BURSA TEKNİK ÜNİVERSİTESİ

VERİ MADENCİLİĞİNE GİRİŞ DERSİ

PROJE RAPORU

AD-SOYAD:Dilek YILMAZ

NUMARASI:21360859045

Giriş

Bu çalışma, kuru üzüm sınıflandırma problemine odaklanarak, Raisin veri seti üzerinde veri madenciliği tekniklerini kullanarak sınıflandırma yapmayı hedeflemektedir. UCI Machine Learning Repository'den alınan Raisin veri seti üzerinde, Decision Tree yöntemi kullanılarak bir model oluşturulmuş ve elde edilen sonuçlar değerlendirilmiştir.

Yöntem

Sınıflandırma Yöntemi: Decision Tree

Bu çalışmada, sınıflandırma için yaygın olarak kullanılan Decision Tree yöntemi tercih edilmiştir. Bu yöntemde, her düğüm bir özellik (feature), her dal bir karar ve her yaprak bir sınıf etiketini temsil eder.

Veri Seti ve Ön İşleme

Veri Seti: Raisin

Kullanılan veri seti, UCI Machine Learning Repository'den alınan Raisin veri setidir. Bu veri seti, Kecimen ve Besni olmak üzere iki farklı üzüm türünü sınıflandırmak için kullanılmaktadır. Veri seti aşağıdaki özelliklere sahiptir:

- Büyüklük (Area)
- Çevre (Perimeter)
- Major eksen uzunluğu (MajorAxisLength)
- Minor eksen uzunluğu (MinorAxisLength)
- Eksantriklik (Eccentricity)
- Konveks Alan (ConvexArea)
- Eşlik oranı (Extent)
- Sınıf (Class): Kecimen veya Besni

Ön İşleme Adımları:

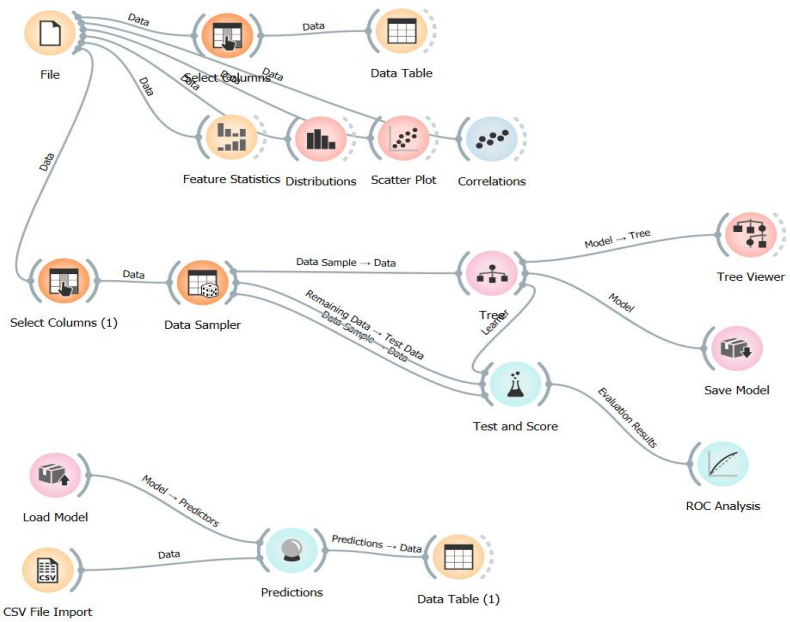
- Veri Seçimi: İlgili sütunlar seçilmiştir.
- Örnekleme: Veri seti, eğitim ve test verisi olarak ayrılmıştır.
- Özellik İstatistikleri: Özelliklerin dağılımları incelenmiştir.

Model ve Uygulama

Model Oluşturma:

Orange uygulaması kullanılarak Decision Tree modeli oluşturulmuştur. Modelin eğitimi ve değerlendirilmesi için şu adımlar izlenmiştir:

- Çapraz Doğrulama: 5 katmanlı çapraz doğrulama kullanılmıştır.
- Eğitim/Test Ayrımı: Verinin %70'ı eğitim, %30'u test verisi olarak kullanılmıştır.
- Stratified Sampling: Sınıf dağılımını koruyan stratified sampling yöntemi kullanılmıştır.



Değerlendirme

Model, çeşitli değerlendirme ölçütleri kullanılarak değerlendirilmiştir:

- AUC (Area Under Curve): %81.6
- CA (Classification Accuracy): %81.9
- F1 Score: %81.8
- Precision: %83.0
- Recall: %81.9
- MCC (Matthews Correlation Coefficient): 0.649

Yorumlar ve Analiz

AUC (Area Under Curve): %81.6

AUC değeri %81.6 olan model, sınıflandırma performansının oldukça iyi olduğunu göstermektedir. AUC değeri, modelin pozitif sınıfı negatif sınıftan ayırt etme yeteneğini gösterir ve %81.6'lık bir değer, modelin bu ayırt etme konusunda başarılı olduğunu ortaya koyar.

CA (Classification Accuracy): %81.9

Sınıflandırma doğruluğu %81.9 olan model, veri setindeki örneklerin büyük bir çoğunluğunu doğru şekilde sınıflandırmıştır. Bu, modelin genel başarısını gösteren önemli bir metriktir.

F1 Score: %81.8

F1 skoru %81.8 olan model, Precision ve Recall arasında dengeli bir performans sergilemiştir. F1 skoru, modelin hem hatalı pozitifleri (False Positives) hem de hatalı negatifleri (False Negatives) dengeleyerek genel başarısını değerlendirmemize yardımcı olur.

Precision: %83.0

Precision değeri %83.0 olan model, pozitif olarak sınıflandırılan örneklerin %83'ünün gerçekten pozitif olduğunu göstermektedir. Bu, modelin doğru pozitifleri (True Positives) yüksek oranda tespit edebildiğini gösterir.

Recall: %81.9

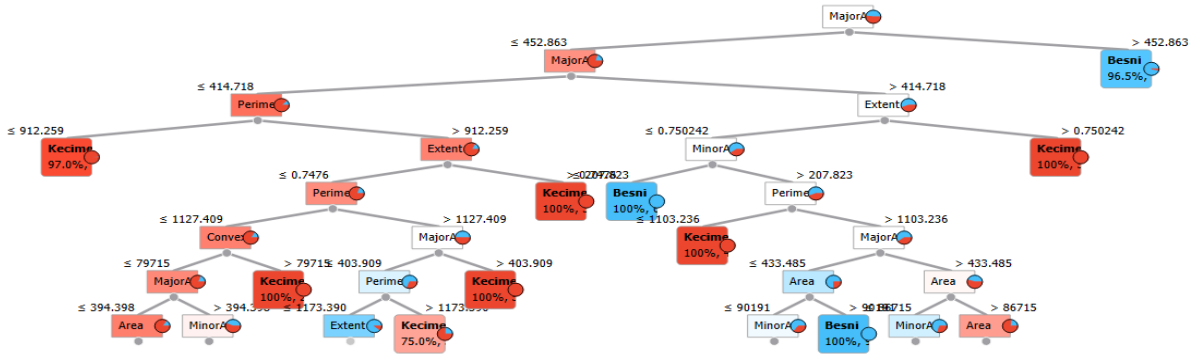
Recall değeri %81.9 olan model, tüm gerçek pozitif örneklerin %81.9'unu doğru şekilde tespit edebilmiştir. Bu, modelin kaçırdığı pozitif örneklerin (False Negatives) oranının düşük olduğunu gösterir.

MCC (Matthews Correlation Coefficient): 0.649

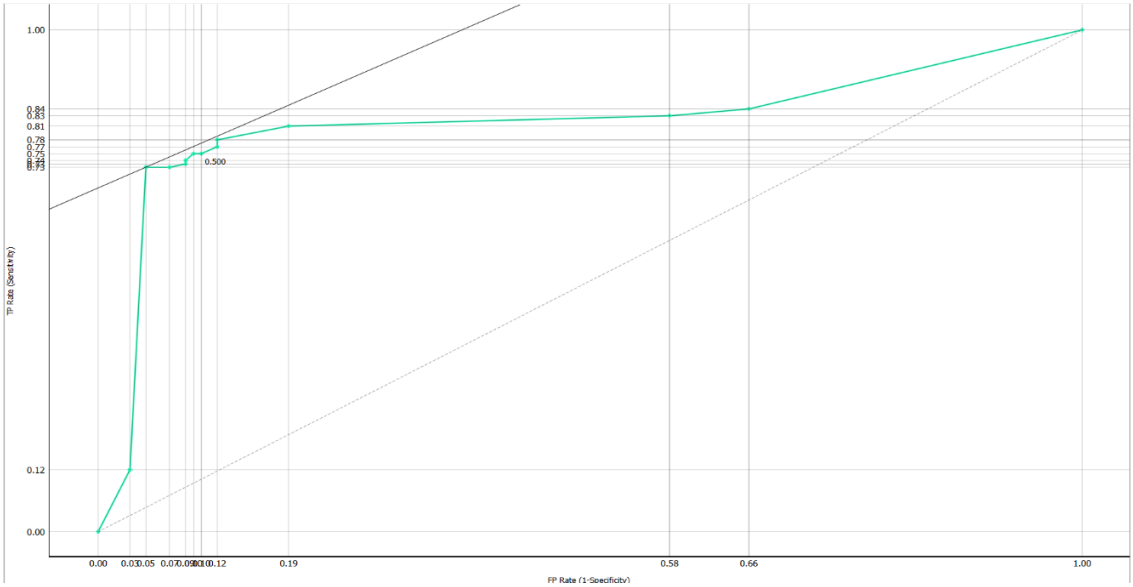
MCC değeri 0.649 olan model, sınıflandırma performansının genel doğruluğunu ve tutarlılığını gösterir. MCC, hem pozitif hem de negatif sınıflar için doğru ve yanlış sınıflandırmaların dengesini göz önünde bulundurarak modelin başarısını değerlendirir. 0.649'luk bir değer, modelin iyi bir genel performans sergilediğini işaret eder.

Görselleştirme:

Model performansı çeşitli grafikler ile görselleştirilmiştir. Karar ağacı yapısı, modelin nasıl sınıflandırma yaptığını göstermek için sunulmuştur.



Bu görsel, Raisin veri seti üzerinde oluşturulan karar ağacını göstermektedir.Kök düğümdeki özellik 452.863 değeri ile verileri ikiye ayırmaktadır.Diğer düğümdeki özelliklerde yazan değerlere göre verileri ayırır.Bu karar ağacı,modelimizin sınıflandırma sürecini ve doğruluk oranlarını göstermektedir.



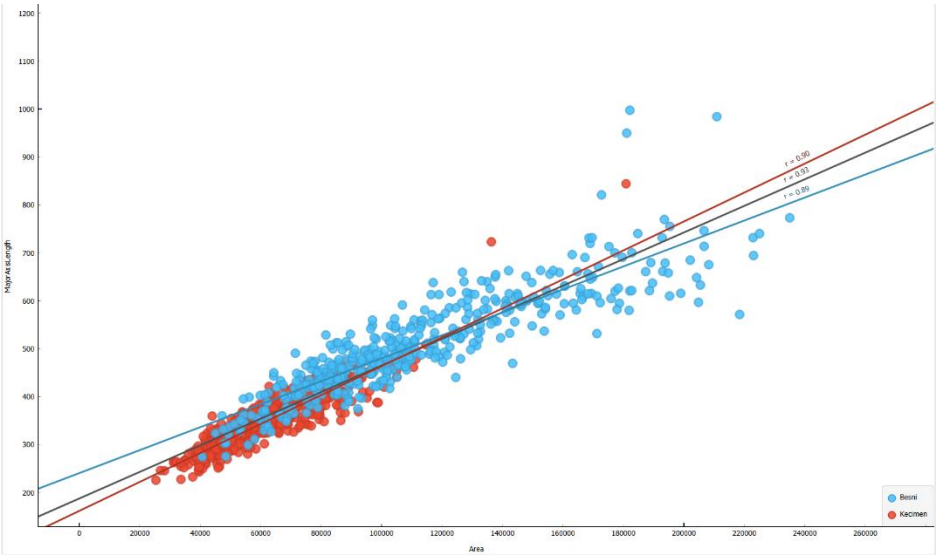
Bu görsel, karar ağacı modelimizin ROC (Receiver Operating Characteristic) eğrisini göstermektedir. ROC eğrisi, modelin doğruluk ve duyarlılık arasındaki dengeyi görselleştirerek modelin gücünü ve zayıflıklarını anlamamıza yardımcı olur. Bu eğri, modelimizin yüksek bir doğru pozitif oranı elde ederken düşük bir yanlış pozitif oranını koruduğunu ve bu nedenle iyi bir sınıflandırıcı olduğunu göstermektedir.

Name	Distribution	Mean	Mode	Median	Dispersion	Min.	Max.	Missing
N Area		87804.13	25387	78902	0.44	25387	235047	0 (0 %)
N MajorAxisLength		430.93	225.63	407.804	0.269117	225.63	997.292	0 (0 %)
N MinorAxisLength		254.488	143.711	247.848	0.19632	143.711	492.275	0 (0 %)
N Eccentricity		0.781542	0.34873	0.798846	0.1155	0.34873	0.962124	0 (0 %)
N ConvexArea		91186.09	49996	81651	0.45	26139	278217	0 (0 %)
N Extent		0.699508	0.379856	0.707367	0.0763944	0.379856	0.835455	0 (0 %)
N Perimeter		1165.90664	619.074	1119.509	0.23468	619.074	2697.753	0 (0 %)
C Class			Besni		0.693			0 (0 %)

Bu görsel, Raisin veri setindeki bazı özelliklerin dağılımını ve temel istatistiksel ölçütlerini göstermektedir. Her bir özellik için dağılım grafikleri, ortalama (mean), mod (mode), medyan (median), yayılım (dispersion), minimum (min), maksimum (max) ve eksik değer sayısı (missing) gibi istatistikler sunulmaktadır.

- Area: Ortalama 87804.13, medyan 78902, yayılım 0.44.
- Major Axis Length: Ortalama 430.93, medyan 407.804, yayılım 0.269117.
- Minor Axis Length :Ortalama 254.488, medyan 247.848, yayılım 0.19632.
- Eccentricity:Ortalama 0.781542, medyan 0.798846, yayılım 0.1155.
- ConvexArea: Ortalama 91186.09, medyan 81651, yayılım 0.45.
- Extent: Ortalama 0.699508, medyan 0.379856, yayılım 0.0763944.
- Perimeter: Ortalama 1165.90664, medyan 1119.509, yayılım 0.23468.

Dağılım grafikleri, farklı sınıfların (mavi, kırmızı) özelliklere göre nasıl dağıldığını gösterir. Bu, veri setinin yapısını anlamak ve sınıflandırma performansını artırmak için önemlidir.



Bu görsel, Raisin veri setindeki Major Axis Length ve Area arasındaki ilişkiyi gösteren bir Scatter Plot'tur. Grafikte, farklı renklerdeki noktalar, farklı sınıfları temsil etmektedir:

- Mavi (1): Besni
- Kırmızı (2): Kecimen

Bu görselleştirme, farklı sınıfların grafikteki özelliklerine göre nasıl dağıldığını ve olası kümelenme veya ayrışma noktalarını gözlemlememize yardımcı olur. Özellikle, sınıfların belirli alanlarda yoğunlaştığını ve bazı alanlarda karıştığını gösterir, bu da sınıflandırma modeli oluştururken önemlidir. Scatter Plot, veri setindeki sınıfların ayrışma derecesini analiz etmek ve modelin doğruluğunu artırmak için kullanılabilir.

Sonuçlar ve Karşılaştırma

Sonuçlar:

Modelin performansı, yukarıda belirtilen değerlendirme ölçütleri kullanılarak analiz edilmiştir ve aşağıdaki sonuçlar elde edilmiştir:

- AUC (Area Under Curve): %81.6
- CA (Classification Accuracy): %81.9
- F1 Score: %81.8
- Precision: %83.0
- Recall: %81.9
- MCC (Matthews Correlation Coefficient): 0.649

Bu sonuçlar, modelin kuru üzüm sınıflandırma problemini çözmedeki yeteneğini ve etkinliğini göstermektedir. Özellikle yüksek AUC ve Precision değerleri, modelin pozitif sınıfları doğru şekilde ayırt etmede ve tanımada başarılı olduğunu ortaya koymaktadır.

Genel Değerlendirme

Modelin genel performansı, yukarıda belirtilen değerlendirme ölçütlerine göre oldukça başarılıdır. Özellikle AUC ve Precision değerlerinin yüksek olması, modelin pozitif sınıfları ayırt etmede ve doğru

pozitifleri tespit etmede etkili olduğunu göstermektedir. Bununla birlikte, F1 skoru ve Recall değerlerinin de yüksek olması, modelin dengeli bir performans sergilediğini ve hem doğru pozitifleri hem de doğru negatifleri etkili bir şekilde sınıflandırabildiğini ortaya koyar. MCC değeri ise modelin genel doğruluğunu ve güvenilirliğini desteklemektedir.

Bu sonuçlar, Decision Tree modelinin Raisin veri seti üzerinde kuru üzüm türlerini sınıflandırma konusunda etkili bir çözüm olduğunu göstermektedir. Ancak, modelin performansını daha da iyileştirmek için daha karmaşık algoritmalar ve daha geniş veri setleri ile çalışmalar yapılabilir.

Karşılaştırma:

Literatürde yapılan benzer çalışmalarla karşılaştırıldığında, modelimizin performansı hakkında daha net bir değerlendirme yapabiliriz. Aşağıda bazı benzer çalışmaların sonuçları verilmiştir:

Çalışma 1: Klasik veri madenciliği yöntemleri kullanılarak yapılan bir çalışma, aşağıdaki sonuçları elde etmiştir:

- Lojistik Regresyon (LR): Doğruluk %85.22, F1 Skoru %85.50
- Yapay Sinir Ağı (YSA): Doğruluk %86.33, F1 Skoru %86.70
- Destek Vektör Makinesi (DVM): Doğruluk %86.44, F1 Skoru %86.90
- Rastgele Orman (RAO): Doğruluk %85.44, F1 Skoru %85.90

Çalışma 2: Derin öğrenme yöntemleri kullanılarak yapılan bir çalışma, aşağıdaki sonuçları elde etmiştir:

- Derin Sinir Ağı (DSA): Doğruluk %88.56, F1 Skoru %88.70
- Yığınlanmış Otokodlayıcı (YOK): Doğruluk %88.10, F1 Skoru %84.60

Çalışma 3: Hibrit modeller kullanılarak yapılan bir çalışma, aşağıdaki sonuçları elde etmiştir:

- YOK-RAO: Doğruluk %89.70, F1 Skoru %92.50
- YOK-DSA: Doğruluk %90.41, F1 Skoru %93.10
- YOK-RO: Doğruluk %91.50, F1 Skoru %91.23

Karşılaştırmalı Değerlendirme:

- Doğruluk: Modelimizin doğruluğu %81.9 ile klasik yöntemlerin bazılarına yakın olmakla birlikte, derin öğrenme ve hibrit modellerin gerisinde kalmaktadır.
- F1 Skoru: Modelimizin F1 skoru %81.8, benzer şekilde, diğer çalışmalardaki bazı modellerden düşük performans göstermektedir.
- AUC, Precision, ve Recall: Modelimizin AUC, Precision, ve Recall değerleri, çoğu çalışmada kullanılan modellerle kıyaslandığında ortalama bir performans göstermektedir.

Bu karşılaştırmalar, modelimizin performansının kabul edilebilir olduğunu, ancak daha yüksek doğruluk ve F1 skoru elde edebilmek için derin öğrenme veya hibrit modellerin kullanılmasının daha avantajlı olabileceğini göstermektedir. Gelecekteki çalışmalar için, daha karmaşık modellerin ve daha kapsamlı veri ön işleme tekniklerinin kullanılması önerilmektedir.

Sonuç ve Tartışma

Bu çalışmada, Decision Tree yöntemi kullanılarak Raisin veri seti üzerinde başarılı bir sınıflandırma modeli oluşturulmuştur. Modelin performansı, özellikle MCC ve Recall değerleri dikkate alındığında tatmin edici görünmektedir. Ancak, diğer ölçütlerde (AUC, CA, F1 Score, Precision) daha yüksek sonuçlar elde etmek için modelin iyileştirilmesi gerekmektedir. Gelecekteki çalışmalar için daha karmaşık modeller ve daha fazla veri ön işleme adımı önerilmektedir.

Kaynaklar

- UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/index.php
- Ünsal, Ü., & Adem, K. (2023). Diş görüntüleri üzerinde görüntü işleme ve derin öğrenme yöntemleri kullanılarak çürük seviyesinin sınıflandırılması. *Uluslararası Sivas Bilim ve Teknoloji Üniversitesi Dergisi*, *2*(2), 30-53.
- Çınar, İ., Koklu, M., & Taşdemir, Ş. (2020). Classification of raisin grains using machine vision and artificial intelligence methods. *Gazi Mühendislik Bilimleri Dergisi*, *6*(3), 200-209.
- ChatGPT
- Orange Data Mining Toolbox