Kalp Krizi Tahmini

1st Hasan Tan
Elektrik ve Elektronik Mühendisliği
TOBB Ekonomi ve Teknoloji Üniversitesi
Ankara, Türkiye
hasan$_t n$@$hotmail.com$

***Özet*—Dünya genelinde Kardiyovasküler Hastalıklar (KVH), yaklaşık olarak her yıl 17.9 milyon yaşamı etkileyerek, tüm ölümlerin %31'ine denk gelen birinci ölüm nedenidir. Bu ölümlerin 4'te 1'i kalp krizleri ve inmelere bağlı olup, bu ölümlerin üçte biri 70 yaş altındaki bireylerde beklenmedik şekilde meydana gelmektedir. Kalp yetmezliği, KVH'lerden kaynaklanan yaygın bir durumdur ve bu veri seti, potansiyel bir kalp hastalığını tahmin etmek için kullanılabilecek 11 özelliği içermektedir.**

Kardiyovasküler hastalığı olan veya hipertansiyon, diyabet, hiperlipidemi gibi bir veya daha fazla risk faktörünün varlığı nedeniyle yüksek kardiyovasküler risk taşıyan kişilerin erken teşhis ve yönetim ihtiyacı bulunmaktadır. Bu noktada, makine öğrenimi modeli büyük yardım sağlayabilir. Bu makalede, kalp krizi riskinin tahmin edilmesi ile ilgili önemli bir konu ele alınmıştır.

I. Giriş

Kardiyovasküler hastalıklar (KVH), dünya genelindeki ölüm nedenleri arasında önde gelen bir tehdittir. Her yıl milyonlarca insan KVH nedeniyle hayatını kaybetmektedir. Bu hastalıkların erken teşhis edilmesi ve etkili bir şekilde yönetilmesi, hastaların yaşam kalitesini artırmak ve yaşam sürelerini uzatmak için kritik bir öneme sahiptir. Son yıllarda makine öğrenimi ve veri analitiği gibi teknolojiler, tıp alanında bu tür sorunların çözümünde önemli bir rol oynamaktadır.

Bu çalışmada, kalp krizi riskinin tahmin edilmesi amacıyla bir makine öğrenimi yaklaşımı ele alınmıştır. Veri setinde yer alan çeşitli klinik özellikler kullanılarak, bireylerin kalp hastalığı riski tahmin edilmeye çalışılmıştır. Veri seti, yaş, cinsiyet, göğüs ağrısı tipi, dinlenme kan basıncı, serum kolesterol seviyesi, açlık kan şekeri, dinlenme elektrokardiyogram sonuçları, maksimum kalp atış hızı, egzersize bağlı anjin varlığı, ST segmenti depresyonu, ST segmenti eğimi gibi 11 farklı özelliği içermektedir. Ayrıca, her bireyin kalp hastalığı varlığını belirten bir çıktı sınıfı bulunmaktadır.

Bu çalışmanın amacı, bu veri setini detaylı bir şekilde incelemek ve makine öğrenimi algoritmalarını kullanarak kalp krizi riski tahmini yapmaktır. Veri keşfi ve analizi (Exploratory Data Analysis - EDA) ile başlayarak, farklı özellikler arasındaki ilişkileri anlamaya çalışacağız. Ardından, eğitim ve değerlendirme aşamalarında çeşitli makine öğrenimi modellerini kullanarak kalp hastalığı tahmini yapacağız.

Bu çalışma, kardiyovasküler hastalıkların erken teşhisine yönelik bir adım olarak, hastaların sağlık durumlarını daha iyi yönetmelerine ve tedaviye erken başlamalarına yardımcı olabilecek önemli bilgiler sunmayı amaçlamaktadır.

II. Veri Özellikleri

Veri setinde bulunan özellikler aşağıda belirtilmiştir:

- **Age**: Hasta yaşını ifade eder (yıl).
- **Sex**: Hasta cinhttps://www.overleaf.com/project/64e208455746f36fd0
gösterir (**M**: Erkek, **F**: Kadın).
- **ChestPainType**: Göğüs ağrısı tipini belirtir (**TA**: Tipik Angina, **ATA**: Atipik Angina, **NAP**: Non-Anginal Ağrı, **ASY**: Asemptomatik).
- **RestingBP**: Dinlenme kan basıncını gösterir (mm Hg).
- **Cholesterol**: Serum kolesterol seviyesini ifade eder (mm/dl).
- **FastingBS**: Açlık kan şekerini belirtir (1: Eğer Açlık Kan Şekeri ¿ 120 mg/dl, 0: Aksi durumda).
- **RestingECG**: Dinlenme elektrokardiyogram sonuçlarını gösterir (**Normal**: Normal, **ST**: ST-T dalga anormalliği (T dalga inversiyonları ve/veya ST yükselmesi veya depresyonu ¿ 0.05 mV), **LVH**: Estes kriterlerine göre muhtemel veya kesin sol ventriküler hipertrofi).
- **MaxHR**: Maksimum kalp atış hızını ifade eder (60 ile 202 arasında sayısal değer).
- **ExerciseAngina**: Egzersiz kaynaklı anjini varlığını belirtir (**Y**: Evet, **N**: Hayır).
- **Oldpeak**: Eski ST segment depresyonunu ifade eder (depresyon olarak ölçülen sayısal değer).
- **ST_Slope**: Tepe egzersiz ST segmentinin eğimini gösterir (**Up**: Yukarı doğru eğim, **Flat**: Düz, **Down**: Aşağı doğru eğim).
- **HeartDisease**: Çıktı sınıfını ifade eder (**1**: Kalp hastalığı, **0**: Normal durum).

III. Gerekli Kütüphaneler ve Veri Setinin Kaynağı

Gerekli kütüphaneler şunlardır:

- `pandas`
- `numpy`
- `matplotlib.pyplot`
- `warnings`
- `plotly.express`
- `plotly.graph_objects`
- `make_subplots` (plotly.subplots)
- `seaborn`
- `LabelEncoder` (sklearn.preprocessing)
- `train_test_split` (sklearn.model_selection)
- `LogisticRegression` (sklearn.linear_model)

- `SVC` (sklearn.svm)
- `RandomForestClassifier` (sklearn.ensemble)
- `GradientBoostingClassifier` (sklearn.ensemble)
- `KNeighborsClassifier` (sklearn.neighbors)
- `DecisionTreeClassifier` (sklearn.tree)
- `confusion_matrix` (sklearn.metrics)

Veri seti Kaggle platformundan "heart.csv" ismi ile indirilmiştir.

IV. Veri Setinin Genel Özellikleri

Veri setinin boyutu 918 satır ve 12 sütundan oluşmaktadır. Bu sütunlar aşağıdaki özellikleri içermektedir:

TABLE I
VERI ÖZELLIKLERI

Özellik	Tip
Age	int64
Sex	object
CP Type	object
RestingBP	int64
Cholesterol	int64
FBS	int64
RestingECG	object
MaxHR	int64
ExerciseAngina	object
Oldpeak	float64
ST Slope	object
HeartDisease	int64

Veri setinde yer alan özelliklerin istatistiksel özet bilgileri aşağıda belirtilmiştir:

TABLE II
İSTATISTIKSEL ÖZET

Özellik	Ortalama	Standart Sapma	Maksimum
Age	53.51	9.43	77
RestingBP	132.40	18.51	200
Cholesterol	198.80	109.38	603
FBS	0.23	0.42	1
MaxHR	136.81	25.46	202
Oldpeak	0.89	1.07	6.20
HeartDisease	0.55	0.50	1

Categorical özelliklerin dağılımı aşağıdaki gibi:

TABLE III
KATEGORIK ÖZELLIK DAĞILIMI

Özellik	En Sık Görülen Değer	Frekans
Sex	M	725
CP Type	ASY	496
RestingECG	Normal	552
ExerciseAngina	N	547
ST Slope	Flat	460

V. Keşifsel Veri Analizi

A. Özniteliklerin Farklı Tür Sayıları

Veri setinde yer alan özelliklerin farklı tür sayıları belirtilmiştir.

- **Age**: 50 farklı tür
- **Sex**: 2 farklı tür
- **ChestPainType**: 4 farklı tür
- **RestingBP**: 67 farklı tür
- **Cholesterol**: 222 farklı tür
- **FastingBS**: 2 farklı tür
- **RestingECG**: 3 farklı tür
- **MaxHR**: 119 farklı tür
- **ExerciseAngina**: 2 farklı tür
- **Oldpeak**: 53 farklı tür
- **ST_Slope**: 3 farklı tür
- **HeartDisease**: 2 farklı tür

B. Çift Değişkenli İlişkiler (PairPlot)

Çift değişkenli ilişkileri incelemek için `seaborn` kütüphanesinin `pairplot` fonksiyonu kullanılarak görsel oluşturulmuştur. Bu görsel, veri setindeki değişkenler arasındaki ilişkileri anlamak, bu ilişkileri görselleştirmek ve dağılımları açısından fikir edinmek için oldukça faydalıdır.

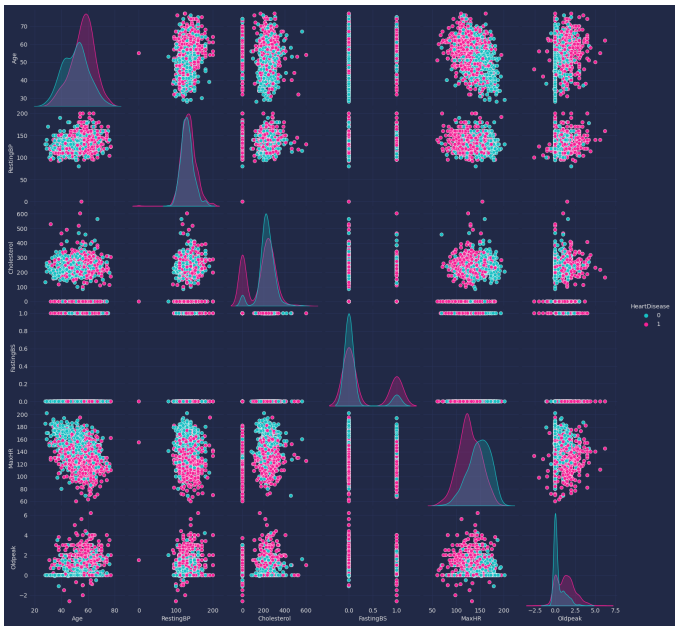

Fig. 1. Çift Değişkenli İlişkileri Gösteren Pairplot

Şekil 1'de gösterilen pairplot görseli, veri setindeki çift değişkenli ilişkileri açıkça göstermektedir. Her bir özelliğin diğer özelliklerle nasıl ilişkilendiğini görmemize yardımcı olur.

C. Cinsiyete Göre Hastalık Dağılımı

Cinsiyetlere göre hastalık dağılımını gösteren grafiği Şekil 2 de görebilirsiniz.

Cinsiyetlere göre hastalık dağılımı grafiği incelendiğinde, hastalık tanısı konulan bireylerin çoğunluğunun erkeklerden oluştuğu görülmektedir. Erkeklerde hastalık sayısı sağlıklı birey sayısını aşarken, kadınlarda hastalık tanısı konulan birey sayısı sağlıklı kadın sayısını geçememektedir. Bu sonuçlar, cinsiyetin kardiyovasküler hastalıklar üzerindeki etkisinin incelenmesi açısından önemlidir.

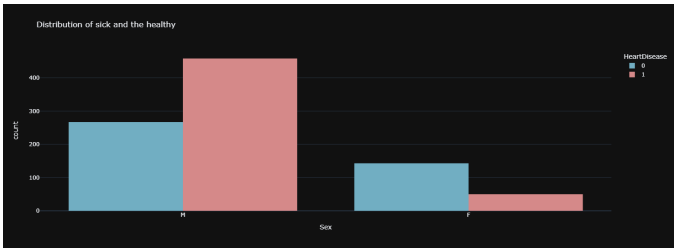

Fig. 2. Cinsiyetlere Göre Hastalık Dağılımı

Bu dağılım, cinsiyet ve hastalık durumu arasındaki ilişkiyi görsel olarak anlamamıza yardımcı olmaktadır. Özellikle erkeklerde hastalık riskinin kadınlara göre daha yüksek olduğu gözlemlenmektedir.

D. En Yaygın Hastalığı Etkileyen Göğüs Ağrısı Türü

Hastalığı etkileyen en yaygın göğüs ağrısı türüne ait dağılımı gösteren grafiği Şekil 3 de inceleyebilirsiniz.

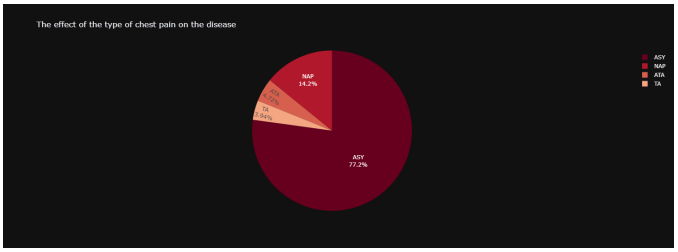

Fig. 3. En Yaygın Hastalığı Etkileyen Göğüs Ağrısı Türü

Şekil 3'de görüldüğü gibi, hastalığı etkileyen en yaygın göğüs ağrısı türü "ASY" (Asymptomatic) olarak belirlenmiştir. Bu sonuçlar, hastalığın belirli semptomlarına sahip olan bireylerin hastalığa yakalanma olasılığının daha yüksek olduğunu göstermektedir.

E. Egzersiz Kaynaklı Anjina Etkisi

Egzersiz kaynaklı anjina (Exercise Angina) durumunun hastalık üzerindeki etkisini gösteren ilişkiyi açıklamak amacıyla aşağıdaki tabloyu inceleyebilirsiniz:

TABLE IV
EGZERSIZ KAYNAKLI ANJINA VE HASTALIK İLIŞKISI

ExerciseAngina	HeartDisease = 0	HeartDisease = 1
N	355	192
Y	55	316

Tablo, egzersiz kaynaklı anjina durumunun hastalık durumu üzerindeki etkisini göstermektedir. Tabloya göre, egzersiz kaynaklı anjina yaşayan bireylerin hastalık durumu ile ilişki olduğu görülmektedir. Egzersiz kaynaklı anjina yaşayan bireylerde hastalık durumu (HeartDisease) daha yüksekken, egzersiz kaynaklı anjina yaşamayan bireylerde hastalık durumu daha düşüktür.

F. Açlık Kan Şekeri ve Hastalık İlişkisi

Veri setindeki bireylerin çoğunluğunun açlık kan şekeri seviyesi 120 mg/dl altındadır. Ancak, açlık kan şekeri seviyesi 120 mg/dl üzerinde olan bireylerde hastalık durumu daha yaygın görülmektedir.

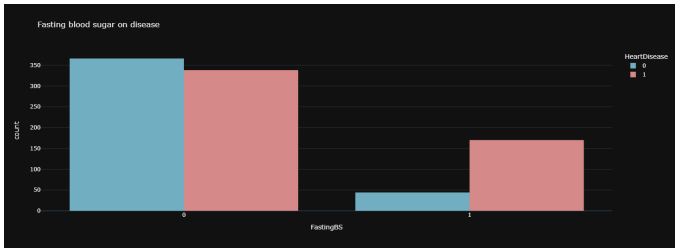

Fig. 4. Açlık Kan Şekeri ve Hastalık İlişkisi

Şekil 4'de gösterilen grafikte, açlık kan şekeri seviyesi 120 mg/dl altında olan bireylerde hastalık durumu ile sağlıklı durumu neredeyse eşittir. Ancak, açlık kan şekeri seviyesi 120 mg/dl üzerinde olan bireylerde hastalık durumu sağlıklı durumdan daha yaygındır.

VI. VERI ÖNIŞLEMESI (DATA PREPROCESSING)

Veri önişlemesinin temel amacı, eldeki verinin istenilen sonuçları almak üzere modele en uygun hale getirilmesini sağlamaktır. Bu aşamada yapılan işlemler arasında veri temizleme, eksik değerleri doldurma, özellik seçimi ve ölçeklendirme yer alabilir. Veri önişlemesi, sonuçta elde edilen Makine Öğrenimi modelinin kalitesini büyük ölçüde etkileyen kritik bir adımdır.

Veri önişlemesi, ham verinin düzensizliğini ve karmaşıklığını azaltarak, veri analizi ve model eğitimi sırasında daha güvenilir sonuçlar elde etmeyi amaçlar. Bu sayede, daha iyi öğrenme sonuçları ve daha iyi genelleme yetenekleri elde edilir.

A. Eksik Veri

Veri seti incelendiğinde eksik veri (null değer) içeren herhangi bir öğenin bulunmadığı görülmektedir. Eksik veriler, veri setindeki özelliklerin veya gözlemlerin eksik veya boş olduğu durumları ifade eder. Bu tür eksiklikler, veri analizi ve modelleme süreçlerini olumsuz etkileyebilir.

B. Yinelenen Veriler

Veri seti incelendiğinde, orijinal veri sayısının aynı olduğu görülmektedir. Bu durum, veri setinde herhangi bir yinelenen (duplicate) verinin bulunmadığını gösterir. Yinelenen veriler, aynı veya çok benzer özelliklere sahip gözlemlerin birden fazla kez veri setinde yer alması durumudur.

Yinelenen veriler, analiz ve modelleme süreçlerini yanıltabilir ve sonuçların güvenilirliğini azaltabilir. Bu nedenle, veri setinde yinelenen verilerin bulunmaması, analizlerin ve öğrenme algoritmalarının daha güvenilir sonuçlar üretmesine olanak tanır.

C. Yaşı Kategorilere Ayırmak

Yaş verisinin kategorilere ayrılmasının kolaylık sağlayabileceği görülmektedir. Yaşı kategorilere ayırmanın avantajlarından biri, verinin daha anlaşılır ve yönetilebilir hale getirilmesidir. Ayrıca, belirli yaş aralıklarındaki hastalık oranlarını incelemek ve karşılaştırmak daha kolay olabilir.

Aşağıda, yaşın kategorilere ayrılması sonucu elde edilen veri ve hastalık oranlarını içeren bir tablo verilmiştir:

TABLE V
YAŞ KATEGORILERINE GÖRE HASTALIK ORANLARI

Yaş Kategorisi	Hasta	Sağlıklı
3	(64.75, 77.0]	0.699029
2	(52.5, 64.75]	0.647196
1	(40.25, 52.5]	0.431973
0	(27.951, 40.25]	0.344086

Bu yöntem, yaş verisini daha anlamlı hale getirmek ve hastalık riskini belirli yaş aralıklarına göre analiz etmek amacıyla kullanılabilir.

D. Kolesterol Sütunu İçin İşlemler

Yüksek kolesterol seviyesi, kan damarlarınızda yağlı birikintilerin oluşmasına neden olabilir. Bu birikintiler zamanla büyüyerek damarlarınızdan yeterli kan akışını geçmesini zorlaştırabilir. Bu birikintiler bazen aniden kırılabilir ve pıhtı oluşturarak kalp krizine yol açabilir. Veriyi eşit olarak üç bölüme bölmek hastalık oranlarını incelemek ve karşılaştırmak açısından kolay olabilir.

Veriyi ve bu işlem sonucunda elde edilen sonuçları aşağıdaki tabloda bulabilirsiniz:

TABLE VI
KOLESTEROL KATEGORILERINE GÖRE HASTALIK ORANLARI

Kolesterol Kategorisi	Hasta	Sağlıklı
(-0.603, 201.0]	0.646341	0.353659
(201.0, 402.0]	0.499133	0.500867
(402.0, 603.0]	0.615385	0.384615

Bu tablo, kolesterol seviyelerinin kategorilere ayrılması sonucu elde edilen hastalık oranlarını göstermektedir. Kolesterol seviyelerine göre hastalık durumu arasındaki ilişkinin incelenmesi, kolesterol seviyelerinin hastalık riski üzerindeki etkisini daha iyi anlamamıza yardımcı olabilir.

E. MaxHR Sütunu İçin İşlemler

MaxHR sütunu, 60 ila 202 arasında bir sayısal değere sahip olmalıdır. Veriyi eşit olarak üç bölüme bölmek suretiyle, her bir bölümdeki ortalama hastalık durumunu tanımlamak kolaylık sağlayabilir.

TABLE VII
MAXHR KATEGORILERINE GÖRE HASTALIK ORANLARI

MaxHR Kategorisi	Hastalık Durumu = 0	Hastalık Durumu = 1
(59.858, 107.333]	0.812500	0.187500
(107.333, 154.667]	0.618538	0.381462
(154.667, 202.0]	0.285714	0.714286

Tabloda görüldüğü gibi, MaxHR değeri düşük olan kategorilerde hastalık durumunun daha yüksek olduğu gözlemlenmektedir. Bu, düşük fiziksel aktivite seviyelerinin ve genel sağlık durumunun hastalık riskini artırabileceğini göstermektedir. Özellikle 59.858 ile 107.333 arasındaki MaxHR değerlerine sahip bireylerde hastalık oranı yüksektir.

F. Korelasyon Matrisi

Veri ön işleme adımları tamamlandıktan sonra korelasyon matrisi çizdirilmiştir. Korelasyon matrisi, veri özellikleri arasındaki ilişkileri gösterir.

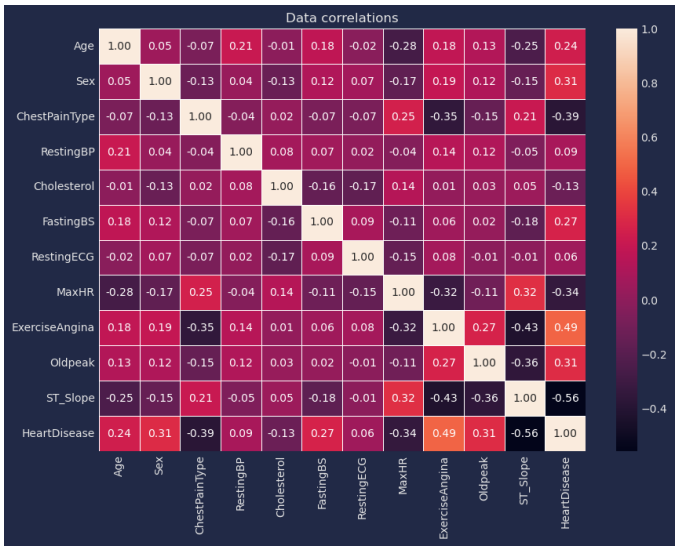

Fig. 5. Veri Özellikleri Korelasyon Matrisi

Korelasyon matrisi görseli, özellikler arasındaki ilişkileri renk skalası ile görselleştirir. Her bir hücredeki renk tonu, iki özellik arasındaki korelasyonun gücünü yansıtır. Pozitif değerler arasındaki ilişki maviye doğru açılırken, negatif değerler kırmızıya doğru açılır.

En etkili 6 özellik aşağıda listelenmiştir:

- Heart Disease
- Exercise Angina
- Oldpeak
- Sex
- Fasting Blood Sugar
- Age

Bu özellikler, veri analizinde en yüksek etkiye sahip oldukları düşünülen özelliklerdir.

Tüm özelliklerin bulunduğu korelasyon matrisinden en etkili 6 özellik alınıp korelasyon matrisi çizdirilmiştir.

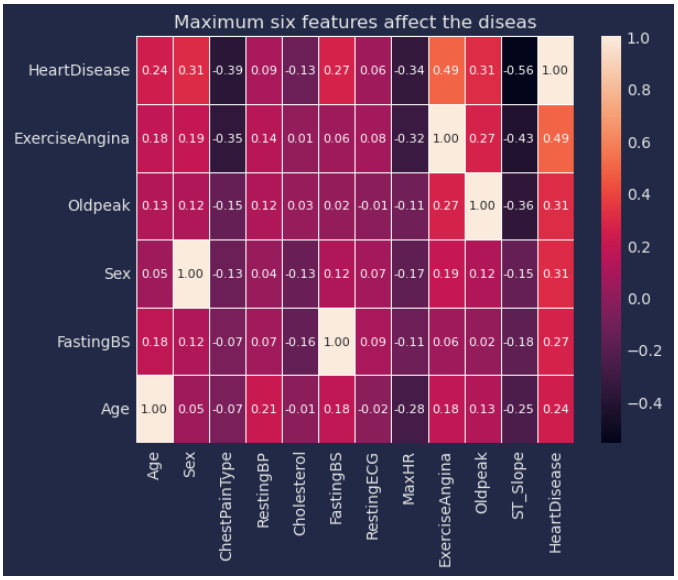

Fig. 6. En Etkili 6 Özellik Korelasyon Matrisi

En etkili 6 özellik seçildikten sonra, bu özellikler arasındaki ilişkilerin daha detaylı bir şekilde incelenmesi amaçlanmıştır. Bu matris, sadece seçilen özelliklerin korelasyonunu gösterir ve daha odaklı bir analiz yapmamıza olanak tanır.

VII. Model Eğitimi

İlk olarak, veri seti tüm modeller için %30'u test verisi olacak şekilde ayrılmıştır.

A. Logistic Regression

1) Logistic Regression Hakkında Kısa Bilgi: Logistic Regression (Lojistik Regresyon), sınıflandırma problemlerinde kullanılan bir istatistiksel yöntemdir. Temel olarak, bağımsız değişkenler ile bağımlı değişken (sınıf etiketi) arasındaki ilişkiyi modellemek için kullanılır. Lojistik Regresyon, bir çıktıyı (örneğin, hasta veya sağlıklı) tahmin etmek amacıyla girdi verileri üzerinde bir olasılık dağılımını tahmin eder. Bu tahmin edilen olasılık değeri daha sonra belirli bir eşik değeri ile karşılaştırılarak son sınıflandırma yapılır.

2) Neden Bu Probleme Uygun?: Logistic Regression, kalp hastalığı gibi sınıflandırma problemleri için uygun bir seçenektir. Bu problemin özelliği, veri özellikleri (yaş, cinsiyet, kan basıncı, vb.) ile sonucu (hasta veya sağlıklı) ilişkilendirmektir. Logistic Regression, bu tür ilişkileri modellemek için idealdir çünkü:

- Lojistik Regresyon, doğrusal ve doğrusal olmayan ilişkileri modelleyebilme yeteneğine sahiptir.
- Lojistik fonksiyon sayesinde tahmin edilen olasılıklar arasında sınırlama yapabilir. Bu, kesin bir sınıflandırma yapmak yerine olasılık değerlerini döndürmesine olanak tanır.
- Veri setindeki özelliklerin etkisi kolayca yorumlanabilir ve modelin nasıl kararlar aldığını anlamak mümkündür.
- Genel olarak hızlı ve hafif bir modeldir, bu da büyük veri setlerinde bile etkili sonuçlar üretebilmesine yardımcı olur.

Bu nedenlerle, Logistic Regression yöntemi, kalp hastalığı gibi sınıflandırma problemleri için uygun bir seçenektir ve genellikle başlangıç noktası olarak tercih edilir.

3) Model Performansı: Eğitilen Logistic Regression modelinin performans sonuçları aşağıdaki gibidir:

- Eğitim Başarısı (Training Score): 0.846
- Test Başarısı (Testing Score): 0.870

Modelin eğitim başarısı %84.6, test başarısı ise %87.0 olarak elde edilmiştir. Modelin test verisinde iyi bir performans göstermesi olumlu bir sonuç olarak değerlendirilebilir.

B. Support Vector Classifier (SVC)

Sıradaki model olarak, veri seti Support Vector Classifier (Destek Vektör Sınıflandırıcı) modeli ile eğitilmiştir.

1) Support Vector Classifier Hakkında Kısa Bilgi: Support Vector Classifier (Destek Vektör Sınıflandırıcı), sınıflandırma problemlerinde kullanılan bir makine öğrenimi algoritmasıdır. Temel olarak, sınırlayıcı hiper düzlemleri kullanarak verileri farklı sınıflara bölmeye çalışır. Destek vektörleri, sınırlayıcı hiper düzlemine en yakın noktalardır ve bu noktaların etrafında en fazla marj ile sınıfları ayırmaya çalışır.

2) Neden Bu Probleme Uygun?: Support Vector Classifier, kalp hastalığı gibi sınıflandırma problemleri için uygun bir seçenektir. Bu tür problemlerde, veri özellikleri arasındaki karmaşık ilişkileri öğrenmek ve sınıfları ayırmak önemlidir. SVC bu noktada uygun bir seçenektir çünkü:

- SVC, doğrusal olmayan ilişkileri modellemek için çekirdek fonksiyonlarını kullanabilir. Bu, verinin karmaşıklığını daha iyi yakalayabilmesine olanak tanır.
- Destek vektörleri kullanarak sınırlayıcı hiper düzlemi optimize ederken, genellemeye yardımcı olur ve overfitting'i engellemeye çalışır.
- Çok boyutlu verilerde iyi performans gösterebilir ve veri setindeki gürültülü veya anormallik içeren örnekleri daha iyi işleyebilir.
- Veri setindeki özelliklerin etkisi hakkında daha fazla bilgi sağlar ve modelin nasıl kararlar aldığını anlamak daha kolaydır.

Bu nedenlerle, Support Vector Classifier yöntemi, karmaşık ilişkileri içeren kalp hastalığı gibi sınıflandırma problemleri için uygun bir seçenektir.

3) Model Performansı: Eğitilen Support Vector Classifier modelinin performans sonuçları aşağıdaki gibidir:

- Eğitim Başarısı (Training Score): 0.871
- Test Başarısı (Testing Score): 0.862

Modelin eğitim başarısı %87.1, test başarısı ise %86.2 olarak elde edilmiştir. Logistic Regression gibi, bu model de test verisinde iyi bir performans sergilemektedir.

C. Random Forest Classifier (RFC)

Sıradaki model olarak, veri seti Random Forest Classifier (Rastgele Orman Sınıflandırıcı) modeli ile eğitilmiştir.

1) Random Forest Classifier Hakkında Kısa Bilgi: Random Forest Classifier, sınıflandırma problemlerinde kullanılan bir ensemble (topluluk) algoritmasıdır. Temel olarak, birden çok karar ağacını bir araya getirerek daha güçlü ve kararlı bir model oluşturur. Her bir ağacın tahminleri toplanarak son tahmin yapılır.

2) Neden Bu Probleme Uygun?: Random Forest Classifier, karmaşık veri setleri ve sınıflandırma problemleri için uygun bir seçenektir. Bu tür problemlerde, birden çok ağacın birleştirilmesi ile daha iyi sonuçlar elde edilebilir. RFC bu noktada uygun bir seçenektir çünkü:

- RFC, overfitting riskini azaltarak genelde daha iyi bir performans sunar.
- Farklı alt özellik kümesi ve alt örneklem kullanarak her ağacın oluşturulması varyansı azaltır.
- RFC, feature importance (özellik önemi) değerlerini sağlayarak hangi özelliklerin modelin tahminlerine en çok katkı sağladığını anlamak için kullanılabilir.

Bu nedenlerle, Random Forest Classifier yöntemi, karmaşık veri setleri ve sınıflandırma problemleri için tercih edilen bir seçenektir.

3) Model Performansı: Eğitilen Random Forest Classifier modelinin performans sonuçları aşağıdaki gibidir:

- Eğitim Başarısı (Training Score): 0.970
- Test Başarısı (Testing Score): 0.848

Modelin eğitim başarısı %97.0, test başarısı ise %84.8 olarak elde edilmiştir. Eğitim verisinde yüksek bir başarım gösterirken, test verisinde bir miktar performans düşüşü gözlemlenmiştir. Bunun sebepleri şunlar olabilir:

- Yüksek Karmaşıklık: Model, çok sayıda karar ağacını birleştiren bir ensemble modeli olduğundan, yüksek karmaşıklık seviyesine sahip olabilir. Bu da aşırı uyum riskini artırabilir.
- Veri Dengesizliği: Test verisi, eğitim verisi ile aynı özelliklere sahip olmayabilir. Veri dengesizliği, modelin doğru sınıflandırma yapmasını zorlaştırabilir.
- Özellik Seçimi: Veri ön işleme ve özellik seçimi doğru yapılmamışsa, modelin veriye uyumu düşük olabilir. Önemli özelliklerin atlanması veya gereksiz özelliklerin dahil edilmesi sonuçları etkileyebilir.
- Hyperparameter Ayarları: Modelin hiperparametreleri doğru şekilde ayarlanmamışsa, performans düşüklüğü gözlemlenebilir. Örneğin, ağaç derinliği, alt özellik kümesi sayısı gibi parametrelerin etkisi önemlidir.

Düşük performansın sebepleri üzerinde detaylı analiz yaparak modelin iyileştirilmesi sağlanabilir. Hyperparameter ayarları, özellik seçimi ve modelin karmaşıklığının gözden geçirilmesi, genel performansı artırabilir.

D. Gradient Boosting Classifier (GBC)

Sıradaki model olarak, veri seti Gradient Boosting Classifier (GBC) modeli ile eğitilmiştir.

1) Gradient Boosting Classifier Hakkında Kısa Bilgi: Gradient Boosting Classifier, ensemble öğrenme yöntemleri arasında yer alan ve ardışık ağırlıklandırılmış modellerin birleştirilmesi ile oluşturulan bir algoritmadır. Zayıf tahmincilerin (örneğin, karar ağaçları) bir araya getirilmesi ile daha güçlü bir tahminci elde edilir.

2) Neden Bu Probleme Uygun?: Gradient Boosting Classifier, karmaşık veri yapılarını ve ilişkilerini öğrenmek için uygun bir seçenektir. Bu tür problemlerde:

- GBC, zayıf tahmincilerin güçlü bir şekilde birleştirilmesi ile modeli güçlendirir. Bu, veri setindeki karmaşık ilişkileri daha iyi anlamasına yardımcı olur.
- Daha önceki tahmincilerin hatalarını düzeltmek için yeni tahminciler eklemesi yapılır, bu da modelin performansını artırabilir.
- GBC, doğrusal olmayan ilişkileri yakalama yeteneği ile veri setindeki gürültüye karşı daha dirençlidir.

Bu nedenlerle, Gradient Boosting Classifier modeli, karmaşık ilişkileri içeren kalp hastalığı gibi sınıflandırma problemleri için uygun bir seçenektir.

3) Model Performansı: Eğitilen Gradient Boosting Classifier modelinin performans sonuçları aşağıdaki gibidir:

- Eğitim Başarısı (Training Score): 0.903
- Test Başarısı (Testing Score): 0.870

Modelin eğitim başarısı %90.3, test başarısı ise %87.0 olarak elde edilmiştir. Bu model, test verisinde yüksek bir performans sergilemektedir.

E. K-Nearest Neighbors (KNN)

Sıradaki model olarak, veri seti K-Nearest Neighbors (KNN) modeli ile eğitilmiştir.

1) K-Nearest Neighbors Hakkında Kısa Bilgi: K-Nearest Neighbors (KNN), sınıflandırma ve regresyon problemlerinde kullanılan bir algoritmadır. Temel fikri, yeni bir veri noktasını çevresindeki komşu veri noktalarının çoğunluğuna göre sınıflandırmaktır.

2) Neden Bu Probleme Uygun?: K-Nearest Neighbors, benzer özelliklere sahip veri noktalarının benzer sonuçlar üretebileceği durumlar için uygun bir seçenektir. Bu tür problemlerde:

- KNN, veri setindeki yapıyı ve komşuluk ilişkilerini kullanarak sınıflandırma yapar.
- Veri noktalarının etrafındaki komşuluk bilgisi, modelin sınıflandırma yaparken dikkate aldığı kritik faktörlerden biridir.
- Veri setindeki anormallikleri veya düzensizlikleri ele alabilir ve esnek bir şekilde öğrenebilir.

Bu nedenlerle, K-Nearest Neighbors yöntemi, benzer özelliklere sahip veri noktalarının sınıflandırılması için kullanılabilecek bir seçenektir.

3) Model Performansı: Eğitilen K-Nearest Neighbors modelinin performans sonuçları aşağıdaki gibidir:

- Eğitim Başarısı (Training Score): 0.882
- Test Başarısı (Testing Score): 0.859

Modelin eğitim başarısı %88.2, test başarısı ise %85.9 olarak elde edilmiştir. Model, test verisinde de iyi bir performans sergilemektedir.

F. Decision Tree Classifier (DT)

Sıradaki model olarak, veri seti Decision Tree Classifier (DT) modeli ile eğitilmiştir.

1) Decision Tree Classifier Hakkında Kısa Bilgi: Decision Tree Classifier, veri kümesindeki özelliklerin değerlerine göre bir karar ağacı oluşturarak sınıflandırma yapar. Karar ağacı, veri kümesini belirli koşullara göre böler ve her bir bölünmüş kısma bir sınıf atar.

2) Neden Bu Probleme Uygun?: Decision Tree Classifier, veri kümesindeki özelliklerin etkisini ve ilişkilerini anlamak için kullanılır. Bu tür problemlerde:

- DT, veri kümesindeki karmaşık ilişkileri ve kararları anlamak için açık ve anlaşılır bir yapı sunar.
- Veri setindeki anormallikleri veya gürültüyü yakalamak için ağacın dallarını incelemek mümkündür.
- Veri setindeki bağımsız değişkenlerin etkilerini değerlendirmek için kullanışlıdır.

Bu nedenlerle, Decision Tree Classifier modeli, veri setindeki ilişkileri anlamak ve açıklamak için uygun bir seçenektir.

3) Model Performansı: Eğitilen Decision Tree Classifier modelinin performans sonuçları aşağıdaki gibidir:

- Eğitim Başarısı (Training Score): 0.970
- Test Başarısı (Testing Score): 0.804

Modelin eğitim başarısı %97.0, test başarısı ise %80.4 olarak elde edilmiştir. Model, eğitim verisine çok iyi uyum sağlamış gibi görünüyor ancak test verisinde düşük bir performans sergiliyor.

G. Model Değerlendirmesi

Altı farklı sınıflandırma modelinin performansı değerlendirildi. Her bir model için Confusion Matrix görselleri birleştirilmiş olarak aşağıda sunulmuştur:

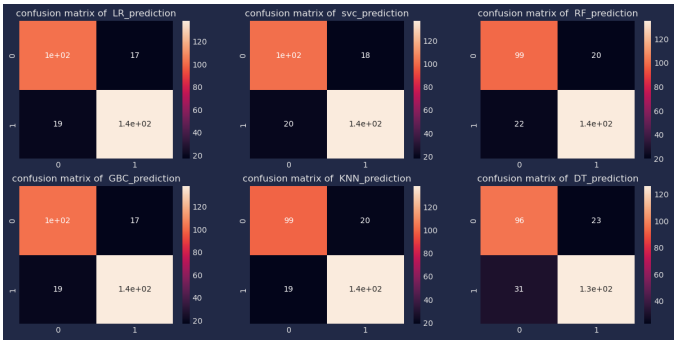

Fig. 7. Altı Farklı Modelin Birleştirilmiş Confusion Matrix

Her bir modelin Confusion Matrix görseli, modelin sınıflandırma performansını göstermektedir. True Positive (TP), True Negative (TN), False Positive (FP) ve False Negative (FN) değerleri ile modellerin doğru ve yanlış sınıflandırmalarını görebiliriz.

Modellerin performans sonuçları tabloda özetlenmiştir:

Model	Başarı Oranı (%)	Eğitim Süresi (s)
GradientBoostingClassifier	85.35	5.96
Support Vector Machines	84.93	2.41
Logistic Regression	84.54	0.64
Random Forest	83.76	12.48
KNN	83.09	1.84
Decision Tree	79.22	0.42

TABLE VIII
FARKLI MODELLERIN BAŞARI ORANLARI VE EĞITIM SÜRELERI

Tabloda görüldüğü üzere, farklı modeller arasında en yüksek başarı oranına sahip olanlar GradientBoostingClassifier ve Support Vector Machines modelleridir, sırasıyla %85.35 ve %84.93 başarı oranına sahiptir. Logistic rregression, KNN ve Random Forest modelleri de benzer şekilde yüksek başarı oranlarına sahiptir. Decision Tree modeli ise diğer modellere kıyasla daha düşük bir başarı oranına sahiptir.

Model değerlendirmesi, farklı algoritmaların performansını karşılaştırarak en iyi seçeneği belirlemeye yardımcı olur. Bu sonuçlar, problem alanına ve veri setine uygun olarak model seçimine yardımcı olabilir.

VIII. Özet

Bu çalışma, kalp hastalığı riskinin tahmin edilmesi amacıyla çeşitli klinik özelliklerin kullanıldığı bir makine öğrenimi yaklaşımını ele almaktadır. Kardiyovasküler hastalıkların erken teşhisinin yaşam kalitesini artırma ve yaşam süresini uzatma açısından kritik öneme sahip olduğu unutulmamıştır. Makine öğrenimi ve veri analitiği teknikleri, tıp alanında bu tür zorlu sorunların çözümünde etkili bir rol oynamaktadır.

Veri Keşfi ve Analizi

- 918 örnekten oluşan veri seti incelenmiş ve verinin genel özellikleri ile dağılımları gözden geçirilmiştir.
- Verinin kategorik ve numerik özellikleri ayrı ayrı analiz edilerek, veri keşfi ve analizi gerçekleştirilmiştir.
- Bu aşamada, özellikler arasındaki ilişkiler ve etkiler gözlemlenmiştir.

Veri Önişleme ve Model Eğitimi

- Numerik veriler farklı aralıklara bölünerek kategorik hâle getirilmiştir.
- Altı farklı sınıflandırma algoritması kullanılarak modeller eğitilmiştir.
- Model performansı sırasıyla aşağıdaki gibidir:
  - Logistic Regression ve Gradient Boosting Classifier: %86.96 doğruluk oranı ile en yüksek performansı sergileyen modellerdir.
  - Support Vector Machines: %86.23 doğruluk oranı ile iyi bir performans göstermiştir.
  - K-Nearest Neighbors: %85.87 doğruluk oranı ile diğer modellere göre biraz daha düşük performans göstermiştir.
  - Random Forest: %84.78 doğruluk oranı ile orta seviyede bir performans sergilemiştir.
  - Decision Tree: %80.43 doğruluk oranı ile en düşük performansa sahip modeldir.

Sonuç ve Amaç

Çalışma, 11 farklı özelliği kullanarak kalp hastalığı riskini tahmin etmeye odaklanmıştır. Bu özellikler arasında yaş, cinsiyet, göğüs ağrısı tipi, kan basıncı, serum kolesterol seviyesi, kan şekeri düzeyi, elektrokardiyogram sonuçları, maksimum kalp atış hızı, egzersize bağlı anjin varlığı, ST segmenti depresyonu ve ST segmenti eğimi bulunmaktadır. Her birey için kalp hastalığı varlığını belirten bir çıktı sınıfı mevcuttur.

Bu çalışmanın amacı, veri setini kapsamlı bir şekilde analiz ederek makine öğrenimi modellerini kullanarak kalp krizi riskini tahmin etmektir. Bu tahminler, hastaların sağlık durumlarını daha iyi yönetmeleri ve tedaviye erken başlamaları için değerli bilgiler sunabilir. Kardiyovasküler hastalıkların erken teşhisine yönelik bu adım, hastaların yaşam kalitesini artırmak ve yaşam süresini uzatmak için önemli bir adım olarak öne çıkmaktadır.

IX. Project Github Link

https://github.com/hasantann/YAP470.git

X. Project Youtube Presentation Link

https://youtu.be/4bvY28TFuYA

Kaynaklar

[1] Fedesoriano. (2021, Temmuz). Heart Failure Prediction Dataset. Kaggle. URL: https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction

[2] Mohamed, W. (2023, Temmuz). Heart Disease EDA with ML. Kaggle. URL: https://www.kaggle.com/code/mohamedwasef/heart-disease-eda-with-ml