HEART DİSEASE UCI CLASSIFICATION

PROBLEMİN AMACI

- Hastarın medikal verilerine dayanarak kalp hastalığı oluğ olmadığını tahmin eden bir makine öğrenmesi modeli geliştirmek amaçlanmaktadır. Erken teshis süreçlerine yardımcı olmak hedeflenmektedir
  - `num` sutunu hedef değişkendir.
  - 0 = hastalık yok, 1–4 = hastalık var şeklindedir. Bu değer, proje kapsamında ikili sınıflandırma için şu şekilde dönüştürülmüştür:
    `df["num"].apply(lambda x: 0 if x == 0 else 1)`
- **Pratikteki Önemi:**
  - Kalp hastalığı dünya genelindeki en yaygın ölümlerden biridir. bu nedenle riskli bireylerin erken tespiti hayat kurtarabilir.

Veri Analizi ve Görselleştirme (EDA)

- Toplam 920 gözlem ve 16 sütun bulunmakta.
- Kategorik ve sayısal değişkenler karışık.
- Eksik veri bulunan sütunlar var: `trestbps, chol, thalach, ca, thal, slope, fbs gibi.`
  - En çok eksik veri `ca` (611 değer) ve `thal` (486) sütunlarındadır. Sayısal eksikler ortalama ile, kategorik eksikler mod ile doldurulmuştur.
- Hedef Değişkenin Dağılımı : %45 hasta yok,%55 hasta var şeklindedir.
- Verideki erkek cinsiyet oranı daha yaygın

Veri Ön İşleme

- `grab_col_names(df, cat_th=10)` fonksiyonu ile numeric columnlar ve kategorik columnlar ayrıldı. Eksik değerleri sonraki aşamada doldurulmak için listelendi.
- `preprocessing(df)` fonksiyonu ile önce eksik değerler mod ve mean ile dolduruldu. sonra `pd.get_dummies`ile kategorik veriler one hot encode edilip data leakage i önelemek amacıyla `drop_first=True)` uygulandı.
- Numeric verilerde çarpıklığı incelemek için grafikler oluşturuldu ve scale tercihi için skew incelendi
- Eğitim/test ayrımı 70-30 olarak seçildi
  - `stratify=y` kullanılarak hedef değişkenin eşit dağılması amaçlandı
- Yalnızca eğitim verisine fit_transform, test verisine transform uygulanacak şekilde scale işlemleri gerçekleştirildi.

4B. Eğer Heart Disease UCI seçildiyse (Sınıflandırma):

- Hedef değişken `y = y.apply(lambda x: 0 if x == 0 else 1)` ile ikili hale getirildi.
- Logistic Regression, Random Forest, KNN, SVM sınıflandırma modelleri kuruldu.
- Başarı metrikleri: Accuracy, Precision, Recall, F1-Score Confusion Matrix ile incelendi.
  - eğitim doğruluğu ve test doğruluğu arasında bir çıkarım yaparak overfitting riskini değerlendirmeye çalışıldı
- ROC-AUC eğrisi oluşturuldu
  - 0.87 lik roc_auc skoru elde edildi

Ekstra

- GridSearchCV ile hiperparametre optimizasyonu yapıldı
  - `best_model_name = max(results.items(), key=lambda x: x[1]['test_accuracy'])[0]` ile en iyi model seçilen Random Forest üzerinden GridSearchCV ile hiper parametre optimizasyonu yapldı.
  - eature Importance görselleştirmesi (özellikle Random Forest ile) yapıldı.
    - `oldpeak, age, thalch, chol, exang_True, ca, trestbps, cp_atypical angina` özelliklerin önemi görüldü

Model Karşılaştırması

Model	Eğitim Doğruluk	Test Doğruluk	Precision (Ort.)	Recall (Ort.)	F1-Score (Ort.)	Overfitting Durumu
Logistic Regression	0.8478	0.8478	0.85	0.84	0.84	☑ Overfitting görünmüyor
Random Forest	1.0000	0.8804	0.89	0.88	0.88	⚠ Overfitting ihtimali var
SVM	0.7562	0.7355	0.75	0.75	0.74	☑ Overfitting görünmüyor
KNN	0.8432	0.8116	0.81	0.81	0.81	☑ Overfitting görünmüyor

- Hangi model daha iyi performans verdi?
  - test_accuracy açısından en yüksek skor random forest da alındı
- Hangi metrik üzerinde öne çıktı
  - her ne kadar accuracy yi ilk etapta incelesek de bu tarz hastalık tespiti gibi amaçlarda f1score ve recall(1) durumları da önemlidir.
- Aşırı öğrenme veya yetersiz öğrenme gözlendi mi?
  - random forest da test_accuracy nin train_accuracy den düşük olduğu ve train_aaccuracy nin 1 oldğu gözlemledndi. Bu overfitting belirtisidir.
  - svc nin çok da iyi sayılamayan bir performans gösterdiği gözlemlendi

Sonuç ve Yorumlar

- Modelin pratik kullanımı hakkında değerlendirme
  - **Genel değerlendirmeden sonra logistic regresyon(recall(1):0.89) ve knn nin en açıklanabilir modeller olduguna karar verildi ***
  - **İlk tercih için Logistic Regresyon daha uygun**
- Daha iyi sonuçlar için neler yapılabilir?
  - cross validation ve hiperparametre ayarlamaları
  - catboost, lightgm modelleri
  - veri çok linear olmamasından kaynaklı svc cok başarı gösterememiş.
  - veri seti boyutu daha fazla olabilirdi
- Veri setiyle ilgili gözlemler

Sütun Adı	Türkçe Adı	Açıklama
id	Kimlik	Her hasta için benzersiz kimlik numarası
age	Yaş	Hastanın yaşı (yıl)
origin	Kaynak / Kaynak Yeri	Çalışmanın yapıldığı yer (veri kaynağı)
sex	Cinsiyet	Cinsiyet (Erkek / Kadın)
cp	Göğüs Ağrısı Tipi	Göğüs ağrısı tipi (tipik angina, atipik angina, non-anginal, vs.)
trestbps	Dinlenme Kan Basıncı	Hastaneye girişteki dinlenme halindeki kan basıncı (mm Hg)
chol	Kolesterol	Serum kolesterol düzeyi (mg/dl)
fbs	Açlık Kan Şekeri	Açlık kan şekeri > 120 mg/dl ise 1, değilse 0
restecg	Dinlenme EKG Sonucu	Dinlenme elektrokardiyografi sonucu (normal, stt abnormality, vs.)
thalach	Maksimum Kalp Atış Hızı	Egzersiz sırasında ulaşılan maksimum kalp atış hızı
exang	Egzersize Bağlı Angina	Egzersiz ile tetiklenen göğüs ağrısı (var / yok)
oldpeak	ST Depresyonu	Egzersizle indüklenen ST segment depresyonu
slope	ST Segment Eğimi	Egzersiz sırasında ST segment eğimi
ca	Renkli Damarlardaki Sayı	Floroskopi ile boyanmış büyük damar sayısı (0-3)
thal	Talasemi Tipi	Talasemi durumu (normal, sabit defekt, geri dönüşümlü defekt)
num	Hedef Değişken	Kalp hastalığı durumu (0 = yok, 1-4 = farklı hastalık seviyeleri)