın , M:Erkek)



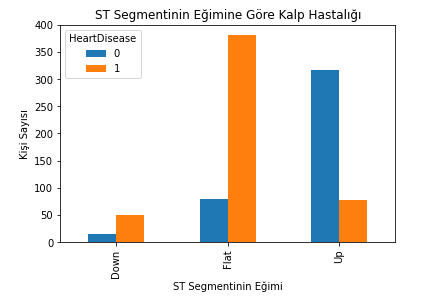
Şekil 1.10 : Göğüs Ağrısı Tipine Göre Kalp Hastalığı Dağılımı



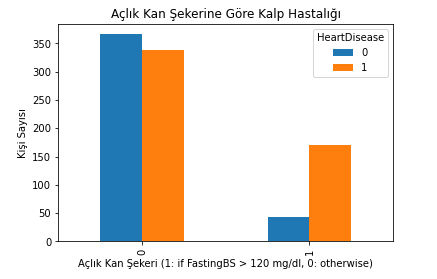
Şekil 1.11 : Dinlenme Kan Basıncına Göre Kalp Hastalığı Dağılımı



Şekil 1.12 : Egzersize Bağlı Anjinaya Göre Kalp Hastalığı Dağılımı

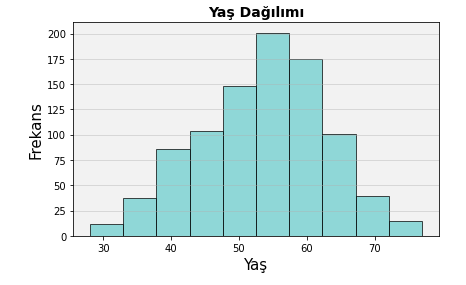


Şekil 1.12 : ST Segmentinin Eğimine Göre Kalp Hastalığı Dağılımı

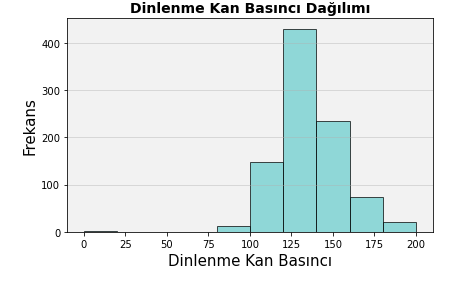


Şekil 1.12 : Açlık Kan Şekerine Göre Kalp Hastalığı Dağılımı

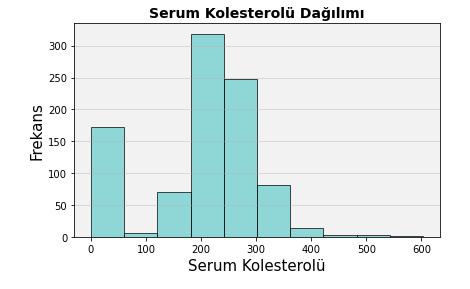
* **Numerik Değişkenlerin Dağılımı**



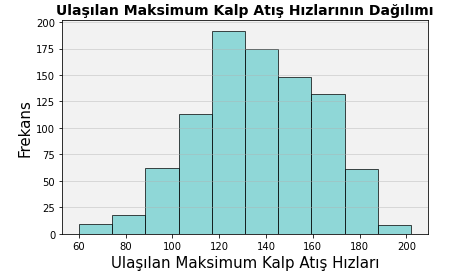
Şekil 1.13 : Yaş Dağılımı



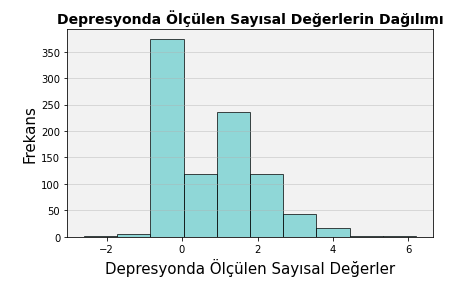
Şekil 1.14 : Dinlenme Kan Basıncı Dağılımı



Şekil 1.15 : Serum Kolesterolü Dağılımı

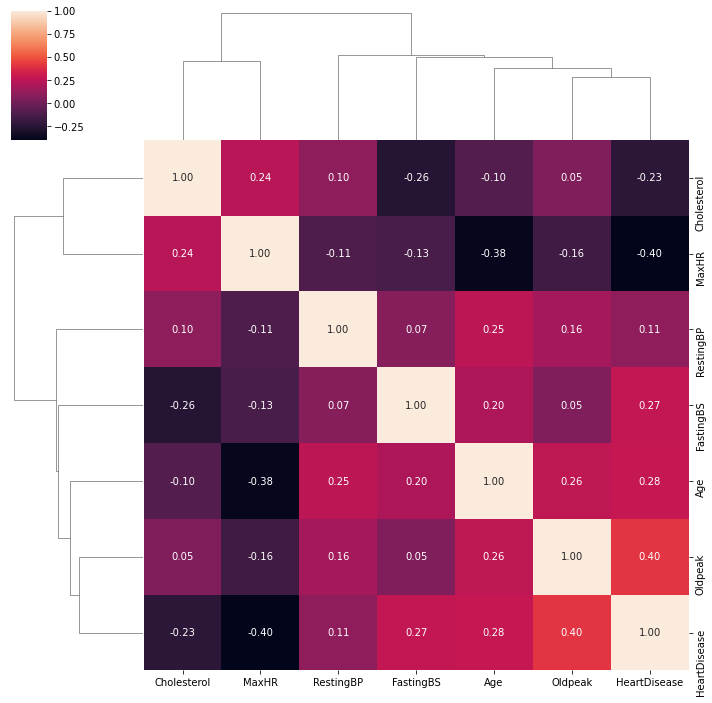


Şekil 1.16 : Ulaşılan Maksimum Kalp Atış Hızlarının Dağılımı



Şekil 1.17 : Depresyonda Ölçlüen Sayısal Değerlerin Dağılımı

* **Korelasyon Matrisi**



Şekil 1.18 : Korelasyon Matrisi

Şekil 1.18’deki korelasyon matrisi, bir verideki değişkenler arasındaki ilişkiyi değerlendirmek için kullanılan bir araçtır. Matris, her bir değişken çifti için korelasyon katsayısını gösterir. Korelasyon katsayısı, -1 ile 1 arasında değer alır. Pozitif bir korelasyon katsayısı, değişkenler arasında doğrusal bir ilişkinin olduğunu gösterirken, negatif bir korelasyon katsayısı, değişkenler arasında ters yönlü bir ilişkinin olduğunu gösterir.

Örnek olarak;

* MaxHR (Ulaşılan maksimum kalp atış hızı) ile HeartDisease (Kalp hastası olup olmadığı) değişkenleri arasında negatif yönlü, doğrusal ve zayıf bir ilişki vardır.

1. ÖZELLİK MÜHENDİSLİĞİ

Özellik mühendisliği, veri biliminde kullanılan bir süreçtir ve veri setinden anlamlı, öngörücü ve daha iyi sonuçlar elde etmek için yeni ve etkili özelliklerin oluşturulmasıyla ilgilidir.

Özellikler, veri setindeki değişkenler veya sütunlar olarak düşünülebilir. Ancak bazen, veri setindeki mevcut özellikler tek başına yeterli olmayabilir veya daha güçlü sonuçlar elde etmek için yeni özelliklere ihtiyaç duyulabilir. İşte bu noktada özellik mühendisliği devreye girer.

Özellik mühendisliği, mevcut özelliklerin dönüşümü, birleştirilmesi veya yeni özelliklerin oluşturulması yoluyla veri setini zenginleştirmeyi amaçlar. Bu, veri setindeki bilgiyi daha iyi yakalamak, modelin performansını artırmak ve daha doğru tahminler yapmak için yapılır.

Özellik mühendisliği aşağıdaki işlemleri içerebilir:

1. Özellik çıkarımı: Mevcut veri setinden yeni özelliklerin çıkarılması. Örneğin, bir tarih sütunu kullanılarak gün, ay veya yıl gibi farklı özelliklerin elde edilmesi.

2. Özellik dönüşümü: Varolan özelliklerin dönüştürülmesi veya ölçeklendirilmesi. Örneğin, logaritmik dönüşüm, normalize etme veya standartlaştırma gibi yöntemler kullanılabilir.

3. Özellik birleştirme: Mevcut özelliklerin birleştirilerek yeni özelliklerin oluşturulması. Örneğin, iki özellik arasındaki çarpım veya toplam gibi işlemlerle yeni bir özellik elde edilebilir.

4. Özellik seçimi: Veri setindeki önemli özelliklerin belirlenmesi ve gereksiz veya tekrarlayan özelliklerin çıkarılması. Bu, modelin karmaşıklığını azaltabilir ve işlem süresini iyileştirebilir.

Özellik mühendisliği, veri setindeki bilgiyi artırmak ve model performansını optimize etmek için kritik bir adımdır. Doğru ve etkili özellikler, daha iyi tahminler yapmak ve daha iyi sonuçlar elde etmek için önemlidir.

* 1. Aykırı Değer Tespiti

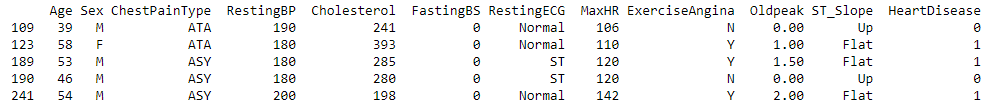
Aykırı değerler (Outlier), istatistiksel analizde kullanılan bir kavramdır ve genellikle veri setinde diğer gözlemlerden önemli ölçüde farklı olan bir noktayı ifade eder. Bir veri setindeki diğer gözlemlerden büyük ölçüde sapma gösteren bir veri noktası, genellikle bir "outlier" olarak adlandırılır.

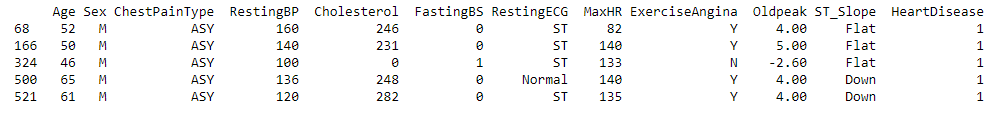
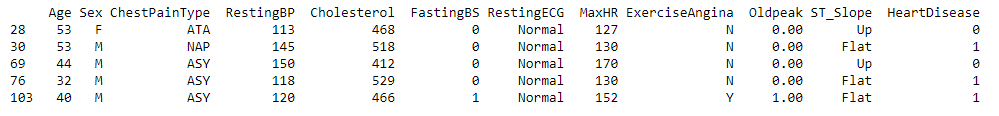
Outlier'lar, veri analizi sürecinde dikkate alınması gereken önemli faktörlerdir. Bu noktalar, veri setinin genel eğilimini veya dağılımını etkileyebilir ve istatistiksel analiz sonuçlarını yanıltabilir. Outlier'lar, veri kaynaklarındaki hatalar, ölçüm yanlışlıkları veya gerçek dünyadaki nadir olayların sonucu olabilir.

Outlier tespiti, istatistiksel yöntemler, grafiksel analiz ve makine öğrenmesi algoritmaları gibi çeşitli teknikler kullanılarak gerçekleştirilebilir. Bu tespit, veri analistleri ve araştırmacılar tarafından, doğru sonuçlar ve güvenilir yorumlar elde etmek için önemli bir adımdır.

Outlier'lar, veri analizinde dikkate alınması gereken istisnai noktalardır. Onları belirlemek ve anlamak, verilerin doğru bir şekilde yorumlanması için önemlidir.

Aşağıda belirtilen gözlemlerde Şekil 1.19’ de gösterilen Kalp Yetmezliği verisinde aykırı değer tespit edilmiştir.





Şekil 1.19 : Aykırı Değerlerin Tespit Edildiği Gözlemler

Aykırı değer tespitleri IQR yöntemine göre gerçekleştirimiştir.

* Aykırı değerler ilk etapta veriden çıkartılmayacaktır. Öncelikle normallik varsayımı kontrolleri yapılacaktır. Daha sonrasında veri modelleme işleminden geçirilip modelin başarı değeri (Accuracy,F1) gözlemlenecektir.
* Daha sonrasında takip eden bölümde alternatif bir model geliştirilecektir. Bu model öncesinde aykırı değerler veriden çıkartılıp normallık varsayımı ve başarı değeri tekrar kontrol edilecektir.
* İki model arasında kıyaslama yapılacaktır ve en iyi model belirlenecektir.
  1. Normallik Testi

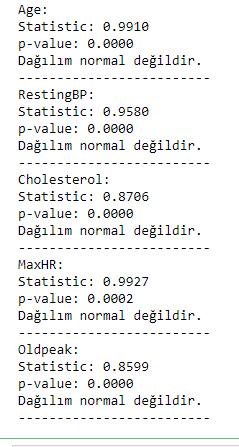
Normallik testi Shapiro-Wilk ile yapılmıştır Shapiro-Wilk testi bir verinin normal dağılım gösterip göstermediğini değerlendirmek için kullanılan istatistiksel bir testtir.

Shapiro-Wilk testi, veri setinin normal dağılımı takip edip etmediğini belirlemek için bir hipotez testidir. Test, verinin normal dağılımdan sapma derecesini ölçer ve p değeri olarak adlandırılan bir istatistik sağlar.

H0 : Veri seti normal dağılıma sahiptir.

H1 : Veri seti normal dağılıma sahip değildir.

Shapiro-Wilk testi, veri setindeki gözlemlerle normal dağılım arasındaki farkı ölçer ve p değeri sağlar. Eğer p değeri belirli bir anlamlılık düzeyinden (genellikle 0.05 olarak kabul edilir) daha küçükse, H0 hipotezi reddedilir ve veri setinin normal dağılımdan sapma gösterdiği sonucuna varılır.



Şekil 1.20 : Shapiro-Wilk Testi

* Şekil 1.20’de ifade edildiği üzere veri normal bir dağılıma sahip değildir. Alternatif modelleme aşaması öncesinde aykırı değerler veriden çıkartılıp verinin normallık varsayımı ve modelin başarı değeri tekrar kontrol edilecektir.
  1. Ki-Kare Testi

Ki-kare testi, iki kategorik değişken arasında bir ilişki olup olmadığını belirlemek için kullanılan bir istatistiksel testtir. Bu test, gözlenen verilerin beklenen verilerle karşılaştırılması yoluyla gerçekleştirilir.

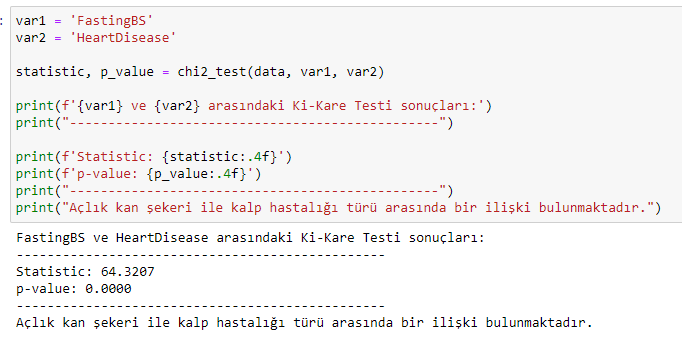
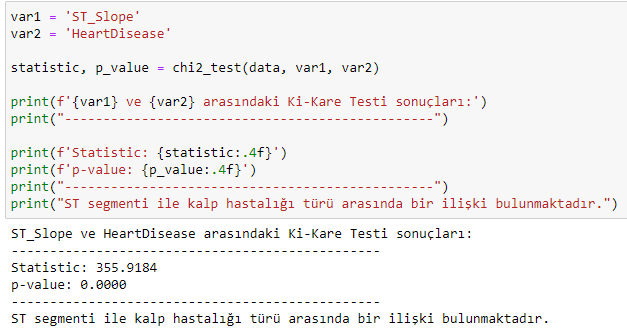
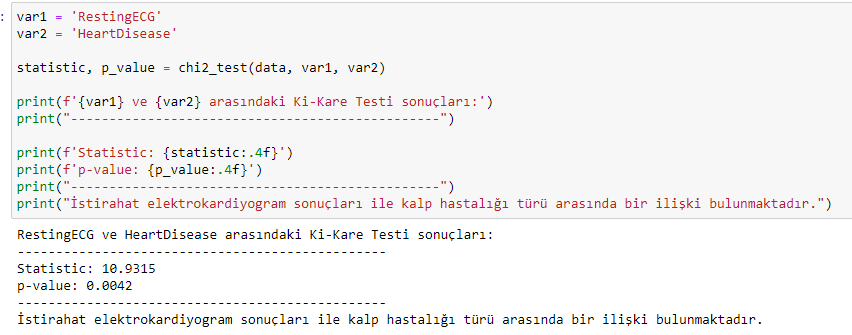
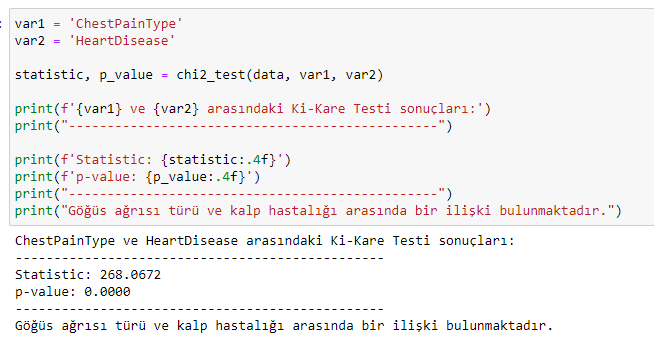
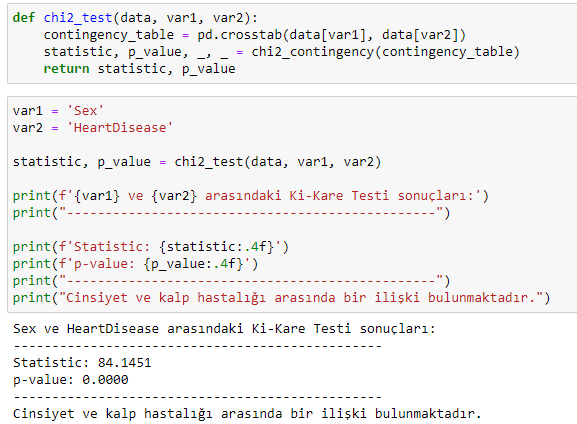
Kalp yetmezliği verisinde, Ki-kare testi, iki veya daha fazla değişken arasında ilişki aramak için bu test kullanılmıştır. Kalp yetmezliği değişkeninin cinsiyet, göğüs ağrısı türü, istirahat elektrokardiyagram sonuçları, ST Segmenti ve açlık kan şekeri özellikleri arasında bir ilişki olup olmadığı araştırılmıştır.

Ki-kare testi, iki değişken arasındaki ilişkinin bağımsızlık hipotezini test eder. Hipotezler şu şekildedir:

- Null hipotezi (H0): İki değişken arasında ilişki yoktur (bağımsızdır).

- Alternatif hipotezi (H1): İki değişken arasında bir ilişki vardır (bağımlıdır).

Ki-kare testi, gözlenen frekanslar ve beklenen frekanslar arasındaki farkları değerlendirerek bir test istatistiği hesaplar. Bu test istatistiği, bir Ki-kare dağılımına karşı test edilir. Eğer test istatistiği, belirli bir anlamlılık düzeyinde (genellikle p < 0.05 olarak kabul edilen) ele alınan eleştirel değerden büyükse, null hipotezi reddederek değişkenler arasında bir ilişki olduğunu söyleyebiliriz.



Yapılan araştırma sonucunda Ki-Kare testi çıktılarında kalp yetmezliği değişkeninin cinsiyet, göğüs ağrısı türü, istirahat elektrokardiyagram sonuçları, ST Segmenti ve açlık kan şekeri özellikleri ile ilişkisi olduğu saptanmıştır.

* 1. Label Encoding İşlemi

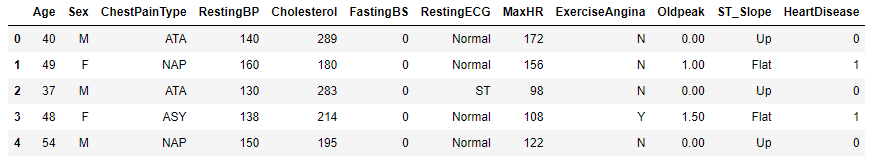
Elbette! Label Encoder, makine öğrenimi ve veri analitiğinde sıkça kullanılan bir veri dönüşüm yöntemidir. Kategorik (nominal veya ordinal) verileri sayısal değerlere dönüştürmek için kullanılır.

Label Encoder, her kategori değerini sırasıyla bir sayısal değerle eşleştirir. Örneğin, "kırmızı", "mavi", "yeşil" gibi kategorik bir değişkeniniz varsa, Label Encoder bu kategorileri sırasıyla 0, 1, 2 gibi sayısal değerlere dönüştürebilir.

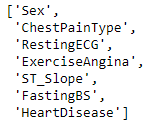
Bu dönüşüm işlemi, makine öğrenimi algoritmalarının kategorik verileri işlemesine yardımcı olur, çünkü çoğu algoritma sayısal verilerle daha iyi çalışır. Label Encoder, kategorik verilerin sıralanma durumuna bağlı olarak ordinal veya nominal olarak kabul edildiğinde kullanılabilir.

Label Encoder, scikit-learn gibi birçok makine öğrenimi kütüphanesinde bulunan bir işlev veya sınıf olarak bulunabilir. Bu kütüphane, label encoding işlemini kolaylaştıran ve veri setinizi sayısal hale getirmenizi sağlayan kullanımı kolay bir API sağlar.

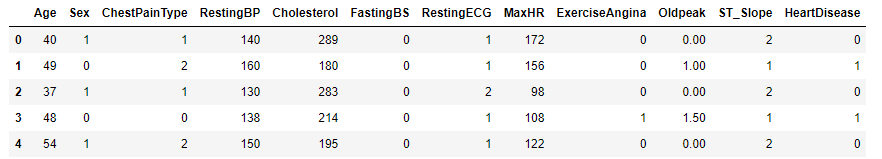
Özetlemek gerekirse, Label Encoder, kategorik verileri sayısal değerlere dönüştürmek için kullanılan bir veri dönüşüm yöntemidir. Bu, makine öğrenimi algoritmalarında kategorik verileri işlemek için yaygın olarak kullanılan bir tekniktir.



Şekil 1.21 : Kalp Yetmezliği Verisi



Şekil 1.22 : Label Encoding İşlemi Uygulanacak Kategorik Değişkenler



Şekil 1.23 : Label Encoding İşlemi Sonucunda Kalp Yetmezliği Verisi

LOJİSTİK REGRESYONU MODELİ

Lojistik regresyon, makine öğrenimi ve istatistikte sınıflandırma problemlerini çözmek için kullanılan bir modeldir. Lojistik regresyon, bir bağımlı değişkenin (sınıflandırılacak olan değişken) bir veya daha fazla bağımsız değişkenle (özellikler) ilişkisini modellemek için kullanılır. Bu model, sonucu 0 veya 1 gibi iki sınıfa ait olabilen verileri tahmin etmek için kullanılır.

* 1. Test – Train

Test ve train kavramları, makine öğrenimi ve model eğitimi sürecinde kullanılan veri kümesinin bölünmesiyle ilgilidir.

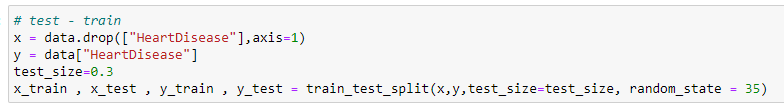
Train (eğitim) veri seti, bir makine öğrenimi modelinin eğitim sürecinde kullanılan veri kümesidir. Model, train veri setindeki örnekleri kullanarak ilişkileri ve desenleri öğrenir. Eğitim veri seti, genellikle modelin performansını değerlendirmek için ayrılan veriye göre daha büyük bir veri setidir.

Test veri seti, eğitim süreci tamamlandıktan sonra modelin performansını değerlendirmek için kullanılan ayrılmış bir veri kümesidir. Test veri seti, daha önce modelin görmediği örnekleri içerir ve modelin bu veriye nasıl genelleme yaptığını değerlendirmek için kullanılır. Test veri seti, genellikle eğitim sürecinde kullanılan veri setinden ayrı bir veri setidir.

Train ve test veri setlerinin ayrılması, makine öğrenimi modellerinin gerçek dünya verileri üzerinde ne kadar iyi performans gösterdiğini belirlemek için önemlidir. Modelin eğitim veri setinde yüksek bir başarı elde etmesi, ancak test veri setinde düşük bir başarı göstermesi durumunda, modelin aşırı uyum (overfitting) sorunu yaşadığı düşünülebilir. Bu durumda model, eğitim veri setine fazla uyum sağlamış ve genelleme yapmakta zorluk çekmektedir.

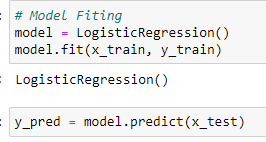
Bu nedenle, train ve test veri setlerinin uygun bir şekilde ayrılması ve modelin test veri setinde iyi performans göstermesi önemlidir. Ayrıca, modelin performansını daha sağlıklı bir şekilde değerlendirmek için cross-validation gibi teknikler de kullanılabilir, bu yöntemde veri seti birden fazla parçaya bölünerek eğitim ve değerlendirme işlemleri tekrarlanır.

Özetlemek gerekirse, train ve test kavramları, makine öğrenimi modellerinin eğitim sürecinde kullanılan veri kümesinin bölünmesiyle ilgilidir. Train veri seti modelin eğitiminde kullanılırken, test veri seti modelin performansının değerlendirilmesinde kullanılır. Bu bölünme, modelin genelleme yeteneğini ve gerçek dünya verileri üzerindeki performansını belirlemek için önemlidir.



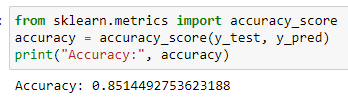
Şekil 1.24 : Kalp Yetmezliği Verisinde Gerçekleştirilen Test – Train İşlemi

* Şekil 1.24’te görülmek üzere veri test – train işleminden geçirilmiştir. Kişinin kalp hastası olup olmadığı yani bağımlı değişken “HeartDisease” y değişkenine, diğer değişkenler ise x dataframe’ne aktarılmıştır.
* Test\_size = 0.3 (%70 train, %30 test) olarak belirlenmiştir.
  1. Model 1



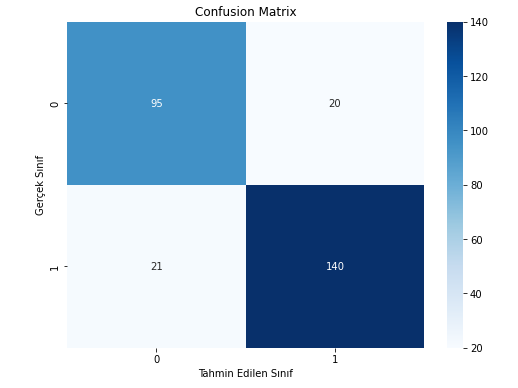
Şekil 1.25 : Lojistik regresyonu modeli

Şekil 1.25’te görülmüş olduğu üzere model train seti üzerinde fit edilmiştir.Ardından test verileri üzerinde tahmin değerleri hesaplanmıştır ve y\_pred değişkenine aktarılmıştır.



Şekil 1.26 : Lojistik regresyonu modeli

Modelde test verileri üzerinden bir tahminleme işlemi gerçekleştirilmiştir ve. Bu tahminleme işleminin sonucunda Şekil 1.26’te görülmüş olduğu üzere modelin başarı değeri (accuracy) %85.14 olarak tespit edilmiştir.



Şekil 1.27 : Confusion Matrix

Şekil 1.27’teki Confusion Matrix, sınıflandırma problemlerinde modelin performansını değerlendirmek için kullanılan bir tablodur. Bu tablo, modelin gerçek ve tahmin edilen sınıflarını görselleştirir ve dört farklı değeri içerir: true positive (TP), true negative (TN), false positive (FP) ve false negative (FN).

TP (doğru pozitif): Modelin doğru bir şekilde bir sınıfı tahmin ettiği durumları temsil eder.

TN (doğru negatif): Modelin doğru bir şekilde bir sınıfı tahmin etmediği durumları temsil eder.

FP (yanlış pozitif): Modelin yanlış bir şekilde bir sınıfı tahmin ettiği durumları temsil eder.

FN (yanlış negatif): Modelin yanlış bir şekilde bir sınıfı tahmin etmediği durumları temsil eder.

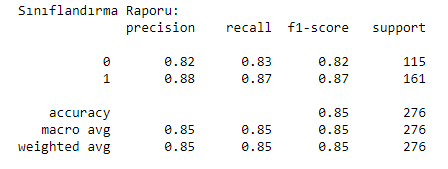
Confusion Matrix, bu dört değeri içeren bir 2x2 tablodur ve genellikle şu şekilde temsil edilir:

|  |  |
| --- | --- |
| Gerçek Pozitif (TP) | Yanlış Pozitif (FP |
| Yanlış Negatif (FN) | Gerçek Negatif (TN) |

Bu matris, modelin sınıflandırma doğruluğunu değerlendirmek için kullanılabilir ve çeşitli performans metriklerinin (örneğin, hassasiyet, özgünlük, doğruluk, F1 skoru) hesaplanmasına yardımcı olur.

Özetlemek gerekirse, Confusion Matrix, bir sınıflandırma modelinin performansını değerlendirmek için kullanılan bir tablodur. Gerçek ve tahmin edilen sınıfları görselleştirir ve doğru/yanlış sınıflandırmaları temsil eden TP, TN, FP ve FN değerlerini içerir. Bu tablo, modelin sınıflandırma doğruluğunu analiz etmek ve performans metriklerini hesaplamak için kullanılır.

* TP (doğru pozitif): 95
* TN (doğru negatif): 140
* FP (yanlış pozitif): 20
* FN (yanlış negatif): 21



Şekil 1.28 : Sınıflandırma Raporu (Classification Report)

Şekil 1.28’teki Classification Report, bir sınıflandırma modelinin performansını değerlendirmek için kullanılan bir rapordur. Bu rapor, sınıflandırma metriklerini (örneğin, hassasiyet, özgünlük, doğruluk, F1 skoru) görselleştirir ve modelin her bir sınıf için performansını ayrıntılı bir şekilde sunar.

Classification Report, genellikle Precision (hassasiyet), Recall (duyarlılık) ve F1 Score gibi metrikleri içerir. Bu metrikler, modelin sınıflandırma doğruluğunu değerlendirmek için kullanılır ve aşağıdaki gibi hesaplanır:

- Precision (Hassasiyet): TP / (TP + FP) formülüyle hesaplanır. Gerçek pozitiflerin (doğru tahmin edilen) toplam pozitif tahminlere oranını temsil eder.

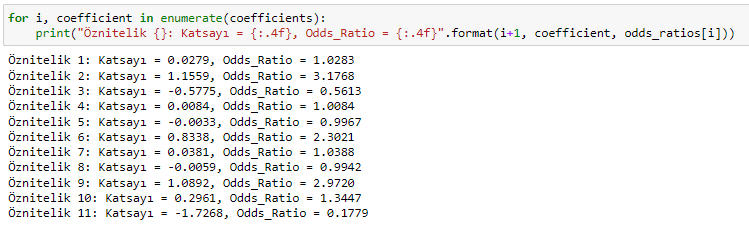
- Recall (Duyarlılık): TP / (TP + FN) formülüyle hesaplanır. Gerçek pozitiflerin toplam gerçek pozitiflere (doğru sınıflandırılan) oranını temsil eder.

- F1 Score: Precision ve Recall metriklerinin harmonik ortalamasıdır ve TP / (TP + ((FN + FP) / 2)) formülüyle hesaplanır. Hem hassasiyeti hem de duyarlılığı göz önünde bulundurarak bir denge sağlar.

Classification Report, bu metrikleri her bir sınıf için ayrı ayrı sunar. Ayrıca, mikro ve makro ortalamalar da raporda yer alabilir. Mikro ortalama, sınıfların toplam performansını hesaplar, makro ortalama ise her sınıfın performansını eşit olarak değerlendirir.

Özetlemek gerekirse, Classification Report, bir sınıflandırma modelinin performansını değerlendiren bir rapordur. Bu rapor, hassasiyet, duyarlılık ve F1 skoru gibi sınıflandırma metriklerini içerir ve modelin her bir sınıf için performansını ayrıntılı bir şekilde sunar.

Modelde test verileri üzerinden bir tahminleme işlemi gerçekleştirilmiştir ve. Bu tahminleme işleminin sonucunda Şekil 1.28’te görülmüş olduğu üzere modelin başarı değerleri (F1 ve Accuracy) %85 olarak tespit edilmiştir.



Şekil 1.29 : Modelin Coefficients(Katsayılar) ve Odds Ratio Değerleri

Şekil 1.29’daki verilere dayanarak, her bir özniteliğin katsayıları ve odds ratio değerlerine bakacak olursak;

1. Öznitelik: Katsayı = 0.0279, Odds\_Ratio = 1.0283

Bu öznitelik için katsayı pozitif olduğu için, değişkenin artışıyla hedef değişkenin olasılığı artma eğilimindedir. Odds ratio değeri 1.0283 olduğundan, bu öznitelik hedef değişkeni üzerinde çok küçük bir etkiye sahip olabilir.

2. Öznitelik: Katsayı = 1.1559, Odds\_Ratio = 3.1768

Bu öznitelik için katsayı oldukça yüksek olduğu için, değişkenin artışı hedef değişkenin olasılığını önemli ölçüde artırır. Odds ratio değeri 3.1768 olduğundan, bu öznitelik hedef değişkeni üzerinde güçlü bir etkiye sahip olabilir.

3. Öznitelik: Katsayı = -0.5775, Odds\_Ratio = 0.5613

Bu öznitelik için katsayı negatif olduğu için, değişkenin artışı hedef değişkenin olasılığını azaltma eğilimindedir. Odds ratio değeri 0.5613 olduğundan, bu öznitelik hedef değişkeni üzerinde zayıf bir etkiye sahip olabilir.

4. Öznitelik: Katsayı = 0.0084, Odds\_Ratio = 1.0084

Bu öznitelik için katsayı ve odds ratio değeri oldukça yakın olduğundan, bu öznitelik hedef değişkeni üzerinde nispeten küçük bir etkiye sahip olabilir.

5. Öznitelik: Katsayı = -0.0033, Odds\_Ratio = 0.9967

Bu öznitelik için katsayı negatif olduğu için, değişkenin artışı hedef değişkenin olasılığını azaltma eğilimindedir. Odds ratio değeri 0.9967 olduğundan, bu öznitelik hedef değişkeni üzerinde çok küçük bir etkiye sahip olabilir.

6. Öznitelik: Katsayı = 0.8338, Odds\_Ratio = 2.3021

Bu öznitelik için katsayı pozitif olduğu için, değişkenin artışı hedef değişkenin olasılığını artırma eğilimindedir. Odds ratio değeri 2.3021 olduğundan, bu öznitelik hedef değişkeni üzerinde orta düzeyde bir etkiye sahip olabilir.

7. Öznitelik: Katsayı = 0.0381, Odds\_Ratio = 1.0388

Bu öznitelik için katsayı pozitif olduğu için, değişkenin artışı hedef değişkenin olasılığını artırma eğilimindedir. Odds ratio değeri 1.0388 olduğundan, bu öznitelik hedef değişkeni üzerinde nispeten küçük bir etkiye sahip olabilir.

8. Öznitelik: Katsayı = -0.0059, Odds\_Ratio = 0.9942

Bu öznitelik için katsayı negatif olduğu için, değişkenin artışı hedef değişkenin olasılığını azaltma eğilimindedir. Odds ratio değeri 0.9942 olduğundan, bu öznitelik hedef değişkeni üzerinde çok küçük bir etkiye sahip olabilir.

9. Öznitelik: Katsayı = 1.0892, Odds\_Ratio = 2.9720

Bu öznitelik için katsayı pozitif olduğu için, değişkenin artışı hedef değişkenin olasılığını artırma eğilimindedir. Odds ratio değeri 2.9720 olduğundan, bu öznitelik hedef değişkeni üzerinde orta düzeyde bir etkiye sahip olabilir.

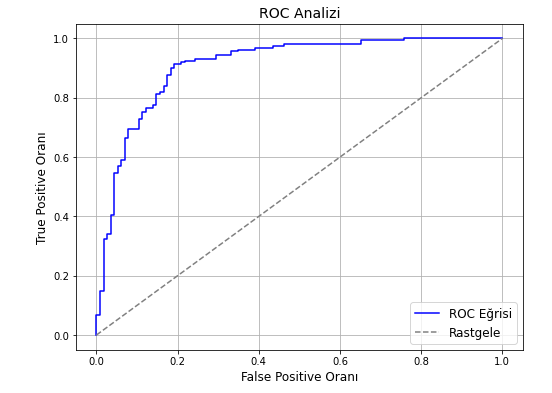
10. Öznitelik: Katsayı = 0.2961, Odds\_Ratio = 1.3447

Bu öznitelik için katsayı pozitif olduğu için, değişkenin artışı hedef değişkenin olasılığını artırma eğilimindedir. Odds ratio değeri 1.3447 olduğundan, bu öznitelik hedef değişkeni üzerinde nispeten küçük bir etkiye sahip olabilir.

11. Öznitelik: Katsayı = -1.7268, Odds\_Ratio = 0.1779

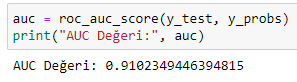
Bu öznitelik için katsayı oldukça negatif olduğu için, değişkenin artışı hedef değişkenin olasılığını önemli ölçüde azaltma eğilimindedir. Odds ratio değeri 0.1779 olduğundan, bu öznitelik hedef değişkeni üzerinde güçlü bir etkiye sahip olabilir.

Lojistik regresyon sonuçlarına dayanarak, özniteliklerin hedef değişken üzerindeki etkilerini değerlendirebiliriz. Katsayılar, değişkenlerin etkisini ve yönünü gösterirken, odds ratio değerleri bu etkilerin büyüklüğünü ifade eder. Ancak, kesin yorumlar yapabilmek için kullanılan veri seti ve analiz yöntemi hakkında daha fazla bilgiye ihtiyaç vardır.



Şekil 1.30 : ROC Eğrisi

AUC (Area Under the Curve), ROC (Receiver Operating Characteristic) eğrisinin altındaki alanı temsil eder. Bu değer, bir sınıflandırma modelinin doğruluğunu ve ayrım yeteneğini ölçmek için kullanılan yaygın bir metriktir. AUC değeri, 0 ile 1 arasında bir değer alır.



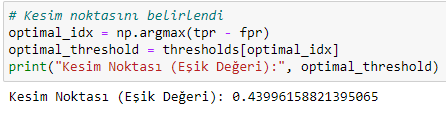
Şekil 1.31 : AUC Değeri

AUC değeri 0.91 olarak tespit edilmiştir. Bu değer oldukça yüksek bir AUC değeridir. Yüksek bir AUC değeri, modelin yüksek bir doğruluk ve ayrım yeteneği olduğunu gösterir. Model, pozitif ve negatif örnekleri doğru şekilde sınıflandırmak için iyi bir performans sergilemektedir.

Bu yüksek AUC değeri, modelin veri setindeki sınıfları iyi bir şekilde ayrıştırabildiğini ve tahminlerinin gerçek değerlere yakın olduğunu göstermektedir. Bu da modelin genel olarak güvenilir ve etkili olduğunu düşündürebilir.

Ancak, AUC değeri tek başına bir modelin performansını tam olarak anlamak için yeterli değildir. Diğer performans metriklerini de dikkate almak önemlidir. Ayrıca, kullanılan veri seti, modelin doğruluğunu ve AUC değerini etkileyebilecek faktörlerden biridir.

Sonuç olarak, 0.91 AUC değeri iyi bir performans gösteren bir sınıflandırma modelini işaret eder. Ancak, modelin performansını daha kapsamlı bir şekilde değerlendirmek için diğer metrikleri ve bağlamı da göz önünde bulundurmanız önemlidir.



Şekil 1.32 : Kesim Noktası

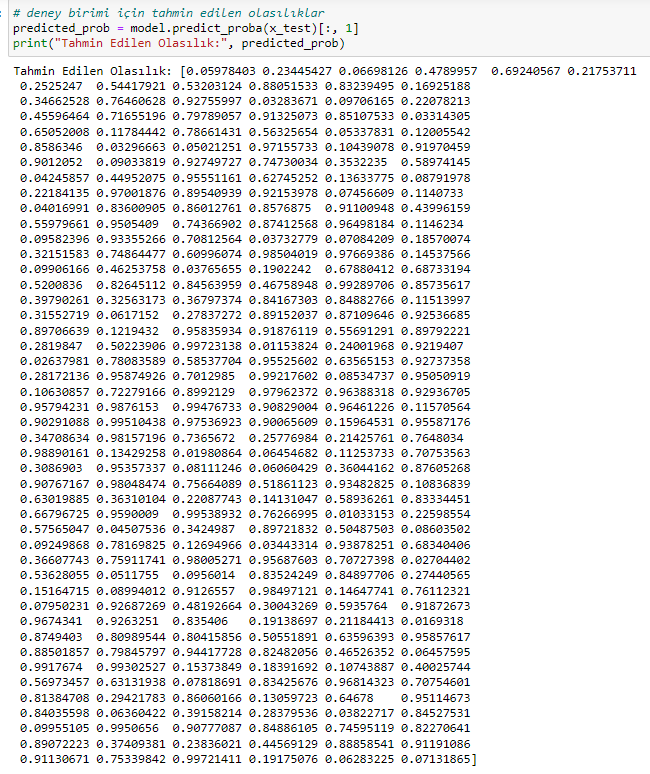
Kesim Noktası veya Eşik Değeri, bir sınıflandırma modelinde pozitif sınıfı ve negatif sınıfı ayırmak için kullanılan bir değerdir. Eşik değeri, modelin tahmin sonuçlarını pozitif veya negatif olarak sınıflandırırken kullanılan bir eşik veya sınır değeridir.

Bu durumda, kesim noktası veya eşik değeri 0.439 olarak belirtilmiş. Bu değer, modelin tahmin sonuçlarını değerlendirirken 0.439 eşik değerinin üzerindeki tahminleri pozitif olarak sınıflandıracağını gösterir.

Eşik değeri, modelin hassasiyeti ve özgüllüğü arasında bir denge oluşturur. Eşik değeri düştükçe, modelin pozitif sınıfı tahmin etme eşiği düşer ve bu da modelin daha fazla pozitif sınıfı yakalayabileceği anlamına gelir. Ancak, aynı zamanda yanlış pozitiflerin sayısı da artabilir.

Eşik değeri seçimi, kullanıcının önceliklerine ve uygulamanın gereksinimlerine bağlıdır. Daha düşük bir eşik değeri, modelin pozitif sınıfı daha iyi yakalamasına olanak tanırken, yanlış pozitiflerin sayısını artırabilir. Daha yüksek bir eşik değeri, modelin pozitif sınıfı daha az yakalamasına ve yanlış pozitiflerin sayısını azaltmasına neden olabilir.

Eşik değerinin seçimi, modelin performansını değerlendirmek ve optimal sonuçlar elde etmek için diğer performans metrikleriyle birlikte dikkate alınmalıdır. Farklı eşik değerleriyle deneme yaparak, işin gereksinimlerine en uygun eşik değerini belirlemek önemlidir.



Şekil 1.33 : Tahmin Edilen Olasılık Değerleri

Elde ettiğiniz tahmin edilen olasılık değerleri, lojistik regresyon modelinizin verilere dayanarak sınıflandırma tahminleri yapması sonucunda elde edilen sonuçlardır. Her bir değer, ilgili deney birimi için pozitif sınıfa ait olma olasılığını temsil eder.

Örneğin, tahmin edilen olasılık değerinizden birini alalım: 0.05978403. Bu değer, ilgili deney biriminin pozitif sınıfa ait olma olasılığının oldukça düşük olduğunu göstermektedir. Bu durumda, modeliniz bu deney birimini negatif sınıfa daha yakın olarak sınıflandıracaktır.

Tahmin edilen olasılık değerleri genellikle eşik değeriyle karşılaştırılarak son sınıflandırma kararı verilir. Örneğin, eşik değeri 0.5 olarak belirlenmişse, 0.05978403 gibi düşük bir olasılık değeri pozitif sınıf yerine negatif sınıfa ait olduğunu düşündürebilir.

Ancak, sınıflandırma kararı eşik değeriyle sınırlı değildir. Eşik değeri, modelin performansını ve gereksinimlerinizi göz önünde bulundurarak seçilmelidir. Eşik değerini düşürerek pozitif sınıfın yakalanma oranını artırabilirsiniz, ancak aynı zamanda yanlış pozitif oranını da artırabilirsiniz. Eşik değerini yükselterek yanlış pozitif oranını azaltabilirsiniz, ancak pozitif sınıfın yakalanma oranı da azalabilir.

Tahmin edilen olasılık değerleri, modelin pozitif ve negatif sınıfları tahmin etme yeteneği hakkında bilgi sağlar. Ancak, sınıflandırma kararı için eşik değeriyle birlikte diğer performans metriklerini de değerlendirmek önemlidir, örneğin, hassasiyet, özgüllük, doğruluk veya alan altında kalan (AUC) değeri gibi metrikler.

Sonuç olarak, tahmin edilen olasılık değerleri, lojistik regresyon modelinizin verilere dayalı olarak pozitif ve negatif sınıfları tahmin etmesi sonucunda elde edilen değerlerdir. Eşik değeriyle birlikte değerlendirilerek sınıflandırma kararlarına yardımcı olurlar.

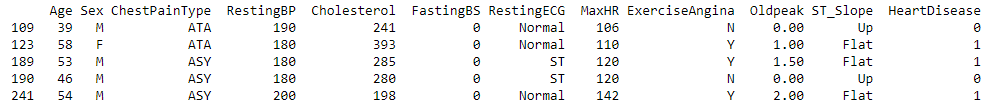
Örneğin, tahmin edilen olasılık değerinizden birini daha alalım: 0.99721411 olan tahmin edilen olasılık değeri, ilgili deney biriminin pozitif sınıfa ait olma olasılığının oldukça yüksek olduğunu göstermektedir. Bu durumda, modeliniz bu deney birimini pozitif sınıfa daha yakın olarak sınıflandıracaktır.

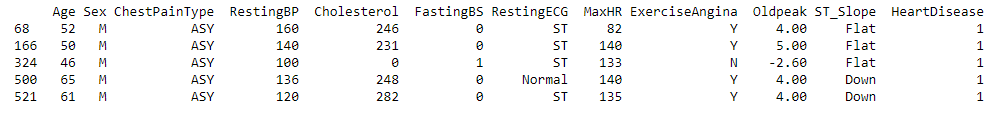
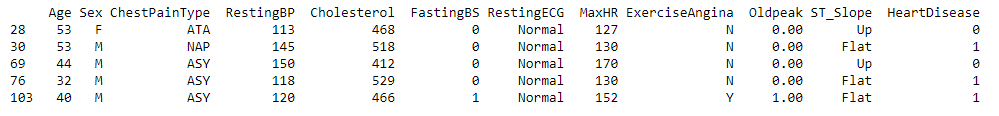
Yüksek bir tahmin edilen olasılık değeri genellikle pozitif sınıfı işaret eder ve modelin bu deney birimini pozitif olarak sınıflandırma güvenini yansıtır. Ancak, her ne kadar 0.99721411 gibi yüksek bir değer pozitif sınıfa ait olma olasılığının yüksek olduğunu gösterse de, mutlaka diğer performans metriklerini ve eşik değerini de göz önünde bulundurmak önemlidir.

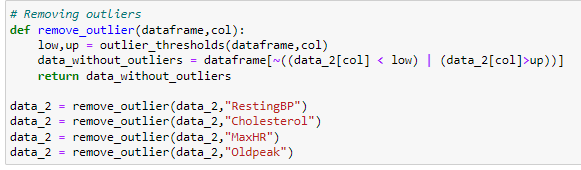
Eşik değeriyle karşılaştırıldığında, bu değer eşik değerinden yüksek olduğu için modeliniz bu deney birimini pozitif sınıfa tahmin etme eğilimindedir. Ancak, sınıflandırma kararı eşik değeri ve diğer metriklerle birlikte değerlendirilmelidir. Eşik değerini düşürmek, pozitif sınıfın yakalanma oranını artırabilir, ancak aynı zamanda yanlış pozitif oranını da artırabilir.

Sonuç olarak, 0.99721411 gibi yüksek bir tahmin edilen olasılık değeri, modelin ilgili deney birimini pozitif sınıfa daha yakın olarak sınıflandırma eğilimini gösterir. Ancak, sınıflandırma kararı için eşik değeri ve diğer performans metriklerini de dikkate almak önemlidir.

* 1. Model 2
* Bu modelde 1. Modelden farklı olarak model kurulum aşamasından önce aykırı değerler veriden çıkartılacaktır.Her bir bağımsız kategorik değişken bağımlı değişken ile ilişkili olduğu için modelleme aşaması için değişken çıkartma gereği duyulmamıştır.
* Daha sonra normallik varsayımı tekrar kontrol edilecektir.
* Lojistik regresyonu modeli kurulacaktır.

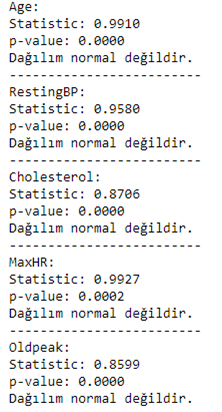






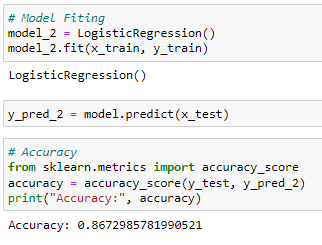
Şekil 1.34 : Aykırı Değerlerin Tespit Edildiği Gözlemler

Şekil 1.34’ teki aykırı değer içeren gözlemler veriden çıkarılmışır.Modelin başarı oranında artış beklenmektedir.



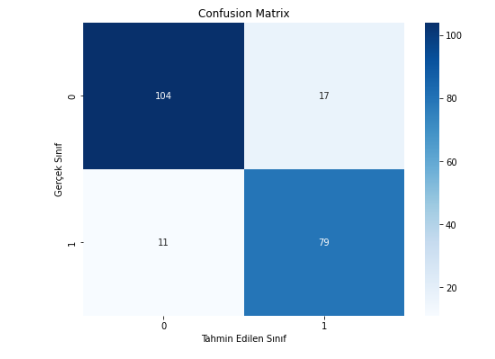
Şekil 1.35 : Shapiro-Wilk Testi

* Şekil 1.35’de ifade edildiği üzere veriden aykırı değerler çıkarılmasına rağmen normallik varsayımı sağlanamamaktadır.



Şekil 1.36 : Lojistik Regresyonu Modeli ve Accuracy Değeri

2.Modelde test verileri üzerinden tekrar bir tahminleme işlemi gerçekleştirilmiştir ve. Bu tahminleme işleminin sonucunda Şekil 1.36’te görülmüş olduğu üzere modelin başarı değeri (accuracy) %86.72 olarak tespit edilmiştir. Accuracy değerinde bir yükseliş söz konusudur.



Şekil 1.37 : Confusion Matrix

* TP (doğru pozitif): 104
* TN (doğru negatif): 79
* FP (yanlış pozitif): 27
* FN (yanlış negatif): 11

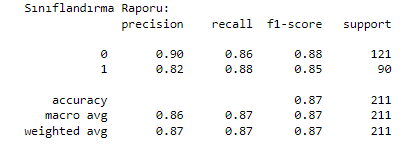
- TP (True Positive): Gerçek pozitif sayısı, modelin pozitif olarak sınıflandırdığı ve gerçekten pozitif olan veri noktalarının sayısıdır. Bu durumda, 104 deney birimi doğru bir şekilde pozitif olarak sınıflandırılmıştır.

- TN (True Negative): Gerçek negatif sayısı, modelin negatif olarak sınıflandırdığı ve gerçekten negatif olan veri noktalarının sayısıdır. Bu durumda, 79 deney birimi doğru bir şekilde negatif olarak sınıflandırılmıştır.

- FP (False Positive): Yanlış pozitif sayısı, modelin pozitif olarak sınıflandırdığı ancak gerçekte negatif olan veri noktalarının sayısıdır. Bu durumda, 27 deney birimi yanlış bir şekilde pozitif olarak sınıflandırılmıştır.

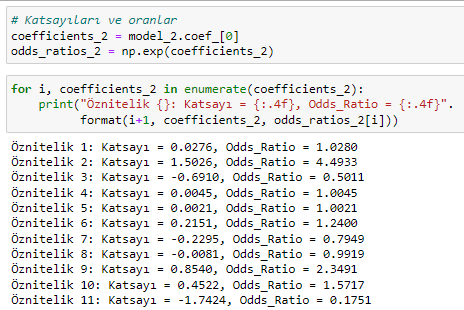
- FN (False Negative): Yanlış negatif sayısı, modelin negatif olarak sınıflandırdığı ancak gerçekte pozitif olan veri noktalarının sayısıdır. Bu durumda, 11 deney birimi yanlış bir şekilde negatif olarak sınıflandırılmıştır.

Şekil 1.37’de Confusion matrix değerlerine bakarak, modelinizin pozitif sınıfı iyi bir şekilde tanımladığını görebiliriz. Ancak, yanlış pozitif ve yanlış negatiflerin de olduğunu gözlemliyoruz. Bu durumda, modelinizin pozitif sınıfı belirlerken bazı hatalar yaptığı ve iyileştirme fırsatları olduğu anlaşılabilir.



Şekil 1.38: Classification Report

Modelde test verileri üzerinden bir tahminleme işlemi gerçekleştirilmiştir ve. Bu tahminleme işleminin sonucunda Şekil 1.38’te görülmüş olduğu üzere modelin başarı değerleri (F1 ve Accuracy) %87 olarak tespit edilmiştir. Modelin başarı metriklerinde bir artış söz konusudur



Şekil 1.39 : Modelin Coefficients(Katsayılar) ve Odds Ratio Değerleri

Şekil 1.39’a göre:

Öznitelik 1: Katsayı = 0.0276, Odds\_Ratio = 1.0280

Bu öznitelik için katsayı pozitif bir değere sahiptir, bu da öznitelik arttıkça hedef değişkenin olasılığının da arttığını gösterir. Odds ratio değeri 1.0280 olduğundan, her bir birimlik artışta hedef değişkenin olasılığının yaklaşık olarak 1.0280 kat arttığını ifade eder.

Öznitelik 2: Katsayı = 1.5026, Odds\_Ratio = 4.4933

Bu öznitelik için katsayı pozitif bir değere sahiptir, bu da öznitelik arttıkça hedef değişkenin olasılığının da arttığını gösterir. Odds ratio değeri 4.4933 olduğundan, her bir birimlik artışta hedef değişkenin olasılığının yaklaşık olarak 4.4933 kat arttığını ifade eder.

Öznitelik 3: Katsayı = -0.6910, Odds\_Ratio = 0.5011

Bu öznitelik için katsayı negatif bir değere sahiptir, bu da öznitelik arttıkça hedef değişkenin olasılığının azaldığını gösterir. Odds ratio değeri 0.5011 olduğundan, her bir birimlik artışta hedef değişkenin olasılığının yaklaşık olarak 0.5011 kat azaldığını ifade eder.

Öznitelik 4: Katsayı = 0.0045, Odds\_Ratio = 1.0045

Bu öznitelik için katsayı pozitif bir değere sahiptir, bu da öznitelik arttıkça hedef değişkenin olasılığının da arttığını gösterir. Odds ratio değeri 1.0045 olduğundan, her bir birimlik artışta hedef değişkenin olasılığının yaklaşık olarak 1.0045 kat arttığını ifade eder.

Öznitelik 5: Katsayı = 0.0021, Odds\_Ratio = 1.0021

Bu öznitelik için katsayı pozitif bir değere sahiptir, bu da öznitelik arttıkça hedef değişkenin olasılığının da arttığını gösterir. Odds ratio değeri 1.0021 olduğundan, her bir birimlik artışta hedef değişkenin olasılığının yaklaşık olarak 1.0021 kat arttığını ifade eder.

Öznitelik 6: Katsayı = 0.2151, Odds\_Ratio = 1.2400

Bu öznitelik için katsayı pozitif bir değere sahiptir, bu da öznitelik arttıkça hedef değişkenin olasılığının da arttığını gösterir. Odds ratio değeri 1.2400 olduğundan, her bir birimlik artışta hedef değişkenin olasılığının yaklaşık olarak 1.2400 kat arttığını ifade eder.

Öznitelik 7: Katsayı = -0.2295, Odds\_Ratio = 0.7949

Bu öznitelik için katsayı negatif bir değere sahiptir, bu da öznitelik arttıkça hedef değişkenin olasılığının azaldığını gösterir. Odds ratio değeri 0.7949 olduğundan, her bir birimlik artışta hedef değişkenin olasılığının yaklaşık olarak 0.7949 kat azaldığını ifade eder.

Öznitelik 8: Katsayı = -0.0081, Odds\_Ratio = 0.9919

Bu öznitelik için katsayı negatif bir değere sahiptir, bu da öznitelik arttıkça hedef değişkenin olasılığının azaldığını gösterir. Odds ratio değeri 0.9919 olduğundan, her bir birimlik artışta hedef değişkenin olasılığının yaklaşık olarak 0.9919 kat azaldığını ifade eder.

Öznitelik 9: Katsayı = 0.8540, Odds\_Ratio = 2.3491

Bu öznitelik için katsayı pozitif bir değere sahiptir, bu da öznitelik arttıkça hedef değişkenin olasılığının da arttığını gösterir. Odds ratio değeri 2.3491 olduğundan, her bir birimlik artışta hedef değişkenin olasılığının yaklaşık olarak 2.3491 kat arttığını ifade eder.

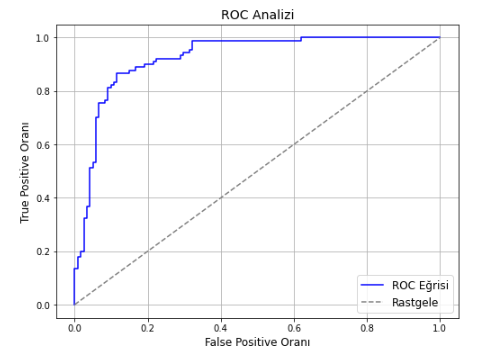
Öznitelik 10: Katsayı = 0.4522, Odds\_Ratio = 1.5717

Bu öznitelik için katsayı pozitif bir değere sahiptir, bu da öznitelik arttıkça hedef değişkenin olasılığının da arttığını gösterir. Odds ratio değeri 1.5717 olduğundan, her bir birimlik artışta hedef değişkenin olasılığının yaklaşık olarak 1.5717 kat arttığını ifade eder.:

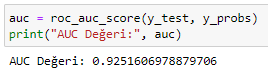
Öznitelik 11: Katsayı = -1.7424, Odds\_Ratio = 0.1751

Bu öznitelik için katsayı negatif bir değere sahiptir, bu da öznitelik arttıkça hedef değişkenin olasılığının azaldığını gösterir. Odds ratio değeri 0.1751 olduğundan, her bir birimlik artışta hedef değişkenin olasılığının yaklaşık olarak 0.1751 kat azaldığını ifade eder.

Bu katsayılar ve odds ratio değerleri, lojistik regresyon modelinizin özniteliklerin hedef değişken üzerindeki etkisini göstermektedir. Pozitif katsayılara sahip öznitelikler, hedef değişkenin artmasına katkıda bulunurken, negatif katsayılara sahip öznitelikler, hedef değişkenin azalmasına katkıda bulunur. Odds ratio değerleri, özniteliklerin bir birimlik değişiminin hedef değişkenin olasılığı üzerindeki etkisini ifade eder.

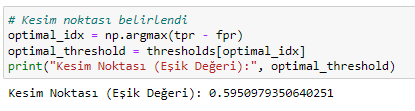


Şekil 1.40 : ROC Eğrisi



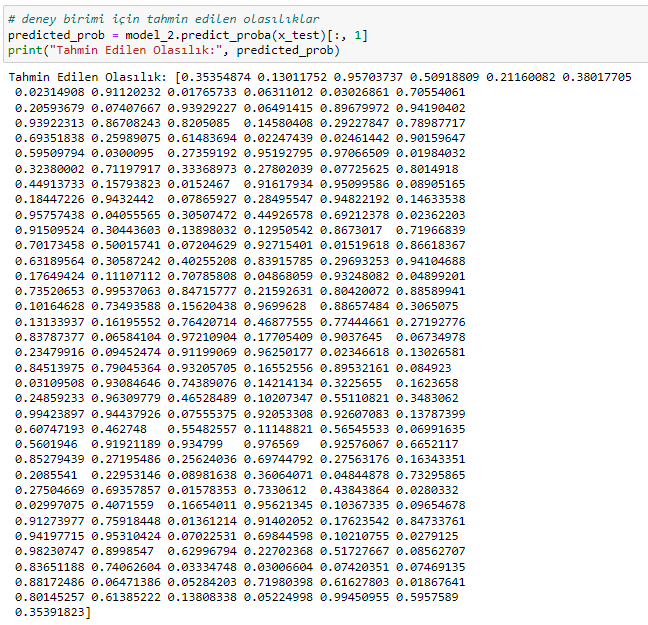
Şekil 1.41 : AUC Değeri

0.9251 AUC değeri, modelin iyi bir performansa sahip olduğunu gösterir. AUC değeri 0 ile 1 arasında bir değer alır, 1'e ne kadar yakınsa, modelin sınıflandırma yeteneği o kadar yüksek demektir. Dolayısıyla, 0.9251 AUC değeri, modelin başarılı bir şekilde sınıflandırma yaptığını ve yüksek bir doğruluk oranına sahip olduğunu göstermektedir.



Şekil 1.42 : Kesim Noktası

Şekil 1.42’de kesim noktası (eşik değeri) 0.595 olarak belirlenmiştir. Bu değer, bir sınıflandırma veya karar verme modelinde kullanılan bir eşik değeridir. Modelin çıktısı bu eşik değerinden büyük veya eşitse, sonuç pozitif olarak sınıflandırılır. Ancak, çıktı bu eşik değerinden küçükse sonuç negatif olarak sınıflandırılır. Yani, bu eşik değeri, modelin pozitif ve negatif sınıfları ayırt etmesinde bir kriter olarak kullanılır.



Şekil 1.43 : Tahmin Edilen Olasılık Değerleri

Şekil 1.43’e göre verilen çıktılar, modelin test veri kümesi için tahmin ettiği olasılık değerleridir. Her bir veri noktası için tahmin edilen olasılık, o noktanın pozitif bir sınıfa ait olma olasılığını temsil eder. Örneğin, ilk veri noktası için tahmin edilen olasılık 0.35354874'tür ve bu, o verinin pozitif sınıfa ait olma olasılığını gösterir.

Genel olarak, tahmin edilen olasılıkların değerleri 0 ile 1 arasında değişir. Daha yüksek olasılık değerleri, veri noktasının pozitif sınıfa ait olduğunu gösterirken, daha düşük değerler negatif sınıfa ait olma olasılığını yansıtır.

SONUÇ

Bu rapor, kalp hastalıklarını tahmin etmek için kullanılabilecek bir makine öğrenimi algoritmalarından, bir sınıflandırma algoritması olan lojistik modelinin sonuçlarını değerlendirmiştir. Veri, 11 özelliği içeriyor ve kalp hastalığı olan veya yüksek kardiyovasküler risk altında olan kişilerin erken tespit ve yönetimine yardımcı olmayı hedeflemiştir.

Sonuçlar, modelin test veri kümesindeki örnekler için tahmin ettiği olasılık değerlerini içermektedir. Bu olasılık değerleri, veri noktalarının kalp hastalığına sahip olma olasılığını temsil etmektedir. Örnek verilerde, tahmin edilen olasılıkların çeşitli değerlerde olduğu görülmektedir.

Bu rapora dayanarak, modelin kalp hastalığını tahmin etmek için kullanılabileceği ve risk altındaki kişilerin erken tespit ve yönetimine yardımcı olabileceği sonucuna varılabilir. Modellerin performansını daha detaylı değerlendirebilmek için Accuracy, F1 gibi değerlendirme metriklerinin de incelenmesi durumunda en iyi lojstik regresyonu modelinin 2. Model olduğu yapılan analizlar sonucunda tespit edilmiştir

Sonuç olarak, bu lojistik regresyonu modellerinin, kardiyovasküler hastalığı olan veya yüksek risk altında olan kişilerin erken teşhis ve yönetimine katkıda bulunabileceği düşünülmektedir. Ancak, bu sonuçların klinik değerlendirme ve uzman görüşleriyle desteklenmesi önemlidir.