Ödev Konusu: Coronary Artery Disease Veri Setinin Ön İşleme, Keşifsel Veri Analizi ve Makine Öğrenimi Modelleri ile Eğitimi
Öğrenci Numarası: Y245012006
Öğrenci Adı Soyadı: Aleyna Barut
Kullanılan Veri Kümesi: Classification of Coronary Artery Disease Veri Kümesi

Coronary Artery Disease (CAD) Sınıflandırma Modeli: Analiz ve Değerlendirme

1. Giriş

1.1. Proje Hedefi
Bu çalışma, Coronary Artery Disease (CAD) (Koroner Arter Hastalığı) tanısının makine öğrenimi algoritmalarıyla sınıflandırılmasını amaçlamaktadır. Sağlık alanında erken teşhis, özellikle ölüm oranlarının azaltılması ve tedavi süreçlerinin iyileştirilmesi açısından kritik öneme sahiptir. Proje kapsamında doğru ve güvenilir bir model geliştirilerek VHD tanısında karar destek sistemlerinin etkinliğini artırmak hedeflenmiştir.

1.2. Veri Kümesi Tanıtımı

Kullanılan veri kümesi, 303 hasta örneğinden (216 KAH hastası, 87 sağlıklı birey) oluşmakta ve toplamda 55 farklı özellik içermektedir. Özellikler şu kategorilere ayrılmaktadır:

- **Demografik Özellikler**
- **Semptomlar ve Muayene Sonuçları**
- **Elektrokardiyogram (EKG) Verileri**
- **Laboratuvar ve Ekokardiyografi**

Veriler hastaların yaş, cinsiyet, kan basıncı, kolesterol seviyesi, egzersiz alışkanlıkları gibi faktörleri içerir. Bu faktörler hastaların sağlık durumunu ve olası kalp hastalığı risklerini tahmin etmek için kullanılır. Veri kümesi eksiksizdir ve herhangi bir eksik değer içermemektedir bu da ön işleme sürecini kolaylaştırmıştır.

2. Veri Ön İşleme

2.1. Veri Kümesinin Özellikleri ve Analizi
Veri kümesindeki özellikler kategorik(categorical), sayısal (numeric) ve sıralı (ordinal) olmak üzere üç ana kategoriye ayrılabilir.

- **Kategorik**: Cinsiyet, diyabet, hipertansiyon gibi niteliksel veriler.
- **Sayısal**: Yaş, kan basıncı gibi sürekli değişken veriler
- **Sıralı**: Egzersiz alışkanlıkları, kalp kapak hastalığı(VHD) gibi belirli bir düzen izleyen veriler

Bu özelliklerin doğru şekilde modellenmesi makine öğrenimi algoritmalarının başarısını doğrudan etkilemektedir.

2.2. Veri Temizleme
Veri kümesinde eksik değerler bulunmamaktadır. Bu nedenle veri temizleme adımına gerek kalmamış ve veri seti doğrudan kullanıma uygun olmuştur.

2.3. Korelasyon Matrisi ve Özellik Seçimi
Veri kümesinde korelasyon matrisi kullanılarak özellikler arasındaki ilişkiler incelenmiştir. Korelasyon analizi, özelliklerin birbirleriyle nasıl etkileşime girdiğini ve modelin doğruluğunu nasıl etkileyebileceğini belirlemede kritik bir adımdır. Özellikle VHD (Kalp Kapak Hastalığı Durumu) ile yüksek korelasyona sahip olan özellikler seçilmiş ve bu özellikler üzerinden modeller eğitilmiştir. Bu özellikler modelin doğruluğunu artıracak şekilde kullanılmıştır.

2.4. Özellik Ölçeklendirme
Makine öğrenimi modellerinin çoğu verilerin belirli bir ölçek aralığında olmasını gerektirir. Bu nedenle StandardScaler kullanılarak tüm sayısal veriler standartlaştırılmış ve modele girdi olarak sunulmuştur. Bu adım, modelin daha hızlı öğrenmesine ve daha iyi genelleme yapmasına yardımcı olmuştur.

2.5. Veri Dengeleme (SMOTE)
Veri kümesinde, kalp hastalığına sahip olmayan bireylerin sayısı hastalığı taşıyanlara göre fazla olduğundan SMOTE (Synthetic Minority Over-sampling Technique) yöntemi ile azınlık sınıfındaki veriler artırılmıştır. Bu modelin her iki sınıfı da daha iyi öğrenmesini sağlamıştır.

3. Modellerin Eğitimi ve Değerlendirilmesi

3.1. Lojistik Regresyon (Logistic Regression)
Lojistik regresyon, sınıflandırma problemlerinde yaygın olarak kullanılan bir yöntemdir. Bu modelde doğruluk oranı %62 olarak elde edilmiştir. Bu model, sınıflandırma doğruluğunu temel alır ancak genellikle doğrusal ilişkilerde daha etkilidir.

3.2. Destek Vektör Makineleri (SVC)
Destek vektör makineleri veri kümesindeki sınıfları ayıran en iyi hipereği bulmaya çalışır. Bu modelde elde edilen doğruluk oranı %54'dür. Özellikle doğrusal olmayan verilerde etkili olduğu için karmaşık sınıflandırma problemlerinde kullanılır.

3.3. Karar Ağaçları (Decision Trees)
Karar ağaçları, veriyi dallara ayırarak sınıflandıran bir modeldir. Bu modelde doğruluk oranı %62'dir. Karar ağaçları kolay yorumlanabilir olmalarıyla dikkat çeker ancak aşırı uyum (overfitting) riski taşırlar.

3.4. Rastgele Orman (Random Forest)
Rastgele Orman, birden fazla karar ağacının birleşiminden oluşan güçlü bir ansambl modelidir. Bu model, %70 doğruluk oranı ile en yüksek performansı sergileyen model olmuştur. Rastgele Orman, yüksek doğruluk oranları ve genelleme gücü ile dikkat çeker.

3.5. Naive Bayes
Naive Bayes, koşullu olasılık temelli bir sınıflandırma modelidir. Bu model, %40 doğruluk

oranı ile diğer modellere göre daha düşük performans sergilemiştir ancak özellikle büyük veri kümelerinde hızlıdır.

3.6. K-Nearest Neighbors (KNN)
KNN, benzerlik esaslı bir modeldir ve sınıflandırma yaparken yakınındaki verileri referans alır. Bu modelde doğruluk oranı %62'dir. KNN her sınıfın komşusundaki örnekleri baz alarak karar verir.

3.7. Özelleştirilmiş KNN ve SVC
Model performanslarını iyileştirmek amacıyla KNN ve SVC modelleri hiperparametre ayarları ile optimize edilmiştir. KNN için doğruluk oranı %63 ve SVC için doğruluk oranı %65 olarak elde edilmiştir.

4. Model Performans Metrikleri

4.1. Karışıklık Matrisleri
Her modelin başarısı karışıklık matrisleri aracılığıyla değerlendirilmiştir. Bu matrisler doğru ve yanlış sınıflandırmaları, doğru pozitif ve doğru negatif değerleri göstermektedir.

4.2. Performans Metrikleri Tablosu
Her modelin performansı doğruluk, duyarlılık, özgüllük, kesinlik ve F1 skoru gibi temel metriklere göre değerlendirilmiştir.

Model	Doğruluk Skoru (Accuracy)
Random Forest	0.704918
Özelleştirilmiş SVC	0.655738
Özelleştirilmiş KNN	0.639344
Logistic Regression	0.622951
KNN	0.622951
Decision Tree	0.622951
SVC	0.540984
Naive Bayes	0.409836

5. Sonuçlar ve İleriye Yönelik Çalışmalar

5.1. Model Başarısı
Elde edilen sonuçlara göre **Rastgele Orman (Random Forest)** algoritması, %70 doğruluk oranı ile en başarılı model olarak öne çıkmıştır. Modelin genelleme başarısı ve doğruluğu, erken teşhis süreçlerinde sağlık sektörüne katkı sağlayabilecek potansiyelde olduğunu göstermektedir.

5.2. Modelin Geliştirilmesi ve Optimizasyonu
Model performansını iyileştirmek için hiperparametre optimizasyonu önemlidir. Grid Search veya Random Search gibi yöntemlerle her modelin en uygun parametreleri belirlenebilir..Ansambl yöntemleri (örneğin XGBoost) kullanılarak birden fazla modelin öngörüleri birleştirilebilir ve daha güçlü sonuçlar elde edilebilir. Bu adımlar modelin genel doğruluğunu ve güvenilirliğini artırır.

5.3. Yeni Özellikler ve Veri Genişletme
Veri kümesine yeni özellikler eklemek ve daha fazla hasta verisi kullanmak modelin genel başarısını artırabilir. Özellikle genetik veriler ve tıbbi geçmiş gibi faktörlerin eklenmesi kalp hastalığının erken tespiti için faydalı olabilir.

6. Sonuç

Bu çalışma Koroner Arter Hastalığı sınıflandırmasında sonuçlar elde edilmiştir. Farklı makine öğrenimi modelleri arasından Rastgele Orman modeli en yüksek doğruluğu elde ederek en başarılı model olmuştur. Bu modelin sağlık sektöründe kullanılabilirliği, erken teşhis ve tedavi süreçlerini iyileştirebilir.