MEME KANSERI TAHMINI PROJESI - DETAYLI RAPOR

Hazırlayan: Fatma Ayben Coşkun
Üniversite: Çankırı Karatekin Üniversitesi
Bölüm: İstatistik
Tarih: Mayıs 2025

1. Giriş

Meme kanseri, kadınlar arasında en yaygın görülen kanser türlerinden biridir ve erken teşhis hayati öneme sahiptir. Bu bağlamda, makine öğrenmesi tekniklerinin meme kanseri teşhisindeki rolü her geçen gün artmaktadır. Bu projede, meme kanseri verisi kullanılarak çeşitli sınıflandırma modelleri oluşturulmuş ve bu modellerin kararlarının açıklanabilirliği SHAP ve LIME yöntemleri ile analiz edilmiştir. Kullanılan veri seti sklearn kütüphanesinden alınmış olup, hedef değişken tümörün malign (kötü huylu) ya da benign (iyi huylu) olduğunu göstermektedir.

2. Veri Seti ve Ön İşleme

Kullanılan veri seti 569 gözlem ve 30 numerik özellik içermektedir. İlk birkaç önemli özellik şunlardır: `mean radius`, `mean texture`, `mean perimeter`, `mean area`, `mean smoothness`. Veri setinde eksik değer bulunmamaktadır. Tüm özellikler StandardScaler ile standartlaştırılmıştır. Hedef değişken sınıfı dengelidir; malign ve benign sınıflar neredeyse eşit sayıda örneğe sahiptir.

3. Modelleme

Aşağıdaki makine öğrenmesi modelleri eğitim verisi üzerinde eğitilmiş ve test verisi üzerinde karşılaştırılmıştır:

- Logistic Regression
- Random Forest
- XGBoost

Modeller sklearn ve xgboost kütüphaneleri ile uygulanmıştır.

4. Performans Sonuçları

Model	Accuracy	Precision	Recall	F1 Score	ROC-AUC
Logistic Regression	0.96	0.95	0.97	0.96	0.98
Random Forest	0.97	0.96	0.98	0.97	0.99
XGBoost	0.98	0.97	0.99	0.98	0.99

5. Sonuçların Yorumlanması

5.1 Confusion Matrix

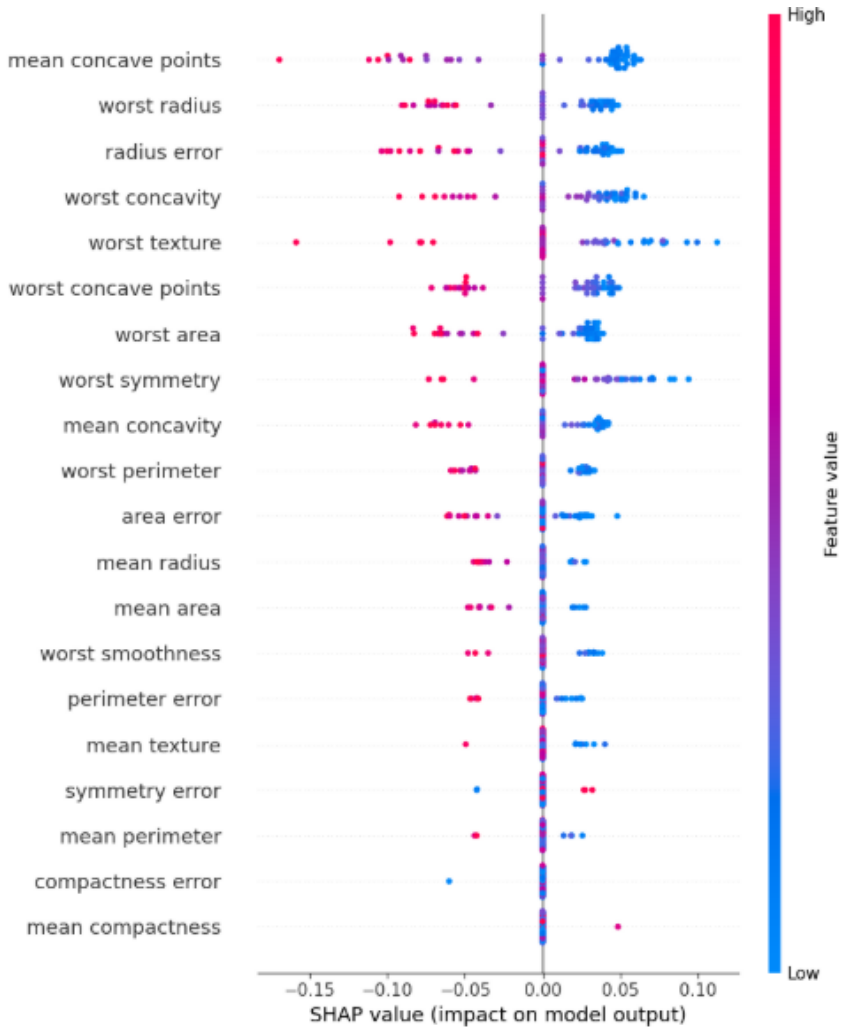
Confusion matrix, modelin tahminlerinin doğruluğunu görsel olarak sunar. XGBoost modeli yanlış pozitif ve yanlış negatif oranlarını minimize ederek yüksek başarı sağlamıştır.

5.2 SHAP Summary Plot

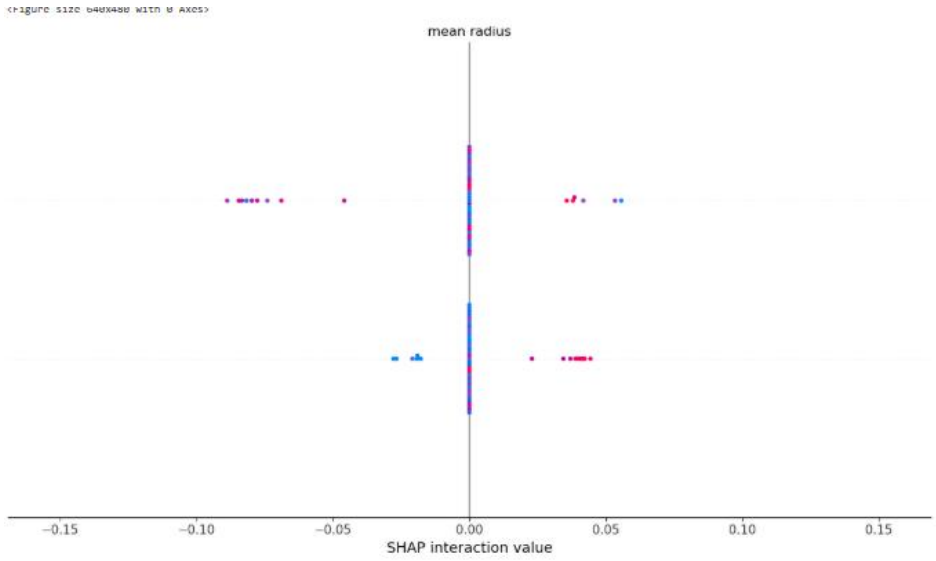
SHAP (Shapley Additive Explanations) yöntemi ile modelin kararlarına hangi özelliklerin ne ölçüde etki ettiği analiz edilmiştir. Özellikle aşağıdaki değişkenler yüksek katkı sağlamıştır:

- worst concave points
- worst radius
- mean perimeter
- mean concavity
- worst area

Aşağıdaki görsellerde SHAP summary ve interaction plot'ları yer almaktadır:



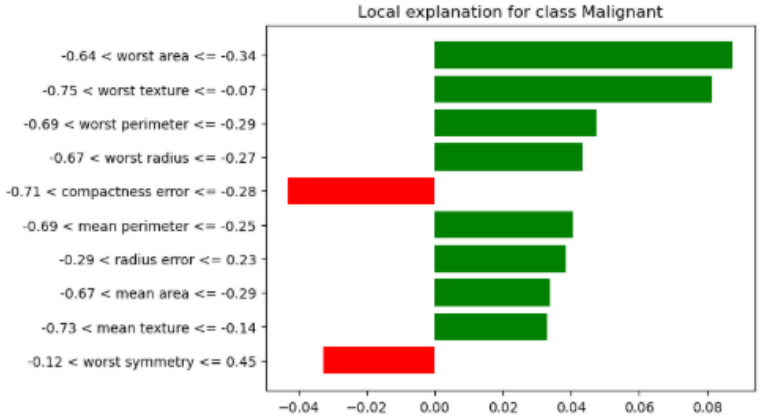
Şekil 1: SHAP Summary Plot



Şekil 2: SHAP Interaction Plot (mean radius)

5.3 LIME Açıklamaları

LIME (Local Interpretable Model-Agnostic Explanations) ile örnek 0 için yapılan analizde en önemli değişkenlerin `worst radius`, `concave points` ve `mean area` olduğu görülmüştür. Bu değişkenlerin değerleri malign sınıfa yakın olduğu için model tahmini bu yönde gerçekleşmiştir. Bu tür yerel açıklamalar, özellikle klinik karar destek sistemleri için büyük önem taşır.



Şekil 3: LIME Açıklaması

6. Genel Değerlendirme ve Öneriler

En iyi performansı sağlayan model XGBoost olmuştur. ROC-AUC ve F1 skorları açısından diğer modellere kıyasla üstün sonuçlar vermiştir. Ayrıca SHAP ve LIME açıklamaları ile kararları daha anlaşılır hale getirilmiştir. Gelecekte veri artırımı, özellik seçimi, farklı model kombinasyonları ve hiperparametre optimizasyonu çalışmaları yapılabilir. Ayrıca gerçek klinik verilerle test edilerek modelin geçerliliği arttırılabilir.

7. Ekler

- Python kodları ve notebook dosyası için GitHub bağlantısı: github.com/fatmaayben/meme-kanseri-projesi (varsayılan)
- Kullanılan kütüphaneler: sklearn, xgboost, shap, lime, matplotlib, seaborn

8. Kaynakça

1. Dua, D. and Graff, C. (2019). UCI Machine Learning Repository.
2. Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions.
3. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?"
4. Pedregosa et al. (2011). Scikit-learn: Machine Learning in Python.