Mahmut Kerem Erden – Meme Kanseri Teşhis Sistemi Projesi

Bu proje, Sanayi ve Teknoloji Bakanlığı'nın Millî Teknoloji Hamlesi kapsamında yürütülen Yapay Zekâ Uzmanlık Programı çerçevesinde, Baykar – Cezeri firması tarafından verilen bir bitirme projesidir.

İçindekiler

Proje Klasörünün Ağaç Yapısı:

.venv: Projede kullanılan Python kütüphanelerinin yer aldığı sanal ortam.

chroma_db: RAG tabanlı chatbot için kullanılan vektör veri tabanı.

data/:

- iyi.csv ve kotu.csv: Model değerlendirmesi için örnek olarak yüklenebilecek CSV dosyaları.
- data.csv: Ham veri seti.
- feature_order.csv: Özellik sırasını korumak için kullanılır.

docs/:

- meme-kanseri-rehberi.pdf: Chatbot'un yanıt üretmesi için kullanılan kaynak PDF.

notebooks/:

- **Proje.ipynb:** Makine öğrenmesi ve derin öğrenme modellerinin oluşturulduğu Jupyter Notebook.

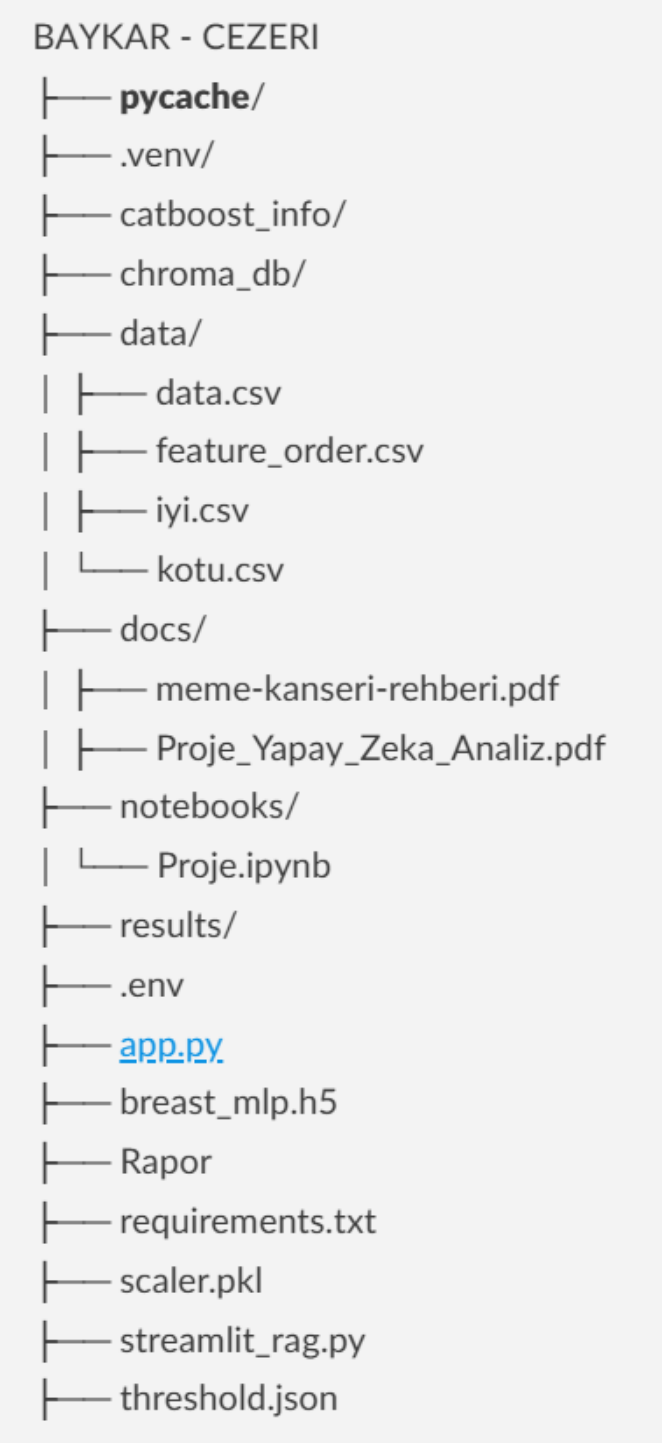
app.py: Streamlit ile yazılmış ana uygulama dosyası; arayüz ve model entegrasyonları burada yapılır.

breast_mlp.h5: Eğitilmiş MLP (derin öğrenme) modelinin ağırlıklarını içeren dosya.

scaler.pkl: Verileri normalize eden StandardScaler nesnesi.

streamlit_rag.py: RAG tabanlı chatbot sistemine ait tüm kodların bulunduğu dosya.

threshold.json: MLP modelinin sınıflandırma eşiğini (threshold) belirten JSON dosyası.



1. PROJE ÖZETİ

Bu proje kapsamında, **Breast Cancer Wisconsin** veri seti kullanılarak meme kanseri teşhisine yönelik **makine öğrenmesi** ve **derin öğrenme** modelleri geliştirilmiştir. Proje sürecinde öncelikle veri seti üzerinde detaylı analizler yapılmış, ardından çeşitli geleneksel sınıflandırma algoritmaları (Random Forest, SVM, Lojistik Regresyon, vb.) ve derin öğrenme tabanlı MLP modeli eğitilerek karşılaştırmalı değerlendirmeler gerçekleştirilmiştir.
Eğitilen modellerin çıktıları, kullanıcı dostu bir arayüz sunan **Streamlit** platformu ile entegre edilerek, teşhis tahmini, veri analizi, model başarımı ve etkileşimli bir **chatbot** üzerinden kullanılabilir hâle getirilmiştir. Proje; hem veri bilimi süreçlerini hem de uygulama geliştirme becerilerini entegre şekilde bir araya getiren kapsamlı bir yapay zekâ çalışmasıdır.

📊 2. VERİ ANALİZİ (Notebook Aşaması – Klasör: notebooks/Proje.ipynb)

Bu bölüm üç alt başlıkta ele alınmaktadır:

1. **Veri Setinin Tanıtımı**
2. **Veri Temizleme ve Ön İşleme**
3. **Görselleştirme ve Keşifsel Veri Analizi (EDA)**

📌 2.1 Veri Setinin Tanıtımı

Bu projede kullanılan veri seti, meme kanseri teşhisine yönelik olarak **Wisconsin Breast Cancer (Diagnostic)** veri setidir. Veri seti, 30 sayısal özellik ve bir hedef değişken (diagnosis) içermektedir.
Her bir satır, bir hastaya ait tümörün çeşitli biyolojik özelliklerini temsil etmektedir. Hedef değişken, tümörün iyi huylu (Benign – B) veya kötü huylu (Malignant – M) olduğunu belirtmektedir.

Veri setindeki başlıca sütunlar şunlardır:

- **radius_mean, texture_mean, perimeter_mean, area_mean, smoothness_mean** … gibi 30 temel özellik
- **id:** Hasta ID'si (analiz dışı bırakılmıştır)
- **diagnosis:** Hedef etiket (M = malign, B = benign)

📌 2.2 Veri Temizleme ve Ön İşleme

Veri seti yüklendikten sonra aşağıdaki işlemler uygulanmıştır:

- **Gereksiz sütunlar silinmiştir:**
  id ve Unnamed: 32 gibi sınıflandırmada anlamlı olmayan sütunlar çıkarılmıştır.
- **Eksik veri kontrolü yapılmıştır:**
  Veri setinde eksik veya boş hücre bulunmadığı tespit edilmiştir.
- **Hedef değişken dönüştürülmüştür:**
  diagnosis sütunu "M" → 1, "B" → 0 olacak şekilde ikili sınıflandırma için sayısal hale getirilmiştir.
- **Özellikler ve hedef değişken ayrılmıştır:**
  X değişkeni tüm giriş özelliklerini, y ise hedef sınıf etiketlerini temsil etmektedir.
- **Eğitim ve test verileri ayrılmıştır:**
  Veri seti %80 eğitim, %20 test olacak şekilde train_test_split() ile ayrılmış ve sınıf dağılımının korunması için stratify=y parametresi kullanılmıştır.

📌 2.3 Keşifsel Veri Analizi (EDA) ve Görselleştirmeler

Veri seti üzerinde ilk olarak dağılım, korelasyon ve boyut indirgeme gibi analizler yapılmıştır:

🔶 Sınıf Dağılımı

- diagnosis sütunundaki örnek sayıları görselleştirilmiştir.
- Veri dengeliye yakın olup, benign sınıf örnekleri daha fazladır.

🔶 Korelasyon Matrisi

- Özellikler arasındaki ilişkiyi görmek için ısı haritası (heatmap) çıkarılmıştır.
- Bazı öznitelikler arasında yüksek korelasyon (örneğin radius, perimeter, area) gözlemlenmiştir.

🔶 Boxplot & Histogram – radius_mean

- radius_mean özelliği için hem kutu grafiği hem histogram çizilmiştir.
- Malign tümörlerde bu değerin genel olarak daha yüksek olduğu görülmüştür.

🔶 PCA – 2 Boyutlu Projeksiyon

- Veriler StandardScaler ile normalize edilmiş, ardından PCA ile iki boyuta indirgenmiştir.
- PCA görselleştirmesinde iki sınıfın ayrılabildiği gözlemlenmiştir, bu da model eğitimi için pozitif bir işarettir.

🤖 3. MODEL GELİŞTİRME (Notebook Aşaması– Klasör: notebooks/Proje.ipynb)

Bu aşamada, meme kanseri teşhisini gerçekleştirmek üzere hem **geleneksel makine öğrenmesi algoritmaları** hem de **derin öğrenme** temelli bir model (MLP) geliştirilmiştir. Tüm modeller, aynı ön işlenmiş veri yapısı üzerinde eğitilmiş ve test verisi üzerinde karşılaştırılmıştır.

📌 3.1 Kullanılan Makine Öğrenmesi Modelleri

Aşağıdaki algoritmalar**, Pipeline** yapısı kullanılarak eğitilmiş ve test edilmiştir:

Model	**Doğruluk (Accuracy)**
Random Forest	**%97.4**
Lojistik Regresyon	%96.5
SVM (Linear Kernel)	%96.5
KNN	%95.6
Naive Bayes	%92.1
XGBoost	**%97.4**
CatBoost	%96.5

Ortak Noktalar:

- Tüm modeller, **StandardScaler** ile normalize edilmiş verilerle eğitilmiştir.
- **train_test_split()** ile ayrılmış test verisi üzerinde doğruluk, sınıflandırma raporu ve karışıklık matrisi değerlendirmeleri yapılmıştır.
- Görselleştirmelerde **sns.heatmap()** (karışıklık matrisi) kullanılarak her modelin performansı net biçimde sunulmuştur.

📌 3.2 Hiperparametre Optimizasyonu

Random Forest modeli için **GridSearchCV** yöntemiyle **hiperparametre optimizasyonu** yapılmıştır.

- n_estimators, max_depth, min_samples_split, min_samples_leaf gibi parametreler denenmiştir.
- **5 katlı çapraz doğrulama** (cv=5) ile en iyi sonuç veren parametre kombinasyonu bulunmuştur.
- Elde edilen en iyi model, test seti üzerinde tekrar değerlendirilmiştir.

Bu süreç, modelin overfitting yapmadan en iyi genel doğruluğa ulaşmasını sağlamıştır.

📌 3.3 Derin Öğrenme Modeli – MLP

Bu projede son adım olarak **çok katmanlı yapay sinir ağı (MLP)** modeli geliştirilmiştir.

🔷 **Mimarisi:**

- **Giriş Katmanı:** 30 giriş özelliği
- **2 Tam Bağlantılı Katman:** 64 ve 32 nöronlu, ReLU aktivasyonlu
- **Batch Normalization** ve **Dropout** katmanları ile düzenleme
- **Çıkış Katmanı:** Sigmoid aktivasyonlu tek nöron (ikili sınıflandırma)

🔷 **Eğitim:**

- Kayıp fonksiyonu: **binary_crossentropy**
- Optimizasyon: **adam**
- Metrikler**: accuracy, AUC**
- Eğitim stratejisi:
    - EarlyStopping (erken durdurma)
    - ReduceLROnPlateau (öğrenme oranı azaltma)
    - class_weight: sınıf dengesizliğine karşı ağırlıklandırma

🔷 **Eşik Belirleme:**

- Tahmin çıktılarındaki olasılıklar için precision_recall_curve kullanılarak optimum eşik (threshold) değeri hesaplanmıştır.
- Bu eşik değeri, **threshold.json dosyasına kaydedilmiştir.**

🔷 **Değerlendirme:**

- Karışıklık matrisi ve sınıflandırma raporu detaylı şekilde çıkarılmıştır.
- **ROC eğrisi ve AUC skoru**, modelin genel başarımını göstermektedir.

📌 3.4 Model Artefaktlarının Kaydedilmesi

Eğitilen derin öğrenme modeli ve destekleyici dosyalar kaydedilmiştir:

Artefakt	Açıklama	Dosya Adı
Model ağırlıkları	MLP modeli (.h5)	breast_mlp.h5
Ölçekleyici nesne	StandardScaler nesnesi (.pkl)	scaler.pkl
Eşik değeri	Belirlenen karar eşiği (.json)	threshold.json
Özellik sırası	Sütun isimlerinin sıralı listesi (.csv)	feature_order.csv

Bu dosyalar, Streamlit uygulamasına doğrudan entegre edilerek tahmin ve arayüz işlemlerinde kullanılmaktadır.

💻 4. Uygulama Geliştirme (Streamlit – Dosya: app.py)

Bu proje kapsamında geliştirilen makine öğrenmesi ve derin öğrenme modellerinin kullanıcı tarafından kolaylıkla kullanılabilmesi amacıyla, Python tabanlı **Streamlit** kütüphanesi ile interaktif bir web arayüzü oluşturulmuştur. Arayüz; **veri yükleme, model tahmini alma, görselleştirme** ve **chatbot** etkileşimi gibi işlevleri içermektedir.

📌 4.1 Uygulamanın Yapısı ve Sekmeler

Arayüz 4 ana sekmeden oluşmaktadır:

1. **Teşhis Tahmini:**
   Kullanıcıyı karşılayan bir mesaj ve proje hakkında temel bilgiler içerir. Kullanıcılar, yeni bir örnek veri yükleyerek (CSV formatında) ya da manuel olarak özellikleri girerek modelden tahmin alabilir. Tahmin sonucunda tümörün iyi huylu (benign) ya da kötü huylu (malign) olma olasılığı gösterilir.
2. **SağlıkGPT Chatbot:**
   Kullanıcılar, meme kanseri hakkında sık sorulan sorular üzerinden örnek mesajlar gönderebilir. Chatbot, bu sorulara kaynaklara dayalı, açıklayıcı ve destekleyici yanıtlar verir.
3. **Veri Analizi:**
   Veri setinin genel özelliklerini ve sınıf dağılımını gösteren grafikler sunar.
4. **Model Değerlendirmesi:**
   Modelin başarı yüzdesi, karışıklık matrisi ve F1 Score gibi sonuçlar görsel olarak sunulmaktadır.

📌 4.2 Teknik Detaylar

- Uygulama Streamlit ile geliştirildi.
- Arayüz Bootstrap teması ile özelleştirildi.
- breast_mlp.h5, scaler.pkl, threshold.json, feature_order.csv dosyaları doğrudan arayüz ile entegre edilmiştir.
- Chatbot, temel LLM entegrasyonu üzerinden çalışmakta; sadece meme kanseriyle sınırlı bilgi sunmaktadır.
- Kullanıcıdan alınan girişler, önceden kaydedilen özellik sırasına göre işlenip modele yönlendirilmiştir.

📌 4.3 Kullanıcı Deneyimi ve Tasarım

- Kullanıcı dostu butonlar, sade ikonlar ve rehber metinlerle desteklenen bir yapı sunulmuştur.
- Her sekmede bilgi kutucukları ile kullanıcı yönlendirilmiştir.
- Sağ alt köşeye sabitlenen "Örnek Sorular" bölümü, kullanıcıya chatbot'u nasıl kullanabileceğini hatırlatır.

📌 4.4 Kullanım Örnekleri

- Kullanıcı 'data' klasörü altında bulunan 'iyi.csv' ve 'kötü.csv' örneklerini yükleyerek model çıktısını test edebilir.
- "Model Değerlendirmesi" sekmesinde modelin başarımı doğrudan görselleştirilmiştir.
- Chatbot'a şu tarz sorular yöneltilebilir:
    - "Meme kanseri nedir?"
    - "Meme kanserinden korunma yolları nelerdir?"
    - "Meme kanserinde belirtiler nelerdir?"

💬 5. LLM Destekli Chatbot (RAG + Gemini 2.5)

Bu projede, kullanıcıların meme kanseriyle ilgili sıkça sorulan sorulara yanıt alabilmesi amacıyla büyük dil modeli (LLM) tabanlı bir chatbot geliştirilmiştir. Chatbot sistemi, **Google Gemini 2.5** modeliyle entegre edilmiş ve **RAG** (Retrieval-Augmented Generation) yaklaşımı kullanılarak bilgi destekli cevaplama sağlanmıştır.

Teknik Özellikler:

- Sadece **meme kanseri** konusuyla sınırlı cevaplar üretir. (PDF dizini – docs/meme-kanseri-rehberi.pdf)
- Kullanıcının mesajı, veri tabanındaki (ChromaDB) bilgilerle eşleştirilerek modele sunulur (RAG).
- Chat geçmişi korunur ve yanıtlar sade, açıklayıcı ve kaynaklı olacak şekilde biçimlendirilir.
- Chatbot arayüzü Streamlit ile entegre şekilde çalışır.

Bu yapı sayesinde uygulama, yalnızca tahmin sunmakla kalmaz, aynı zamanda bilgi verme işlevi de üstlenir.

✅ 6. Sonuç ve Değerlendirme

Bu projede, Wisconsin Meme Kanseri veri seti üzerinde hem makine öğrenmesi hem de derin öğrenme modelleri kullanılarak tümörlerin iyi huylu veya kötü huylu olup olmadığının tahminine yönelik bir teşhis sistemi geliştirilmiştir. Elde edilen sonuçlar, geliştirilen modellerin yüksek doğruluk oranlarına ulaştığını ve özellikle Random Forest, XGBoost ve MLP gibi modellerin %97'nin üzerinde başarı sağladığını göstermiştir.

Projede ayrıca geliştirilen Streamlit tabanlı arayüz sayesinde kullanıcılar:

- Veri yükleyerek veya manuel giriş yaparak tahmin alabilmekte,
- Görselleştirmelerle veriyi analiz edebilmekte,
- Chatbot üzerinden konuya dair sorularına yanıt alabilmektedir.

Bu çok yönlü yapı, yalnızca bir sınıflandırma modeli sunmakla kalmamış; kullanıcı etkileşimi, veri görselleştirme ve bilgi destekli yanıt mekanizmalarını da içeren bütüncül bir yapay zekâ uygulamasına dönüşmüştür.

Gelecekte bu sistem; daha büyük veri setleri, farklı kanser türleri ve klinik verilerle genişletilerek, sağlık alanında daha etkili karar destek sistemlerine dönüştürülebilir.