Meme Kanseri Tahmini Projesi Detaylı Dokümantasyonu

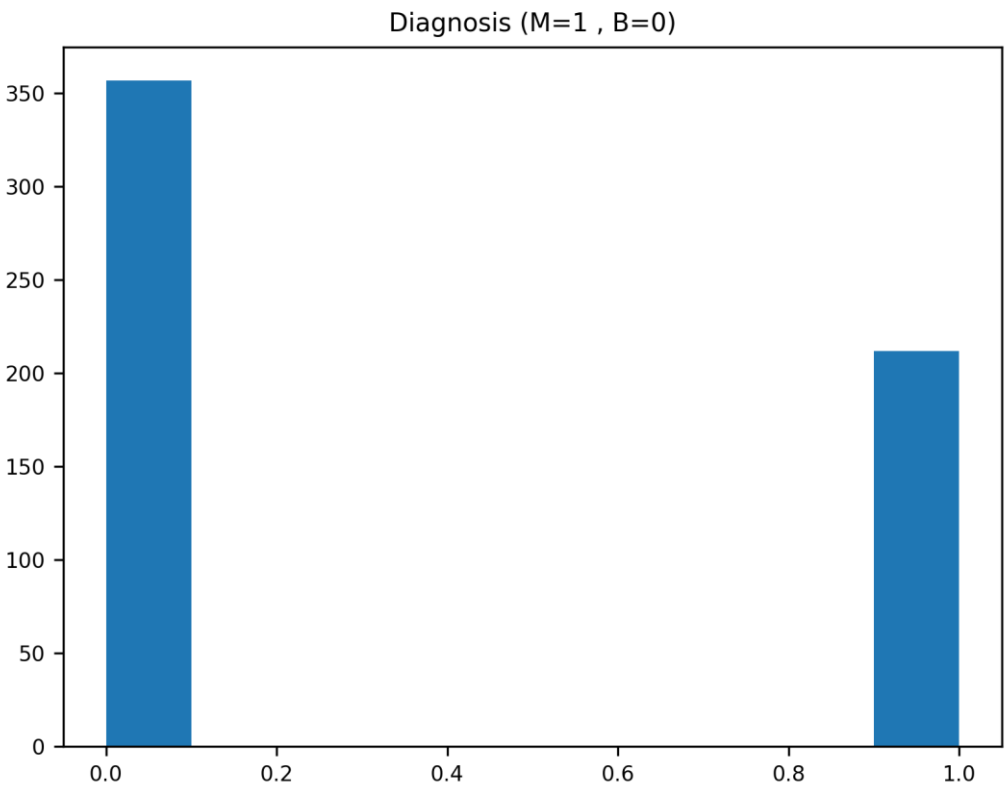
1. Giriş

Bu doküman, meme kanseri teşhisi için kullanılan makine öğrenmesi modellerini detaylandırır. Proje kapsamında, üç farklı makine öğrenmesi modeli kullanılmıştır: Random Forest, XGBoost ve CatBoost. Modellerin performansları karşılaştırılarak en iyi model belirlenmiştir.

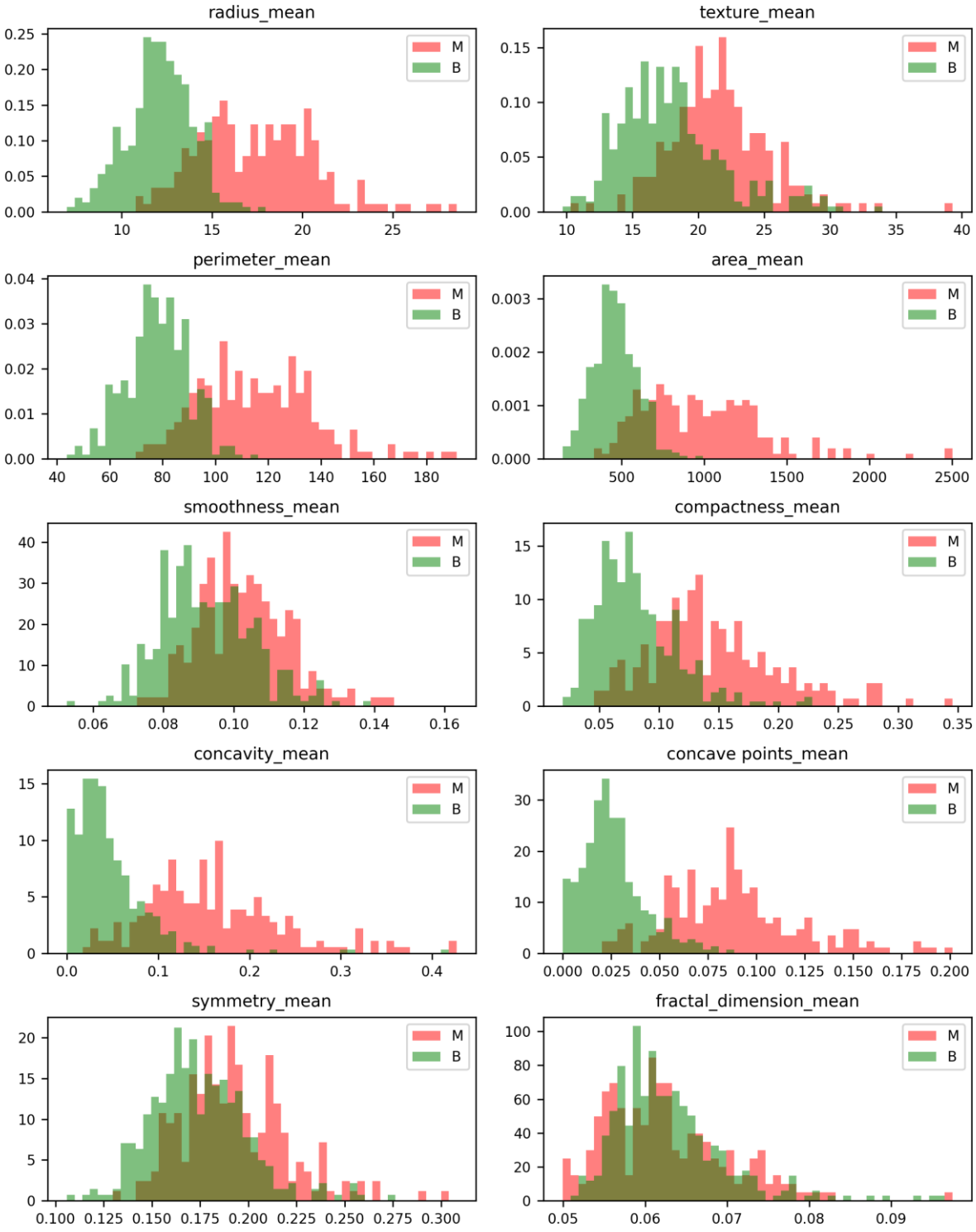
2. Veri Seti ve Özellikleri

Veri seti, meme kanseri teşhisi için çeşitli özellikler içermektedir ve Wisconsin Breast Cancer Dataset kullanılmıştır. Özellikler şunlardır:

- Kimlik numarası
- Tanı (M = kötü huylu, B = iyi huylu)



- 30 adet özellik (örneğin, yarıçap, doku, çevre, alan, düzgünlük, kompaktlık, içbükeylik, simetri, fraktal boyut)



3. Veri Hazırlığı

Veri seti, makine öğrenmesi modellerine uygun hale getirilmesi için bazı ön işlemlerden geçirilmiştir:

1. **Etiketleme ve Dönüştürme:** Tanı etiketleri (M ve B), **LabelEncoder** kullanılarak sayısal değerlere dönüştürülmüştür.
2. **Ölçeklendirme:** Özellikler, **StandardScaler** kullanılarak standartlaştırılmıştır. Bu adım, her bir özelliğin ortalamasını 0 ve standart sapmasını 1 olacak şekilde ölçeklendirir.

4. Modelleme Süreci

4.1 XGBoost (Extreme Gradient Boosting)

XGBoost, gradient boosting framework'ü kullanarak hızlı ve performanslı tahminler yapabilen bir modeldir. Bu modelin avantajları arasında yüksek doğruluk, hız ve overfitting'i azaltma yeteneği bulunmaktadır.

Modelin Eğitimi:

- **Başlangıç Modeli:** Varsayılan parametrelerle XGBoost modeli eğitilmiştir.
- **Hiperparametre Optimizasyonu:** GridSearchCV kullanılarak en iyi hiperparametreler belirlenmiştir. Hiperparametreler arasında max_depth, n_estimators, subsample ve learning_rate bulunmaktadır.

En İyi Parametreler:

- max_depth: 5
- n_estimators: 100
- subsample: 0.8
- learning_rate: 0.02

Model Performansı:

- Doğruluk: %96.50

4.2 Random Forest

Random Forest, birden fazla karar ağacı oluşturup bu ağaçların çıktılarının ortalamasını alarak tahmin yapan bir topluluk öğrenme yöntemidir. Çok sayıda ağacın kullanılması, modelin overfitting yapma riskini azaltır ve genel performansını artırır.

Modelin Eğitimi:

- **Başlangıç Modeli:** Varsayılan parametrelerle Random Forest modeli eğitilmiştir.

- **Hiperparametre Optimizasyonu:** GridSearchCV kullanılarak en iyi hiperparametreler belirlenmiştir. Hiperparametreler arasında max_depth, max_features ve n_estimators bulunmaktadır. **En İyi Parametreler:**
    - max_depth: 10
    - max_features: 8
    - n_estimators: 100

Model Performansı:

- Doğruluk: %96.50

4.3 CatBoost

CatBoost, özellikle kategorik verilerle iyi performans gösteren bir gradient boosting algoritmasıdır. CatBoost, kategorik özellikleri ön işleme gerek kalmadan kullanabilme yeteneğine sahiptir ve bu nedenle veri ön işleme sürecini basitleştirir.

Modelin Eğitimi:

- **Başlangıç Modeli:** Varsayılan parametrelerle CatBoost modeli eğitilmiştir.
- **Hiperparametre Optimizasyonu:** GridSearchCV kullanılarak en iyi hiperparametreler belirlenmiştir. Hiperparametreler arasında depth, iterations ve learning_rate bulunmaktadır.

En İyi Parametreler:

- depth: 3
- iterations: 500
- learning_rate: 0.02

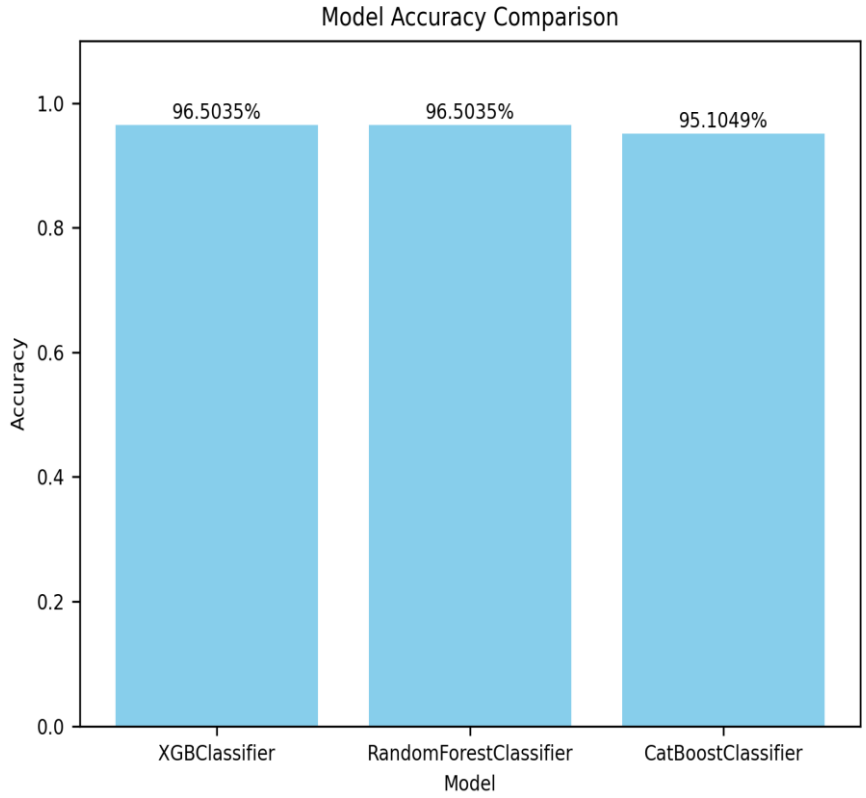
Model Performansı:

- Doğruluk: %95.10

5. Model Karşılaştırması

Modellerin performansları, doğruluk oranlarına göre karşılaştırılmıştır. Her bir modelin doğruluk oranı aşağıda belirtilmiştir:

- **XGBoost:** %96.50
- **Random Forest:** %96.50
- **CatBoost:** %95.10



6. Sonuç ve Değerlendirme

Bu projede, meme kanseri tahmini için üç farklı makine öğrenmesi modeli kullanılmış ve değerlendirilmiştir. XGBoost ve Random Forest modelleri, benzer yüksek doğruluk oranlarına ulaşmış olup, CatBoost modeli ise biraz daha düşük doğruluk oranına sahiptir.

6.1 Veri Setinin Detaylı Özellikleri

- **Yarıçap:** Tüm hücrelerin ortalama yarıçapı.
- **Doku:** Hücrelerin dokusunun ortalama değerleri.
- **Çevre:** Hücrelerin çevresinin ortalama değerleri.
- **Alan:** Hücrelerin alanının ortalama değerleri.
- **Düzgünlük:** Hücrelerin düzgünlük oranlarının ortalama değerleri.
- **Kompaktlık:** Hücrelerin kompaktlık oranlarının ortalama değerleri.
- **İçbükeylik:** Hücrelerin içbükeylik oranlarının ortalama değerleri.
- **Simetri:** Hücrelerin simetri oranlarının ortalama değerleri.
- **Fraktal Boyut:** Hücrelerin fraktal boyut oranlarının ortalama değerleri.

6.2 Performans Ölçütleri ve Değerlendirme

Modellerin doğruluk oranları, doğru tahminlerin toplam tahminlere oranı olarak hesaplanmıştır. Bu ölçüt, modelin genel performansını değerlendirmek için kullanılmıştır. Doğruluk oranları aşağıdaki gibidir:

- XGBoost ve Random Forest modelleri: %96.50
- CatBoost modeli: %95.10

6.3 Hiperparametrelerin Anlamı

- **max_depth:** Karar ağaçlarının maksimum derinliği. Daha derin ağaçlar, daha karmaşık karar sınırları oluşturabilir ancak overfitting riskini artırabilir.
- **n_estimators:** Toplam ağaç sayısı. Daha fazla ağaç, modelin daha iyi genelleştirmesine yardımcı olabilir ancak eğitim süresini uzatabilir.
- **subsample:** Her bir ağacın oluşturulması sırasında kullanılacak örnekleme oranı. Bu oran, modelin overfitting yapmasını engellemeye yardımcı olur.
- **learning_rate:** Öğrenme hızı. Daha düşük bir öğrenme hızı, daha yavaş ancak daha genel bir öğrenme süreci sağlar.
- **max_features:** Her bir ağacın oluşturulması sırasında kullanılacak maksimum özellik sayısı. Bu parametre, modelin varyansını ve overfitting riskini azaltır.

Bu dokümantasyon, meme kanseri tahmini için kullanılan modellerin eğitim, optimizasyon ve karşılaştırma süreçlerini kapsamlı bir şekilde açıklamaktadır. Her bir modelin avantajları ve performansları detaylı olarak ele alınmıştır.