MİPOWER: WOMEN EMPOWERMENT PROGRAMI BİTİRME PROJESİ

MAKİNE ÖĞRENİMİ İLE KALP RAHATSIZLIĞI TAHMİNİ

Şevval MERTOĞLU

Her yıl, kardiyovasküler hastalıklar dünya genelinde yaklaşık 17.9 milyon can kaybına neden olmakta ve bu ölüm oranları, meme kanseri, prostat kanseri ve bağırsak kanseri gibi diğer kanser türlerine kıyasla daha yüksek seviyede seyretmektedir. Özellikle kalp yetmezliği yaşayan bireylerde, sağ kalım oranlarını tahmin etmek son derece önem arz etmektedir. Sağ kalım tahmini, hastalığın erken aşamada teşhis edilmesine ve temel risk faktörlerinin belirlenmesine önemli bir katkı sağlar. Tıp alanındaki araştırmalar, son 25 yılda bilgisayar bilimlerinin hızlı ilerlemesi sayesinde, makine öğrenimi ve yapay zeka teknikleriyle birleştirilerek daha etkili hale gelmiştir. Bu çalışmada, kalp hastalıklarının tespiti için bir veri seti üzerinde çeşitli makine öğrenimi için K-nn, Destek vektör Makineleri ve Yapay Sinir Ağları kullanılmıştır ve elde edilen Doğruluk Değerleri karşılaştırılmıştır.Veri seti analizi, hangi verilerin kalp hastalıklarının belirtilerini doğru bir şekilde işaret edebileceğini belirlemiştir.

Anahtar Kelimeler: Makine öğrenmesi, Kalp hastalığı tespiti, Medikal veri analizi, Kalp yetmezliği, Sınıflandırma.

1.Problem Tanımı

Son yıllarda bilişim teknolojilerinin her sektörde yaygın bir şekilde kullanıldığı bilinmektedir. Özellikle sağlık sektöründe hastalıkların belirlenmesinde makine öğrenmesi tekniklerinin kullanımı her geçen gün artmaktadırKonu canlılar ve canlıların sağlığı olduğunda hastalıkların önceden belirlenerek erken teşhis edilmesi tedavide başarım oranını yükseltmekte ve hayat kurtarmaktadır. Erken teşhis aşamasında bilişim teknolojilerinin gözde alanlarından olan makine öğrenme teknikleri faydalı ve başarılı sonuçlar vermektedir.

Özellikler: Yaş, Cinsiyet, Göğüs Ağrısı Tipi, BP (kan basıncı), Kolesterol seviyesi, FBS'nin 120'nin üzerinde olması (açlık kan şekeri), EKG Sonuçları (elektrokardiyogram sonuçları), Max HR (maksimum kalp atış hızı), Egzersiz Anjina durumu, ST Depresyonu (EKG'de ST segmentinin depresyonu), ST'nin Eğimi (EKG'de ST segmentinin eğimi), Damar Sayısı Fluroskopi (floroskopide görülen damar sayısı) ve Talyum Stres testi sonuçlarıdır.Her değişkenin kalp hastalığı riski ile nasıl ilişkili olduğunu anlamak, bu veri setine dayanarak daha iyi tahminler yapmanıza yardımcı olacaktır. Örneğin, yaş, arteriyel tıkanıklıklarla ilgili olarak birinin kalp krizi veya felç geçirme olasılığını etkileyen bir değişkendir - bu nedenle, yaşın bağımsız bir faktör olup olmadığını veya diğer faktörlerin bireysel bir hastanın durumunda olasılıkları artırıp artıramayacağını not etmek önemlidir.

Bu çalışmada, kalp rahatsızlığını tespit etmek için Predicting Heart Disease Using Clinical Variables örnek veri seti üzerinde makine öğrenmesi tekniği uygulanmıştır. Random Forest Classifier Algoritması kullanılmış ve kalp rahatsızlığı olan bireyler tespit edilmeye çalışılmıştır.

2.Literatür Taraması

Literatür taraması kapsamında Kalp rahatsızlığı alanında makine öğrenme algoritma çalışmaları aşağıda yer alan başlıklar halinde verilmiştir:

1.Genetik Algoritma Yaklaşımıyla Öznitelik Seçimi Kullanılarak Makine Öğrenmesi Algoritmaları ile Kalp Hastalığı Tahmini

[1]Vatansever B. & Aydın H.(2021), Araştırmalarında elde ettikleri sonuçlarda kalp hastalığı tahminin GA yaklaşımı ile öznitelik seçimi yapılması durumunda daha yüksek doğruluk oranının elde edildiği görülmüştür. Çalışmalarının GA ile öznitelik seçimi yapılarak MÖ ile kalp hastalığınınım tahmin edilmesinde katkı sağlamışlardır. Çalışmalarında MÖ algoritmalarından K-En Yakın Komşu (K-EYK), Lojistik Regresyon (LR), Karar Ağacı (KA), Rastgele Orman (RO), Naive Bayes (NB) ve Destek Vektör Makinesi (DVM) algoritmaları ile 3 (üç) farklı grupta toplamda 28 (yirmi sekiz) deney gerçekleştirmişlerdir.

2. Kalp yetmezliği riskinin makine öğrenmesi yöntemleri ile analiz edilmesi

[2] Bilekyiğit, S. (2022),Bu çalışmasında açık erişime sahip UCI ve Kaggle veritabanından alınan Kalp Yetmezliği, Cleveland kalp hastalığı, Statlog kalp hastalığı ve Framingham kalp hastalığı veri setleri kullanmıştır. KNN, Karar Ağaçları, Navie Bayes, Destek Vektör Makineleri, Rastgele Orman ve Lojistik Regresyon makine öğrenmesi algoritmaları 5 çapraz doğrulama yöntemi kullanılarak veri setlerine uygulanmıştır.

3.Kardiyovasküler Hastalık Tahmininde Makine Öğrenmesi Sınıflandırma Algoritmalarının Karşılaştırılması

[3]Kaba G. & Bağdatlı Kalkan (2022), Çalışmalarında Kardiyovasküler Hastalığın erken teşhisine katkı sağlamak için makine öğrenmesi algoritmaları ile çalışmada kullanılan veriler üzerinde en başarılı sınıflandırma tahminini yapan algoritmaya ulaşmayı hedeflenmiştir.

Naive Bayes, Lojistik Regresyon, Rastgele Orman, K-En Yakın Komşu ve Destek Vektör Makineleri olmak üzere beş farklı makine öğrenmesi yöntemi kullanılarak performansları karşılaştırılmıştır. En başarılı performansı veren yöntem tespit edilmiştir.

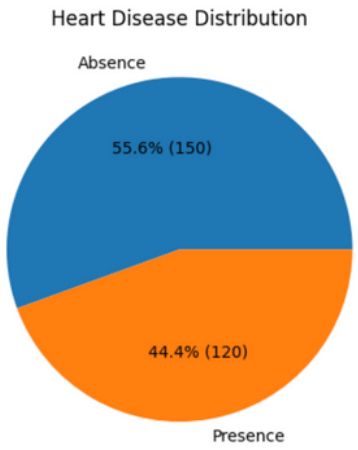
4.Makine Öğrenimi Modelleri Kullanılarak Kırılganlık Derecesi Tahmini Prediction of Frailty Grade Using Machine Learning Models

[4] Erdaş, Ç. & Ölçer D. (2022),. Bu çalışmalarında, kalp yetmezliği hastalığına sahip olan hastaların mortalite hayatta kalma durumlarını tahmin etmeye yönelik makine öğrenme tabanlı bir sistem önerilmektedir. Böylelikle mortalite ihtimali olan kişiler tespit edilerek, daha etkili ve yakından takip ile hastaların hayatta kalma ihtimalleri artırılması amaçlamıştır.

5.Kalp Yetmezliği Hastalarının Sağ Kalımlarının Sınıflandırma Algoritmaları ile Tahmin Edilmesi

[5] AKTAŞ POTUR, E., & ERGİNEL, N. (2021),Bu çalışmada kalp yetmezliği hastalarının sağ kalımlarının tahmin edilmesi amacıyla Naive Bayes, lojistik regresyon, çok katmanlı algılayıcı, destek vektör makineleri ve J48 karar ağacı sınıflandırma yöntemleri WEKA'da bulunan InfoGainAttributeEval, CfsSubsetEval ve ReliefFAttributeEval öznitelik seçim yöntemleri kullanılarak değerlendirme ölçütleri açısından karşılaştırılmıştır. Değerlendirme ölçütü olarak doğru sınıflandırma oranı, F-ölçütü ve Kappa istatistiği metrikleri kullanılmıştır. En yüksek sınıflandırma başarısına sahip sınıflandırıcı %90 doğru sınıflandırma oranı ile çok katmanlı algılayıcı olmuştur.

3.Veri Toplama ve Ön İşleme



Şekil 1. Sağlıklı ve Kalp Hastası Birey Dağılımı

Bu çalışmamızda, Predicting Heart Disease Using Clinical Variables(Kaggle, 2019), Veri Setini kullanarak projemizi Uyguladık. Veri analizi için Python kodu ve website oluşturmak için FLASK freamwork'ünü kullandık.

Veri setimiz içinde 13 farklı özellik ve Kalp Rahatsızlığı riskini Belirten Özellik de dahil olmak üzere toplam 14 sütun yer almaktadır.(Index sütunuyla birlikte 15). Toplam 270 birey sayısı bulunmaktadır.

Veri seti içerisinde bulunan hasta verilerine ait toplam hasta ve sağlıklı birey sayısı Şekil 1'de gösterilmiştir. Şekil 1'de görüldüğü üzere toplam 270 birey içerisinden 120 hastalık belirtisi gösteren birey, 150 kişi ise sağlıklı birey olarak sınıflandırılmaktadır.

Verilerin Analiz İçin Hazırlanması

Veri setindeki değerler incelenmiş ve eksik değer bulunmamıştır. Aykırı değerler için ise bir değişiklik yapılmamıştır.

Veriler kategorik ve numerik değişkenler olarak ayrılmıştır. Başlangıçta doğru ayrılmayan değişkenler yeniden düzenlenmiştir. Örneğin, yaş değişkeni yaş aralıklarına göre gruplandırılmıştır:

```
# Yaş aralıklarını tanımlama
bins = [0, 18, 30, 50, 60, 100]
labels = ['0-18', '18-30', '30-50', '50-60','60+',]
```

```
# Yaş sütununu kategorik hale getirir
df['Age'] = pd.cut(df['Age'], bins=bins, labels=labels, right=False)
```

ST depresyon değeri de benzer şekilde kategorilere ayrılmıştır:

```
st_depression_bins = [0, 0.5, 1, 2, 100]  # 0'dan 100'e kadar olan değerler için
st_depression_labels = ['Normal (0.0-0.5 mm)','Orta (0.5-1 mm)', 'Yüksek (1-2 mm)', 'Ciddi (>2 mm)']
```

- Target kolonu için Encoding (Label Encoding) uygulanmıştır.
- Aykırı değeleri değiştirmediğimiz için RobustScaler ölçeklendirme yöntemini uyguladık.
- Kategorik Değerler için One-Hot Encoding uyguladık.

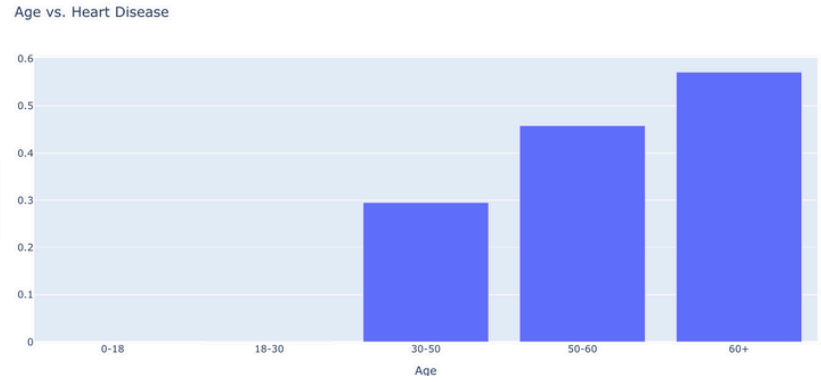
4.Veri Analizi ve Görselleştirme

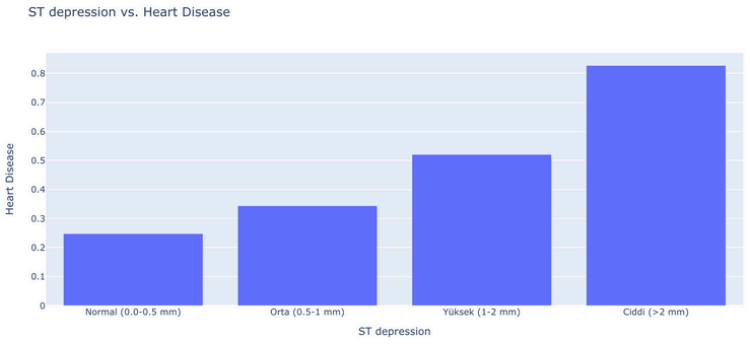
4.1 Temel İstatistiksel Analizler

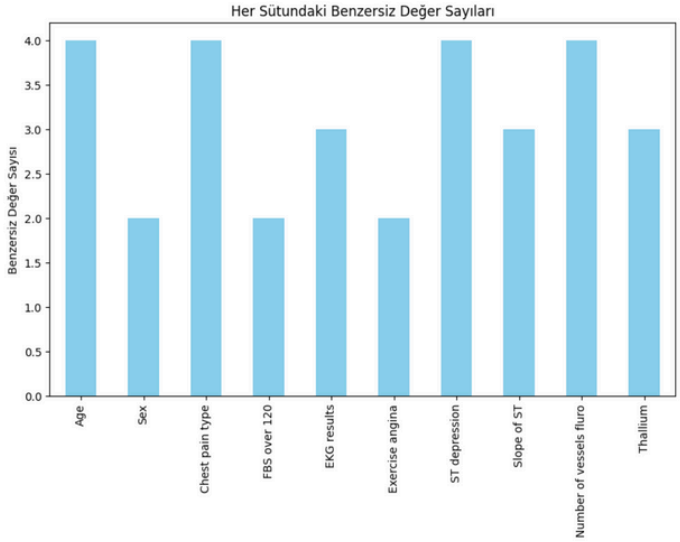
Veri setinin temel istatistiksel özellikleri incelenmiştir. Yaş, cinsiyet, kan basıncı... gibi özelliklerin ortalama, medyan ve dağılım değerleri hesaplanmıştır.

4.2 Verinin Görselleştirilmesi

Veriler, çeşitli grafikler ve görselleştirme teknikleri kullanılarak analiz edilmiştir. Önemli bulgular ve ilişkiler grafiklerle ifade edilmiştir. Kategorik değişkenler ve sayısal değişkenler ayrı ayrı görselleştirilmiştir. Korelasyon Analizi görselleştirilmiştir. Ayrıca hedef değişkenin (kalp rahatsızlığı) diğer değişkenlerle olan ilişkisi incelenmiştir. Aykırı değerler IQR Methodu ile görselleştirilmiştir. Ayrıca FLASK kullanarak oluşturduğumuz website de görseller aktarılmıştır.







5.Modelleme ve Makine Öğrenimi

5.1 Uygun Makine Öğrenimi Modellerinin Seçilmesi

Proje kapsamında, sınıflandırma problemi için çeşitli makine öğrenimi modelleri değerlendirilmiştir.Sınıflandırma problemi için uygun olan Random Forest Classifier modeli tercih edilmiştir.

5.2 Verinin Eğitim ve Test Setlerine Bölünmesi

Veri seti, %80 eğitim ve %20 test setlerine bölünmüştür. Bu sayede modelin performansı bağımsız veri üzerinde değerlendirilmiştir.

5.3 Modelin Uygulanması ve Sonuçların Karşılaştırılması

Random Forest Classifier modeli, belirli hiperparametreler ile eğitilmiştir:

```
# RandomForestClassifier modelini daha az karmaşık hale getirmek için parametreler
model = RandomForestClassifier(max_depth=10, min_samples_split=10, min_samples_leaf=4, n_estimators=100, random_state=42)
model.fit(X_train, y_train)
```

6.Model Değerlendirme ve İyileştirme

6.1 Model Performansının Değerlendirilmesi

Modelin performansı, confusion matrix ve diğer sınıflandırma parametreleri kullanılarak değerlendirilmiştir. Başlangıçta %85 doğruluk elde edilmesine rağmen, modelin ezberleme yaptığı (overfitting) tespit edilmiştir.

6.2 Modelin İyileştirilmesi

Modelin ezberleme yapmasını önlemek amacıyla, hiperparametreler yeniden düzenlenmiş ve model daha az karmaşık hale getirilmiştir. İyileştirilmiş model ile %83 doğruluk elde edilmiştir, bu da modelin genel performansının dengelendiğini göstermektedir.

7.Sonuçlar ve Raporlama

7.1 Elde Edilen Sonuçların Özeti

Geliştirilen model önceki yapılan çalışmaya göre, kalp rahatsızlığı tahmininde %83 doğruluk oranı ile başarılı sonuçlar vermiştir. Modelin performansı, eğitim ve test setlerinde benzer doğruluk değerleri göstermiştir, bu da modelin genelleme yeteneğinin iyi olduğunu göstermektedir. Ezberleme yapma sorunu çözülmüştür.

7.2 Proje Sürecinde Karşılaşılan Zorluklar

Proje sürecinde, veri setindeki eksik verilerin kontrol edilmesi, kategorik ve numerik değişkenlerin doğru şekilde ayrılması gibi zorluklarla karşılaşılmıştır. Bu zorluklar, uygun veri işleme teknikleri ve model iyileştirmeleri ile aşılmıştır.

7.3 Gelecek Çalışmalar İçin Öneriler

Gelecek çalışmalar için, farklı makine öğrenimi algoritmalarının denenmesi ve modelin daha geniş veri setleri ile eğitilmesi önerilmektedir. Ayrıca, modelin gerçek dünya uygulamaları için daha fazla optimize edilmesi gerekmektedir.

Kaynaklar

[1] Algoritma, G., Öznitelik, Y., Kullanılarak, S., Öğrenmesi, M., Ile Kalp, A., Tahmini, H., Aydin, H., Üniversitesi, İ. A., Çetinkaya, A., & Üniversitesi, G. (2021). Heart Disease Prediction with Machine Learning Algorithm Using Feature Selection by Genetic Algorithm. Researchgate.Net, 2(2), 67–80. https://doi.org/10.53525/jster.1005934

[2] Bilimleri, M., Dalı Bilgisayar, A., & Programı, M. (2022). Kalp yetmezliği riskinin makine öğrenmesi yöntemleri ile analiz edilmesi. http://earsiv.kmu.edu.tr/xmlui/handle/11492/6597

[3] Gamze, K., Bilimleri, S. K.-İ. T. Ü. F., & 2022, undefined. (2022). KARDİYOVASKÜLER HASTALIK TAHMİNİNDE MAKİNE ÖĞRENMESİ SINIFLANDIRMA ALGORİTMALARININ KARŞILAŞTIRILMASI. Dergipark.Org.TrK Gamze, SB KalkanİstanbulTicaret Üniversitesi Fen Bilimleri Dergisi, 2022•dergipark.Org.Tr, 21(42), 183–193. https://doi.org/10.55071/ticaretfbd.1145660

[4] Modelleri, M. Ö., Kırılganlık, K., Tahmini, D., Erdaş, Ç. B., Ölçer, D., Mühendisliği, B., Fakültesi Başkent, M., & Ankara, Ü. (2022). Makine Öğrenimi Modelleri Kullanılarak Kırılganlık Derecesi Tahmini Prediction of Frailty Grade Using Machine Learning Models. https://www.biyoklinikder.org/TIPTEKNO22_Bildiriler/071.pdf

[5] Potur, E., Dergisi, N. E.-A. B. ve T., & 2021, undefined. (2021). Kalp yetmezliği hastalarının sağ kalımlarının sınıflandırma algoritmaları ile tahmin edilmesi. Dergipark.Org.TrEA Potur, N ErginelAvrupa Bilim ve Teknoloji Dergisi, 2021•dergipark.Org.Tr, 24, 112–118. https://doi.org/10.31590/ejosat.902357