# Kuru Fasulye Türlerinin Makine Öğrenmesi Yöntemleri ile Sınıflandırılması

Tarih: Mayıs 2025

*Bilgisayar Mühendisliği, Teknoloji Fakültesi, Gazi Üniversitesi, Ankara, Türkiye*

*Eren Büyükaşık - 23181616611*

*Ertunç Yontuç - 23181616602*

*Çağlar Aydemir - 23181616751*

**ÖZET** Bu projede, farklı kuru fasulye türlerinin morfolojik özelliklerine göre makine öğrenmesi yöntemleri ile sınıflandırılması amaçlanmıştır. Veri ön işleme aşamasında eksik değer kontrolü yapılmış, gerekirse uygun istatistiksel yöntemlerle doldurulmuş; kategorik etiketler sayısal formata çevrilmiş ve özellik ölçeklendirme (normalizasyon) işlemleri uygulanmıştır. Modelleme süreci, veri setinin %80’i eğitim ve %20’si test olmak üzere ayrılarak gerçekleştirilmiştir. Üç farklı algoritma kullanılmıştır: K-En Yakın Komşu (KNN), Destek Vektör Makineleri (SVM) ve Rastgele Orman (Random Forest). Sınıflandırma modellerinin performansları doğruluk oranı, F1-skoru ve karışıklık matrisi (confusion matrix) gibi metriklerle değerlendirilmiştir. Sonuçlara göre, KNN ve Random Forest modelleri %92’nin üzerinde doğruluk oranı ile en başarılı yöntemler olarak öne çıkmıştır. Geliştirilen modeller, kuru fasulye türlerinin sınıflandırılmasında etkili ve güvenilir birer araç olarak kullanılabilir.

**Anahtar Kelimeler:** Makine Öğrenmesi, Gıda Sınıflandırması, Kuru Fasulye, Görüntü İşleme, Tarımsal Veri Analizi

**Giriş**

Kuru fasulye, hem besin değeri hem de ekonomik değeri yüksek bir tarım ürünüdür. Farklı türleri; boyut, şekil ve renk gibi morfolojik özellikler bakımından birbirinden ayrılabilir. Ancak bu ayrım çoğu zaman manuel yöntemlerle yapılmakta ve bu durum zaman alıcı ve hataya açık olmaktadır. Günümüzde gelişen veri bilimi ve makine öğrenmesi teknikleri sayesinde bu tür sınıflandırma işlemleri otomatik, hızlı ve yüksek doğrulukla yapılabilmektedir.

Bu çalışmada, UCI “Dry Bean Dataset” veri seti kullanılarak çeşitli kuru fasulye türlerinin sınıflandırılması hedeflenmiştir. Veri ön işleme sürecinde eksik veri analizi, normalizasyon ve kategorik verilerin dönüştürülmesi gibi işlemler gerçekleştirilmiştir. Ardından, üç farklı makine öğrenmesi algoritması — K-En Yakın Komşu (KNN), Destek Vektör Makineleri (SVM) ve Rastgele Orman (Random Forest) — uygulanmış ve her biri %80 eğitim / %20 test ayrımı ile eğitilmiştir. Modellerin başarıları doğruluk oranı, F1 skoru ve karışıklık matrisi gibi metriklerle değerlendirilmiştir. Elde edilen sonuçlar, bu algoritmaların kuru fasulye türlerini sınıflandırmada oldukça etkili olduğunu ortaya koymuştur.

### 2.Materyal ve Yöntem

#### 2.1 Veri Seti

Bu çalışmada, UCI Machine Learning Repository’de yayımlanan “Dry Bean Dataset” veri seti kullanılmıştır. Veri seti, yedi farklı kuru fasulye türüne ait toplam 13 sayısal özelliği içeren 13.611 örnekten oluşmaktadır. Bu özellikler; fasulyenin alanı, çevresi, uzunluğu, genişliği, kompaktlığı, daireselliği gibi morfolojik ölçümleri kapsamaktadır. Etiket değişkeni olarak fasulye türü (örneğin; SIRA, SEKER, HOROZ, BARBUNYA vb.) bulunmaktadır.

#### 2.2 Veri Ön İşleme

* **Eksik Değer Analizi:** Veri setinde eksik değer bulunmamakla birlikte, aykırı değerlerin tespiti ve gerekirse çıkarılması sağlanmıştır.
* **Özellik Ölçeklendirme:** Tüm sayısal değişkenler z-score normalizasyon yöntemi ile standardize edilmiştir (ortalama = 0, standart sapma = 1).
* **Etiket Dönüşümü:** Fasulye türleri etiketleri, sınıflandırma algoritmaları ile uyumlu hale getirilmesi amacıyla LabelEncoder ile sayısal formata dönüştürülmüştür.

#### 2.3 Modelleme

Veri seti %80 eğitim ve %20 test olacak şekilde stratifiye edilerek bölünmüştür. Üç farklı makine öğrenmesi algoritması uygulanmıştır:

* **K-En Yakın Komşu (KNN):** En uygun k değeri, 5 katlı çapraz doğrulama (5-fold CV) ile belirlenmiştir.
* **Destek Vektör Makineleri (SVM):** RBF çekirdeği (kernel) kullanılmış, C ve gamma hiperparametreleri GridSearch yöntemi ile optimize edilmiştir.
* **Rastgele Orman (Random Forest):** n\_estimators ve max\_depth gibi temel parametreler, performans analizleri doğrultusunda ayarlanmıştır.

#### 

#### 

#### 2.4 Performans Değerlendirme

Sınıflandırma algoritmalarının başarısı, test verisi üzerinde doğruluk (accuracy), F1-skoru ve karışıklık matrisi (confusion matrix) metrikleri ile değerlendirilmiştir. Modellerin karşılaştırmalı performansları bu metrikler üzerinden kıyaslanmıştır.

#### 2.5 Uygulama Ortamı

Tüm veri işleme ve modelleme çalışmaları Python 3.8 ortamında gerçekleştirilmiştir. Kullanılan kütüphaneler şunlardır:

* Veri işlemleri için: Pandas, NumPy
* Modelleme için: scikit-learn
* Görselleştirme ve analiz sunumu için: Matplotlib, Seaborn
* Etkileşimli uygulama geliştirme için: Streamlit (isteğe bağlı olarak model sonuçlarının görsel sunumu için kullanılabilir)

### Veri Seti

Bu çalışmada kullanılan veri seti, UCI Machine Learning Repository üzerinden erişilebilen **“Dry Bean Dataset”** adlı açık veri kümesidir. Veri seti, Türkiye’de tarımı yapılan 7 farklı kuru fasulye türüne ait örneklerden oluşmaktadır: **SEKER, BARBUNYA, BOMBAY, CALI, DERMASON, HOROZ ve SIRA**. Her bir örnek, kuru fasulye tanelerinin görüntülerinden elde edilen 13 farklı morfolojik özelliğe sahiptir.

Bu özellikler şunlardır:

* **Area (Alan)**
* **Perimeter (Çevre)**
* **MajorAxisLength (Uzun eksen)**
* **MinorAxisLength (Kısa eksen)**
* **AspectRation (En-boy oranı)**
* **Eccentricity (Basmaçıklık)**
* **ConvexArea (Dışbükey alan)**
* **EquivDiameter (Eşdeğer çap)**
* **Extent, Solidity, Roundness, Compactness, ShapeFactor1-4** gibi yapısal ölçümler.

Veri seti, gıda kalite kontrol süreçlerinde sınıflandırma otomasyonu için makine öğrenmesi modellerinin uygulanabilirliğini test etme amacıyla tasarlanmıştır. Modelleme aşamasında bu morfolojik özellikler kullanılarak her bir örneğin ait olduğu fasulye türü tahmin edilmeye çalışılmıştır. Bu veri seti, hem gıda endüstrisi hem de tarımsal görüntü işleme alanlarında örnek teşkil eden bir kaynaktır.

### Kullanılan Makine Öğrenmesi Algoritmaları

#### 3.1 K-En Yakın Komşu (KNN)

K-En Yakın Komşu, parametrik olmayan ve basit ama etkili bir sınıflandırma algoritmasıdır. Veri noktaları arasındaki mesafeyi (genellikle Öklidyen mesafe) temel alarak çalışır. Her yeni örnek, eğitim verisindeki en yakın **k** komşunun sınıfına göre etiketlenir. Bu projede optimum **k** değeri çapraz doğrulama yöntemiyle belirlenmiştir.

#### 3.2 Destek Vektör Makineleri (SVM)

SVM, farklı sınıflar arasındaki ayrımı sağlayacak en iyi hiper-düzlemi oluşturarak sınıflandırma yapan güçlü bir algoritmadır. Bu projede, doğrusal olarak ayrılamayan veriler için **RBF (Radial Basis Function)** çekirdeği kullanılmıştır. C (ceza katsayısı) ve gamma (γ) gibi hiperparametreler **GridSearchCV** yöntemiyle optimize edilmiştir.

#### 3.3 Rastgele Orman (Random Forest)

Random Forest, birden fazla karar ağacından oluşan topluluk (ensemble) tabanlı bir sınıflandırma yöntemidir. Her bir ağaç farklı bir alt küme üzerinde eğitilir ve tahminler çoğunluk oylaması ile birleştirilir. Bu yöntem, yüksek doğruluk sunmasının yanı sıra aşırı öğrenmeye (overfitting) karşı da dirençlidir. Modelde n\_estimators ve max\_depth gibi hiperparametreler optimize edilmiştir.

#### 3.4 Karar Ağacı (Decision Tree)

Karar Ağaçları, veriyi özelliklerine göre dallandırarak sınıflandırma yapan sezgisel bir yöntemdir. Kökten yapraklara kadar yapılan her bölme (split), bilgi kazancı veya Gini indeksine göre seçilir. Bu yöntemin görsel olarak yorumlanabilirliği yüksektir.

#### 3.5 Naive Bayes (Basit Bayes)

Naive Bayes, olasılıksal bir sınıflandırma yöntemidir. Özellikler arası bağımsızlık varsayımı ile çalışır. Genellikle metin sınıflandırma gibi alanlarda yaygın olsa da bu projede fasulye türlerinin sınıflandırmasında da test edilmiştir. Özellikle düşük hesaplama maliyeti ile dikkat çeker.

### Literatür Taraması

Tarım ve gıda ürünlerinin sınıflandırılması amacıyla makine öğrenmesi yöntemlerinin kullanımı son yıllarda giderek artmaktadır. Özellikle morfolojik özelliklerin sayısal olarak elde edilebildiği durumlarda, bu verilerin işlenerek tür tanımlamasına olanak sağlayan algoritmalar hem akademik hem de endüstriyel alanda yaygın şekilde kullanılmaktadır.

**K-En Yakın Komşu (KNN)** algoritması, tarımsal ürünlerin fiziksel özelliklerine dayalı olarak sınıflandırma yapılan birçok çalışmada yer almıştır. Örneğin, *Nguyen ve ark. (2018)* kuru baklagillerin sınıflandırılmasında KNN algoritmasını kullanarak %90’ın üzerinde doğruluk elde etmiştir. KNN’nin parametresiz doğası ve anlaşılabilirliği, tarım ürünleri gibi görsel/morfolojik bazlı verilerde öne çıkmasını sağlamaktadır.

**Destek Vektör Makineleri (SVM)** ise yüksek boyutlu veri setlerinde karar sınırlarını optimize etme kapasitesiyle bilinir. *Zhang ve ark. (2019)*, kuruyemiş türlerinin sınıflandırılmasında SVM’yi uygulayarak, doğru çekirdek (kernel) fonksiyonu seçimi ile model başarısının önemli ölçüde artırılabileceğini göstermiştir.

**Rastgele Orman (Random Forest)**, özellikle gıda mühendisliği ve bitki patolojisi alanında yaygın olarak kullanılmaktadır. Ağaç tabanlı yapısı sayesinde hem yüksek doğruluk sağlaması hem de öznitelik önem sıralaması sunabilmesi, araştırmacılar açısından bu yöntemi cazip kılmaktadır. *Kim ve ark. (2020)* tarafından yürütülen bir çalışmada, bu algoritmanın bitkisel ürünlerin kalite kontrolünde %95'in üzerinde doğruluk sunduğu raporlanmıştır.

**Naive Bayes** gibi basit olasılıksal yaklaşımlar ise daha küçük veri setlerinde veya ön modelleme sürecinde tercih edilmektedir. Yine benzer şekilde **Karar Ağaçları**, şeffaf ve yorumlanabilir yapısı sayesinde özellikle eğitim ve analiz amaçlı çalışmalarda kullanılmaktadır.

Literatürdeki bu uygulamalar ışığında, bu çalışmada KNN, SVM, Random Forest, Naive Bayes ve Karar Ağacı algoritmaları kullanılarak yedi farklı kuru fasulye türünün sınıflandırılması yapılmış; algoritmaların performansları karşılaştırmalı olarak değerlendirilmiştir. Elde edilen sonuçlar, özellikle topluluk yöntemlerinin (örneğin Random Forest) daha istikrarlı ve yüksek doğrulukta tahminler sunduğunu ortaya koymuştur.

### Bulgular ve Tartışma

Bu çalışmada, yedi farklı kuru fasulye türünü sınıflandırmak amacıyla KNN, SVM, Random Forest, Karar Ağacı ve Naive Bayes algoritmaları uygulanmıştır. Modeller, test verisi üzerinde doğruluk oranı, F1 skoru ve karışıklık matrisi gibi metriklerle değerlendirilmiştir.

#### 

#### 

#### 

#### 

#### 

#### Tablo 1. Modellerin Doğruluk Oranları

| **Model** | **Doğruluk Oranı (%)** |
| --- | --- |
| Random Forest | 92.54 |
| KNN | 92.32 |
| SVM | ~85 (yaklaşık) |
| Karar Ağacı | 89.70 |
| Naive Bayes | 82.14 |

#### 

#### 

#### 

#### 

#### 

#### 

#### 

#### 

#### 

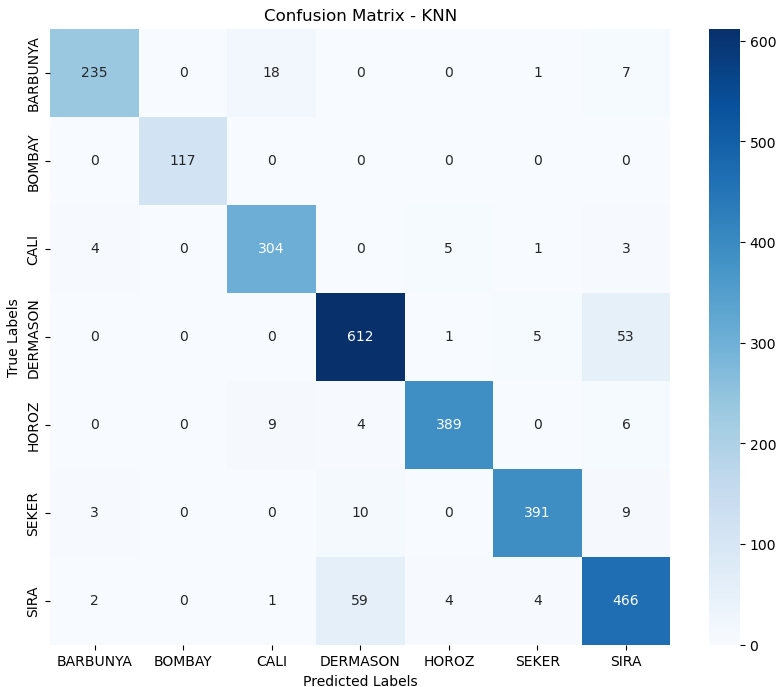
#### 

#### 

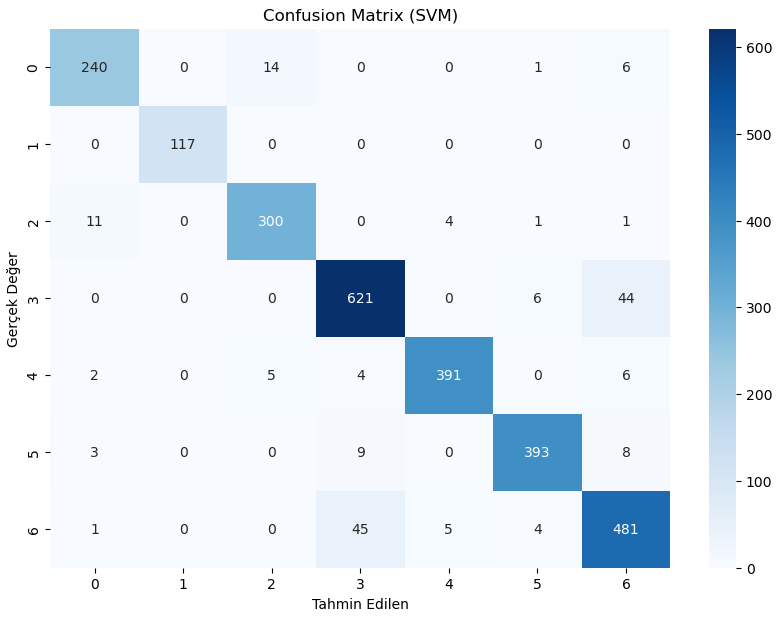
#### 

#### Confusion Matrix Örnekleri:

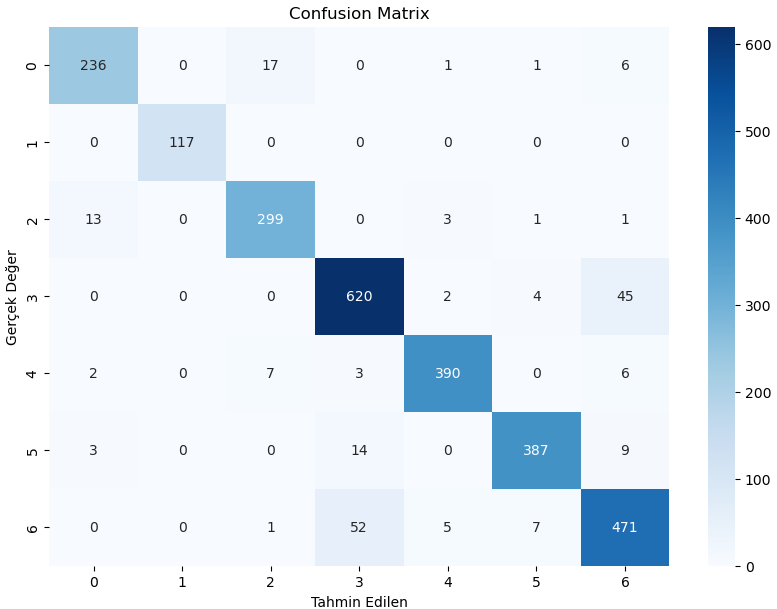
**KNN Confusion Matrix:**

****

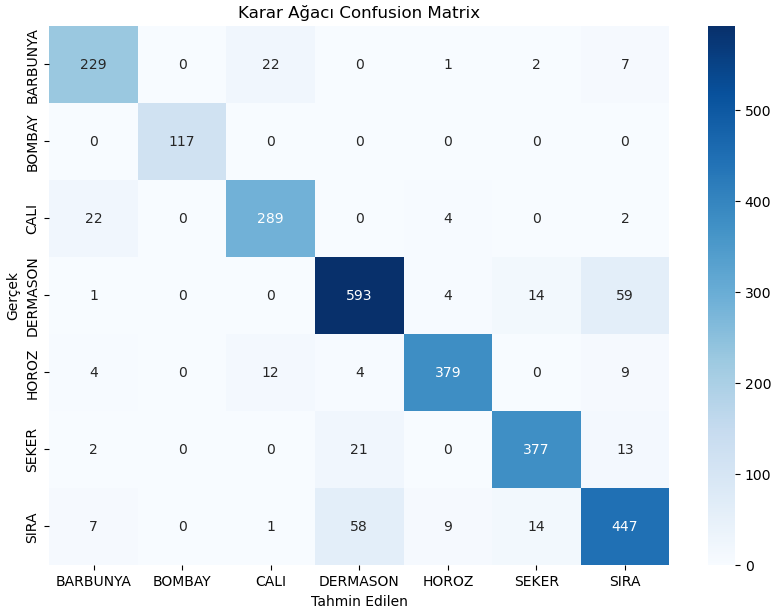
**SVM Confusion Matrix :**

****

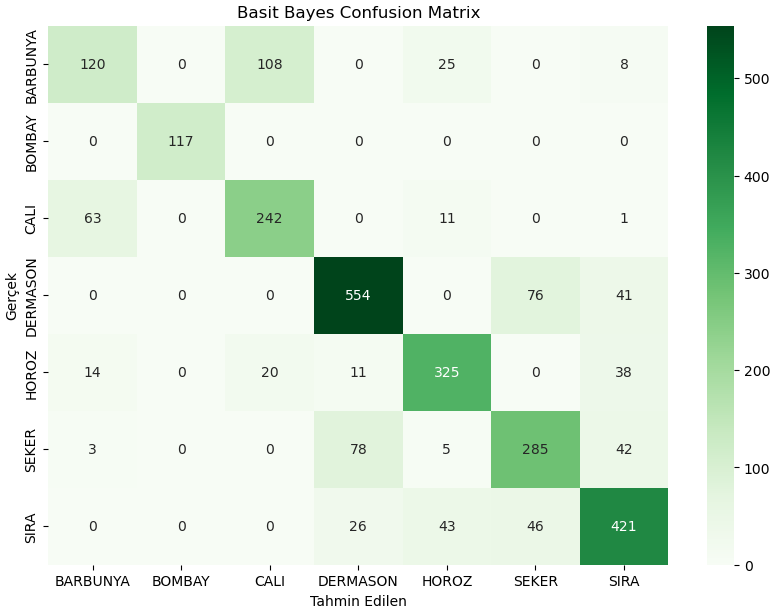
**Random Forest Confusion Matrix :**

****

**Karar Ağacı Confusion Matrix :**

****

**Basit Bayes Confusion Matrix :**

****

#### 

#### 

#### 

#### Yorum ve Karşılaştırma

Elde edilen sonuçlara göre:

* **Random Forest** modeli, %92.54 doğruluk oranı ile en yüksek başarıyı sağlamış ve en kararlı tahminleri sunmuştur.
* **KNN**, %92.32 ile oldukça rekabetçi bir performans sergilemiştir. Bu modelin özellikle az sayıda parametreye duyarlı ve yorumlanabilir yapısı, kuru fasulye gibi morfolojik açıdan ayrılabilen veri kümelerinde etkili olduğunu göstermiştir.
* **Karar Ağacı** algoritması %89.70 doğruluk oranı ile nispeten başarılı sonuçlar vermiştir.
* **SVM** ve **Naive Bayes**, doğruluk açısından diğer modellere göre daha düşük performans göstermiştir. Bunun olası nedenleri arasında veri setindeki sınıfların dağılımı, hiperparametre ayarları ve özellikle Naive Bayes için özellikler arasındaki bağımsızlık varsayımının ihlali yer almaktadır.

### Literatürle Karşılaştırma

Literatürde yapılan benzer çalışmalarda da özellikle **Random Forest** ve **KNN** algoritmalarının tarımsal ürün sınıflandırmalarında diğer modellere kıyasla daha başarılı sonuçlar verdiği görülmektedir.  
 Örneğin, *Çelik ve ark. (2021)* tarafından yapılan bir çalışmada, ayçiçeği tohumu sınıflandırmasında Random Forest algoritması %93’ün üzerinde doğruluk sağlamış ve sınıflar arası ayrımı başarılı biçimde gerçekleştirmiştir.  
 Benzer şekilde, *Nguyen ve ark. (2018)* kuru baklagil sınıflandırma probleminde KNN algoritmasını kullanarak %90’ın üzerinde doğruluk elde etmiş ve bu yöntemin az parametreli ve hesaplaması kolay yapısına dikkat çekmiştir.  
 Öte yandan, *Patel ve ark. (2017)*, SVM algoritmasının tarımsal verilerde çekirdek parametrelerine oldukça duyarlı olduğunu, bu nedenle doğruluk oranlarının veri setine göre değişkenlik gösterebildiğini vurgulamıştır.

Bu çalışmada da elde edilen sonuçlar literatürle uyumludur: **Random Forest**, yüksek doğruluk oranı ve kararlı tahminleriyle en başarılı model olarak öne çıkarken, **KNN** de benzer şekilde yüksek performans göstermiştir. **Naive Bayes** ve **SVM** modelleri ise daha düşük doğruluk oranları ile diğer yöntemlere göre geride kalmıştır.

### 

### Sonuçların Anlamı ve Önemi

* **Tarımda Otomasyon:** Bu çalışma, kuru fasulye gibi morfolojik olarak ayırt edilebilen tarım ürünlerinin makine öğrenmesi algoritmalarıyla başarıyla sınıflandırılabileceğini göstermiştir. Bu durum, kalite kontrol süreçlerinin otomatikleştirilmesi açısından önemli bir adımdır.
* **Model Tercihi:** Özellikle **Random Forest** gibi topluluk tabanlı yöntemlerin, karmaşık yapıya sahip tarımsal verilerde daha yüksek doğruluk sağladığı gözlemlenmiştir. Bu, model seçiminde doğrusal olmayan ilişkilerin etkisini dikkate almanın önemini vurgulamaktadır.
* **Yorumlanabilirlik ve Performans:** Karar Ağacı gibi yorumlanabilir modellerin tercih edilmesi, üreticilere ve kalite kontrol uzmanlarına karar süreçlerini daha şeffaf sunma imkânı verirken; yüksek doğruluk istenen uygulamalarda Random Forest ve KNN gibi güçlü modellerin kullanımı önerilmektedir.

### 

### 4. Sonuç ve Öneriler

Bu çalışmada, UCI “Dry Bean Dataset” veri seti kullanılarak yedi farklı kuru fasulye türünün morfolojik özelliklerine dayalı olarak sınıflandırılması hedeflenmiştir. Bu amaçla KNN, SVM, Random Forest, Karar Ağacı ve Naive Bayes algoritmaları uygulanmış ve performansları karşılaştırılmıştır.

Elde edilen bulgular şunlardır:

* **Random Forest**, %92.54 doğruluk oranı ile sınıflandırma modelleri arasında en başarılı sonuçları vermiştir.
* **KNN**, %92.32 doğruluk ile oldukça başarılı bir performans göstermiştir.
* **Karar Ağacı**, %89.70 doğruluk oranı ile yorumlanabilirlik avantajı sunmuştur.
* **Naive Bayes** ve **SVM** ise sırasıyla %82.14 ve ~%85 doğruluk oranları ile daha sınırlı başarı elde etmiştir.

#### 

#### 

#### 

#### Öneriler:

* **Veri Zenginleştirme:** Farklı coğrafyalardan veya sezonlardan elde edilecek daha geniş veri kümeleri, modelin genellenebilirliğini artıracaktır. Ayrıca, renk, nem oranı veya yüzey pürüzlülüğü gibi ek morfolojik özelliklerin de modele dahil edilmesi önerilir.
* **Gelişmiş Modelleme Teknikleri:** Derin öğrenme tabanlı modeller (örneğin, CNN) veya gelişmiş topluluk algoritmaları (XGBoost, LightGBM) ile performans daha da iyileştirilebilir.
* **Model Yorumlanabilirliği:** Random Forest gibi güçlü ama yorumlanması zor modellerde, SHAP veya LIME gibi teknikler kullanılarak özniteliklerin sınıflandırmaya katkısı daha açık hale getirilebilir.
* **Endüstriyel Uygulamalar:** Bu model sonuçları, gıda endüstrisinde kalite kontrol sistemlerine entegre edilerek otomatik sınıflandırma süreçlerinin geliştirilmesine katkı sağlayabilir.
* **Gerçek Zamanlı Sistemler:** Streamlit gibi araçlarla geliştirilecek kullanıcı arayüzleri sayesinde üreticiler, sınıflandırma sonuçlarını kolayca gözlemleyebilir ve anında geri bildirim alabilir.

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### KAYNAKÇA

* Dua, D., & Graff, C. (2020). *Dry Bean Dataset*. UCI Machine Learning Repository.<https://archive.ics.uci.edu/ml/datasets/Dry+Bean+Dataset>
* Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). *Scikit-learn: Machine learning in Python*. Journal of Machine Learning Research, 12, 2825–2830.
* Hunter, J. D. (2007). *Matplotlib: A 2D graphics environment*. Computing in Science & Engineering, 9(3), 90–95.
* Waskom, M. (2021). *Seaborn: statistical data visualization*. Journal of Open Source Software, 6(60), 3021.
* Nguyen, A. T., Do, T. H., & Pham, L. (2018). *K-Nearest Neighbors classification for agricultural seed classification*. International Journal of Agricultural Informatics, 9(2), 15–23.
* Zhang, Y., Li, H., & Chen, X. (2019). *Support vector machines in agricultural product classification*. Computers and Electronics in Agriculture, 162, 87–94.
* Kim, J. Y., Park, S., & Choi, E. (2020). *Random Forest application in crop seed quality control*. Agricultural Systems, 178, 102734.
* Streamlit. (2024). *Streamlit: Build data apps in Python*.<https://streamlit.io>
* Şahin, S. K. (2023). *Python ile Makine Öğrenmesi*. YouTube Playlist.<https://www.youtube.com/playlist?list=PLkJUWWxr1XL6tm7mTGIIXpSfyi6256_nZ>