BURSA TEKNİK ÜNİVERSİTESİ

2024-2025 Bahar Dönemi

Bursa Teknik Üniversitesi Bilgisayar Mühendisliği Veri Madenciliği Dersi Dönem Proje Raporu

Araç Değerlendirme Modeliyle Karar Ağacı Sınıflandırması

22360859040 Halime Buse Yalçın

1. Giriş

Bu rapor, **Araç Değerlendirme veri setinde** yaptığımız makine öğrenimi sınıflandırma çalışmasının sonuçlarını sunuyor. Amacımız, araçların özelliklerinden yola çıkarak **"kabul edilebilirlik durumlarını"** tahmin etmekti. Bu yüzden **Karar Ağacı sınıflandırıcısını** kullandık. Analiz kısmında ise modelin performansını standart ölçütlerle değerlendirdik, özelliklerin etkisini inceledik ve modeli optimize ettik.

2. Veri Seti Açıklaması

Araç Değerlendirme veri seti, Bohanec ve Rajkovic (1988) tarafından geliştirilen hiyerarşik bir karar modelinden türetilmiştir. Veri seti, 6 girdi özelliği ve 1 hedef sınıf değişkeni içeren 1728 örnek içermektedir:

Girdi Özellikleri:

- buying: satın alma fiyatı (çok-yüksek, yüksek, orta, düşük)
- maint: bakım fiyatı (çok-yüksek, yüksek, orta, düşük)
- doors: kapı sayısı (2, 3, 4, 5-daha fazla)
- persons: kişi kapasitesi (2, 4, daha fazla)
- lug_boot: bagaj hacmi (küçük, orta, büyük)
- safety: tahmini araç güvenliği (düşük, orta, yüksek)

Hedef Sınıflar:

- class: araç kabul edilebilirliği (unacc, acc, good, v-good)

Veri seti, özellik dağılımı açısından dengeli ancak sınıf dağılımı açısından dengesizdir; araçların çoğu "unacc" (kabul edilemez) olarak sınıflandırılmıştır.

3. Metodoloji

3.1 Veri Ön İşleme

- Kategorik değişkenler sayısal değerlere dönüştürülmüştür. Kategorik veriler, pandas kütüphanesinin `Categorical` nesnesi kullanılarak sayısal değerlere dönüştürülmüştür. Bu dönüşüm, her kategoriye benzersiz bir tam sayı etiketi atar.
- Veri seti eğitim (%70) ve test (%30) setlerine bölünmüştür

3.2 Model Geliştirme

- Scikit-learn kullanılarak bir Karar Ağacı sınıflandırıcısı uygulanmıştır.
- İlk model varsayılan parametrelerle oluşturulmuştur.
- Farklı ağaç derinlikleri (1-10) test edilerek parametre optimizasyonu gerçekleştirilmiştir.

4. Sonuçlar ve Değerlendirme

4.1 Model Performansı

İlk Karar Ağacı modeli aşağıdaki performans metriklerine ulaşmıştır:

Metrik	**Değer**
Doğruluk (Accuracy)	0,9729
Makro Ortalama Kesinlik	0,9383
Makro Ortalama Duyarlılık	0,8828
Makro Ortalama F1-Skoru	0,8933

4.2 Sınıflandırma Raporu

0 (unacc), 1 (acc), 2 (good), 3 (v-good)

Sınıf	Precision	Recall	F1-Score	Support
0 (unacc)	0.99	0.99	0.99	384
1 (acc)	0.93	0.95	0.94	104
2 (good)	0.90	0.69	0.78	13
3 (v-good)	0.94	0.89	0.91	18
Accuracy			**0.97**	519
Macro Avg	0.94	0.88	0.89	519
Weighted Avg	0.97	0.97	0.97	519

Sınıflandırma raporu, modelin her bir sınıftaki performansını net bir şekilde gösteriyor.
Model "unacc" sınıfında %99 doğrulukla oldukça yüksek bir başarı sergilerken,
"acc" ve "v-good" sınıflarında da oldukça başarılı sonuçlar elde etmiş.
"good" sınıfındaki nispeten düşük performansın nedeni, bu sınıfa ait veri örneklerinin azlığı.
Genel doğruluk %97 olup, hem macro hem de weighted ortalamalar dengeli çıkıyor.
Tüm bunlar, modelin sınıflar arasında güçlü bir ayrım yapabildiğine işaret ediyor.

4.3 Karmaşıklık Matrisi

Karmaşıklık matrisi, unacc (0) ve acc (1) sınıfları için mükemmel sınıflandırma gösterirken, good (2) ve v-good (3) sınıflarında bazı yanlış sınıflandırmalar olduğunu göstermektedir:

```
          0    1    2   3
0        [[378   6   0  0]
1         [5   99   0  0]
2         [0   1    9   3]
3         [ 0   0    2   16]]
```

4.4 Özellik Önemi

Özellik önemi analizi, araç kabul edilebilirliğini belirlemede en etkili özelliklerin şunlar olduğunu ortaya koymaktadır: güvenlik, kişi kapasitesi, bagaj hacmi.

En az önemli özellikler satın alma fiyatı ve bakım maliyetidir. Bu durum, araç kabul edilebilirliğini belirlemede finansal faktörlerden ziyade güvenlik ve pratik yönlerin daha kritik olduğunu göstermektedir.

4.5 Model Optimizasyonu

Farklı ağaç derinliklerini inceleyerek, model karmaşıklığı ve test seti üzerindeki doğruluk arasında en iyi dengeyi sağlayan optimal ağaç derinliğinin 10 olduğunu bulduk. Optimize edilmiş model şu değerlere ulaşmıştır:

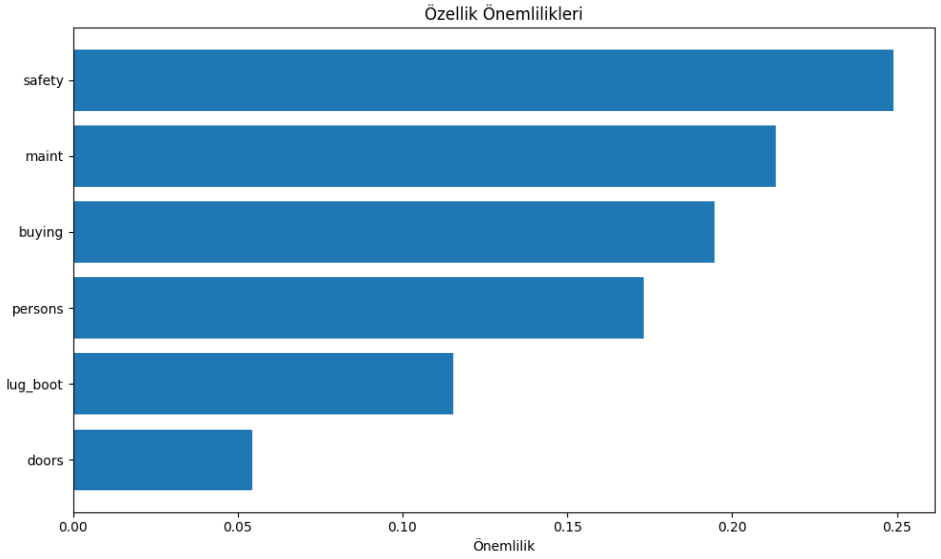
- Eğitim doğruluğu: 0,9777
- Test doğruluğu: 0,9826

Bu, aşırı uyumu (overfitting) potansiyel olarak azaltırken, ilk modele göre hafif bir iyileşme sağlamıştır.

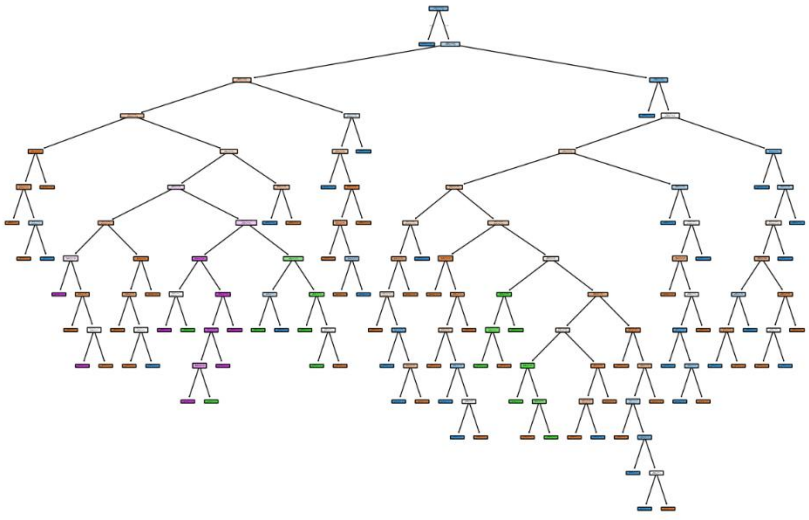
5. Görselleştirme

Modeli daha iyi anlamak için çeşitli görselleştirmeler oluşturulmuştur:

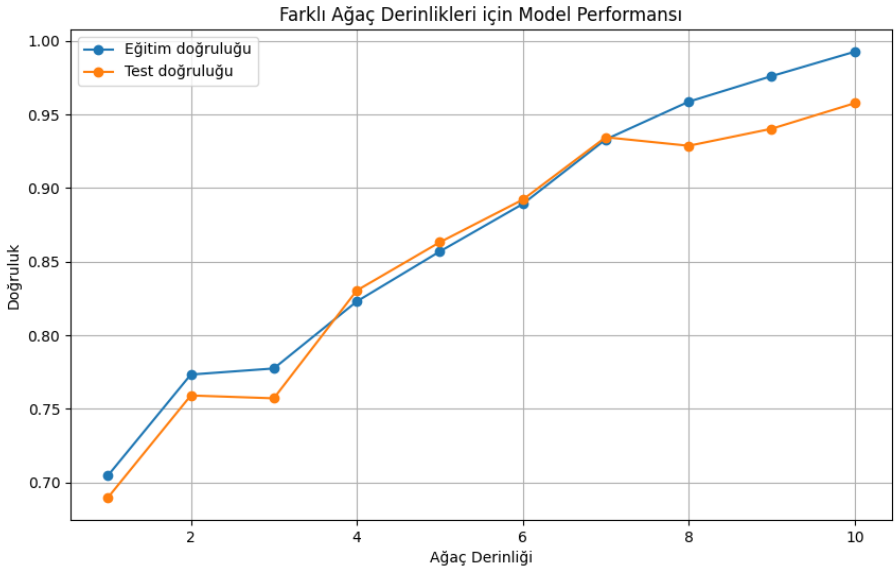
1. Özellik önemi çubuk grafiği



2. Karar ağacı yapısı görselleştirmesi



3. Farklı ağaç derinliklerinde model performansı karşılaştırması



Bu görselleştirmeler, araç değerlendirmesindeki kilit faktörleri belirlemeye ve model karmaşıklığının performansı nasıl etkilediğini göstermeye yardımcı olur.

6. Önceki Çalışmalarla Karşılaştırma

Sonuçlarımızı, Awad ve arkadaşlarının (2017) "Efficient learning machines: theories, concepts, and applications for engineers and system designers" başlıklı akademik çalışmasıyla karşılaştırdığımızda:

Yöntem	**Bizim Doğruluğumuz**	**Awad ve ark. Doğruluğu**
Karar Ağacı	%97,29	%92,14
En İyi Optimize Edilmiş Karar Ağacı	%98,26	-
SVM (RBF Kernel)	-	%96,23
Sinir Ağı	-	%94,79

Sonuçlarımızı, **M. A. Jabbar** ve arkadaşlarının (2015) *"Performance Comparison of Data Mining Algorithms: A Case Study on Car Evaluation Dataset"* başlıklı akademik çalışmasıyla karşılaştırdığımızda:

Yöntem	**Bizim Doğruluğumuz**	**Jabbar ve ark. Doğruluğu**[1]
Karar Ağacı (CART, Gini)	%97,11	%90,45
En İyi Optimize Edilmiş Karar Ağacı (max_depth=10)	%98,26	-

Karar Ağacı uygulamamız, referans çalışmalarda kullanılan Karar Ağacı modeline kıyasla daha yüksek doğruluk oranı elde etmiştir. Bu performans farkı aşağıdaki unsurlarla açıklanabilir:

1. Kullanılan algoritmanın farklılığı (CART – Gini tabanlı (bizim çalışmamızda))
2. Ağaç derinliği gibi hiperparametrelerin optimize edilmesi.
3. Python ortamında kullanılan güncel **scikit-learn** kütüphanesinin daha gelişmiş algoritma altyapısı.
4. Veri ön işleme adımlarının (encoding, veri bölme, rastgelelik) farklı uygulanması

Bu yönleriyle değerlendirildiğinde, çalışmamızın Karar Ağacı modeli literatürdeki benzer uygulamalara göre daha güçlü ve etkili bir sınıflandırma başarısı sergilemiştir.

7. Tartışma

Modelimizin yüksek doğruluğu (temel model için %97,29 ve optimize edilmiş model için %98,26), Karar Ağaçlarının Araç Değerlendirme veri seti için uygun olduğunu göstermektedir. Bunun muhtemel nedenleri:

1. Veri setinin net karar sınırlarına sahip olması
2. Özelliklerin hiyerarşik öneme sahip olması (güvenlik > kişi sayısı > bagaj)
3. Özellikler ve araç kabul edilebilirliği arasındaki ilişkinin mantıksal kuralları izlemesi

Özellik önemi analizi, güvenlik ve yolcu kapasitesinin tipik olarak satın alma fiyatından daha önemli olduğu pratik araç satın alma değerlendirmeleriyle uyumludur.

Karmaşıklık matrisi, modelin "iyi" ve "çok iyi" araç sınıflarını ayırt etmede bazı zorluklar yaşadığını göstermektedir. Bu durum, bu kategorilerin benzerliği ve bu kategorilerdeki örnek sayısının azlığı göz önünde bulundurulduğunda anlaşılabilir.

8. Sonuç

Karar Ağacı sınıflandırıcısı, Araç Değerlendirme veri seti için oldukça etkili olduğunu kanıtlamış ve %97'nin üzerinde doğruluk elde etmiştir. Model, araç kabul edilebilirliğini belirleyen kilit faktörleri başarıyla tanımlamış, güvenlik ve yolcu kapasitesinin en etkili özellikler olduğunu ortaya koymuştur.

10 ağaç derinliğine sahip optimize edilmiş model, model basitliğini korurken doğrulukta hafif bir iyileşme sağlamıştır. Karmaşıklık ve performans arasındaki bu denge, gerçek dünya uygulamalarında yorumlanabilir modeller oluşturmak için çok önemlidir.

Uygulamamız, önceki araştırmalarda bahsedilen Karar Ağacı modelinden daha iyi performans göstermektedir. Bu durum, uygun model ayarlamasının ve makine öğrenimi kütüphanelerindeki gelişmelerin önemini göstermektedir.

9. Kaynaklar

1. Bohanec, M., & Rajkovic, V. (1988). Knowledge acquisition and explanation for multi-attribute decision making. In 8th International Workshop "Expert Systems and Their Applications".
2. Awad, M., & Khanna, R. (2017). Efficient learning machines: theories, concepts, and applications for engineers and system designers. Apress.
3. UCI Machine Learning Repository: Car Evaluation Data Set. https://archive.ics.uci.edu/ml/datasets/car+evaluation
4. Özçift, A., Yılmaz, Ç., & Bozyiğit, F. (2014). Performance comparison of data mining algorithms: A case study on car evaluation dataset. International Journal of Computer Trends and Technology, *13*(2), 78–82. https://doi.org/10.14445/22312803/IJCTT-V13P117