logo, metin, simge, sembol, kırpıntı çizim içeren bir resim

Açıklama otomatik olarak oluşturuldu

T.C.

ONDOKUZ MAYIS ÜNİVERSİTESİ LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ BİLGİSAYAR MÜHENDİSLİĞİ

ÖRÜNTÜ TANIMA-BM605

Şarap Kalitesi Veri Seti ile Analiz Raporu

FIRAT KAAN BİTMEZ - 23281855

SAMSUN, 2024-2025 Eğitim Öğretim Yılı Güz Yarıyılı

**1) Veri Kümesi Ne Zaman ve Hangi Makale/Çalışma Kapsamında Oluşturulmuştur?**

Bu çalışmada kullandığımız veri kümesi, **Paulo Cortez** ve arkadaşları tarafından **6 Ekim 2009** tarihinde oluşturulmuştur. Veri seti, "Decision Support Systems" dergisinde yayımlanan "Modeling wine preferences by data mining from physicochemical properties" adlı makale kapsamında toplanmıştır. Çalışma, Portekiz’in kuzeybatı bölgesinde üretilen **Vinho Verde** şaraplarına yönelik fiziksel ve kimyasal testlerin sonuçlarını kullanarak şarap kalitesini modellemeyi amaçlamaktadır. Bu çalışma kapsamında, şarapların tat profillerini tahmin etmek ve sertifikasyon sürecinde oenologların (oenolog:şarap bilimcisi) karar verme süreçlerine yardımcı olmak için destek vektör makineleri, yapay sinir ağları ve regresyon yöntemleri kullanılmıştır.

**2) Veri Kümesi Kaç Sınıftan Oluşmaktadır?**

Veri kümesinde yer alan şaraplar **kalite(quailty)** açısından sınıflandırılmıştır ve hedef değişken (quality), 0 ile 10 arasında bir değere sahiptir. Ancak pratikte veri setlerinde genellikle **3 ile 9** arasında kalite değerleri bulunur.

* **Kırmızı Şarap Sınıfları**: 3, 4, 5, 6, 7, 8
* **Beyaz Şarap Sınıfları**: 3, 4, 5, 6, 7, 8, 9

Kırmızı şarap veri setinde 6 sınıf, beyaz şarap veri setinde ise 7 sınıf bulunmaktadır.

**3) Veri Kümesi Dağılımı Nasıl? (Dengeli/Dengesiz)**

**metin, ekran görüntüsü, diyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Açıklama otomatik olarak oluşturuldu**

Veri setindeki sınıfların dağılımı dengeli değildir. Genel olarak düşük kaliteli şaraplar daha fazla sayıda örneğe sahipken, yüksek kaliteli şaraplar daha azdır. Bu, veri setinin sınıf dengesizliği olduğunu göstermektedir. Bu dengesizlik, sınıflandırma algoritmalarının performansını etkileyebilir.

**Sınıfların Quailty’e Göre Dağılımları**

|  |  |  |
| --- | --- | --- |
| Quality | Red Wine Sınıf Dağılımı | White Wine Sınıf Dağılımı |
| 3 | 10 | 20 |
| 4 | 53 | 163 |
| 5 | 681 | 1457 |
| 6 | 638 | 2198 |
| 7 | 199 | 880 |
| 8 | 18 | 175 |
| 9 |  | 5 |

Bu dağılımlar bize veri setinin sınıflar arasında dengesiz olduğunu, özellikle orta kaliteye sahip şarapların daha fazla olduğunu göstermektedir. Bu dengesizlik, makine öğrenmesi algoritmalarının daha düşük performans göstermesine neden olabilir. Düşük kaliteli şarapların daha fazla örneğe sahip olması, modelin bu sınıfları daha doğru tahmin etmesine yol açarken, yüksek kaliteli şaraplar için tahmin doğruluğu azalabilir. Bu durumu düzeltmek için sınıf ağırlıklandırma gibi yöntemler kullanılabilir.

**4) Veri Kümesi İçerisinde Kaç Tane Örnek Yer Almaktadır?**

Veri setinde toplam:

* **Kırmızı Şarap** veri setinde **1.599** örnek bulunmaktadır.
* **Beyaz Şarap** veri setinde **4.898** örnek bulunmaktadır.

Her iki veri setinde de her örnek bir şarap numunesini temsil etmektedir.

Toplamda 6.497 örnekten oluşan veri kümesi, şarap kalitesiyle ilgili oldukça geniş bir veri sağlamaktadır. Bu örneklerin büyük bir kısmı orta kalite şaraplara aittir. Bu durum, modelin orta kalite şaraplar için yüksek performans göstermesine olanak tanırken, nadir bulunan yüksek kalite şaraplar için zorluk yaratabilir. Ayrıca, veri setinin boyutu, makine öğrenmesi modellerinin eğitim sürecini hızlandırmak ve modelin genelleme kapasitesini artırmak için yeterli büyüklüktedir.

**5) Veri Kümesinde Boş/Eksik/Yanlış Değer Var mı?**

Veri kümesinde herhangi bir eksik veya boş değer bulunmamaktadır. Bu durum, veri analiz sürecinde herhangi bir ön işleme gerek olmadığını ve tüm özelliklerin tam olduğunu göstermektedir. Eksik veri olup olmadığını kontrol etmek için isnull().sum() metodu kullanılarak her iki veri seti de incelenmiştir ve tüm sütunlarda "0" eksik değer görülmüştür.

**Özellikler**: Sabit asidite, uçucu asidite, sitrik asit, artan şeker, kloridler, serbest sülfür dioksit, toplam sülfür dioksit, yoğunluk, pH, sülfatlar, alkol, kalite.

|  |  |  |
| --- | --- | --- |
| Features | Red Wine Eksik Veri Kontrolü | White Wine Eksik Veri Kontrolü |
| fixed acidity | 0 | 0 |
| volatile acidity | 0 | 0 |
| citric acid | 0 | 0 |
| residual sugar | 0 | 0 |
| chlorides | 0 | 0 |
| free sulfur dioxide | 0 | 0 |
| total sulfur dioxide | 0 | 0 |
| density | 0 | 0 |
| pH | 0 | 0 |
| sulphates | 0 | 0 |
| alcohol | 0 | 0 |
| quality | 0 | 0 |

**6) Veri Kümesinde Kaç Adet Öznitelik Vardır? Bu Öznitelikler Hangi Yapıdadır?**

Her iki veri setinde de 12 adet fiziksel ve kimyasal özellik bulunmaktadır. Bu öznitelikler ve yapılarına dair bilgiler aşağıdaki gibidir:

* **Öznitelikler**:
  1. Sabit Asidite (fixed acidity) - Sayısal (float)
  2. Uçucu Asidite (volatile acidity) - Sayısal (float)
  3. Sitrik Asit (citric acid) - Sayısal (float)
  4. Artık Şeker (residual sugar) - Sayısal (float)
  5. Kloridler (chlorides) - Sayısal (float)
  6. Serbest Sülfür Dioksit (free sulfur dioxide) - Sayısal (float)
  7. Toplam Sülfür Dioksit (total sulfur dioxide) - Sayısal (float)
  8. Yoğunluk (density) - Sayısal (float)
  9. pH - Sayısal (float)
  10. Sülfatlar (sulphates) - Sayısal (float)
  11. Alkol (alcohol) - Sayısal (float)
  12. Kalite (quality) - Sayısal (int)

Tüm öznitelikler sayısal verilerdir ve ölçülebilir fizikokimyasal testler sonucunda elde edilmiştir.

**7) Temel İstatistiksel Özellikler**

Her iki veri setinin temel istatistiksel özelliklerini analiz ettik. Bu analizde, ortalama (mean), standart sapma (std), minimum ve maksimum değerler bulunmuştur.

# Gerekli kütüphanelerin yüklenmesi

import pandas as pd

# Red ve White wine veri setlerinin yolları

red\_wine\_path = r'/mnt/data/winequality-red.csv'

white\_wine\_path = r'/mnt/data/winequality-white.csv'

# Veri setlerinin okunması

red\_wine = pd.read\_csv(red\_wine\_path, sep=';')

white\_wine = pd.read\_csv(white\_wine\_path, sep=';')

# Red Wine için istatistiksel özet

print("\nRed Wine Dataset - Statistical Summary:")

print(red\_wine.describe())

# White Wine için istatistiksel özet

print("\nWhite Wine Dataset - Statistical Summary:")

print(white\_wine.describe())

Veri kümesinin temel istatistiksel özellikleri, her bir öznitelik için ortalama, standart sapma, minimum ve maksimum değerler üzerinden analiz edilmiştir. Kırmızı ve beyaz şarap veri setlerinin istatistiksel özetleri şu şekildedir:

* **Kırmızı Şarap**:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Özellik | Ortalama | Standart Sapma | Min | Max |
| fixed acidity | 8.32 | 1.74 | 4.60 | 15.90 |
| volatile acidity | 0.53 | 0.18 | 0.12 | 1.58 |
| citric acid | 0.27 | 0.19 | 0.00 | 1.00 |
| residual sugar | 2.54 | 1.41 | 0.90 | 15.50 |
| chlorides | 0.08 | 0.05 | 0.01 | 0.61 |
| free sulfur dioxide | 15.87 | 10.46 | 1.00 | 72.00 |
| total sulfur dioxide | 46.47 | 32.90 | 6.00 | 289.00 |
| density | 0.9967 | 0.0019 | 0.990 | 1.0037 |
| pH | 3.31 | 0.15 | 2.74 | 4.01 |
| sulphates | 0.66 | 0.17 | 0.33 | 2.00 |
| alcohol | 10.42 | 1.07 | 8.40 | 14.90 |
| quality | 5.64 | 0.81 | 3 | 8 |

* **Beyaz Şarap**:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Özellik | Ortalama | Standart Sapma | Min | Max |
| fixed acidity | 6.85 | 0.84 | 3.80 | 14.20 |
| volatile acidity | 0.28 | 0.10 | 0.08 | 1.10 |
| citric acid | 0.33 | 0.12 | 0.00 | 1.66 |
| residual sugar | 6.39 | 5.07 | 0.60 | 65.80 |
| chlorides | 0.05 | 0.02 | 0.01 | 0.35 |
| free sulfur dioxide | 35.31 | 17.01 | 2.00 | 289.00 |
| total sulfur dioxide | 138.36 | 42.50 | 9.00 | 440.00 |
| density | 0.9940 | 0.0030 | 0.987 | 1.039 |
| pH | 3.19 | 0.15 | 2.72 | 3.82 |
| sulphates | 0.49 | 0.11 | 0.22 | 1.08 |
| alcohol | 10.51 |  |  |  |

Veri setinin temel istatistiksel özellikleri incelendiğinde, "alkol" ve "sülfatlar" gibi özniteliklerin kaliteyle daha güçlü ilişkiler gösterdiği görülmektedir. Özellikle, kırmızı ve beyaz şaraplarda alkol oranı arttıkça kalite puanlarının da artış gösterdiği dikkat çekmektedir. Standart sapma değerleri, bazı özniteliklerde büyük bir varyans olduğunu göstermektedir, bu da bazı özniteliklerin farklı kalite sınıflarında daha belirleyici olabileceğini işaret etmektedir.

**8) Box Plot Grafikleri ile Özniteliklerin Yorumlanması**

import seaborn as sns

import matplotlib.pyplot as plt

# Red wine için box plot

plt.figure(figsize=(10, 6))

sns.boxplot(data=red\_wine)

plt.title('Red Wine Özniteliklerinin Box Plot Grafiği')

plt.xticks(rotation=90)

plt.show()

# White wine için box plot

plt.figure(figsize=(10, 6))

sns.boxplot(data=white\_wine)

plt.title('White Wine Özniteliklerinin Box Plot Grafiği')

plt.xticks(rotation=90)

plt.show()

**diyagram, metin, ekran görüntüsü, tasarım içeren bir resim

Açıklama otomatik olarak oluşturuldudiyagram, metin, ekran görüntüsü, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu**

Box plot grafiklerinden elde edilen bulgular, veri setinde bazı uç değerlerin var olduğunu ortaya koymaktadır. Özellikle "alkol" ve "sülfatlar" özniteliklerinde tespit edilen bu uç değerler, model performansını olumsuz etkileyebilir. Bu tür uç değerlerin modele dahil edilip edilmeyeceği veya bu değerlere yönelik veri dönüşümleri (örneğin, logaritmik dönüşüm) yapılması gerektiği değerlendirilmelidir. Modelin performansını artırmak için uç değerlerin etkisini minimize etmek önemlidir.

**9) Özniteliklerin Histogram Grafiklerimetin, diyagram, plan, harita içeren bir resim

Açıklama otomatik olarak oluşturuldumetin, diyagram, harita, plan içeren bir resim

Açıklama otomatik olarak oluşturuldu**

Özniteliklerin histogram grafikleri çizilmiş ve her bir özniteliğin veri setindeki dağılımı incelenmiştir. Histogram grafikleri, veri setindeki değerlerin nasıl dağıldığını görmek için kullanılır. Özellikle alkol ve asidite gibi özniteliklerin normal dağılımdan sapmaları analiz edilmiştir.

# Red wine histogram

red\_wine.hist(bins=15, figsize=(10, 8))

plt.suptitle('Red Wine Özniteliklerinin Histogramı')

plt.show()

# White wine histogram

white\_wine.hist(bins=15, figsize=(10, 8))

plt.suptitle('White Wine Özniteliklerinin Histogramı')

plt.show()

**10) Scatter Plot ile Veri Kümesinin Görselleştirilmesi**

# Red Wine Scatter Plot

sns.pairplot(red\_wine, diag\_kind="kde")

plt.suptitle('Red Wine Scatter Plot Grafikleri')

plt.show()

# White Wine Scatter Plot

sns.pairplot(white\_wine, diag\_kind="kde")

plt.suptitle('White Wine Scatter Plot Grafikleri')

plt.show()

**metin, kalıp, desen, düzen, tipografi içeren bir resim

Açıklama otomatik olarak oluşturuldumetin, kalıp, desen, düzen, el yazısı içeren bir resim

Açıklama otomatik olarak oluşturuldu**

Scatter plot grafikleri, özniteliklerin birbirleriyle olan ilişkisini incelemek için kullanılmıştır. Özellikle **alkol** ile **kalite** arasındaki ilişki dikkat çekicidir. Alkol oranı arttıkça kalite puanlarının da arttığı gözlemlenmiştir. Aşağıda bu grafikle ilgili bir örnek yer almaktadır.

**11) Scatter Plot ile Özniteliklerin Sınıf Ayrımını Göstermemetin, el yazısı, kalıp, desen, düzen içeren bir resim

Açıklama otomatik olarak oluşturuldumetin, el yazısı, ekran görüntüsü, kalıp, desen, düzen içeren bir resim

Açıklama otomatik olarak oluşturuldu**

Scatter plot ile farklı öznitelikler arasındaki sınıf ayrımını inceleyeceğiz. Örneğin, alcohol ile quality arasındaki ilişkiyi gözlemleyeceğiz.

sns.scatterplot(x='alcohol', y='quality', data=red\_wine)

plt.title('Red Wine - Alcohol ve Quality İlişkisi')

plt.show()

sns.scatterplot(x='alcohol', y='quality', data=white\_wine)

plt.title('White Wine - Alcohol ve Quality İlişkisi')

plt.show()

**12) Violin Plot ile Sınıf Yoğunluklarının Görselleştirilmesi**

Violin plot ile her bir sınıfın yoğunluğunu göstereceğiz.

diyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, ekran görüntüsü, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldudiyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, çizgi, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu

# Red wine violin plot

plt.figure(figsize=(10, 6))

sns.violinplot(x='quality', y='alcohol', data=red\_wine)

plt.title('Red Wine - Quality ve Alcohol Violin Plot')

plt.show()

# White wine violin plot

plt.figure(figsize=(10, 6))

sns.violinplot(x='quality', y='alcohol', data=white\_wine)

plt.title('White Wine - Quality ve Alcohol Violin Plot')

plt.show()

**13) Sınıflandırma Algoritmaları ile Deneyler**

Veri kümesinde sınıflandırma algoritmaları kullanarak sonuçları değerlendireceğiz.

Aşağıdaki beş farklı algoritma her iki veri seti için uygulanmıştır:

* **KNN (K-Nearest Neighbors):** Verilere yakın olan komşuları baz alarak sınıflandırma yapar.
* **SVM (Support Vector Machines):** Verileri maksimum marj ile ayıran bir sınır çizer.
* **Naive Bayes:** Her bir özniteliğin sınıftan bağımsız olduğunu varsayarak olasılık hesaplamaları yapar.
* **Random Forest:** Birden fazla karar ağacından oluşan bir modeldir. Her bir ağaç için rastgele öznitelikler ve örnekler kullanılır.
* **Decision Tree:** Karar ağacı kullanarak veriyi sınıflandırır.

Her algoritmanın her iki veri seti için sonuçlarını ve bu sonuçlar üzerinden yorumları şu şekilde açıklayabiliriz:

**Kırmızı Şarap (Red Wine) Sonuçları:**

* **KNN Accuracy**: %48.54
  + KNN algoritması, kırmızı şarap veri setinde %48.54 oranında bir doğruluk elde etmiştir. Bu düşük sonuç, veri setindeki dengesizlik ve özniteliklerin dağılımından kaynaklanıyor olabilir.
* **SVM Accuracy**: %50.42
  + SVM algoritması, KNN'e göre bir miktar daha iyi performans göstermiştir ancak genel doğruluğu sınırlı kalmıştır.
* **Naive Bayes Accuracy**: %54.17
  + Naive Bayes algoritması, bağımsızlık varsayımına dayanmasına rağmen kırmızı şarap veri setinde %54.17 doğruluk oranı elde etmiştir. Bu sonuç, özellikle öznitelikler arasında güçlü ilişkiler bulunan veri setlerinde Naive Bayes'in sınırlamalarını göstermektedir.
* **Random Forest Accuracy**: %65.83
  + Random Forest algoritması, en yüksek doğruluğu sağlamış olup, %65.83 doğrulukla diğer algoritmaların çok üzerinde performans göstermiştir. Bu durum, Random Forest'in birden fazla karar ağacı kullanarak daha karmaşık veri yapısını daha iyi modelleyebilmesinden kaynaklanmaktadır.
* **Decision Tree Accuracy**: %56.25
  + Decision Tree algoritması, Naive Bayes'ten biraz daha iyi sonuç vermiştir. Ancak tek bir karar ağacına dayanması nedeniyle Random Forest kadar yüksek performans göstermemiştir.

**Beyaz Şarap (White Wine) Sonuçları:**

* **KNN Accuracy**: %47.55
  + KNN algoritması, beyaz şarap veri setinde de düşük performans sergilemiştir. Veri setindeki sınıf dağılımı ve yakınlık hesaplamalarının etkisi burada da görülmektedir.
* **SVM Accuracy**: %45.92
  + SVM, beyaz şarap veri setinde %45.92 doğruluk oranı ile en düşük performansı göstermiştir. Özellikle sınıf dengesizliği, bu algoritmanın düşük performans göstermesine yol açabilir.
* **Naive Bayes Accuracy**: %42.45
  + Naive Bayes, beyaz şarap veri setinde en düşük doğruluğu elde etmiştir. Veri setindeki öznitelikler arası bağımsızlık varsayımı bu veri seti için geçerli olmadığından, performansı düşmüştür.
* **Random Forest Accuracy**: %67.55
  + Random Forest, %67.55 doğrulukla beyaz şarap veri setinde en yüksek performansı göstermiştir. Birden fazla ağaç ve rastgele öznitelikler kullanarak karmaşık veri yapısını daha iyi modelleyebilmiştir.
* **Decision Tree Accuracy**: %58.78
  + Decision Tree, %58.78 doğruluk oranı ile Random Forest'in gerisinde kalmış ancak diğer algoritmalara kıyasla daha iyi performans sergilemiştir.

Random Forest, sınıf dengesizliği ve karmaşık veri yapısında diğer algoritmalara göre daha yüksek performans sergilemiştir. Ayrıca sınıf ağırlıklandırma veya SMOTE gibi veri dengeleme teknikleri ile doğruluğun daha da artırılması mümkündür.

Bu bölümde, kırmızı ve beyaz şarap veri setleri üzerinde çeşitli sınıflandırma algoritmalarını kullanarak yapılan deneyler açıklanmaktadır. Algoritmaların doğrulukları, precision, recall, ve F1-score değerleri değerlendirilmiş olup her iki veri seti için karşılaştırmalar yapılmıştır.

# **Red Wine Sonuçları**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sınıf | Precision | Recall | F1-Score | Destek (Support) |
| 3 | 0.0 | 0.0 | 0.0 | 1.0 |
| 4 | 0.071 | 0.059 | 0.065 | 17.0 |
| 5 | 0.659 | 0.626 | 0.642 | 195.0 |
| 6 | 0.560 | 0.580 | 0.570 | 200.0 |
| 7 | 0.448 | 0.492 | 0.469 | 61.0 |
| 8 | 0.200 | 0.167 | 0.182 | 6.0 |
| Genel Doğruluk | 0.5625 | 0.5625 | 0.5625 | 0.5625 |
| Macro Avg | 0.323 | 0.320 | 0.321 | 480.0 |
| Weighted Avg | 0.563 | 0.562 | 0.562 | 480.0 |

# **White Wine Sonuçları**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sınıf | Precision | Recall | F1-Score | Destek (Support) |
| 3 | 0.0 | 0.0 | 0.0 | 7.0 |
| 4 | 0.220 | 0.275 | 0.244 | 40.0 |
| 5 | 0.630 | 0.631 | 0.631 | 426.0 |
| 6 | 0.637 | 0.623 | 0.630 | 668.0 |
| 7 | 0.525 | 0.521 | 0.523 | 280.0 |
| 8 | 0.386 | 0.449 | 0.415 | 49.0 |
| Genel Doğruluk | 0.5878 | 0.5878 | 0.5878 | 0.5878 |
| Macro Avg | 0.400 | 0.417 | 0.407 | 1470.0 |
| Weighted Avg | 0.591 | 0.588 | 0.589 | 1470.0 |

**KAYNAKLAR**

Kaynak makale: **Paulo Cortez, Antonio Cerdeira, Fernando Almeida, Telmo Matos, Jose Reis**, *Modeling wine preferences by data mining from physicochemical properties*, Decision Support Systems, Elsevier, 2009.