VERİ ANALİZİ DERSİ – DÖNEM PROJESİ

Akciğer Kanseri Tahmini (Survey Lung Cancer)

Muhammet KUTLU 23430070059 BST3

1. GİRİŞ

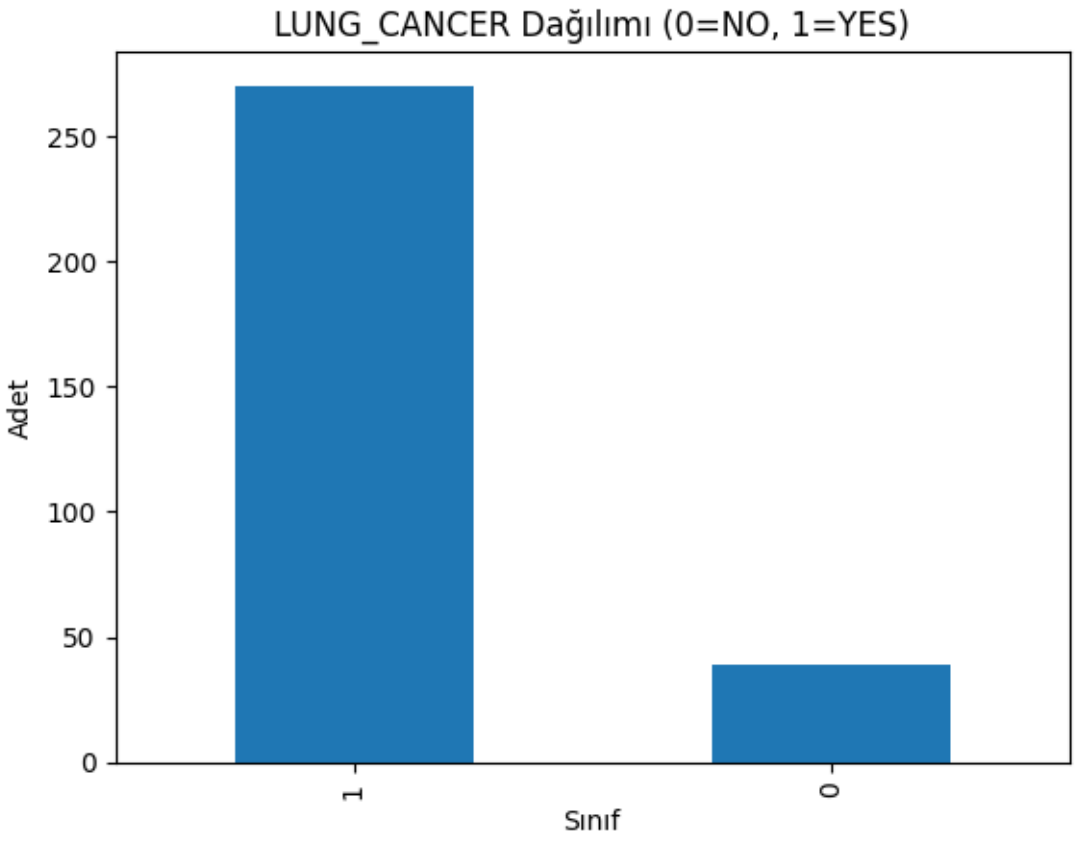
Bu projede Kaggle üzerinde paylaşılan “Survey Lung Cancer” veri seti kullanılmıştır. Veri seti; bireylerin demografik bilgileri (AGE, GENDER) ile çeşitli semptom ve alışkanlık değişkenlerini (ör. SMOKING, COUGHING, WHEEZING, SHORTNESS_OF_BREATH, ALCOHOL_CONSUMING vb.) içeren anket tabanlı kayıtlar sunmaktadır.

Çözülmek istenen problem, hedef değişken olan LUNG_CANCER (YES/NO) bilgisini kullanarak bir bireyin akciğer kanseri sınıfını tahmin etmektir. Bu nedenle problem ikili sınıflandırma (binary classification) olarak ele alınmıştır.

2. VERİ ANALİZİ (EDA)

2.1 Veri setinin genel özellikleri

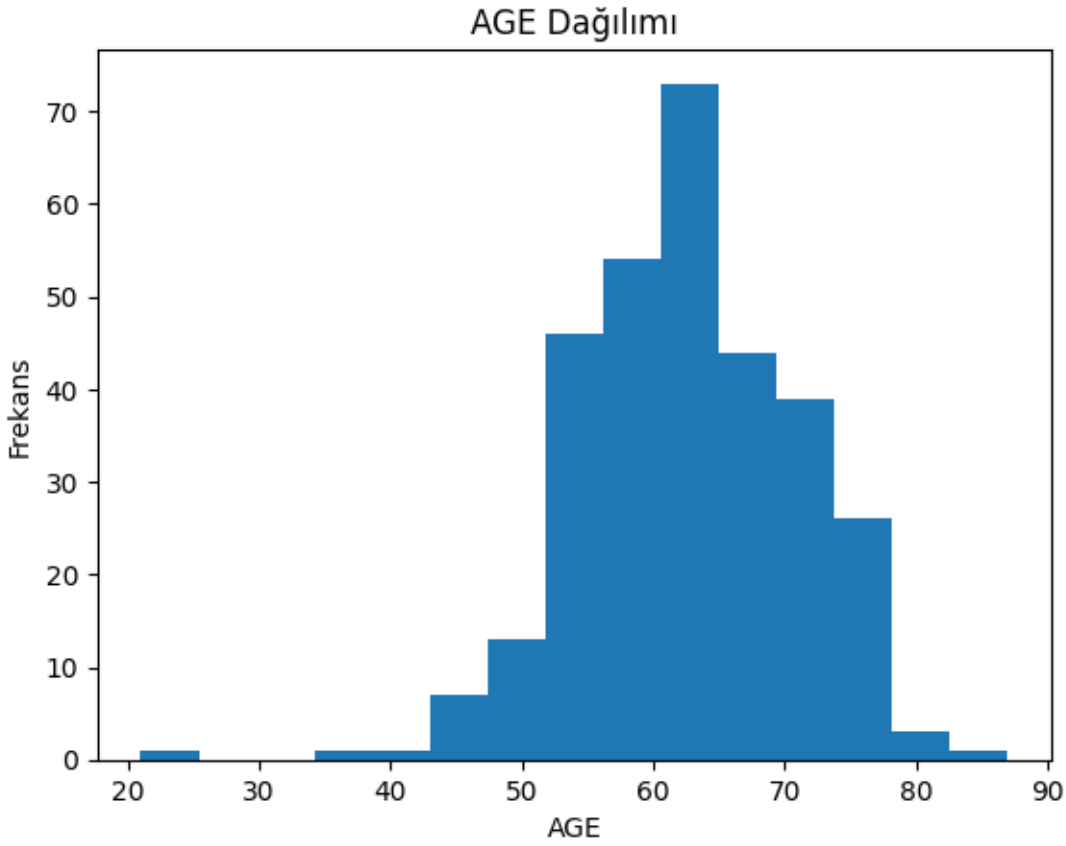
- Veri boyutu: 309 satır, 16 sütun.
- Hedef değişken: LUNG_CANCER (YES/NO) -> (1/0).
- Sınıf dağılımı: 1 (YES)=270, 0 (NO)=39. Veri seti dengesizdir; bu nedenle değerlendirmede yalnızca accuracy değil, ROC-AUC ve confusion matrix de dikkate alınmıştır.



Şekil 1. LUNG_CANCER sınıf dağılımı (0=NO, 1=YES).

2.2 Yaş (AGE) dağılımı

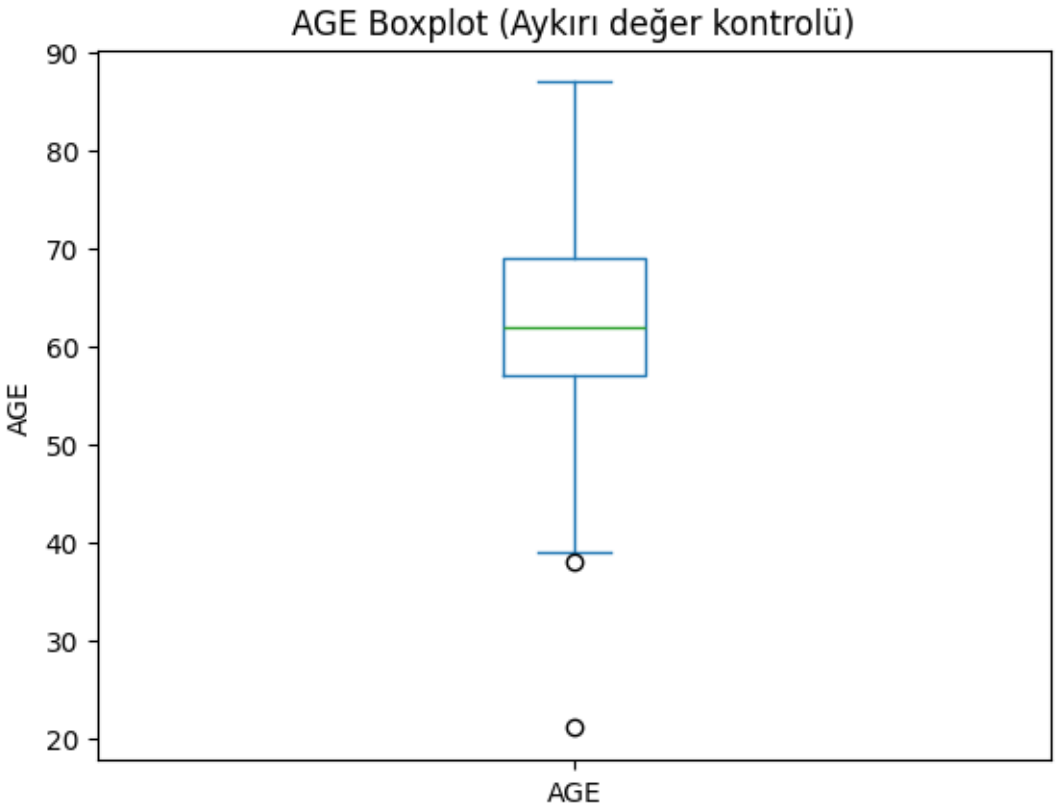
AGE değişkeninin histogramı incelendiğinde gözlemlerin büyük kısmının orta–ileri yaş aralığında yoğunlaştığı görülmektedir.



Şekil 2. AGE dağılımı (histogram).

2.3 Aykırı değer analizi

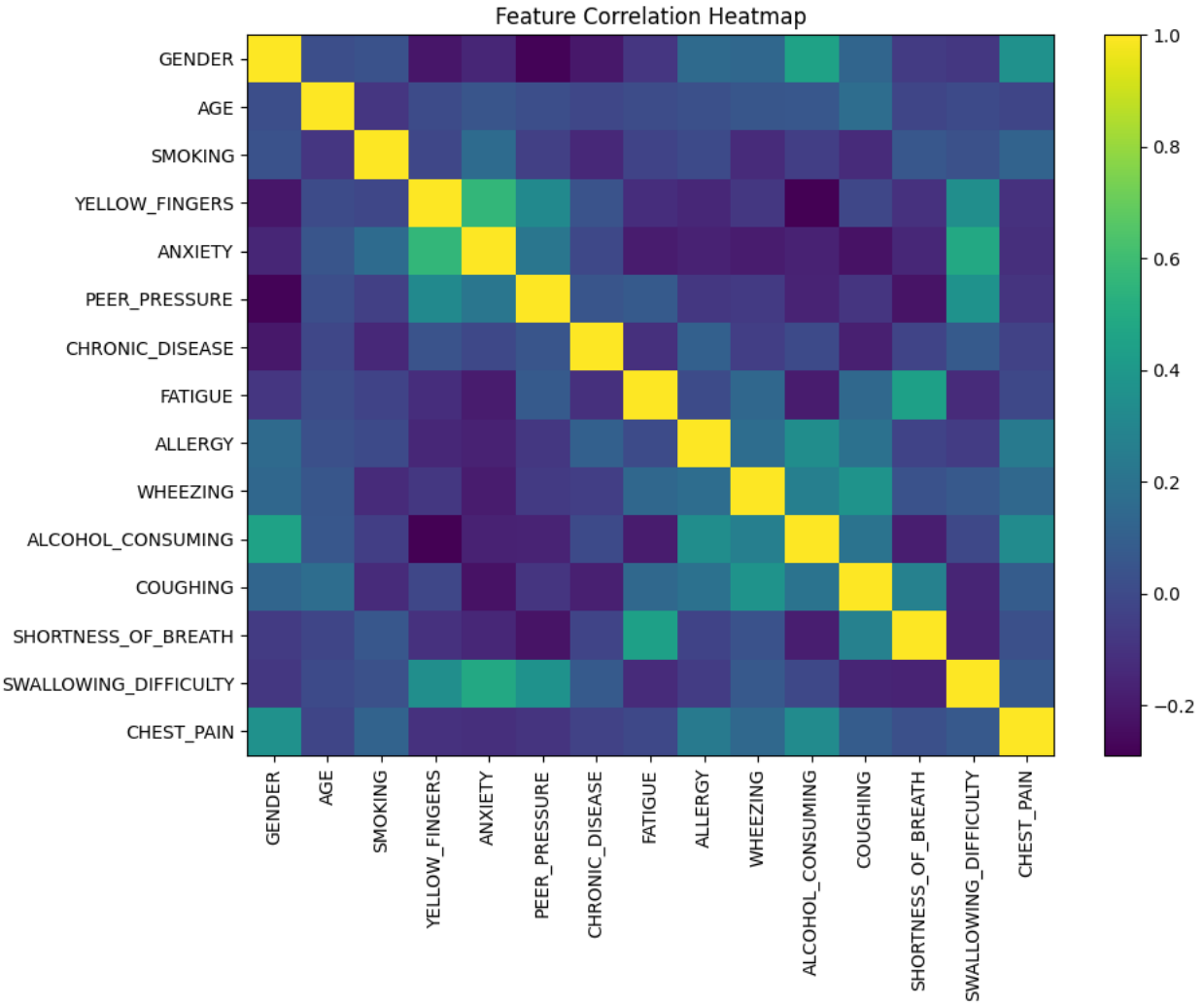
AGE değişkeni için IQR yöntemi ile aykırı değer analizi yapılmıştır. Hesaplanan sınırlar yaklaşık olarak lower=39 ve upper=87 olup 2 adet aykırı değer gözlenmiştir.



Şekil 3. AGE boxplot (aykırı değer kontrolü).

2.4 Korelasyon analizi

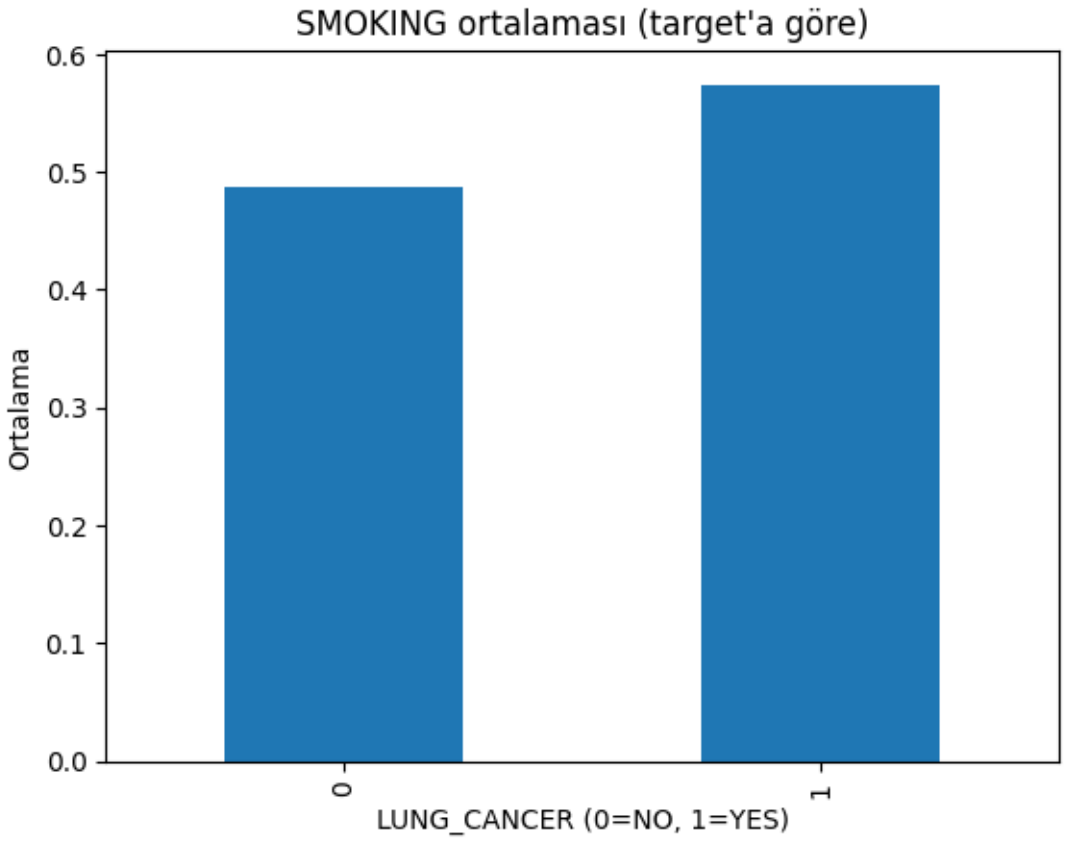
Özellikler arasındaki ilişkileri görmek amacıyla korelasyon matrisi görselleştirilmiştir. Genel olarak değişkenler arasında düşük/orta düzey ilişkiler gözlenmiştir.



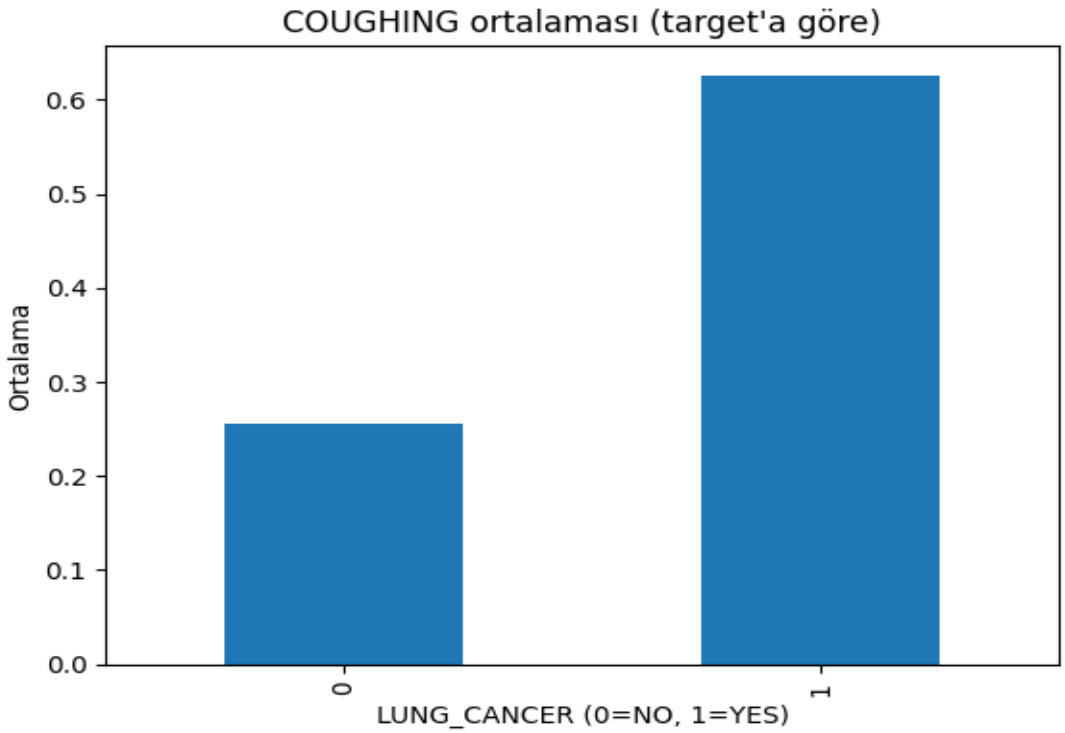
Şekil 4. Özellik korelasyon ısı haritası (heatmap).

2.5 Target'a göre karşılaştırmalar

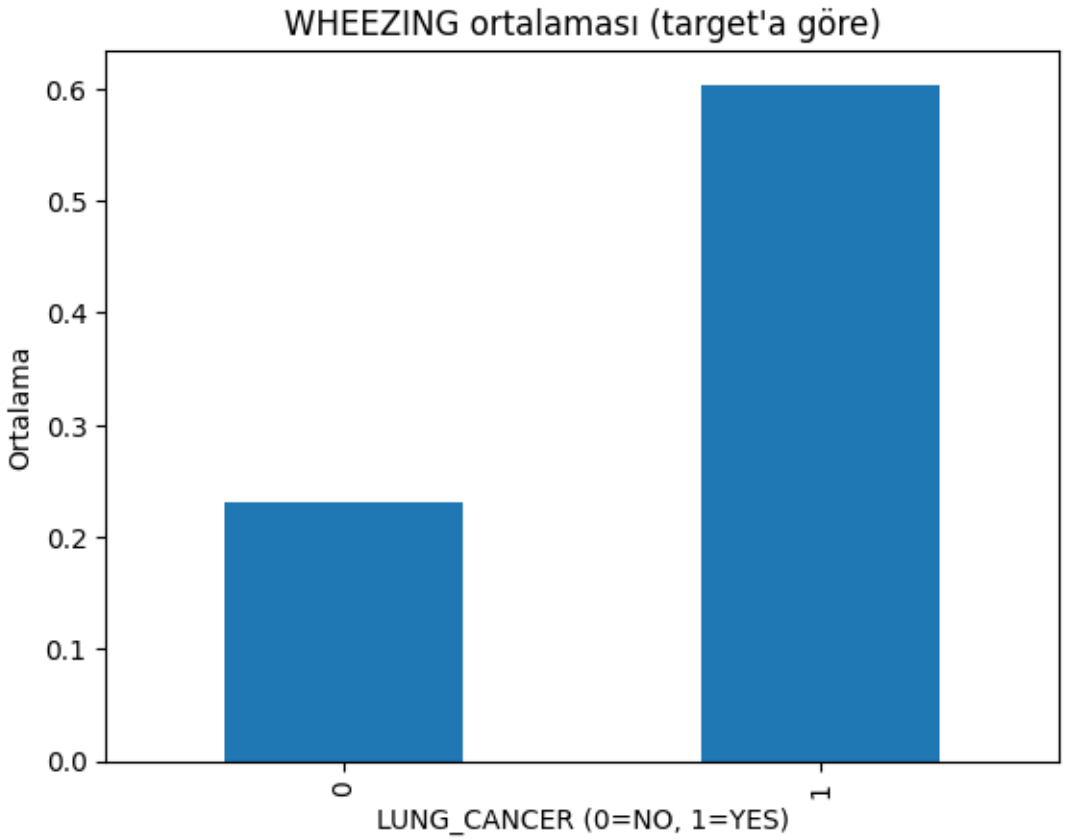
Bazı semptom ve alışkanlık değişkenleri hedef sınıfa (LUNG_CANCER) göre karşılaştırılmıştır. Grafiklerden, LUNG_CANCER=1 sınıfında SMOKING, COUGHING, WHEEZING ve SHORTNESS_OF_BREATH değişkenlerinin ortalamalarının daha yüksek olma eğiliminde olduğu gözlenmiştir.



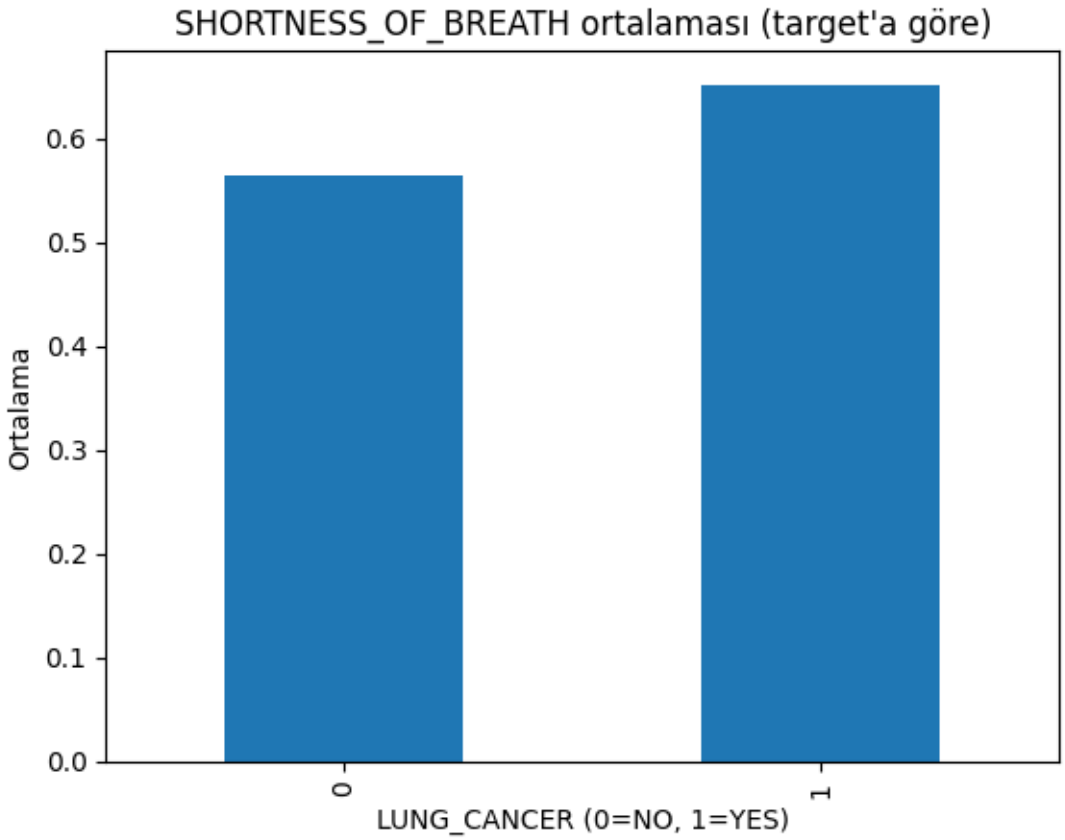
Şekil 5. SMOKING ortalaması (target'a göre).



Şekil 6. COUGHING ortalaması (target'a göre).



Şekil 7. WHEEZING ortalaması (target'a göre).



Şekil 8. SHORTNESS_OF_BREATH ortalaması (target'a göre).

3. YÖNTEM

3.1 Ön işleme adımları

- LUNG_CANCER değişkeni YES/NO'dan 1/0 formatına dönüştürülmüştür.
- GENDER değişkeni MALE/FEMALE'den 1/0 formatına kodlanmıştır.
- Modelleme aşamasında olası eksik değer durumlarına karşı pipeline içinde median imputation uygulanmıştır.
- Logistic Regression modeli için StandardScaler ile ölçekleme yapılmıştır.
- Veri seti %80 eğitim ve %20 test olarak ayrılmış; sınıf dengesizliğine karşı stratify=y kullanılmıştır.

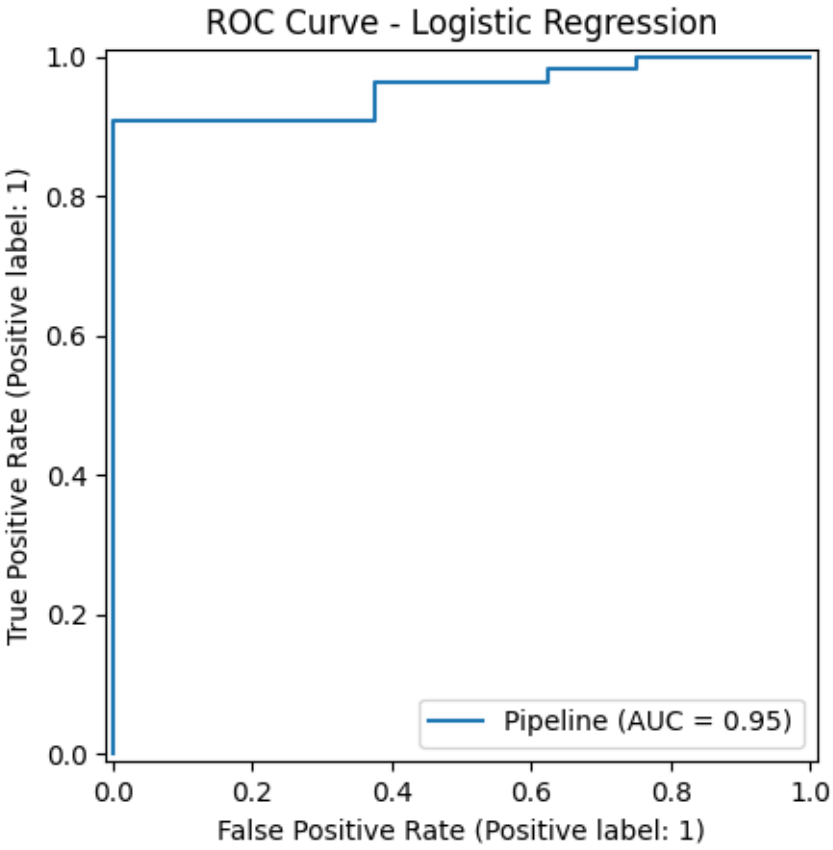
3.2 Kullanılan algoritmalar ve seçilme gerekçeleri

- Logistic Regression: Basit, hızlı ve yorumlanabilir bir baseline model olarak seçilmiştir.
- Random Forest: Doğrusal olmayan ilişkileri yakalayabilen, genelde güçlü performans veren ve feature importance sağlayan ağaç tabanlı bir yöntemdir.

4. SONUÇLAR VE BULGULARIN YORUMLANMASI

4.1 Logistic Regression performansı

- Accuracy: 0.871
- ROC-AUC: 0.954
- Confusion Matrix: [[8, 0], [8, 46]]



Şekil 9. ROC eğrisi - Logistic Regression.

Modelin ROC-AUC değerinin yüksek olması sınıfları ayırmada başarılı olduğunu göstermektedir. Bununla birlikte veri dengesizliği nedeniyle yalnızca accuracy ile yorum yapmak yeterli değildir; confusion matrix ve sınıf bazlı metrikler birlikte değerlendirilmelidir.

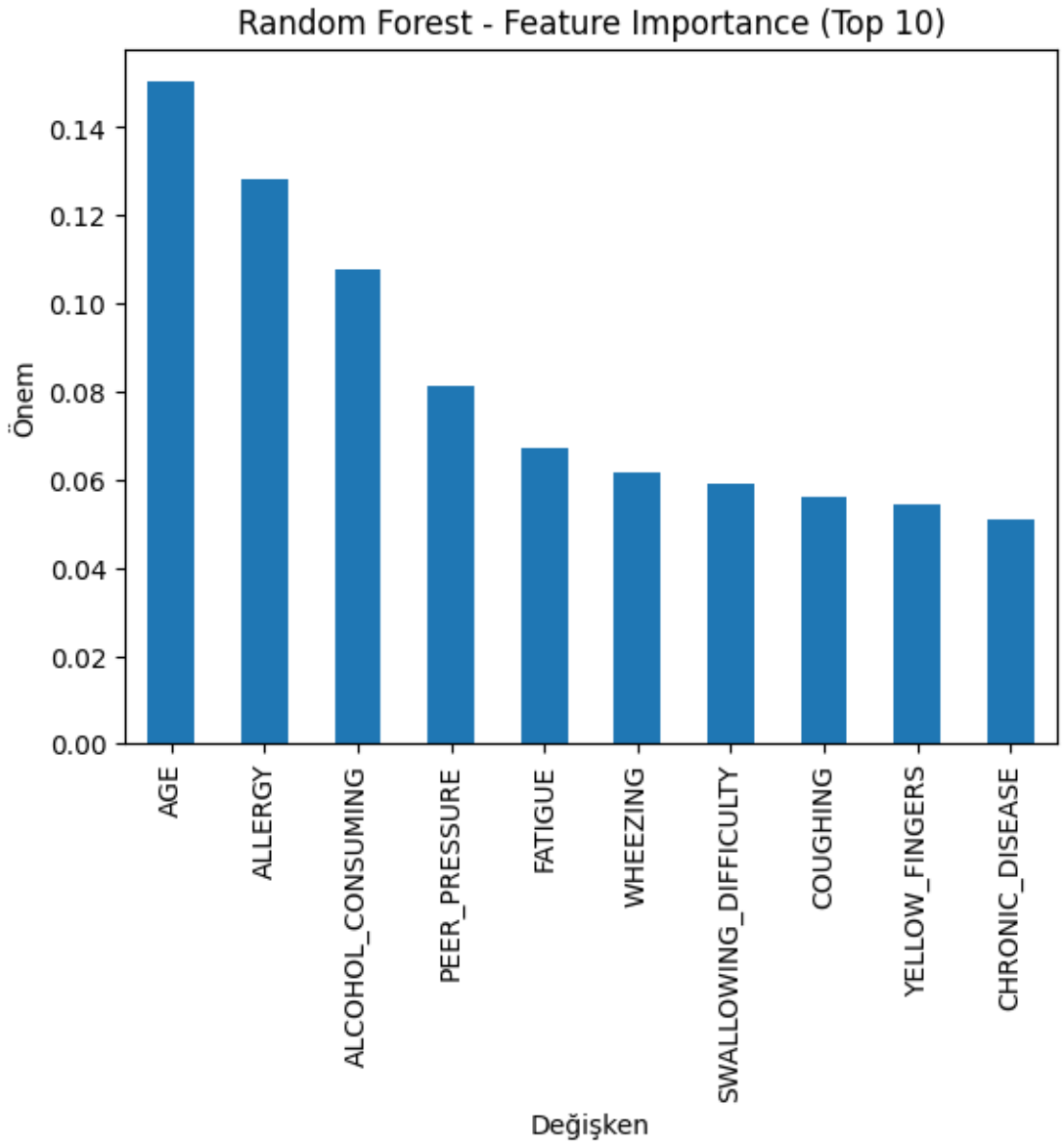
4.2 Random Forest performansı

- Accuracy: 0.887
- ROC-AUC: 0.946
- Confusion Matrix: [[4, 4], [3, 51]]

Random Forest modeli accuracy açısından Logistic Regression'a göre biraz daha yüksek sonuç vermiştir. Dengesiz veri yapısı nedeniyle pozitif ve negatif sınıflardaki hata türleri (false positive / false negative) ayrıca dikkate alınmalıdır.

4.3 Feature Importance (yorumlanabilirlik)

Random Forest modelinden elde edilen feature importance sonuçlarına göre en önemli değişkenler arasında AGE, ALLERGY, ALCOHOL_CONSUMING, PEER_PRESSURE ve FATIGUE gibi özellikler öne çıkmıştır. Bu değişkenler, modelin karar mekanizmasında daha yüksek ağırlığa sahiptir.



Şekil 10. Random Forest - Feature Importance (Top 10).

5. GITHUB LİNKİ

GitHub Repo:

https://github.com/mkutluwork-maker/AKCIGER-KANSERI-VERI-ANALIZI