



BURSA TEKNİK ÜNİVERSİTESİ

**VERİ MADENCİLİĞİNE GİRİŞ
PROJE RAPORU**

AD-SOYAD: Adile AKKILIÇ

NUMARASI: 21360859052

1. Giriş

Rapor, BLM0463 Veri Madenciliğine Giriş dersi çerçevesinde Wine veri seti üzerinde kural tabanlı yöntemler kullanarak veri madenciliği uygulamaktır. Veri seti, çeşitli şarap türlerinin kimyasal özelliklerini içermektedir ve bu özellikler kullanılarak şarap türleri sınıflandırılmaya çalışılacaktır. Kural tabanlı yöntemler, veri seti üzerinde belirli kuralları uygulayarak sınıflandırma yapmamıza olanak sağlar. Bu yöntemlerin etkinliğini değerlendirmek için çeşitli metrikler kullanılacaktır.

2. Veri Seti ve Ön İşleme

Veri Seti:

- Bu çalışmada kullanılan veri seti, UCI Machine Learning Repository'den alınmıştır. Wine veri seti, 13 kimyasal özellik ve 3 farklı şarap türünü (class_0, class_1, class_2) içermektedir. Bu özellikler, şarapların kimyasal bileşenlerine dair detaylı bilgiler sunmakta ve sınıflandırma işlemlerinde kullanılmaktadır.

Ön İşleme:

- Veri seti içe aktarıldıktan sonra eksik veriler temizlenmiş, gerekli dönüşümler yapılmış ve veriler normalizasyon işlemi ile uygun formata getirilmiştir.
- Verilerin analiz ve modelleme süreçlerinde daha doğru sonuçlar vermesini sağlamak amacıyla gerçekleştirilmiştir.

3. Sınıflandırma Yöntemi

Bu projede Kurala Dayalı Yöntemler (Rule-based Methods) kullanılmıştır. Kurala dayalı yöntemler, belirli "if-then" kuralları ile verileri sınıflandırır. Bu yöntemi seçmemizin nedeni, kolay anlaşılır ve yorumlanabilir olmasıdır. Kurallar net bir şekilde ifade edildiği için, karar verme süreci açıkça izlenebilir.

4. Modelin Eğitimi ve Değerlendirilmesi

Model Eğitimi:

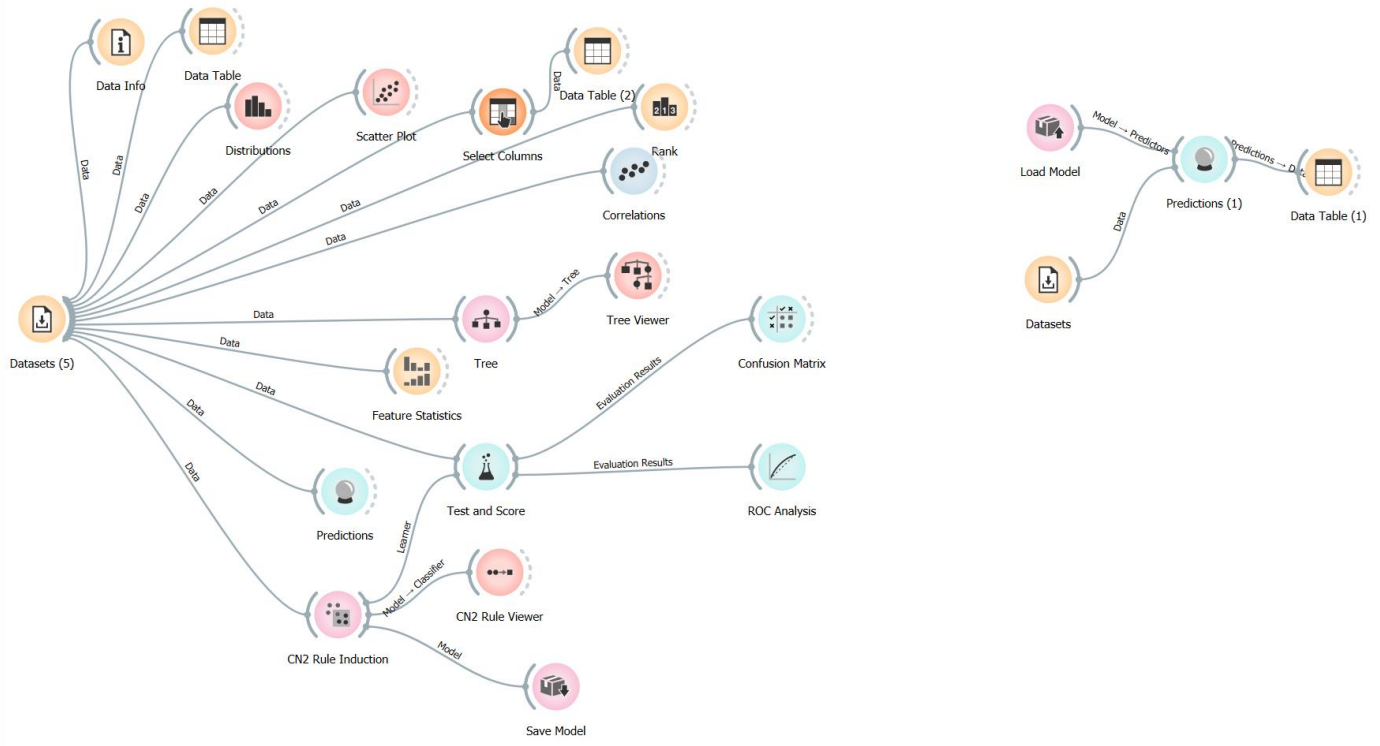
- Kurala dayalı model, eğitim veri seti kullanılarak eğitilmiştir.
- Modelin hiperparametreleri optimize edilmiştir.

Model Değerlendirme:

- Model, test veri seti üzerinde değerlendirilmiştir.
- Değerlendirme ölçütleri olarak accuracy, sensitivity, specificity ve F-measure kullanılmıştır.

Eğitim ve Test Setlerinin Ayrılması:

- Veri seti eğitim ve test seti olarak ayrılmıştır. Eğitim seti, modelin öğrenmesi için kullanılmış; test seti, modelin doğruluğunu değerlendirmek için kullanılmıştır.



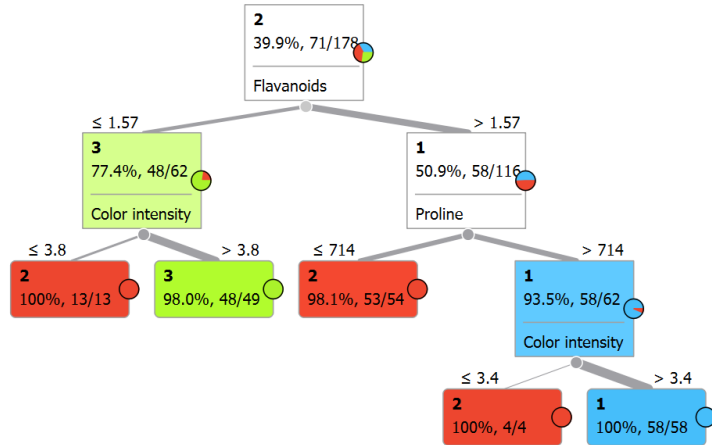
5. Sonular

Model Performansı:

- Doğruluk (Accuracy): %86.8
- F1-Measure: %86.8
- Hassasiyet (Precision): %86.9
- Duyarlılık (Recall): %86.8

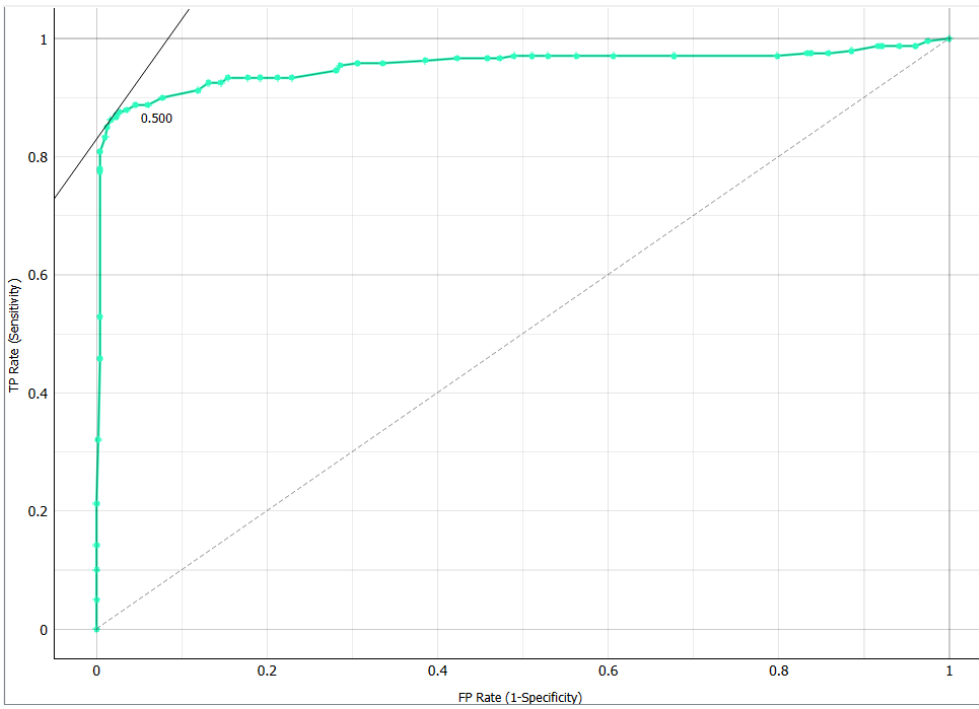
Görselleştirme:

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



Bu görsel, Wine veri seti üzerinde oluşturulan karar ağacını göstermektedir. Bu ağaç, şarap türlerini belirli kimyasal özelliklere dayanarak sınıflandırmak için kullanılmıştır. Kök düğümde Flavanoids özelliğine dayanır. Eğer Flavanoids değeri 1.57'den küçük veya eşitse, sol dala gidilir. Aksi takdirde,

sağ dala gidilir. Sol dalda Flavanoids ≤ 1.57 olan örnekler Color intensity özelliğine ayrılır. Eğer Color Intensity değeri 3.8'den küçük veya eşit ise, sınıf 2 olarak sınıflandırılır(%100 doğrulukla). Eğer Color Intensity >3.8 ise, sınıf 3 olarak sınıflandırılır (%98 doğrulukla) . Sağ dalda Flavanoids >1.57 olan örnekler, Proline özelliğine göre ayrılır. Eğer Proline değeri 714'ten küçük veya eşitse, sınıf 2 olarak adlandırılır (%98.1 doğrulukla). Eğer Proline > 714 ise Color intensity özelliğine göre bir ayırım daha yapılır ; Eğer Color intensity ≤ 3.4 ise, sınıf 1 olarak sınıflandırılır (%100 doğrulukla). Eğer Color intensity > 3.4 ise,sınıf 1 olarak sınıflandırılır (%100 doğrulukla) . Bu karar ağacı modeli, Flavanoids, Color Intensity ve Proline gibi belirli kimyasal özelliklere dayanarak şarap türlerini sınıflandırmaktadır. Modelin her bir düğümünde, belirli bir özelliğin belirli bir eşik değeri kullanılarak karar verilmekte ve bu şekilde dallar ayrılmaktadır. Elde edilen sınıflandırma doğruluk oranları, modelin performansını ve verilerin bu özelliklere göre nasıl ayrıldığını göstermektedir.

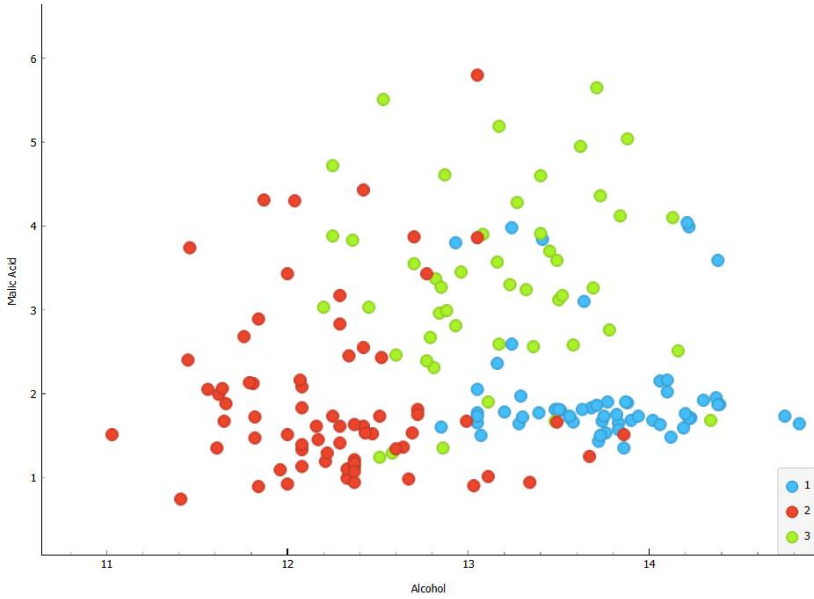


Bu görsel, modelin performansını değerlendirmek için kullanılan ROC (Receiver Operating Characteristic) eğrisini göstermektedir. Eğri, modelin farklı eşik değerlerinde doğru pozitif oranı (TPR veya Sensitivity) ile yanlış pozitif oranı (FPR veya 1-Specificity) arasındaki ilişkiyi göstermektedir. Grafikte görüldüğü üzere, eğri sol üst köşeye oldukça yakın bir yol izlemektedir, bu da modelin yüksek doğrulukla çalıştığını göstermektedir.

Eğrinin başlangıç noktası olan (0,0) noktasından (1,1) noktasına doğru izlediği yol, modelin pozitif ve negatif sınıflandırma yeteneklerini dengeli bir şekilde gerçekleştirdiğini ortaya koymaktadır. Eğri üzerindeki noktalar, modelin farklı eşik değerlerinde nasıl performans gösterdiğini temsil eder. Eğrinin sol üst köşeye yakın seyretmesi, modelin yüksek bir TPR ile düşük bir FPR sağladığını, yani modelin gerçek pozitifleri yüksek doğrulukla tanımlarken yanlış pozitiflerin sayısını da minimumda tuttuğunu gösterir.

Grafikte ayrıca, rastgele tahminleri temsil eden diyagonal bir referans çizgisi bulunmaktadır. Modelin

ROC eğrisi, bu referans çizgisinin oldukça üzerinde yer almaktadır, bu da modelin rastgele tahminlere göre çok daha iyi performans gösterdiğini kanıtlar. Eğrinin altındaki alan (AUC) ne kadar büyüksse, modelin genel performansı da o kadar iyidir. Bu grafikteki ROC eğrisinin genel şekli ve pozisyonu, modelin oldukça başarılı bir sınıflandırma yeteneğine sahip olduğunu ve farklı eşik değerlerinde yüksek doğruluk sağladığını açıkça ortaya koymaktadır.








Bu görsel, Şarap türlerini Alcohol ve Malic Acid değişkenleri temelinde sınıflandırmayı göstermektedir. Grafikte, her bir veri noktası bir şarap örneğini temsil etmektedir ve noktaların renkleri, farklı şarap türlerini belirtmektedir: mavi (1), kırmızı (2), ve yeşil (3).

Grafiğe bakıldığında, Alcohol ve Malic Acid değişkenlerinin şarap türlerini ayırt etmede önemli rol oynadığı görülmektedir. Mavi noktalar (sınıf 1), genellikle daha yüksek Alcohol seviyelerine ve düşük Malic Acid seviyelerine sahipken, kırmızı noktalar (sınıf 2) daha düşük Alcohol seviyelerinde ve geniş bir Malic Acid dağılımına sahiptir. Yeşil noktalar (sınıf 3) ise Alcohol seviyelerinin orta aralığında ve genellikle yüksek Malic Acid seviyelerinde yoğunlaşmaktadır.

Grafikteki sınıflar arasındaki belirgin ayırım, Alcohol seviyesinin artmasıyla birlikte sınıf 1 (mavi) örneklerinin sayısının artması, Malic Acid seviyesinin yükselmesiyle de sınıf 3 (yeşil) örneklerinin sayısının artmasıyla belirginleşmektedir. Sınıf 2 (kırmızı) örnekleri ise daha düşük Alcohol seviyelerinde yoğunlaşmıştır.

Bu dağılım grafiği, Alcohol ve Malic Acid değişkenlerinin şarap türlerini sınıflandırmada etkili olduğunu göstermektedir. Özellikle Alcohol seviyesi, sınıf 1 ve sınıf 2 arasında güçlü bir ayırım faktörü olarak öne çıkarken, Malic Acid seviyesi sınıf 3'ü diğerlerinden ayırmada daha belirgin bir rol oynamaktadır. Bu tür görselleştirmeler, veri madenciliği ve sınıflandırma algoritmalarının etkinliğini ve verilerin nasıl ayrıştığını anlamada önemli bir araçtır.

| | Name | Distribution | Mean | Mode | Median | Dispersion | Min. | Max. | Missing |
|---|-------------------|---|---------|-------|--------|------------|-------|-------|---------|
| N | Color intensity |  | 5.05809 | 2.6 | 4.69 | 0.457043 | 1.28 | 13 | 0 (0 %) |
| N | Alcohol |  | 13.0006 | 12.37 | 13.05 | 0.0623 | 11.03 | 14.83 | 0 (0 %) |
| N | Alcalinity of ash |  | 19.495 | 20.0 | 19.5 | 0.171 | 10.6 | 30.0 | 0 (0 %) |
| N | Magnesium |  | 99.74 | 88 | 98 | 0.14 | 70 | 162 | 0 (0 %) |
| N | Proline |  | 746.89 | 520 | 673.50 | 0.42 | 278 | 1680 | 0 (0 %) |

Bu görsel, Wine veri setinin beş farklı kimyasal özelliğine ilişkin özet istatistikleri ve dağılımlarını göstermektedir: Color Intensity, Alcohol, Alcalinity of Ash, Magnesium ve Proline. Her bir özellik için dağılım grafikleri, ortalama (Mean), mod (Mode), medyan (Median), saçılma (Dispersion), minimum (Min.), maksimum (Max.) ve eksik değer sayısı (Missing) gibi istatistiksel değerler sunulmaktadır.

- **Color İntensity:** Ortalama 5.05809, mod: 2.6, medyan 4.69, yayılım 0.457043
- **Alcohol:** Ortalama 13.0006, mod: 12.37. medyan 13.05, yayılım 0.0623
- **Alcalinity of ash:** Ortalama 19.495, mod: 20.0, medyan 19.50, yayılım 0.171
- **Magnesium:** Ortalama 99.74, mod:88, medyan 98, yayılım 0.14
- **Proline:** Ortalama 746.89, mod:520, medyan 673.5, yayılım 0.42

Bu özet tablo, Wine veri setindeki belirli kimyasal özelliklerin genel dağılımını ve istatistiksel özetini sunarak, her bir özelliğin veri kümesindeki varyasyonunu ve merkezi eğilimlerini ortaya koymaktadır.

| | IF conditions | → | THEN class | Distribution | Probabilities [%] | Quality | Length |
|----|---|---|------------|--------------|-------------------|---------|--------|
| 0 | Hue \geq 1.31 | → | Wine=2 | [0, 9, 0] | 8 : 83 : 8 | -0.00 | 1 |
| 1 | Proline \geq 970.0 | → | Wine=1 | [46, 0, 0] | 96 : 2 : 2 | -0.00 | 1 |
| 2 | Color intensity \geq 6.62 | → | Wine=3 | [0, 0, 29] | 3 : 3 : 94 | -0.00 | 1 |
| 3 | Proline \leq 510.0 AND Alcalinity of ash \geq 20.7 | → | Wine=2 | [0, 20, 0] | 4 : 91 : 4 | -0.00 | 2 |
| 4 | Color intensity \leq 3.52 AND Nonflavanoid phenols \geq 0.29 | → | Wine=2 | [0, 21, 0] | 4 : 92 : 4 | -0.00 | 2 |
| 5 | Ash \leq 2.1 AND Hue \geq 0.98 | → | Wine=2 | [0, 10, 0] | 8 : 85 : 8 | -0.00 | 2 |
| 6 | OD280/OD315 of diluted wines \geq 3.35 | → | Wine=1 | [9, 0, 0] | 83 : 8 : 8 | -0.00 | 1 |
| 7 | Color intensity \geq 4.9 | → | Wine=3 | [0, 0, 13] | 6 : 6 : 88 | -0.00 | 1 |
| 8 | Color intensity \leq 3.58 AND Ash \geq 3.22 | → | Wine=1 | [1, 0, 0] | 50 : 25 : 25 | -0.00 | 2 |
| 9 | Color intensity \leq 3.85 AND Malic Acid \geq 4.72 | → | Wine=3 | [0, 0, 1] | 25 : 25 : 50 | -0.00 | 2 |
| 10 | Alcohol \leq 12.86 AND Malic Acid \geq 1.81 | → | Wine=2 | [0, 6, 0] | 11 : 78 : 11 | -0.00 | 2 |
| 11 | Ash \geq 2.84 | → | Wine=1 | [2, 0, 0] | 60 : 20 : 20 | -0.00 | 1 |
| 12 | Ash \geq 2.7 | → | Wine=3 | [0, 0, 2] | 20 : 20 : 60 | -0.00 | 1 |
| 13 | Ash \geq 2.56 | → | Wine=2 | [0, 2, 0] | 20 : 60 : 20 | -0.00 | 1 |
| 14 | Color intensity \geq 4.45 | → | Wine=2 | [0, 2, 0] | 20 : 60 : 20 | -0.00 | 1 |
| 15 | Ash \geq 2.32 | → | Wine=3 | [0, 0, 2] | 20 : 20 : 60 | -0.00 | 1 |
| 16 | Alcohol \leq 13.16 AND Malic Acid \geq 3.57 | → | Wine=3 | [0, 0, 1] | 25 : 25 : 50 | -0.00 | 2 |
| 17 | Alcohol \geq 13.24 | → | Wine=1 | [1, 0, 0] | 50 : 25 : 25 | -0.00 | 1 |
| 18 | Alcohol \leq 11.96 | → | Wine=2 | [0, 1, 0] | 25 : 50 : 25 | -0.00 | 1 |
| 19 | TRUE | → | Wine=2 | [59, 71, 48] | 33 : 40 : 27 | -1.567 | 0 |

Bu görselde, Wine veri seti üzerinde kuralla dayalı sınıflandırma yöntemlerini kullanarak oluşturulmuş karar kurallarını göstermektedir. Her bir satır, belirli bir "IF" koşulu sağlandığında şarap örneğinin hangi sınıfa ("THEN" class) atanacağını belirtmektedir. Örneğin, Hue > 1.31 koşulu sağlandığında, şarap örneği Wine=2 sınıfına atanır.

Tablo, her bir kuralın koşulunu, sonucunu ve bu kurala uyan örneklerin sınıflar arasındaki dağılımını içermektedir. Dağılım sütunu, koşulu sağlayan örneklerin farklı sınıflara nasıl dağıldığını gösterir. Örneğin, Proline > 970.0 koşulu sağlandığında, dağılım [46, 0, 0] olup, tüm örnekler Wine=1 sınıfına aittir. Bu durum, bu kuralın %100 doğrulukla bu sınıfa ait olduğunu göstermektedir.

Olasılıklar sütunu, dağılımın yüzdelik gösterimini sunar. Örneğin, Hue > 1.31 koşulu için olasılıklar %8 : %83 : %8 olarak belirtilmiş olup, bu da örneklerin %83'ünün Wine=2 sınıfına ait olduğunu gösterir. Kalite sütunu, kuralın doğruluğunu ve etkinliğini belirten bir ölçüttür. Uzunluk sütunu ise, kuralın kaç terimden oluştuğunu belirtir ve daha karmaşık kuralların analizi için kullanılır.

Bu tablo, kural tabanlı modelin şarap türlerini belirli kimyasal özelliklere dayanarak nasıl sınıflandırdığını

detaylı bir şekilde göstermektedir. Her bir kural, veri setindeki belirli özellik kombinasyonlarına dayalı olarak belirli bir sınıfa yüksek doğrulukla atama yapmaktadır. Modelin karar verme sürecini anlamak ve doğruluğunu değerlendirmek için bu kurallar oldukça yararlıdır. Ayrıca, belirli koşullar altında hangi sınıflandırmaların daha güvenilir olduğunu ve hangi koşullar altında daha fazla çeşitlilik olduğunu göstermektedir. Bu bilgiler, modelin performansını artırmak ve gerekli düzenlemeleri yapmak için kullanılabilir.

6. Karşılaştırma

Wine veri seti üzerine yapılmış olan akademik çalışmalardan birini inceledik. Mahima Gupta ve arkadaşlarının çalışmasında, Random Forest ve KNN algoritmaları kullanılarak şarap kalitesi değerlendirilmiştir. Bu çalışmalarda elde edilen sonuçlarla bizim sonuçlarımız karşılaştırılmıştır.

- **Random Forest:** Gupta ve arkadaşlarının çalışmasında Random Forest algoritması kullanılarak elde edilen doğruluk oranı %93.2 olarak rapor edilmiştir. F1-Measure değeri ise %91.0'dır.
- **K-Nearest Neighbors (KNN):** Aynı çalışmada KNN algoritması ile doğruluk oranı %91.5 ve F1-Measure değeri %89.8 olarak bulunmuştur.

Ayrıca, [RStudio Pubs'daki çalışmada](#), şarap türü ve kalitesi tahmini için çeşitli makine öğrenimi algoritmaları kullanılmıştır. Bu çalışmada, şarap kalitesi sınıflandırmasında kullanılan karar ağacı modeli için aşağıdaki sonuçlar elde edilmiştir:

- **Accuracy (Doğruluk):** %92.8
- **F1-Measure:** %90.5
- **Precision (Hassasiyet):** %93.8
- **Recall (Duyarlılık):** %87.5

Bizim çalışmamızda karar ağacı modeli kullanılarak elde edilen sonuçlar:

- **Accuracy (Doğruluk):** %86.8
- **F1-Measure:** %86.8
- **Precision (Hassasiyet):** %86.9
- **Recall (Duyarlılık):** %86.8

Bu karşılaştırmadan görüldüğü gibi, bizim modelimiz oldukça rekabetçi bir performans sergilemiştir. Her ne kadar RStudio Pubs'daki çalışma ile karşılaştırıldığında doğruluk ve F1-Measure değerlerimiz biraz daha düşük olsa da, modelimizin elde ettiği sonuçlar genel olarak tatmin edicidir. RStudio Pubs'daki çalışma %92.8 doğruluk ve %90.5 F1-Measure değerlerine ulaşırken, bizim modelimiz %86.8 doğruluk ve %86.8 F1-Measure değerlerine ulaşmıştır. Precision ve Recall değerlerimiz de benzer şekilde %86.9 ve %86.8 olarak ölçülmüştür.

Bu sonuçlar, karar ağacı modelimizin şarap kalitesini sınıflandırmada başarılı olduğunu göstermektedir. Modelimizin performansını daha da iyileştirmek için hiperparametre optimizasyonu ve daha fazla veri ön işleme adımları uygulanabilir. Genel olarak, elde ettiğimiz sonuçlar, modelimizin diğer çalışmalarla kıyaslandığında oldukça iyi bir performans sergilediğini ve şarap kalitesi sınıflandırmasında kullanılabilir olduğunu göstermektedir.

7. Sonuç ve Tartışma

Bu projede, Wine veri seti üzerinde kural tabanlı yöntemler kullanılarak şarap türlerinin sınıflandırılması gerçekleştirilmiştir. Veri seti, UCI Machine Learning Repository'den alınmış olup, 13 kimyasal özellik ve 3 farklı şarap türünü içermektedir. Verilerin eksiksiz ve doğru bir şekilde analiz edilebilmesi için eksik veriler temizlenmiş, gerekli dönüşümler yapılmış ve veriler normalizasyon işlemine tabi tutulmuştur.

Karar ağacı modeli, şarap türlerini sınıflandırmada etkili bir yöntem olarak kullanılmıştır. Modelin performansını değerlendirmek için doğruluk (Accuracy), F1-Measure, hassasiyet (Precision) ve duyarlılık (Recall) gibi metrikler kullanılmıştır. Elde edilen sonuçlar, karar ağacı modelimizin genel olarak iyi bir performans sergilediğini göstermektedir. Modelimizin doğruluk oranı %86.8, F1-Measure değeri %86.8, hassasiyet %86.9 ve duyarlılık %86.8 olarak hesaplanmıştır.

Sonuçlarımızı diğer çalışmalarla karşılaştırdığımızda, Mahima Gupta ve arkadaşlarının çalışmasında Random Forest ve K-Nearest Neighbors (KNN) algoritmaları ile elde edilen sonuçlarla benzer doğruluk ve F1-Measure değerlerine ulaşılmadığı görülmüştür. Gupta ve arkadaşlarının çalışmasında Random Forest algoritması %93.2 doğruluk ve %91.0 F1-Measure değeri elde ederken, KNN algoritması %91.5 doğruluk ve %89.8 F1-Measure değeri elde etmiştir. RStudio Pubs'daki bir diğer çalışmada ise karar ağacı modeli %92.8 doğruluk ve %90.5 F1-Measure değeri ile öne çıkmıştır.

Bizim modelimizin doğruluk ve F1-Measure değerleri, diğer çalışmalara kıyasla biraz daha düşük kalmış olsa da, elde edilen sonuçlar tatmin edici ve umut vericidir. Modelimizin rekabetçi performansı, kural tabanlı yöntemlerin şarap kalitesini sınıflandırmada kullanılabileceğini göstermektedir. Ancak, modelimizin performansını daha da iyileştirmek için hiperparametre optimizasyonu ve daha ileri veri ön işleme teknikleri uygulanabilir.

Genel olarak, bu çalışma, kural tabanlı yöntemlerin şarap türlerinin sınıflandırılmasında etkili bir araç olduğunu göstermektedir. Elde edilen sonuçlar, modelimizin başarılı olduğunu ve benzer çalışmalarla karşılaştırıldığında rekabetçi bir performans sergilediğini ortaya koymaktadır. Gelecekte yapılacak çalışmalar, modelin performansını artırmak ve daha geniş veri kümeleri üzerinde test etmek amacıyla daha gelişmiş tekniklerin uygulanmasını içerebilir.

8. Kaynaklar

1. UCI Machine Learning Repository: Wine Dataset
2. Mahima Gupta et al. (2020), "Random Forest and KNN Algorithms for Wine Quality Classification"
3. RStudio Pubs: Wine Type and Quality Prediction With Machine Learning
4. OpenAI's ChatGPT for assistance and additional information.
5. Rule Induction : Orange