**Meme Kanseri Teşhis ve Tahmin Platformu Proje Raporu**

**1. Giriş**

**Bu rapor, Wisconsin Meme Kanseri veri seti kullanılarak geliştirilmiş teşhis ve tahmin platformunun kapsamlı değerlendirmesini sunar. Amaç; tıbbi karar destek mekanizması olarak kullanılabilecek, aynı zamanda eğitim ve araştırma amaçlı deneysel bir araç olarak işlev gören modüler bir web uygulaması ortaya koymaktır.**

**2. Proje Amacı**

* **Karar Destek: Klinik uzmanlara, tümörlerin iyi veya kötü huylu olma olasılıklarını sayısal ve görsel olarak sunmak.**
* **Eğitim & Araştırma: Akademik kullanıcıların hem kod hem de grafiklerle modelleri incelemesine imkân tanımak.**
* **Teknoloji Entegrasyonu: Streamlit ile etkileşimli UI, scikit-learn ve TensorFlow/Keras ile modüler modelleme.**

**3. Kullanılan Teknolojiler**

|  |  |
| --- | --- |
| **Kategori** | **Teknoloji / Kütüphane** |
| **Programlama** | **Python 3.11** |
| **Web Uygulaması** | **Streamlit** |
| **Veri İşleme** | **pandas, NumPy** |
| **Görselleştirme** | **Matplotlib, Seaborn, Plotly** |
| **Makine Öğrenmesi** | **scikit-learn, XGBoost** |
| **Derin Öğrenme** | **TensorFlow/Keras, PyTorch-TabNet** |
| **Açıklanabilirlik** | **SHAP** |
| **Test Otomasyonu** | **pytest** |
| **Konteynerizasyon** | **Docker, Docker-Compose** |

**4. Veri Hazırlama**

1. **Yükleme & Ön Temizleme: Veri analizi ve modelleme süreçlerinde gereksiz veya alakasız sütunların bulunması, model performansını düşürebileceği için CSV dosyalarından öncelikle anlam ifade etmeyen ya da analizlerde kullanılmayacak sütunlar çıkarılır. Ayrıca veri tipi hataları ve tutarsızlıkları düzeltilerek verinin bütünlüğü sağlanır.**
2. **Eksik Değer Doldurma: Eksik değerler modellerin performansını doğrudan etkilediği için dikkatlice ele alınır. Sürekli değişkenlerde medyan kullanılarak dağılımın uç değerlerden etkilenmesi azaltılırken, kategorik değişkenlerde ise mod tercih edilerek kategorik bütünlük korunur.**
3. **Ölçeklendirme: Farklı ölçeklerdeki özelliklerin model üzerinde baskın olmaması için ölçeklendirme yapılır. Kullanıcı, uygulama içinde interaktif olarak Min–Max veya StandardScaler seçenekleri arasından seçim yapabilir. Bu seçenekler, modelin doğruluk ve tutarlılığını iyileştirmek için uygulanmaktadır.**
4. **Train/Test Bölme: Eğitim ve test verilerinin dağılımlarının benzer olmasını sağlamak amacıyla, sınıflandırma problemlerinde özellikle önemli olan stratify parametresi kullanılır. Böylece eğitim ve test setleri sınıf dağılımlarını koruyarak bölünür, bu da modelin genel performansını daha gerçekçi bir şekilde ölçmeye imkân tanır.**
5. **Yükleme & Ön Temizleme: CSV’den gereksiz sütunlar atılır, veri tipi uyumsuzlukları düzeltilir.**
6. **Eksik Değer Doldurma: Sürekli değişkenlerde medyan; kategoriklerde mod.**
7. **Ölçeklendirme: Kullanıcı tercihine bağlı olarak Min–Max veya StandardScaler.**
8. **Train/Test Bölme: Stratify parametresi ile sınıf dengesini koruyarak böler.**

**5. Keşifsel Veri Analizi (EDA)**

* **İstatistiksel Özet: Ortalama, medyan, çeyrek değerler.**
* **Dağılım Analizleri: Histogram ve kutu grafiği ile uç değer belirleme.**
* **Korelasyon: Isı haritası ve yüksek korelasyonlu özniteliklerin bar grafiği ile görselleştirme.**

**6. Özellik Mühendisliği**

* **SelectKBest: Bu yöntem, özelliklerin hedef değişken ile ilişkisini istatistiksel testlerle ölçerek en etkili özellikleri seçer. Veri setinde çok sayıda özellik bulunduğunda, gereksiz veya düşük etkili özellikleri elemeyi sağlar, böylece modellerin daha az gürültüyle ve daha iyi performansla çalışmasına olanak tanır.**
* **PCA: Başlıca bileşen analizi (PCA), veri setindeki özellikler arasındaki korelasyonları azaltmak ve boyutsal karmaşıklığı düşürmek için kullanılır. Bu yöntem, özelliklerin kombinasyonlarından yeni bileşenler oluşturarak verinin varyansını en iyi şekilde açıklayan bileşenleri bulur. Böylece, çok sayıda özellik ile oluşabilecek aşırı öğrenme sorunlarını azaltır ve modellerin daha hızlı ve etkin çalışmasını sağlar.**

**Alternatif yöntemler olarak Recursive Feature Elimination (RFE) ve Embedded yöntemler (örneğin Random Forest feature importance) de düşünülebilir, ancak PCA ve SelectKBest basitlikleri, yorumlanabilirlikleri ve hesaplama etkinlikleri nedeniyle tercih edilmiştir.**

**7. Makine Öğrenmesi Yaklaşımları**

**Bu projede kullanılan modeller; Lojistik Regresyon, Karar Ağaçları, Random Forest, SVM, Gradient Boosting ve XGBoost'tur. Modellerin seçiminde temel olarak veri setinin boyutu, modelin yorumlanabilirliği ve beklenen performans kriterleri dikkate alınmıştır. Örneğin, küçük ve yorumlanabilir bir model gerektiğinde lojistik regresyon tercih edilirken, yüksek doğruluk ve karmaşık ilişkileri modelleyebilme ihtiyacı durumunda XGBoost veya Gradient Boosting modelleri kullanılmıştır. Ayrıca hiperparametre optimizasyonu, GridSearchCV ve RandomizedSearchCV yöntemleri ile en uygun model parametrelerinin belirlenmesi için yapılmıştır. Değerlendirme aşamasında ise çapraz doğrulama, ROC AUC, Precision–Recall analizleri ve karışıklık matrisi gibi metrikler kullanılarak modellerin genel performansı karşılaştırmalı olarak değerlendirilmiştir.**

* **Modeller: Lojistik Regresyon, Karar Ağaçları, Random Forest, SVM, Gradient Boosting, XGBoost.**
* **Hiperparametre Optimizasyonu: GridSearchCV ve RandomizedSearchCV panel aracılığıyla.**
* **Değerlendirme: Çapraz doğrulama, ROC AUC, Precision–Recall, karışıklık matrisi analizleri.**

**8. Derin Öğrenme Yaklaşımları**

* **MLP & TabNet: Farklı mimariler ile karşılaştırmalı eğitim.**
* **Eğitim Kontrolleri: EarlyStopping, ReduceLROnPlateau ile aşırı öğrenme önleme.**
* **Eğitim Grafikleri: Eğitim/validasyon kayıp ve doğruluk eğrileri.**

**9. Tahmin ve Model Açıklanabilirliği**

* **Interaktif Girdi: Slider ve manuel veri panelleri.**
* **Tahmin Sonucu: Olasılık, sınıf kararı ve SHAP değerleri ile öznitelik etkisi.**
* **Görsel Rapor: Gauge grafikleri ve özet tablolar.**

**10. Sonuç ve Gelecek Çalışmalar. Sonuç ve Gelecek Çalışmalar**

**10.1 Sonuç**

**Bu projede hem geleneksel makine öğrenmesi hem de derin öğrenme yaklaşımlarıyla meme kanseri teşhisinde yüksek performans elde edilmiştir. Test seti sonuçlarına göre:**

* **XGBoost modeli: Yaklaşık %98–99 doğruluk, ROC AUC ≈ 0.99**
* **MLP (Keras): %95–96 doğruluk**
* **TabNet: %97–98 doğruluk**

**SHAP analizleri, en fazla etki eden özniteliklerin hücre tekstürü, hücre alanı ve kenar pürüzlülüğü olduğunu göstermiştir. Bu sayede klinik kullanıcılar modelin karar mekanizmasını da görselleştirebilir. Ayrıca Streamlit arayüzü, model parametrelerinin dinamik değişimine ve anlık geri bildirim alınmasına imkân tanımıştır.**

**10.2 Gelecek Çalışmalar**

* **Veri Çeşitliliği: Çok merkezli veri setleri ile genelleme kabiliyetinin artırılması.**
* **Otomatik Optimizasyon: Optuna veya Hyperopt entegrasyonu ile hiperparametre aramalarının otomasyonu.**
* **Model Versiyonlama: MLflow veya DVC ile deney takibi ve model yönetimi.**
* **API Servisleri: FastAPI veya Flask tabanlı gerçek zamanlı REST endpoint’leri.**
* **Kullanıcı Raporlama: Dinamik PDF/PowerPoint raporu ve e-posta bildirim modülleri.**
* **Mobil Erişim: React Native veya Flutter ile mobil uygulama geliştirme.**
* **Bulut Altyapısı: Sunucusuz mimari (AWS Lambda) veya Kubernetes tabanlı ölçeklendirme.**