T.C.
ÜSKÜDAR
ÜNİVERSİTESİ

DRY BEAN DATASET ÜZERİNDE MAKİNE ÖĞRENMESİ PİPELİNE UYGULAMASI:

VERİ ÖN İŞLEME, BOYUT İNDİRGEME, MODELLEME VE ROC ANALİZİ

234329054 - VAHAP AYZET

1.İçindekiler

2. Projenin Amacı

Bu projenin amacı, gerçek bir veri kümesi olan **Dry Bean Dataset** üzerinden tam bir **Makine Öğrenmesi Pipeline'ı** uygulayarak öğrenciye uçtan uca bir ML deneyimi kazandırmaktır. Bu kapsamda:

- Eksik veri ve aykırı değer işlemleri gerçekleştirilmiş,
- Sayısal ve kategorik değişkenler uygun şekilde dönüştürülmüş,
- Boyut indirgeme teknikleri (PCA ve LDA) uygulanmış,
- Birden fazla makine öğrenmesi modeli değerlendirilmiş,
- Nested cross-validation ile modellerin genellenebilirliği ölçülmüş,
- ROC eğrileri çizilerek en iyi modelin görsel olarak da doğruluğu test edilmiştir.

Bu süreç, makine öğrenmesinde model başarısını etkileyen tüm kritik aşamaları kapsar.

3. Veri Kümesi ve Açıklaması

Çalışmada kullanılan veri seti, UCI Machine Learning Repository'de yer alan **Dry Bean Dataset**'tir. Veri seti, yedi farklı kuru fasulye türüne (Barbunya, Bombay, Cali, Dermason, Horoz, Seker, Sira) ait çeşitli morfolojik özellikleri içermektedir.

Toplamda **13.611 örnek** ve **16 sayısal özellik** bulunmaktadır. Her örnek, aşağıdaki kategorilere ayrılmıştır:

- Area, Perimeter, MajorAxisLength, MinorAxisLength, AspectRatio, Eccentricity, ConvexArea, EquivDiameter, Extent, Solidity, Roundness, Compactness, ShapeFactor1-4
- Class (etiket – fasulye türü)

Veri seti dengeli olup, çok sınıflı sınıflandırma problemi olarak ele alınmıştır.

4. Veri Ön İşleme

Bu aşamada, veri kümesi üzerinde temel temizleme ve hazırlık işlemleri yapılmıştır. Süreç dört ana başlık altında yürütülmüştür:

4.1 Eksik Verilerle Başa Çıkma

Veri setinde başlangıçta eksik veri bulunmamakla birlikte, proje kapsamında belirlenen üç sütuna manuel olarak eksik veriler eklenmiştir:

- Area ve Perimeter sütunlarına %5 oranında

- Eccentricity sütununa %35 oranında eksik değer (NaN) eklenmiştir

Eksik değerlerle başa çıkma stratejileri:

- %5 eksik veri içeren Area ve Perimeter sütunları, ilgili sütunun **ortalama değeriyle** doldurulmuştur.
- %35 eksik veri içeren Eccentricity sütunu için satır bazlı silme yöntemi uygulanmıştır (bilgi kaybını minimize etmek için sütun silinmemiştir).

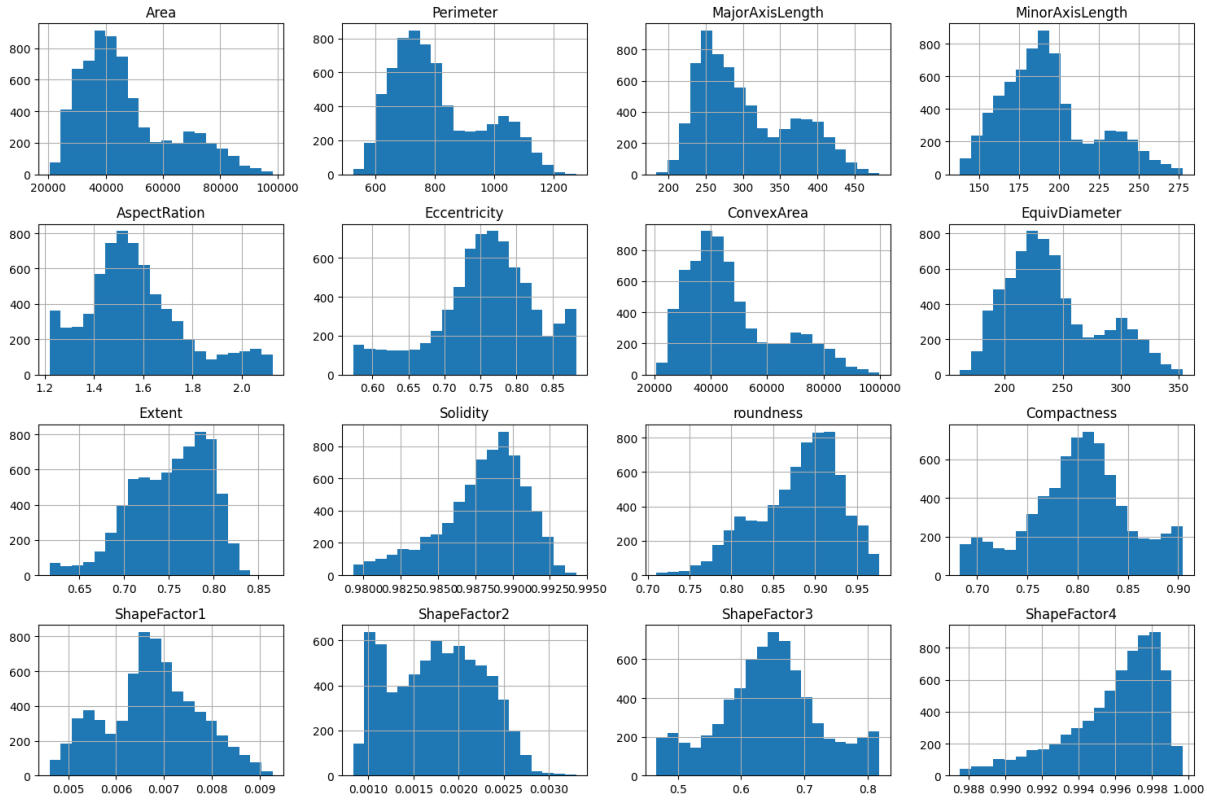
4.2 Aykırı Değerlerin Tespiti ve İşlenmesi

Aykırı değerler, **IQR (Interquartile Range)** yöntemi ile tespit edilmiştir. Sayısal sütunlar için:

- Q1 (1. çeyrek), Q3 (3. çeyrek) ve IQR hesaplanmış,
- Alt ve üst sınır dışındaki değerler aykırı olarak belirlenmiştir.

Tespit edilen aykırı değerler, ilgili gözlemlerle birlikte veri setinden çıkarılmıştır.

Aşağıda, aykırı değerler temizlendikten sonra özelliklerin dağılımı verilmiştir.



4.3 Özellik Ölçekleme (Feature Scaling)

Tüm sayısal özellikler, **StandardScaler** ile ölçeklenmiştir. Bu yöntem:

- Ortalamayı 0'a, standart sapmayı 1'e çeker

- Özellikle mesafe tabanlı modeller (Logistic Regression, LDA, PCA) için gereklidir

Bu sayede tüm özelliklerin katkısı eşitlenmiş ve modelleme adımında öğrenme dengesi sağlanmıştır.

4.4 Kategorik Verilerin Kodlanması

Veri setinde tek kategorik sütun olan Class sütunu:

- **LabelEncoder** ile numerik değerlere çevrilmiştir (örnek: 'CALI' → 1, 'SIRA' → 6).
- OVA (One-vs-All) tabanlı ROC analizinde kullanılmak üzere ayrıca **One-hot encoding** ile binarize edilmiştir.

5. Özellik Seçimi ve Boyut İndirgeme

Veri seti yüksek boyutlu bir yapıya sahip olduğundan, boyut indirgeme teknikleriyle daha kompakt ve anlamlı temsiller elde edilmiştir. Bu amaçla iki yöntem uygulanmıştır:

5.1 Principal Component Analysis (PCA)

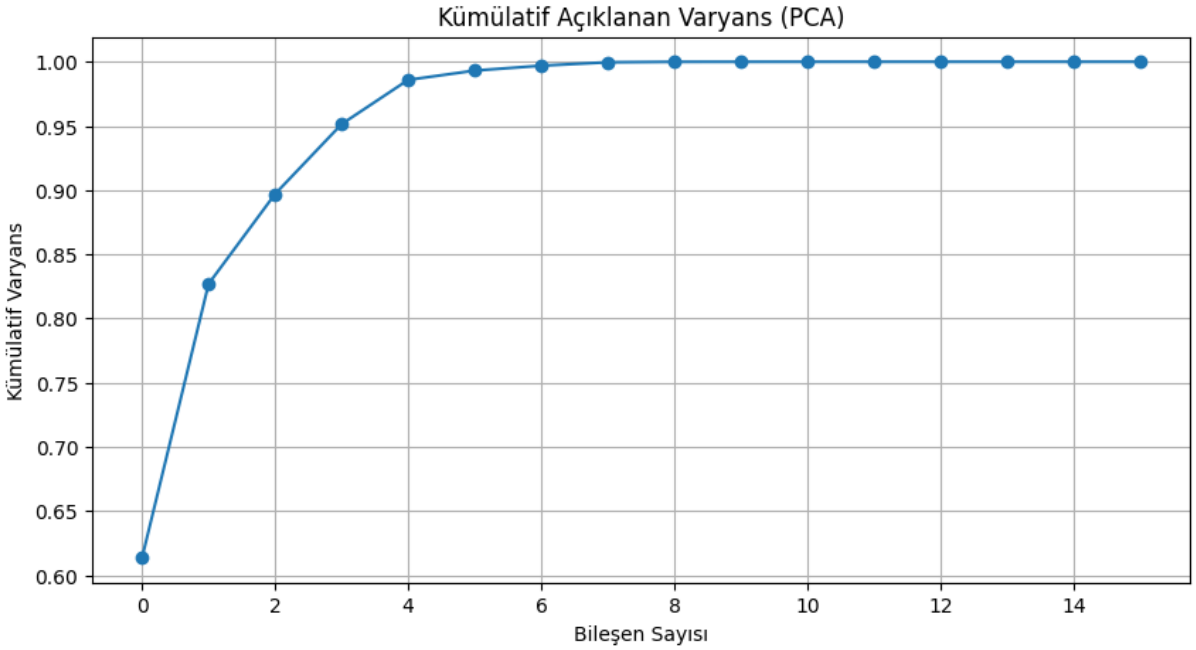
PCA, varyansı maksimize ederek yeni öznitelikler oluşturan bir boyut indirgeme yöntemidir. Uygulama adımları:

- Ölçeklenmiş sayısal veriler üzerinde PCA uygulanmıştır.
- Açıklanan varyans oranlarına göre ilk **4 bileşen**, toplam varyansın %98'inden fazlasını açıklamaktadır.
- Bu nedenle modelleme sürecinde ilk 4 bileşen seçilmiştir.

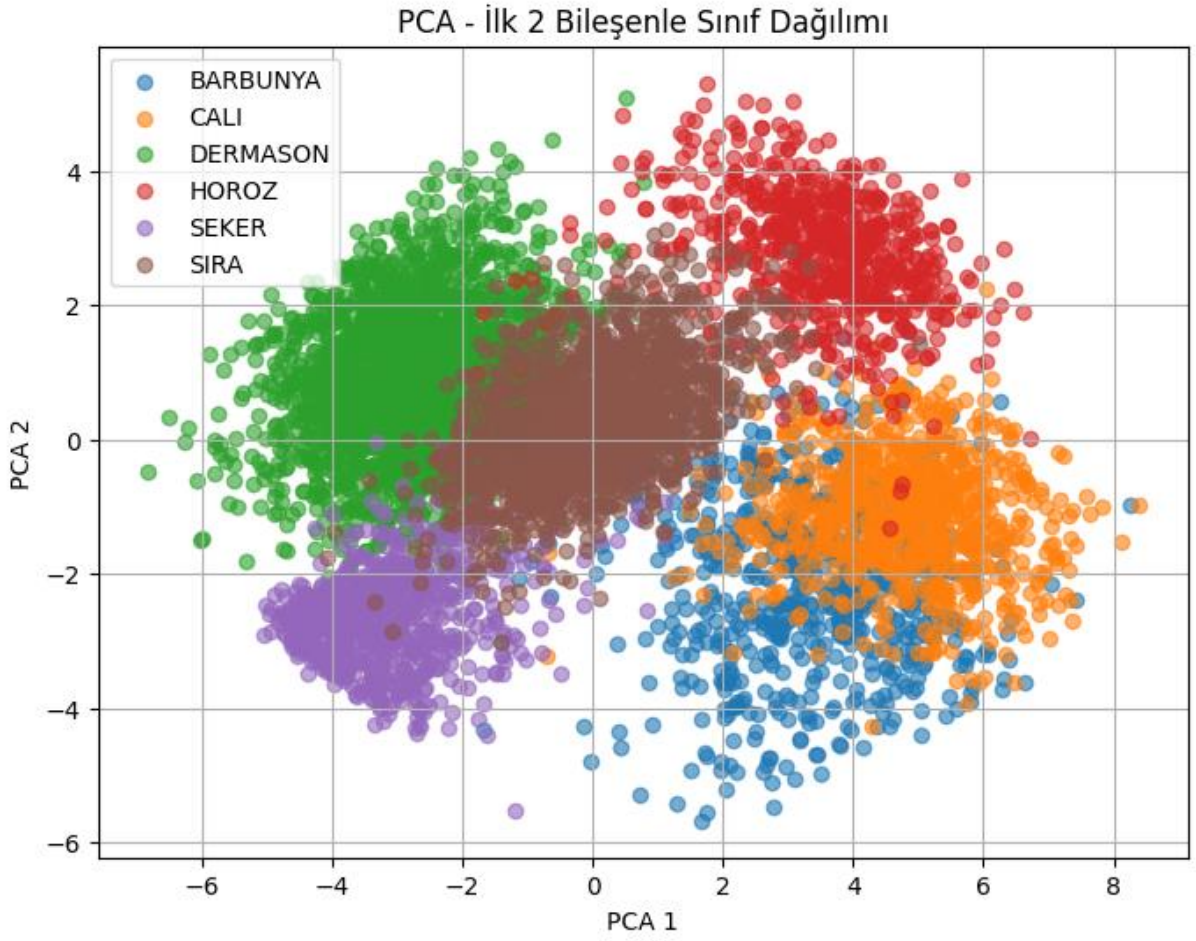
Ek olarak:

- İlk iki PCA bileşeni kullanılarak 2D bir görselleştirme yapılmıştır.
- Görselleştirmede sınıflar kısmen ayrışmıştır, ancak bazı sınıflar (örneğin CALI ve SIRA) iç içe geçmiştir.

PCA uygulaması sonucunda elde edilen kümülatif varyans grafiği aşağıda gösterilmiştir. İlk 4 bileşen toplam varyansın %98'inden fazlasını açıklamaktadır.



İlk iki PCA bileşeniyle oluşturulan iki boyutlu dağılımda sınıflar kısmen ayrışmıştır.

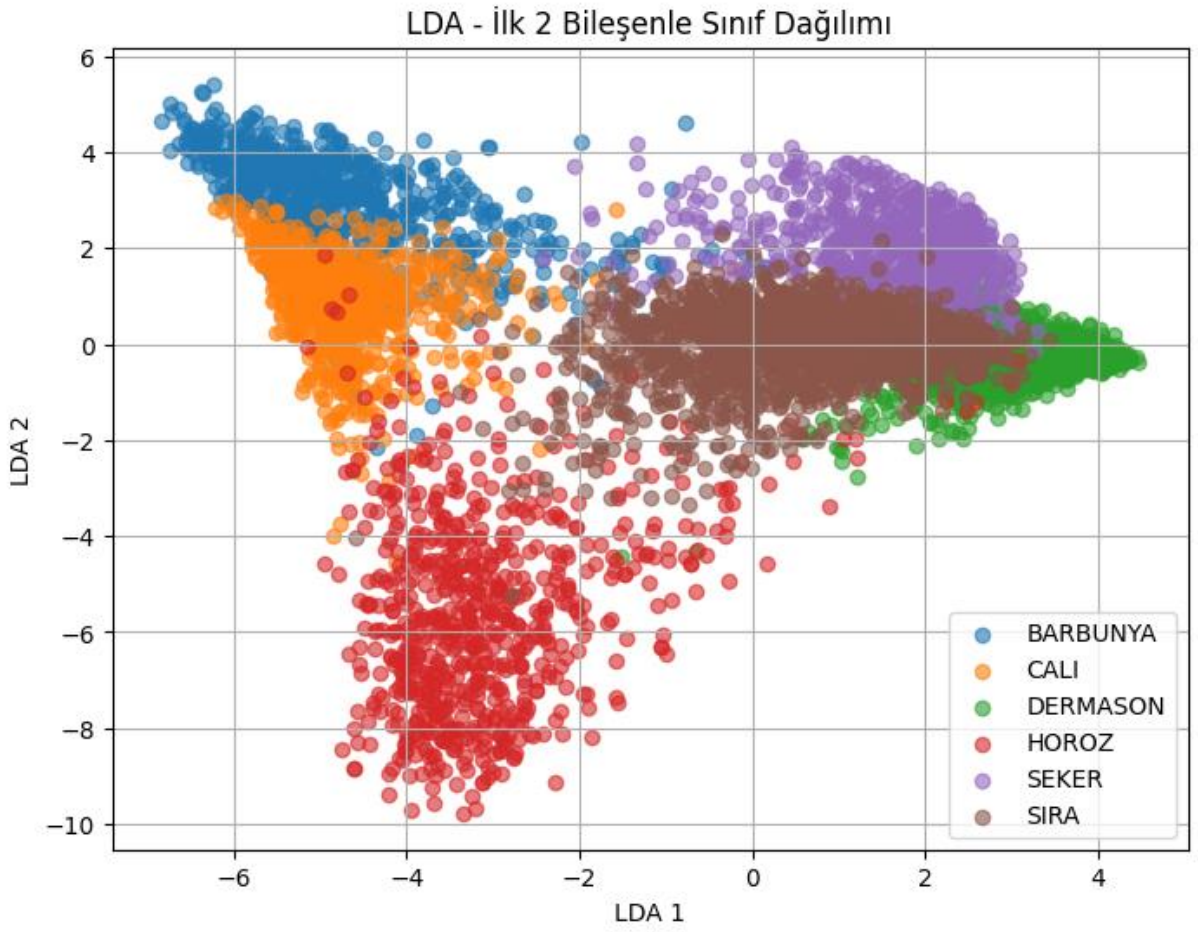


5.2 Linear Discriminant Analysis (LDA)

LDA, sınıflar arası ayrımı maksimize eden bir boyut indirgeme yöntemidir. PCA'den farklı olarak sınıf etiketlerini dikkate alır.

- Veri seti 6 sınıflı olduğundan, LDA ile maksimum 5 bileşen elde edilebilir.
- Çalışmada ilk **2 LDA bileşeni** kullanılarak görselleştirme yapılmıştır.
- Görselleştirmede sınıfların net bir şekilde ayrıştığı gözlemlenmiştir.

LDA temsili, PCA'ya göre sınıflar arası ayrımı daha iyi sağladığı için bazı modellerde daha yüksek başarı elde edilmiştir.



LDA ile oluşturulan iki boyutlu temsil sınıflar arasında daha belirgin ayrım sağlamıştır.

6. Modelleme ve Değerlendirme

Bu projede, model başarısını genellenebilir şekilde ölçmek için **Nested Cross-Validation** yapısı kullanılmıştır.

6.1 Nested Cross-Validation Yapısı

- **Dış döngü (outer loop):** 5 katmanlı (5-fold)
- **İç döngü (inner loop):** 3 katmanlı (3-fold), GridSearchCV ile hiperparametre optimizasyonu

Her dış döngüde:

- Eğitim ve test seti yeniden oluşturulmuştur
- İç döngüdeki hiperparametre aramaları sadece eğitim seti üzerinde yapılmıştır

Bu yapı sayesinde **model overfitting'i önlenmiş** ve **doğru karşılaştırma ortamı** sağlanmıştır.

6.2 Kullanılan Modeller

Her veri temsili için aşağıdaki 5 model uygulanmıştır:

- Logistic Regression
- Decision Tree
- Random Forest
- XGBoost
- Naive Bayes

6.3 Karşılaştırmalı Metrik Sonuçları

Aşağıdaki tablolar, tüm modellerin Accuracy, Precision, Recall ve F1-score metriklerini göstermektedir.

Model	Ham Veri (%)	PCA Temsili (%)	LDA Temsili (%)
Logistic Reg.	**91.21**	86.94	88.51
Decision Tree	88.44	85.90	**88.43**
Random Forest	91.02	88.31	89.73
XGBoost	**91.42**	88.28	89.72
Naive Bayes	87.99	86.46	**88.76**

6.4 ROC Eğrileri ve AUC Skorları

En başarılı model olan **XGBoost (Ham Veri)** için, sınıf bazlı **ROC eğrileri (One-vs-All)** çizilmiştir. AUC skorları aşağıdaki gibidir:

Sınıf	AUC Skoru
BARBUNYA	1.00
CALI	0.99
DERMASON	0.99
HOROZ	0.99
SEKER	1.00
SIRA	0.98

7. Sonuç ve Yorumlar

Bu projede, Dry Bean Dataset üzerinde uçtan uca bir makine öğrenmesi süreci gerçekleştirilmiş; veri ön işleme, boyut indirgeme, modelleme ve değerlendirme adımları sistematik biçimde uygulanmıştır.

En Başarılı Model

Nested cross-validation sonuçlarına göre:

- En yüksek doğruluk, F1 ve ROC-AUC skorları **XGBoost (Ham Veri)** modelinde elde edilmiştir.
- Bu model, özellikle sınıflar arası net ayrımı sayesinde %91.42 doğruluk ve %92.06 F1 skoruna ulaşmıştır.
- ROC eğrileri, modelin tüm sınıflarda güçlü tahmin yeteneğine sahip olduğunu göstermektedir (AUC skorları ≈ 0.99–1.00 aralığında).

Boyut İndirgeme Karşılaştırması

- **LDA**, düşük boyutlu (2D) temsiliyle beklenenden daha güçlü sonuçlar vermiştir ve çoğu modelde PCA'dan daha iyi performans göstermiştir.
- **PCA**, varyansı iyi korumasına rağmen sınıf ayrımını optimize etmediğinden performansı düşürmüştür.
- Ancak LDA temsili, model karmaşıklığını azaltarak daha hızlı eğitim süresi ve daha kompakt modeller sunmuştur.

Genel Değerlendirme

Bu çalışma sonucunda:

- Boyut indirgeme tekniklerinin model performansı üzerindeki etkisi net şekilde gözlemlenmiştir.

- Ensemble yöntemlerin (Random Forest, XGBoost) genellikle daha istikrarlı ve yüksek doğruluk verdiği görülmüştür.
- Nested cross-validation ile genellenebilirlik test edilmiş ve güvenilir sonuçlar elde edilmiştir.