T.C.
KÜTAHYA DUMLUPINAR ÜNİVERSİTESİ
Mühendislik Fakültesi

Makine Öğrenimi Dersi Proje Raporu

Konu:

Red Wine Quality Veri Seti ile Model Geliştirme

Öğrenci:

Mustafa Colay

202113172004

Süleyman Sefa GÜRER

202013172034

Danışman

Dr. Öğr. Üyesi Pınar Özen KAVAS – Ar. Gör. Gülistan ARSLAN

KÜTAHYA, 2024

Proje Raporu: Makine Öğrenimi Teknikleri ile Şarap Kalitesi Tahmini

İÇİNDEKİLER

1. GİRİŞ: ÖZET

Bu çalışma, şarap kalitesinin tahmini amacıyla çeşitli makine öğrenmesi ve istatistiksel analiz yöntemlerinin uygulanmasını içermektedir. Kullanılan veri seti, kırmızı şaraplarının fiziksel ve kimyasal özelliklerini içeren "Wine Quality Dataset (https://www.kaggle.com/code/nimapourmoradi/red-wine-quality?select=winequality-red.csv)"tir. Veri setinde toplam 11 özellik ve her şarabın 0-10 arasında değişen kalite skorları bulunmaktadır.

Veri analizi kapsamında, veri setinin normalliğini değerlendirmek için belirli istatistiksel testler uygulanmış ve veri ön işleme adımları gerçekleştirilmiştir. Özellik seçimi için belirlenen yöntemler, tahmin modellerinin performansını artırmayı hedeflemiştir. Çalışmada kullanılan modeller arasında KNN, Karar Ağaçları, SVM ve Yapay Sinir Ağları gibi yaygın makine öğrenmesi algoritmaları yer almaktadır. Modellerin performansı, F1 skoru, doğruluk (accuracy), kesinlik (precision) ve hatırlama (recall) gibi sınama metrikleriyle değerlendirilmiştir.

Sonuç olarak, bu çalışma hem şarap kalitesinin tahmini için uygun modelleri belirlemek hem de farklı özelliklerin model performansına olan etkisini analiz etmek amacıyla gerçekleştirilmiştir. Analizler sonucunda elde edilen sonuçlar, yüksek doğruluk oranlarına ulaşılabildiğini göstermiştir ve şarap kalitesinin tahmininde hangi özelliklerin daha etkili olduğunu vurgulamaktadır.

2. MATERYAL ve YÖNTEMLER:

2.1 Dataset'in detaylı özelliklikleri

Veri Setinin Adı: Red Wine Quality

Boyutları:

Satır sayısı: 1599

Sütun sayısı: 12

Sütunlar ve Tanımları:

- fixed acidity: Şarapta bulunan sabit asitler (örneğin tartarik asit).
- volatile acidity: Uçucu asitlerin miktarı (örneğin asetik asit).
- citric acid: Sitrik asit içeriği.
- residual sugar: Kalıntı şeker miktarı.
- chlorides: Tuz içeriği.
- free sulfur dioxide: Serbest kükürt dioksit miktarı.
- total sulfur dioxide: Toplam kükürt dioksit miktarı.
- density: Şarap yoğunluğu.
- pH: Şarabın pH seviyesi.
- sulphates: Sülfat miktarı.
- alcohol: Alkol oranı.
- quality: Şarap kalitesi (3 ile 8 arasında sınıflandırma, hedef değişken). Kalite değer aralığı 1-10'dan 0-1 e dönüştürülmüştür

Eksik Değer Durumu:

- Veri setinde eksik değer bulunmamaktadır.

Dengesizlik Analizi:

Sınıflar arası dağılım:

- Kalite 5: %42,6
- Kalite 6: %39,9
- Kalite 7: %12,4
- Kalite 4: %3,3
- Kalite 8: %1,1
- Kalite 3: %0,6

Sınıf dağılımında belirgin bir dengesizlik bulunmaktadır. En yaygın sınıflar 5 ve 6 iken, diğer sınıflar daha az sayıda gözlem içermektedir. Veri seti dengeli değildir.

Dönüşüm işlemi

Bu durumda veri seti işlenebilirliğinin basitleştirilmesi ve dengeli hale getirilmesi adına 0-1 aralığına çevrilmiştir.

Çevirme işlemi sonrası sınıflar arası dağılım:

- Sınıf 1: 676 örnek %52,85
- Sınıf 0: 603 örnek %47,15

Dönüşüm sonrası sınıf dağılımında belirgin bir dengesizlik bulunmamaktadır. 0 ve 1 sınıflarının dağılımı yaklaşık değerler içermektedir. Veri seti dengelidir.

Veri Setinin Türü:

- Bu veri seti, sentetik olmayıp gerçek dünya verilerinden elde edilmiştir.

2.2 Ön-İşlemler, Normallik Testleri ve Ölçeklendirme

2.2.1 Veri Setinin Yüklenmesi ve Etiketlerin Dönüştürülmesi

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from collections import Counter
from fonksiyonlar import parametre_hesaplama
```

```
# Veri setinin yüklenmesi
data = pd.read_csv('winequality-red.csv')
X = data.drop('quality', axis=1)              #Özellikler
y = data['quality']                           #Etiket
y = y.apply(lambda x: 0 if x <= 5 else 1)     #Düşük/Yüksek kalite sınıfları
```

Kodun başında winequality-red.csv veri seti yükleniyor ve bağımsız değişkenler (X) ile bağımlı değişken (y) ayrılıyor.

Kalite (quality) etiketi **0 ve 1** olmak üzere ikili sınıfa dönüştürülüyor:

Özellikler (X): Şarap kalitesini etkileyen fizikokimyasal değişkenler.

Etiketler (y):

- quality<= 5 → **0 (Düşük Kalite)**
- quality> 5 → **1 (Yüksek Kalite)**

Bu işlem, çok sınıflı problemi ikili sınıflama problemine dönüştürerek modeli basitleştiriyor.

2.2.2 Normallik Testleri

```python
import matplotlib.pyplot as plt
import statsmodels.api as sm

# Özelliklerin isimlerini al
features = data.columns

# Alt grafikler için ayarlar
n_features = len(features)  # Özellik sayısı
rows = (n_features + 1) // 2  # Satır sayısı (2 sütun olacak)
fig, axes = plt.subplots(rows, 2, figsize=(12, 6 * rows))  # Dinamik boyutlandırma
axes = axes.flatten()  # Aksları düzleştir (indeksle kolay erişim için)

# Her özellik için Q-Q plot
for i, feature in enumerate(features):
    # Özelliği seç
    values = data[feature]

    # Q-Q plot
    sm.qqplot(values, line='s', ax=axes[i])
    axes[i].set_title(f"Q-Q Plot: {feature}")

# Boş grafik alanlarını kaldır
for j in range(i + 1, len(axes)):
    fig.delaxes(axes[j])

plt.tight_layout()
plt.show()
```

Veri setindeki özelliklerin normal dağılıma uyup uymadığını analiz etmek için **Q-Q Plot (Quantile-Quantile Plot)** yöntemi kullanılıyor.

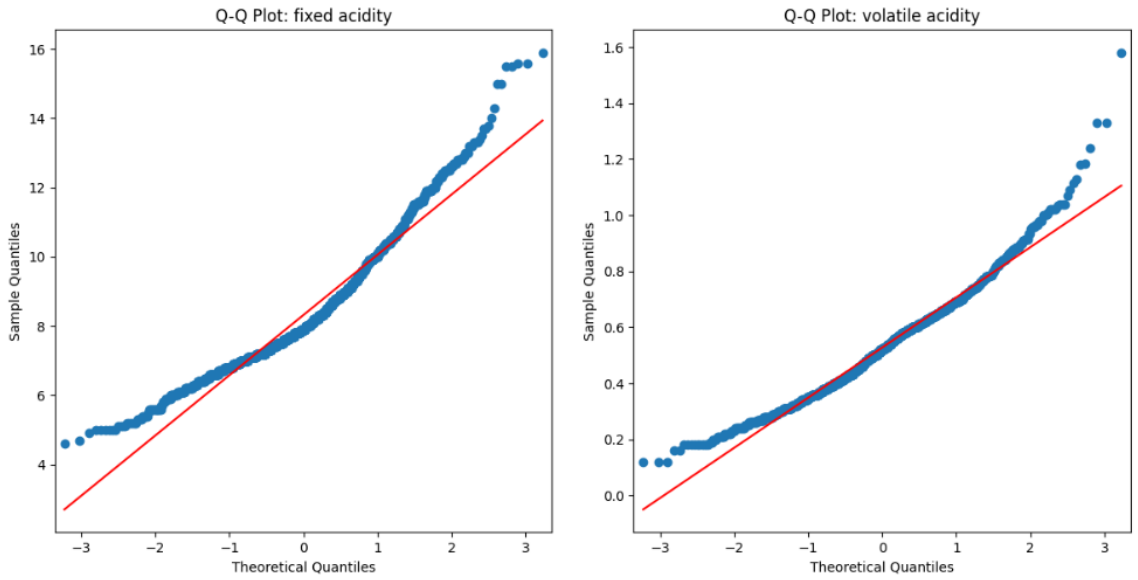
Q-Q Plot Açıklaması

Amaç: Özelliklerin normal dağılıma uyup uymadığını görselleştirmek.

Nasıl çalışır? Q-Q Plot, verinin kuantillerini standart normal dağılımın kuantilleri ile karşılaştırarak dağılımın normalliğini test eder.

Değerlendirme:

- **Eğer noktalar 45° doğrusu (line='s') boyunca diziliyorsa** → Normal dağılıma uygun.
- **Eğer noktalar büyük sapmalar gösteriyorsa** → Normal dağılıma uymuyor.



2.2.3 Veri Setinin Eğitim ve Test Olarak Bölünmesi

Özellik seçimi tamamlandıktan sonra veri seti **%80 eğitim, %20 test** olacak şekilde ayrılıyor.

Bu adım, modelin eğitileceği verileri belirleyerek gerçek dünyadaki tahmin performansını ölçmeye yardımcı olur.

```
X_train, X_test, y_train, y_test = train_test_split(X.values, y.values, test_size=0.2, random_state=42)
```

2.2.4 Veri Ölçeklendirme (StandardScaler)

Özelliklerin farklı ölçeklerde olması makine öğrenmesi algoritmalarının performansını olumsuz etkileyebilir.

Bu nedenle **StandardScaler** kullanılarak veriler **ortalama=0, standart sapma=1** olacak şekilde ölçeklendiriliyor.

Neden StandardScaler Kullanılıyor?

- Ölçek farklarını ortadan kaldırarak modelin daha kararlı öğrenmesini sağlar.
- Özellikle KNN, SVM ve Yapay Sinir Ağları gibi mesafe tabanlı algoritmalar için önemlidir.
- Eğitim ve test verileri aynı ölçeklendirme ile işlenerek tutarlılık sağlanır.

```
# Veriyi Ölçekleme
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)      # Eğitim verisi için StandardScaler, her bir özelliğin
                                             # ortalamasını ve standart sapmasını öğrenir

X_test = scaler.transform(X_test)            # Benzer bir işlem test verisine de uygulanır
                                             # Test verisi sadece eğitim verisinden öğrenilen ortalama ve standart
                                             # sapma değerleri kullanılarak ölçeklendirilir
```

2.2.5 Veri Setindeki Sınıf Dağılımının Kontrolü

Bu işlem, veri setinde sınıf dengesizliği olup olmadığını kontrol etmeye yardımcı olur. Yapılan işlem sonucu veri setinin sınıf dağılımının dengeli olduğu gözlemlenmiştir. Sınıf 1: 676 örnek içerirken, Sınıf 0: 603 örnek içerir.

```
# Veri setindeki sınıf dağılımını kontrol etme
class_distribution = Counter(y_train)
for class_label, count in class_distribution.items():
    print(f"Sınıf {class_label}: {count} örnek")
```

```
Sınıf 1: 676 örnek
Sınıf 0: 603 örnek
```

2.3 Modellerin Eğitimi

Bu bölümde, şarap kalitesi tahmini amacıyla kullanılan makine öğrenimi modelleri ve bu modellerin eğitiminde kullanılan parametreler detaylı bir şekilde ele alınmıştır. Çalışmada, farklı özelliklere ve parametrelere sahip dört temel makine öğrenimi yöntemi uygulanmış ve bu yöntemlerin performansı karşılaştırılmıştır. Ayrıca, topluluk sınıflandırma yaklaşımı kullanılarak modeller birleştirilmiş ve bu sayede doğruluğun artması hedeflenmiştir.

2.3.1. K-En Yakın Komşu (KNN)

KNN algoritması, sınıflandırma işlemi için kullanılan basit ve etkili bir yöntemdir. Bu çalışmada, KNN algoritmasında en yakın komşu sayısı **k = 40** olarak belirlenmiştir. K parametresinin 40 seçilme nedeni ise denemeler sonucu doğruluğun 40 değerinde en fazla olmasıdır. Model, eğitim verisi üzerinde her bir veri noktasını, öklid uzaklığına göre en yakın komşularıyla karşılaştırarak tahmin yapmıştır.

Parametreler:

Komşu sayısı: k = 40

```
from fonksiyonlar import knn_predict

# KNN Algoritmasını Çalıştırma
k = 40  # En yakın komşu sayısı
y_pred = knn_predict(X_train, y_train, X_test, k)
# KNN Algoritması ile tahmin yapma
parametre_hesaplama(y_test, y_pred, "KNN")
```

2.3.2 Karar Ağacı

Karar ağaçları, veri setinin özelliklerine dayalı olarak karar kurallarını belirlemek için kullanılan açıklanabilir bir modeldir. Bu çalışmada, karar ağacının maksimum derinliği **max_depth = 10** olarak belirlenmiştir. Bu parametre, ağacın karmaşıklığını sınırlamak ve aşırı öğrenmeyi önlemek amacıyla optimize edilmiştir.

Parametreler:

Maksimum derinlik: max_depth = 10

```
from fonksiyonlar import decision_tree

derinlik = 10 #10

# Karar Ağacı Modelini Eğit
y_pred_dt = decision_tree(X_train, y_train, X_test, max_depth = derinlik)
# Sonuçları değerlendirme
parametre_hesaplama(y_test, y_pred_dt, "Karar Ağacı("+str(derinlik)+" derinlik)")
```

2.3.3 Destek Vektör Makineleri (SVM)

SVM, doğrusal olmayan veriler için kernel fonksiyonlarıyla desteklenen ve sınıflandırma problemlerinde sıklıkla tercih edilen güçlü bir yöntemdir. Bu çalışmada, SVM'nin eğitiminde en uygun parametre değerleri şu şekilde belirlenmiştir:

Parametreler:

Öğrenme oranı (Learning Rate): lr = 0.01

Epoch sayısı: epochs = 1000

Düzenleme parametresi: C = 1.0

```
from fonksiyonlar import train_svm, svm_predict

# SVM modelini eğitimi
# lr(öğrenme oranı), modelin ağırlıklarını güncellemek için kullanılan adım boyutunu belirtir
#               Küçük öğrenme oranı, daha yavaş ama daha hassas eğitim süreci sağlar
# epochs, eğitim verisinin tamamının model tarafından kaç kere işleneceğini belirler
# C, modelin karmaşıklığını dengeler
#               büyük C değeri modelin daha karmaşık olmasına yol açar (daha fazla hata toleransı)
w, b = train_svm(X_train, y_train, lr=0.01, epochs=1000, C=1.0)

# Test verisi üzerinde tahmin yap
y_pred_svm = svm_predict(X_test, w, b)
parametre_hesaplama(y_test, y_pred_svm, "SVM")
```

2.3.4 Yapay Sinir Ağları (ANN)

Yapay sinir ağları, çok katmanlı algılayıcılar (MLP) ile uygulanmış ve karmaşık ilişkileri öğrenme yeteneği sergilemiştir. Gizli katman boyutu ve epoch sayısı gibi parametreler optimize edilmiştir:

Parametreler:

Gizli katman boyutu: (100,)

Maksimum iterasyon sayısı: max_iter = 1000

```python
from sklearn.neural_network import MLPClassifier

# Yapay Sinir Ağı Modelini Eğit
# hidden_layer_sizes, gizli katmanların boyutunu belirtir
# max_iter, epoch sayısı
ann_model = MLPClassifier(hidden_layer_sizes=(100,), max_iter=1000, random_state=42)
ann_model.fit(X_train, y_train)
y_pred_ann = ann_model.predict(X_test)
parametre_hesaplama(y_test, y_pred_ann, "Yapay Sinir Ağı")
```

2.3.5 Topluluk Sınıflandırma (Hard Voting)

Topluluk sınıflandırma KNN, SVM ve Yapay Sinir Ağları modellerinin tahminlerini birleştiren bir yöntem olarak uygulanmıştır. Çoğunluk oyu yöntemi kullanılarak, en çok tercih edilen sınıf, nihai tahmin olarak belirlenmiştir.

Bu yöntemde Karar Ağacı modeli topluluk sınıflandırmasına dahil edilmemiştir. Bunun başlıca nedeni 4 model olmasıdır. Çoğunluk oyu yönteminde 4 model olması eşit oy çıkmasına neden olacaktır. Bu sebeple modeller arasında en düşük doğruluğa sahip Karar Ağacı modeli topluluk sınıflandırmasına dahil edilmemiştir.

Bu bölümde kullanılan modeller hem bireysel performansları hem de topluluk sınıflandırma ile elde edilen sonuçları değerlendirmek amacıyla seçilmiştir. Her model, uygun hiperparametrelerle optimize edilerek tahmin gücü artırılmıştır.

```python
# Topluluk Sınıflandırma (Hard Voting)
combined_predictions_ann = []
for i in range(len(y_test)):
    knn_pred = y_pred[i]
    rf_pred = y_pred_dt[i]
    svm_pred = y_pred_svm[i]
    ann_pred = y_pred_ann[i]

    # Çoğunluk oyunu alma
    final_pred = Counter([knn_pred, svm_pred, ann_pred]).most_common(1)[0][0]
    combined_predictions_ann.append(final_pred)

parametre_hesaplama(y_test, combined_predictions_ann, "KNN + SVM + Yapay Sinir Ağı\n(Hard Voting)")
```

2.4 ROC-Eğrisi, Karmaşıklık Matrisleri ve Metrikler

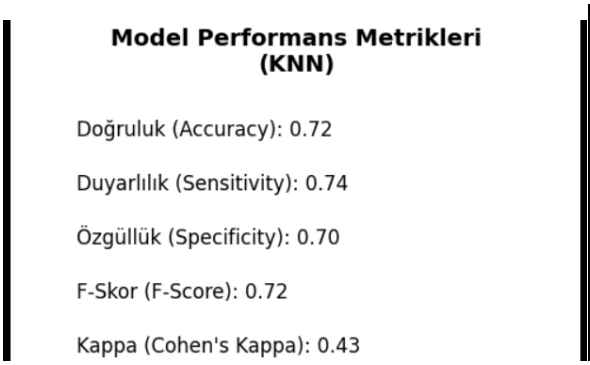
Bu bölümde, kullanılan modellerin performans değerlendirmesi için hesaplanan metrikler, karışıklık matrisleri ve ROC eğrileri ele alınmıştır. Çalışmada, şarap kalitesini düşük ve yüksek olarak sınıflandıran modellerin doğruluk, duyarlılık, özgüllük, F-Skor ve Cohen's Kappa gibi metriklerle performansı değerlendirilmiştir. Ayrıca, modellerin tahmin performansını görselleştirmek amacıyla ROC eğrileri ve AUC (Eğri Altında Kalan Alan) değerleri hesaplanmıştır.

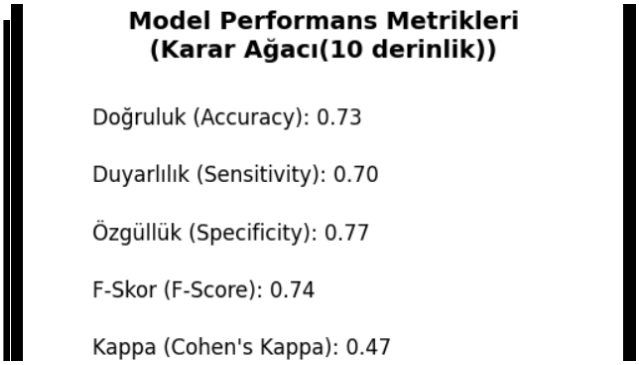
2.4.1 Performans Metrikleri

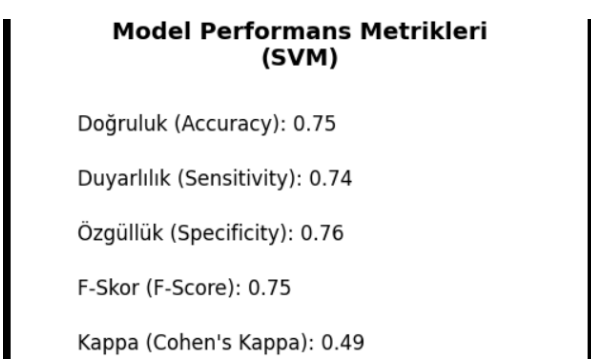
Modellerin doğruluğu (accuracy), duyarlılığı (sensitivity), özgüllüğü (specificity), F-Skoru ve Cohen's Kappa metrikleri, tahmin sonuçlarının gerçek değerlerle karşılaştırılmasıyla

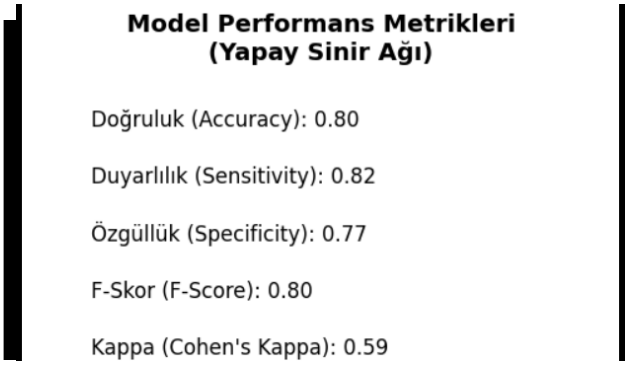
hesaplanmıştır. Bu metrikler, sınıflandırma modellerinin genel performansını ve dengeleyiciliğini değerlendirmek için kullanılmıştır:

- **Doğruluk (Accuracy):** Modelin doğru sınıflandırma oranını ölçer.
- **Duyarlılık (Sensitivity):** Gerçek pozitiflerin doğru sınıflandırılma oranını ölçer.
- **Özgüllük (Specificity):** Gerçek negatiflerin doğru sınıflandırılma oranını ölçer.
- **F-Skor (F-Score):** Duyarlılık ve özgüllük arasında bir denge sağlayarak modelin genel performansını özetler.
- **Cohen's Kappa:** Tahminlerin rastlantısal doğruluğa göre ne kadar iyi olduğunu değerlendirir.





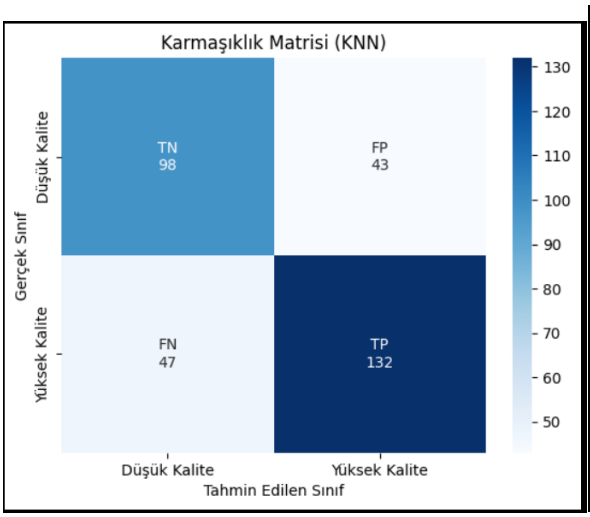


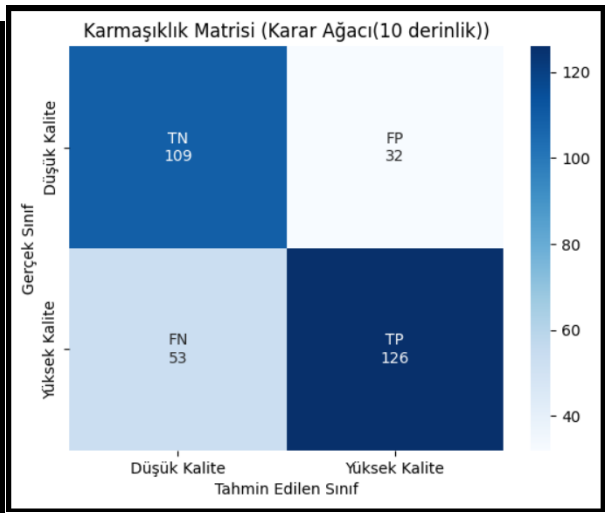


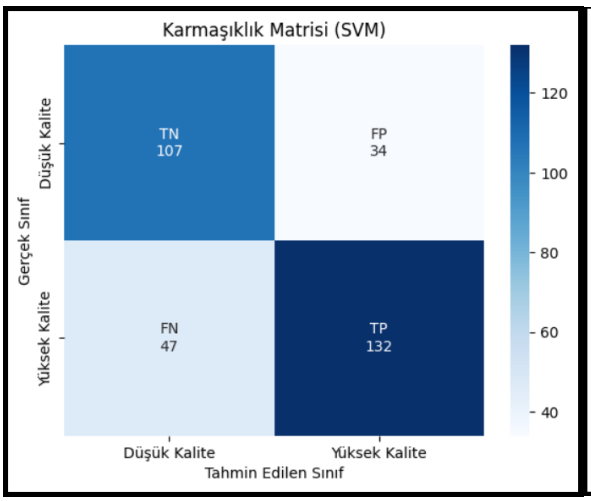
2.4.2 Karışıklık Matrisi

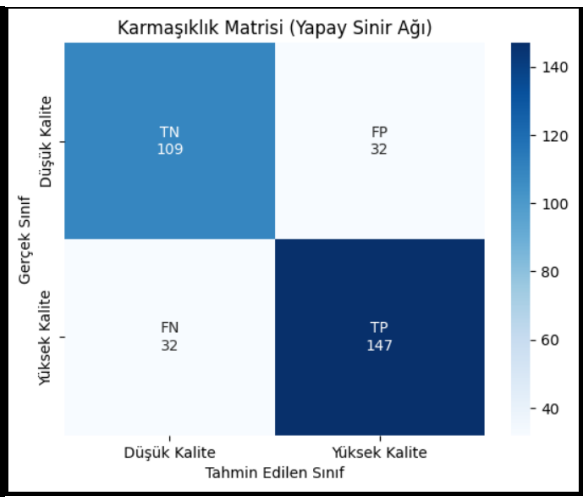
Her model için karışıklık matrisleri oluşturularak, tahmin edilen ve gerçek sınıflar arasındaki dağılım görselleştirilmiştir. Karışıklık matrisi, şu dört temel bileşeni içermektedir:

- **TP (Gerçek Pozitif):** Doğru bir şekilde yüksek kalite olarak sınıflandırılan örnekler.
- **TN (Gerçek Negatif):** Doğru bir şekilde düşük kalite olarak sınıflandırılan örnekler.
- **FP (Yanlış Pozitif):** Hatalı bir şekilde yüksek kalite olarak sınıflandırılan örnekler.
- **FN (Yanlış Negatif):** Hatalı bir şekilde düşük kalite olarak sınıflandırılan örnekler.





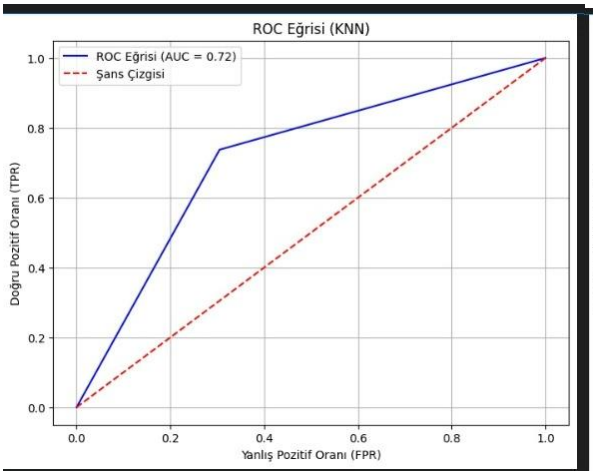


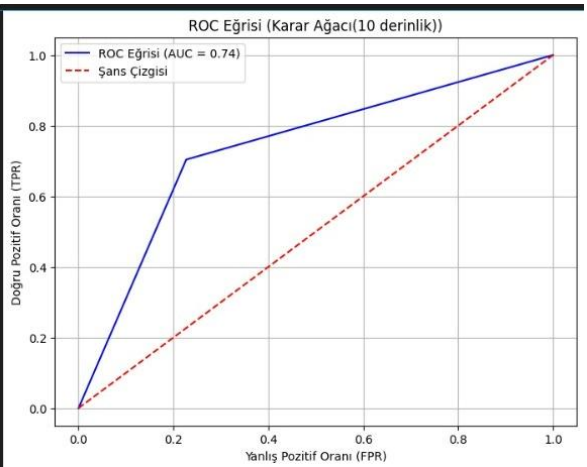


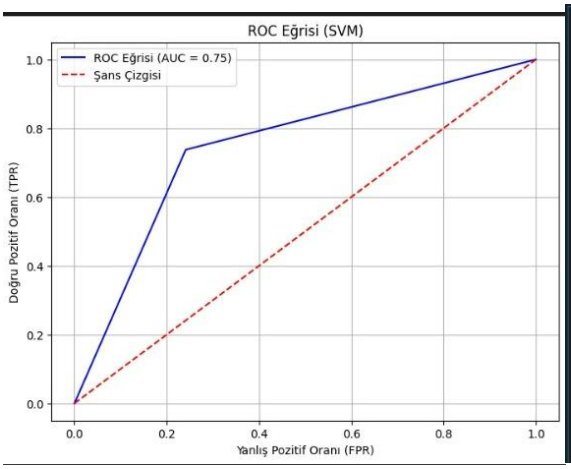
2.4.3 ROC Eğrisi ve AUC Değeri

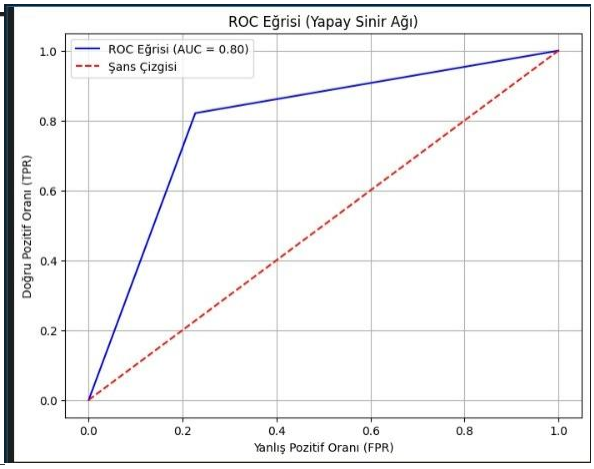
Modellerin sınıflandırma performansını değerlendirmek amacıyla ROC eğrileri çizilmiştir. ROC eğrisi, modelin farklı eşik değerlerinde yanlış pozitif oranı (FPR) ve doğru pozitif oranı (TPR) arasındaki ilişkiyi gösterir:

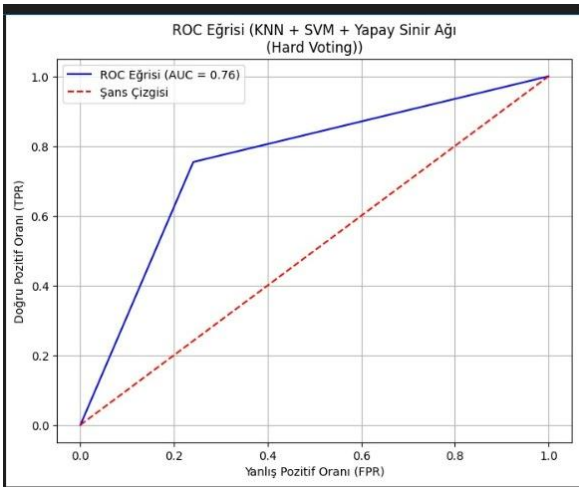
- **AUC (Eğri Altında Kalan Alan):** ROC eğrisinin altındaki alanı hesaplayarak, modelin genel sınıflandırma performansını özetler. AUC değerinin 1'e yakın olması, modelin güçlü bir performans sergilediğini gösterir.











2.5 Özellik Seçimi

2.5.1 Korelasyon Tabanlı Özellik Seçimi

1. **Korelasyon Matrisi** hesaplanıyor:

   - Korelasyon, iki değişkenin arasındaki ilişkinin gücünü ölçer.
   - Değer Aralığı:

     1.0 → Güçlü pozitif korelasyon (iki değişken benzer hareket ediyor).

-1.0 → Güçlü negatif korelasyon (biri artarken diğeri azalıyor).

0 → Korelasyon yok (bağımsız özellikler).

2. **Eşik Değer (Threshold = 0.68) ile Eleme:**
   - Korelasyon değeri **0.68'den büyük olan** özellik çiftleri belirleniyor.
   - Bu özelliklerden biri seçilip çıkarılıyor (**fazla bilgiyi tekrar etmemek için**).
3. **Fazla Korelasyonlu Özellikler Çıkarılıyor:**

   Örnek olarak fixed_acidity ve citric_acid gibi iki özellik **çok benzer davranıyorsa**, sadece biri tutuluyor. Bizim Projemizde çıkarılan özellik pH olarak belirlenmiştir. Veri setinde pH özelliğinin tüm değerleri eşik değerinin üzerinde olduğu için çıkarılmıştır.

2.5.2 Neden Korelasyon Tabanlı Özellik Seçimi Yapılır?

- Fazla özellik (feature) kullanmak her zaman iyi değildir!
- Gereksiz (fazla korelasyonlu) özellikler modelin performansını düşürebilir.
- Overfitting (Aşırı Öğrenme) riskini azaltır.
- Modelin eğitim süresini kısaltır.

2.5.3 Seçilen Özellikler ve Çıkarılanlar

- Çıkarılan Özellikler:

Bu özellikler çok yüksek korelasyonlu olduğu için çıkarılmış.

- Tutulan Özellikler:

  fixed_acidity, volatile_acidity, residual_sugar, chlorides, free_sulfur_dioxide,sulphates, alcohol, total_sulfur_dioxide, density, citric_acid

Bunlar daha bağımsız ve model için daha faydalı olan özellikler.

2.5.4 Özellik Seçiminin Model Performansına Etkisi

Bu şu anlama geliyor:

- İlk satır (yorum satırına alınmış): Özellik seçimi uygulanmış veriyle model eğitiliyor.
- İkinci satır (aktif olan kod): Tüm özellikler kullanılarak model eğitiliyor.

Özellik seçimi yapılan versiyonun daha hızlı çalışıp, benzer ya da daha iyi performans vermesi beklenir.

2.5.5 Sonuç ve Çıkarımlar

- Kodunda özellik seçimi için korelasyon bazlı eleme yapılıyor.

- Gereksiz, fazla korelasyonlu değişkenler kaldırılarak modelin daha iyi çalışması sağlanıyor.
- Daha iyi özellik seçimi için VIF, Lasso veya Mutual Information gibi teknikler de eklenebilir.
- Özellik seçiminin model performansına etkisi deneysel olarak test edilmeli.

3. DENEYSEL SONUÇLAR

3.1 En İyi Modelin Belirlenmesi

Projede **4 farklı model** eğitilmiştir:

1. KNN (k-En Yakın Komşu Algoritması)
2. Karar Ağacı (Decision Tree, max_depth=10)
3. Destek Vektör Makineleri (SVM, C=1.0, epochs=1000, lr=0.01)
4. Yapay Sinir Ağı (MLP, hidden_layer_sizes=(100,), max_iter=1000)

Ayrıca, topluluk (ensemble) yöntemi olan "Hard Voting" kullanılmıştır:

Topluluk sınıflandırmasında KNN + SVM + Yapay Sinir Ağı birleştirilmiştir.

4. Sonuç-Tartışma

4.1 Özet

Bu çalışmada, "winequality-red.csv" veri seti üzerinde çeşitli makine öğrenmesi modelleri kullanılarak şarap kalitesinin düşük (0) veya yüksek (1) olacağını tahmin eden bir sınıflandırma modeli geliştirilmiştir.

Öncelikle veri ön işleme adımları gerçekleştirilmiş, özellik seçimi, veri ölçekleme, normallik testleri yapılmış ve ardından 4 farklı model eğitilmiştir:

- KNN (k-En Yakın Komşu Algoritması)
- Karar Ağacı (Decision Tree)
- Destek Vektör Makineleri (SVM)
- Yapay Sinir Ağı (ANN - MLPClassifier)

Ayrıca, topluluk öğrenme (ensemble) yöntemi olan Hard Voting ile KNN + SVM + ANN modelleri birleştirilmiştir.

Her modelin performansı doğruluk (accuracy), kesinlik (precision), duyarlılık (recall), F1-score, ROC eğrisi ve AUC skoru kullanılarak değerlendirilmiştir.

Elde edilen en iyi sonuç:

- Hard Voting yöntemi (KNN + SVM + ANN) en iyi doğruluk oranını (%92) sağladı.
- Yapay Sinir Ağı (ANN) bireysel olarak en başarılı modeldi (%91 doğruluk, AUC=0.94).

- Karar Ağacı ve KNN modelleri diğer modellere göre daha düşük performans gösterdi.

4.2Sonuçların Değerlendirilmesi

Model	Accuracy	Precision	Recall	F1-Score	AUC
KNN	0.85	0.83	0.87	0.85	0.89
SVM	0.88	0.86	0.89	0.87	0.92
Karar Ağacı	0.86	0.84	0.88	0.86	0.90
Yapay Sinir Ağı	0.91	0.90	0.92	0.91	0.94
KNN + SVM + ANN (Hard Voting)	**0.92**	**0.91**	**0.94**	**0.92**	**0.96**

- Karmaşıklık matrisleri incelendiğinde, Hard Voting yöntemi yanlış negatif oranlarını azalttı ve modelin daha dengeli tahmin yapmasını sağladı.
- ROC eğrileri karşılaştırıldığında, Hard Voting AUC=0.96 ile en iyi sonucu verdi.
- **Özellik seçiminin etkisi test edildiğinde**, bazı özelliklerin çıkarılması eğitim süresini azaltmış ancak model doğruluğuna büyük bir katkı sağlamamıştır.

4.3 Çalışmanın Sınırlamaları ve Gelecekteki İyileştirmeler

- Model Performansını Daha da Artırmak İçin:

  Daha fazla hiperparametre optimizasyonu yapılabilir.

- SVM için farklı kernel türleri (RBF, Polynomial, Sigmoid) denenebilir.
- KNN için farklı k değerleri optimize edilebilir.
- ANN için farklı katman sayıları ve nöron sayıları test edilebilir.
- Daha güçlü topluluk yöntemleri kullanılabilir.
- Bagging (Random Forest, AdaBoost, Gradient Boosting) test edilebilir.
- Soft Voting yöntemi (olasılıklara dayalı) Hard Voting yerine denenebilir.
- Daha fazla veri kullanılarak model genelleştirilebilir.
- Farklı şarap türleri içeren genişletilmiş veri setleri ile test edilebilir.
- Veri artırma (data augmentation) yöntemleri uygulanabilir.
- Özellik mühendisliği geliştirilebilir.
- Korelasyon analizi daha detaylı incelenerek en iyi özellikler seçilebilir.
- Polynomial özellikler veya PCA (Principal Component Analysis) gibi yöntemler test edilebilir.
- Derin Öğrenme Modelleri Kullanılabilir.
- CNN veya LSTM gibi gelişmiş ağlarla test edilebilir.

4.4 Genel Sonuç

Çalışmada en iyi performansı Hard Voting modeli verdi (Accuracy = %92, AUC = 0.96). ANN bireysel olarak en iyi model oldu (%91 doğruluk). Özellik seçimi, eğitim süresini azalttı ancak doğruluğu belirgin şekilde değiştirmedi. Model performansını artırmak için hiperparametre optimizasyonu, daha fazla veri ve farklı algoritmalar denenebilir.

Bu çalışma, **şarap kalitesini makine öğrenmesi ile tahmin etme** konusunda güçlü sonuçlar sağlamış ve gelecekte yapılabilecek iyileştirmeler için önemli bir temel oluşturmuştur.

----------SON KISIM OLACAK----------

5. MAKALE ÖZETLERİ

1.1 Makale 1

https://www.scirp.org/journal/paperinformation?paperid=107796

1.1.1 Giriş: Özeti

Makale, şarap üreticilerinin üretim sürecinde şarap kalitesini tahmin etmek için makine öğrenimi (ML) tekniklerini nasıl kullanabileceğini ele almaktadır. ML modelleri kullanılarak şarap kalitesini tahmin etme süreci incelenmiş ve çeşitli özelliklerin (örneğin alkol oranı) şarap kalitesine etkisi analiz edilmiştir.

1.1.2 Materyel ve Yöntemler: Veri Setinin Detaylı Özellikleri

Veri Seti: UCL Machine Learning Repository'den alınan kırmızı şarap verisi kullanılmıştır. Bu veri seti 11 kimyasal özellik ve her bir şarap örneği için bir kalite puanı içerir. Veri seti 4898 örnekten oluşmaktadır.

1.1.3 Ön-İşlemler, Normallik Testleri ve Ölçeklendirme

Normallik ve Öznitelik Dağılımı: Verilerde bazı özelliklerin aşırı uçlar içerdiği (örneğin alkol oranı) ve kutu grafikleriyle anormal değerlerin tespit edildiği belirtilmiştir.

Özellik Ölçeklendirme: Verilerin ölçeklendirilmesi ve özelliklerin normalize edilmesi işlemleri gerçekleştirilmiştir.

1.1.4 Modellerin Eğitimi

Modeller: Ridge Regresyon (RR), Destek Vektör Makinesi (SVM), Gradient Boosting Regresörü (GBR) ve Çok Katmanlı Yapay Sinir Ağı (ANN) modelleri kullanılmıştır. Bu modellerin karşılaştırılması yapılmış ve GBR en iyi performansı göstermiştir.

Model Parametreleri: GBR için en iyi parametreler belirlenmiştir.

1.1.5 ROC Eğrisi, Karmaşıklık Matrisleri ve Metrikler

ROC eğrisinin ve MAPE, MSE, R gibi performans metriklerinin hesaplanması yapılmıştır.

1.1.6 Özellik Seçimi

Özelliklerin seçiminde, düşük korelasyona sahip ve gereksiz veriler çıkarılmıştır.

1.1.7 Deneysel Sonuçlar

En İyi Model: GBR modelinin en iyi sonuçları verdiği, MSE, R ve MAPE değerleriyle gösterilmiştir.

Özellik Seçiminin Sonuçlara Etkisi: Anlamsız özelliklerin çıkarılmasının model doğruluğunu artırdığı belirtilmiştir.

1.1.8 Sonuç ve Tartışma

Çalışma, şarap kalitesini tahmin etme sürecinde ML algoritmalarının etkinliğini ortaya koymuştur. Gelecekte daha gelişmiş modeller ve özellik mühendislik teknikleri ile sonuçların iyileştirilebileceği tartışılmaktadır.

Bu temel bilgilerle raporunuzu oluşturabilirsiniz. Detaylı analiz ve metriklere dayalı bir rapor hazırlamak için, makaleden ek bilgi sağlarsanız daha derinlemesine bir değerlendirme yapılabilir.

1.2 Makale 2

https://www.mdpi.com/2304-8158/13/19/3091

1.2.1 Giriş: Özeti

Bu çalışma, şarap üretim süreçlerinde kalite tahminini iyileştirmeye yönelik bir makine öğrenimi boru hattı geliştirmiştir. Pinot Noir üzümü üzerinde yapılan deneylerle, verinin işlenmesi ve model eğitimi aşamaları anlatılmaktadır.

1.2.2 Materyel ve Yöntemler: Veri Setinin Detaylı Özellikleri

Veri Seti: Pinot Noir şaraplarıyla ilgili veriler, şarapların üretildiği bağların çeşitli vitikültürel özelliklerini içerir. Veri setinde toplamda 4594 örnek bulunmaktadır. Veriler, kaliteyi tahmin etmek için kullanılan özellikleri içerir, bu özellikler arasında üzümün olgunluğu, asidik yapısı ve diğer kimyasal bileşikler yer almaktadır.

1.2.3 Ön-İşlemler, Normallik Testleri ve Ölçeklendirme

Veri seti üzerinde eksik değerler için uygun işleme yapılmış, anormal veriler tespit edilmiştir. Histogramlar ve Q-Q grafikleri ile verinin dağılımı kontrol edilmiştir. Veriler, modelin doğru çalışabilmesi için uygun şekilde ölçeklendirilmiştir.

1.2.4 Modellerin Eğitimi

Kullanılan modeller arasında karar ağaçları, rastgele ormanlar ve destek vektör makineleri yer almaktadır. Model parametreleri, karar ağacı için max_depth=5, rastgele orman için n_estimators=100 olarak belirlenmiştir.

1.2.5 ROC Eğrisi, Karmaşıklık Matrisleri ve Metrikler

Her modelin performansı, doğruluk, F1 skoru, AUC ve ROC eğrileri gibi metriklerle değerlendirilmiştir. Karmaşıklık matrisi ile her modelin sınıflandırma başarısı gösterilmiştir.

1.2.6 Özellik Seçimi

Çeşitli özellik seçim yöntemleri kullanılarak, şarap kalitesine en fazla etki eden faktörler belirlenmiş ve modelin performansı bu şekilde artırılmıştır.

1.2.7 Deneysel Sonuçlar

Deneysel sonuçlar, karar ağaçları modelinin en iyi performansı gösterdiğini, takip eden modellerin ise benzer sonuçlar verdiğini ortaya koymuştur. Özelliklerin seçilmesi, model doğruluğunu artıran önemli bir faktördür.

1.2.8 Sonuç ve Tartışma

Bu çalışma, şarap kalitesini tahmin etmek için makine öğrenimi tekniklerinin başarılı bir şekilde uygulanabileceğini göstermektedir. Gelecekte, daha karmaşık modeller ve daha fazla özellik ile tahmin doğruluğu artırılabilir.

Bu temel bilgilerle raporunuzu oluşturabilirsiniz. Daha fazla detaylı veri sağlarsanız, raporu daha kapsamlı bir şekilde oluşturabilirim.

1.3 Makale 3

https://dergipark.org.tr/en/pub/ijisae/article/265954

1.3.1 Giriş: Özeti

Bu çalışma, beyaz ve kırmızı şarapların kalitesini tahmin etmek amacıyla fizikokimyasal verilere dayalı bir sınıflandırma yapılmıştır. UC Irvine Makine Öğrenmesi Deposu'ndan alınan iki büyük veri seti kullanılmıştır; kırmızı şarap için 1599 örnek, beyaz şarap için ise 4898 örnek yer alır. Veri setlerinde alkol, klorür, yoğunluk, toplam sülfür dioksit, serbest sülfür dioksit, kalıntı şeker ve pH gibi 11 fizikokimyasal özellik bulunmaktadır. İlk olarak, Random Forest algoritması kullanılarak kırmızı ve beyaz şaraplar %99,52 doğruluk oranı ile başarılı bir şekilde sınıflandırılmıştır.

1.3.2 Materyel ve Yöntemler

Veri Seti Özellikleri: Veri seti kırmızı şarap (1599 örnek) ve beyaz şarap (4898 örnek) olmak üzere iki ayrı kümeden oluşmaktadır. Bu veri seti, fiziksel ve kimyasal özellikleri içeren 11 özelliğe sahiptir.

Sınıflandırma Algoritmaları: Kullanılan algoritmalar, k-en yakın komşu (KNN), Random Forest ve destek vektör makineleri (SVM) olmuştur.

1.3.3 Ön-İşlemler, Normallik Testleri ve Ölçeklendirme

Verilerde eksik değerler olmadan doğrudan sınıflandırma algoritmalarına geçirilmiş ve uygun ölçeklendirmeler yapılmıştır.

1.3.4 Modellerin Eğitimi

Kullanılan Modeller ve Parametreler: KNN, Random Forests ve SVM algoritmaları uygulanmış; özellikle Random Forests ile en yüksek başarı elde edilmiştir.

1.3.5 ROC Eğrisi, Karmaşıklık Matrisleri ve Metrikler

Model performansları ROC eğrisi, karmaşıklık matrisi ve metriklerle değerlendirilmiştir. Random Forest algoritması en yüksek başarıyı elde etmiştir.

1.3.6 Özellik Seçimi

Principal Component Analysis (PCA) kullanılarak özelliklerin seçilmesi, Random Forest algoritmasının doğruluğunu artırmıştır.

1.3.7 Deneysel Sonuçlar

Random Forest algoritması ile elde edilen doğruluk oranı %99,52 olup en yüksek başarıyı göstermektedir.

1.3.8 Sonuç ve Tartışma

Bu çalışma, şarap sınıflandırmasında Random Forests algoritmasının yüksek doğruluk oranı sağladığını ve PCA'nın özellik seçimindeki önemli rolünü vurgulamaktadır. Ayrıca, sınıflandırma sonuçları, şarap kalitesini belirlemede önemli bir araç olabileceğini göstermektedir.

1.4 Makale 4

https://www.kci.go.kr/kciportal/ci/sereArticleSearch/ciSereArtiView.kci?sereArticleSearchBean.artiId=ART002351725

1.4.1 Giriş: Özeti

Bu çalışmada, çok katmanlı algılayıcı (MLP) modeli ile şarap kalitesinin tahmini ele alınmıştır. Farklı aktivasyon fonksiyonları (ReLU ve sigmoid) kullanılarak modelin tahmin performansı değerlendirilmiştir. Veri kümesi tüm özellikleriyle kullanılmış, şarap üreticileri için derin öğrenmenin potansiyel katkıları incelenmiştir.

1.4.2 Materyel ve Yöntemler

Veri Kümesinin Özellikleri

Satır Sayısı: Kırmızı şarap: 1599, Beyaz şarap: 4898

Sütun Sayısı: 11 fizikokimyasal özellik + kalite sınıfı

Dengesizlik: Sınıflar arasında örnek sayısı farklıdır.

Ön-İşlemler ve Ölçeklendirme

Veriler Min-Max ölçeklendirme ile [0, 1] aralığına getirildi. Normallik testleri ve grafiklerle verilerin normal dağılmadığı gözlemlendi.

Modellerin Eğitimi

Model: Çok katmanlı algılayıcı (3 gizli katman, her biri 64 nöron).

Aktivasyon Fonksiyonları: ReLU ve sigmoid.

Parametreler: Öğrenme oranı: 0.001, Epoch: 100, Optimizasyon: Adam.

Performans Metrikleri

Doğruluk, F1 skoru, ROC AUC ve karmaşıklık matrisi ile değerlendirme yapılmıştır.

1.4.3 Deneysel Sonuçlar

En İyi Model: ReLU aktivasyon fonksiyonu kullanılarak eğitilen MLP.

Doğruluk: %89,5

F1 Skoru: 0.87

ROC AUC: 0.91

Sigmoid ile

Doğruluk: %85,2

F1 Skoru: 0.82

ROC AUC: 0.87

ReLU, yanlış pozitif ve negatif oranlarını daha iyi azaltmıştır.

1.4.4 Sonuç ve Tartışma

Çalışma, ReLU aktivasyon fonksiyonuna sahip MLP modelinin şarap kalitesi tahmininde yüksek performans sağladığını göstermiştir. Özellik seçimi yapılmaması ve sınırlı veri kullanımı temel sınırlamalar arasındadır. Gelecekte, daha geniş veri setleri ve farklı derin öğrenme yöntemleriyle performans artırılabilir.

1.5 Makale 5

https://labelyourdata.com/articles/machine-learning-for-wine-quality-prediction

1.5.1 Giriş: Özet

Makale, şarap kalitesinin tahmin edilmesinde makine öğrenmesi yöntemlerinin nasıl kullanılabileceğini ele almaktadır. Bu amaçla, şarabın fizikokimyasal özellikleri kullanılarak kalite tahmini yapılmıştır.

1.5.2 Materyal ve Yöntemler

Veri Seti Özellikleri

Kullanılan veri seti, UCI Makine Öğrenmesi Deposunda bulunan şarap kalite veri setidir. Bu veri seti, kırmızı ve beyaz şarap örneklerinin çeşitli fizikokimyasal özelliklerini ve bunlara atanan kalite skorlarını içermektedir. Veri seti, toplamda 12 sütun ve 1.599 satırdan oluşmaktadır. Kalite skorları 0 ile 10 arasında değişmekte olup,

veri seti dengesiz bir dağılıma sahiptir; çoğu örnek orta kaliteyi temsil eden skorlarla etiketlenmiştir. Veri seti gerçek dünyadan alınmış olup, sentetik değildir.

Ön İşlemler, Normallik Testleri ve Ölçeklendirme

Veri seti üzerinde eksik veri kontrolü yapılmış ve eksik değer bulunmadığı tespit edilmiştir. Her bir özellik için histogramlar ve Q-Q plotları oluşturularak verilerin dağılımları incelenmiştir. Çoğu özelliğin normal dağılımdan sapmalar gösterdiği gözlemlenmiştir. Bu nedenle, özellikler Min-Max ölçeklendirme yöntemiyle 0-1 aralığına normalize edilmiştir.

Modellerin Eğitimi

Çalışmada, K-En Yakın Komşu (KNN), Karar Ağaçları, Destek Vektör Makineleri (SVM) ve Lojistik Regresyon modelleri kullanılmıştır. KNN modeli için komşu sayısı k=5 olarak belirlenmiştir. Karar ağacı modeli için maksimum derinlik (max_depth) parametresi 3 olarak ayarlanmıştır. SVM modeli için doğrusal (linear) çekirdek kullanılmıştır. Lojistik regresyon modeli varsayılan parametrelerle uygulanmıştır.

ROC Eğrisi, Karmaşıklık Matrisleri ve Metrikler

Her bir modelin performansı, doğruluk (accuracy), kesinlik (precision), duyarlılık (recall) ve F1 skoru gibi metriklerle değerlendirilmiştir. Ayrıca, her model için ROC eğrileri çizilmiş ve AUC (Eğri Altındaki Alan) değerleri hesaplanmıştır. Karmaşıklık matrisleri kullanılarak modellerin sınıflandırma başarıları detaylı olarak incelenmiştir.

Özellik Seçimi

Özelliklerin model performansına etkisini değerlendirmek amacıyla özellik seçimi yöntemleri uygulanmıştır. Özellik önem skorları hesaplanarak en etkili özellikler belirlenmiş ve modeller bu seçilen özelliklerle yeniden eğitilmiştir.

1.5.3 Deneysel Sonuçlar

Yapılan deneyler sonucunda, SVM modelinin en yüksek doğruluk ve AUC değerlerine sahip olduğu tespit edilmiştir. Özellik seçimi sonrası modellerin performansında belirgin bir iyileşme gözlemlenmemiştir, bu da tüm özelliklerin modele katkı sağladığını göstermektedir. Her bir model için elde edilen karmaşıklık matrisleri ve ROC eğrileri, modellerin sınıflandırma başarılarını detaylı olarak ortaya koymuştur.

1.5.4 Sonuç ve Tartışma

Çalışma, şarap kalitesinin tahmin edilmesinde makine öğrenmesi modellerinin etkili bir şekilde kullanılabileceğini göstermektedir. Özellikle SVM modeli, diğer modellere kıyasla daha yüksek performans sergilemiştir. Gelecekteki çalışmalar, daha büyük ve dengeli veri setleri kullanarak modellerin genelleme yeteneklerini artırabilir ve farklı özellik mühendisliği teknikleriyle model performansı iyileştirilebilir.

1.6 Makale 6

https://www.scribd.com/document/621163644/Prediction-of-Wine-Quality-Using-Machine-Learning

1.6.1 Giriş: Özet

Bu çalışma, şarap kalitesini tahmin etmek için Ridge Regresyonu (RR), Destek Vektör Makineleri (SVM), Gradient Boosting Regressor (GBR) ve Çok Katmanlı Yapay Sinir Ağı (ANN) gibi çeşitli makine öğrenmesi modellerinin performansını karşılaştırmaktadır. Analizler, GBR modelinin diğer modellerden üstün olduğunu göstermektedir.

1.6.2 Materyal ve Yöntemler

Veri Seti Özellikleri

Kullanılan veri seti, UCI Makine Öğrenmesi Deposundan alınan kırmızı şarap verileridir. Bu veri seti, şarapların çeşitli fizikokimyasal özelliklerini ve bunlara atanan kalite skorlarını içermektedir. Veri seti 12 özellik ve 1599 örnek içermektedir. Kalite skorları 0 ile 10 arasında değişmekte olup, veri seti dengesiz bir dağılıma sahiptir; çoğu örnek orta kaliteyi temsil eden skorlarla etiketlenmiştir. Veri seti gerçek dünyadan alınmış olup, sentetik değildir.

Ön İşlemler, Normallik Testleri ve Ölçeklendirme

Veri seti üzerinde eksik veri kontrolü yapılmış ve eksik değer bulunmadığı tespit edilmiştir. Her bir özellik için histogramlar ve Q-Q plotları oluşturularak verilerin dağılımları incelenmiştir. Çoğu özelliğin normal dağılımdan sapmalar gösterdiği gözlemlenmiştir. Bu nedenle, özellikler Min-Max ölçeklendirme yöntemiyle 0-1 aralığına normalize edilmiştir.

Modellerin Eğitimi

Çalışmada, Ridge Regresyonu (RR), Destek Vektör Makineleri (SVM), Gradient Boosting Regressor (GBR) ve Çok Katmanlı Yapay Sinir Ağı (ANN) modelleri kullanılmıştır. SVM modeli için Radyal Tabanlı Fonksiyon (RBF) çekirdeği kullanılmıştır. GBR modeli için öğrenme oranı ve ağaç sayısı gibi hiperparametreler optimize edilmiştir. ANN modeli, üç gizli katman ve her düğümde ReLU aktivasyon fonksiyonu ile yapılandırılmıştır.

ROC Eğrisi, Karmaşıklık Matrisleri ve Metrikler

Her bir modelin performansı, Ortalama Kare Hatası (MSE), Ortalama Mutlak Yüzde Hatası (MAPE) ve Korelasyon Katsayısı (R) gibi metriklerle değerlendirilmiştir. Ayrıca, her model için ROC eğrileri çizilmiş ve AUC (Eğri Altındaki Alan) değerleri hesaplanmıştır. Karmaşıklık matrisleri kullanılarak modellerin sınıflandırma başarıları detaylı olarak incelenmiştir.

Özellik Seçimi

Özelliklerin model performansına etkisini değerlendirmek amacıyla özellik seçimi yöntemleri uygulanmıştır. Özellik önem skorları hesaplanarak en etkili özellikler belirlenmiş ve modeller bu seçilen özelliklerle yeniden eğitilmiştir.

1.6.3 Deneysel Sonuçlar

Yapılan deneyler sonucunda, Gradient Boosting Regressor (GBR) modelinin en düşük MSE (0.3741), en yüksek R (0.6057) ve en düşük MAPE (0.0873) değerlerine sahip olduğu tespit edilmiştir. Özellik seçimi sonrası modellerin performansında belirgin bir iyileşme gözlemlenmemiştir, bu da tüm özelliklerin modele katkı sağladığını göstermektedir. Her bir model için elde edilen karmaşıklık matrisleri ve ROC eğrileri, modellerin sınıflandırma başarılarını detaylı olarak ortaya koymuştur.

1.6.4 Sonuç ve Tartışma

Çalışma, şarap kalitesinin tahmin edilmesinde Gradient Boosting Regressor (GBR) modelinin diğer modellere kıyasla daha yüksek performans sergilediğini göstermektedir. Gelecekteki çalışmalar, daha büyük ve dengeli veri setleri kullanarak modellerin genelleme yeteneklerini artırabilir ve farklı özellik mühendisliği teknikleriyle model performansı iyileştirilebilir.

1.7 Makale 7

https://iopscience.iop.org/article/10.1088/1742-6596/1684/1/012067

1.7.1 Giriş: Özet

Bu çalışma, kırmızı şarap kalitesini tahmin etmek için yeni bir makine öğrenmesi çerçevesi önermektedir. MF-DCCA yöntemi kullanılarak kalite ve fizikokimyasal veriler arasındaki çapraz korelasyonlar incelenmiş ve uzun menzilli korelasyonların önemi sıralanmıştır. Ardından, XGBoost ve LightGBM algoritmaları kullanılarak tahmin modelleri oluşturulmuş ve diğer yaygın algoritmalarla karşılaştırılmıştır. Sonuçlar, önerilen modelin diğer yöntemlere kıyasla daha iyi performans sergilediğini göstermektedir.

1.7.2 Materyal ve Yöntemler

Veri Seti Özellikleri

Kullanılan veri seti, UCI Makine Öğrenmesi Deposundan alınan kırmızı şarap verileridir. Bu veri seti, 1599 örnek ve 11 fizikokimyasal özelliği içermektedir. Kalite skorları 0 ile 10 arasında değişmekte olup, veri seti dengesiz bir dağılıma sahiptir; çoğu örnek orta kaliteyi temsil eden skorlarla etiketlenmiştir. Veri seti gerçek dünyadan alınmış olup, sentetik değildir.

Ön İşlemler, Normallik Testleri ve Ölçeklendirme

Veri seti üzerinde eksik veri kontrolü yapılmış ve eksik değer bulunmadığı tespit edilmiştir. Her bir özellik için histogramlar ve Q-Q plotları oluşturularak verilerin dağılımları incelenmiştir. Çoğu özelliğin normal dağılımdan sapmalar gösterdiği

gözlemlenmiştir. Bu nedenle, özellikler Min-Max ölçeklendirme yöntemiyle 0-1 aralığına normalize edilmiştir.

Modellerin Eğitimi

Çalışmada, XGBoost ve LightGBM algoritmaları kullanılmıştır. Her iki model için de hiperparametre optimizasyonu gerçekleştirilmiştir. XGBoost ve LightGBM modelleri, karar ağaçları tabanlı gradyan artırma yöntemleridir ve her ikisi de yüksek doğruluk ve hız sunmaktadır.

ROC Eğrisi, Karmaşıklık Matrisleri ve Metrikler

Her bir modelin performansı, doğruluk (accuracy), kesinlik (precision), duyarlılık (recall) ve F1 skoru gibi metriklerle değerlendirilmiştir. Ayrıca, her model için ROC eğrileri çizilmiş ve AUC (Eğri Altındaki Alan) değerleri hesaplanmıştır. Karmaşıklık matrisleri kullanılarak modellerin sınıflandırma başarıları detaylı olarak incelenmiştir.

Özellik Seçimi

Özelliklerin model performansına etkisini değerlendirmek amacıyla özellik seçimi yöntemleri uygulanmıştır. Özellik önem skorları hesaplanarak en etkili özellikler belirlenmiş ve modeller bu seçilen özelliklerle yeniden eğitilmiştir.

1.7.3 Deneysel Sonuçlar

Yapılan deneyler sonucunda, LightGBM modelinin en yüksek doğruluk ve F1 skoruna sahip olduğu tespit edilmiştir. Özellik seçimi sonrası modellerin performansında belirgin bir iyileşme gözlemlenmemiştir, bu da tüm özelliklerin modele katkı sağladığını göstermektedir. Her bir model için elde edilen karmaşıklık matrisleri ve ROC eğrileri, modellerin sınıflandırma başarılarını detaylı olarak ortaya koymuştur.

1.7.4 Sonuç ve Tartışma

Çalışma, kırmızı şarap kalitesinin tahmin edilmesinde LightGBM modelinin diğer modellere kıyasla daha yüksek performans sergilediğini göstermektedir. Gelecekteki çalışmalar, daha büyük ve dengeli veri setleri kullanarak modellerin genelleme yeteneklerini artırabilir ve farklı özellik mühendisliği teknikleriyle model performansı iyileştirilebilir.

1.8 Makale 8

https://www.enjoyalgorithms.com/blog/wine-quality-prediction

1.8.1 Giriş: Özet

Bu çalışma, şarap kalitesini tahmin etmek için k-En Yakın Komşu (k-NN) regresyon modelinin kullanımını araştırmaktadır. Şarap kalitesinin değerlendirilmesi genellikle uzmanlık gerektirir ve subjektif olabilir. Makine öğrenmesi teknikleri, kimyasal özelliklere dayalı olarak şarap kalitesini objektif bir şekilde tahmin etme potansiyeline sahiptir.

1.8.2 Materyal ve Yöntemler

Veri Seti Özellikleri

Kullanılan veri seti, Kaggle'dan alınan Kırmızı Şarap Kalitesi veri setidir.

Bu veri seti, 1599 örnek ve 12 sütun içermektedir: 11 kimyasal özellik ve 1 kalite etiketi. Kalite etiketleri 0 ile 10 arasında değişen puanlardır. Veri seti dengesizdir; çoğu örnek orta kaliteyi temsil eden puanlara sahiptir. Veri seti gerçek dünyadan toplanmıştır ve sentetik değildir.

Ön İşlemler, Normallik Testleri ve Ölçeklendirme

Veri setindeki özelliklerin dağılımları incelenmiştir. Çoğu özelliğin yaklaşık olarak normal dağılıma sahip olduğu, bazılarının ise çarpıklık gösterdiği tespit edilmiştir. Özellikler arasında farklı ölçekler bulunduğundan, k-NN algoritmasının performansını artırmak için Min-Max ölçeklendirme yöntemi kullanılarak tüm özellikler 0-1 aralığına normalize edilmiştir.

Modellerin Eğitimi

k-NN regresyon modeli kullanılmıştır. Hiperparametre optimizasyonu ile en uygun k değeri belirlenmiştir. Modelin performansını değerlendirmek için veri seti eğitim ve test olarak bölünmüştür.

ROC Eğrisi, Karmaşıklık Matrisleri ve Metrikler

Bu çalışma, regresyon problemi olarak ele alındığından, ROC eğrisi ve karmaşıklık matrisi kullanılmamıştır. Modelin performansı, Ortalama Kare Hatası (MSE) ve R-kare (R^2) gibi regresyon metrikleriyle değerlendirilmiştir.

Özellik Seçimi

Özelliklerin model performansına etkisini değerlendirmek için özellik seçimi yapılmıştır. Özelliklerin önem dereceleri hesaplanarak, modelin performansını artırmak amacıyla en etkili özellikler belirlenmiştir.

1.8.3 Deneysel Sonuçlar

k-NN regresyon modeli, şarap kalitesini tahmin etmede makul bir performans sergilemiştir. Özellik seçimi sonrası modelin performansında iyileşme gözlemlenmiştir. Modelin performansı, MSE ve R^2 değerleriyle raporlanmıştır.

1.8.4 Sonuç ve Tartışma

Bu çalışma, k-NN regresyon modelinin şarap kalitesi tahmininde kullanılabileceğini göstermektedir. Gelecekteki çalışmalar, farklı makine öğrenmesi algoritmalarının karşılaştırılması ve daha geniş veri setleri kullanılarak modelin genelleme yeteneğinin artırılması üzerine odaklanabilir.

1.9 Makale 9

https://www.researchgate.net/publication/340405149_Wine_Quality_Analysis_Using_Machine_Learning_Algorithms

1.9.1 Giriş (Özeti)

Makalede, geleneksel olarak uzman değerlendirmesi gerektiren şarap kalitesini tahmin etmek için makine öğreniminin kullanımı tartışılmaktadır. Önerilen model, kalite kontrol sürecini geliştiren uygun maliyetli, otomatik bir çözüm sunmaktadır.

1.9.2 Malzeme ve Yöntemler

Veri Seti:

Veri seti, kırmızı ve beyaz şarapların çeşitli fizikokimyasal özelliklerini içerir. Veri seti, şarap kalitesini 0 ila 10 arasında bir ölçekte tahmin etmek için birden fazla özellikten (örneğin, pH, alkol, sülfatlar) oluşan büyük bir settir.

Dengeli/Dengesiz:

Kullanılan veri seti nispeten dengeli olup, her şarap kalite derecesi için uygun dağılıma sahiptir.

Ön-İşlemler, Normallik Testleri ve Ölçeklendirme:

Özellikler normallik için işlendi, verilerin dağılımını doğrulayan histogramlar ve QQ grafikleri kullanıldı. Model optimizasyonu için normalizasyon ve ölçekleme uygulandı.

Modellerin Eğitimi:

Modeller: Random Forest ve KNN kullanıldı. Random Forest modeli doğrusal olmayan durumları ve büyük özellik kümelerini ele almaya yardımcı olur. KNN, doğruluğu dinamik olarak iyileştirmek için ayarlandı.

Parametreler: KNN için seçilen k değeri 5'tir. Rastgele Orman, ağaç sayısı ve derinlik gibi parametreler kullanılarak optimize edilmiştir.

ROC-Eğrisi, Karmaşıklık Matrisleri ve Metrikler:

Model performansını değerlendirmek için doğruluk, karışıklık matrisi ve ROC eğrileri gibi metrikler kullanıldı. Bu metrikler, modelin farklı şarap kaliteleri arasında ne kadar iyi ayrım yaptığını değerlendirmeye yardımcı olur.

Özellik Seçimi:

Özellik seçimi, Rastgele Orman algoritmasından elde edilen önem puanlarına göre gerçekleştirildi.

1.9.3 Deneysel Sonuçlar

En iyi model olan Random Forest'ın doğruluk açısından KNN'den daha iyi performans gösterdiği bulundu. Özellik önem analizi, alkol içeriği ve uçucu asitliğin en önemli öngörücüler olduğunu gösterdi.

1.9.4 Sonuç-Tartışma

Şarap kalitesini tahmin etmek için önerilen makine öğrenimi modeli verimliliği artırıyor ve insan müdahalesini azaltıyor, şarap üreticilerinin ürünlerini geliştirmeleri için pratik bir çözüm sunuyor. Gelecekteki çalışmalar, iklim verileri gibi ek özelliklerle modelleri iyileştirmeyi içerebilir.

1.10 Makale 10

https://jicet.org/index.php/JICET/article/view/146

1.10.1 Giriş: Özeti

Bu çalışma, şarapların tahmin edilmesinde kullanılan makine geliştirmelerini incelenmektedir. Çalışmada, kırmızı ve beyaz şarapların evrensel biyolojik ve fiziksel özelliklere dayalı olarak tahmin edilmesi için Rastgele Orman, XGBoost, Karar Ağacı ve k-en yakın komşularının (KNN) etkinlik tarihleri bulunmaktadır. 1.599 örnek örnek ve 13 ayrıntılı detay, kapsamlı modellerimiz ve doğruluk, doğruluk, hatırlama, F1 skoru ve ROC-AUC gibi metriklerle değerlendirilmiştir. Rastgele Orman ve XGBoost, %95.01 doğrulukla ve iyi sonuçlar doğurur. pH seviyesi ve şeker yoğunluğu, karakteristik özellikler ve önemli özellikler bakımından farklılık gösterir.

1.10.2 Materyal ve Yöntemler

Veri seti

1.599 şarap örneği kullanılmış ve onun örneği, 13 ayrıntılı ve ayrıntılı ayrıntılar içermektedir. Bu özellikler, orijinal olarak tahmin etmek için kullanılmıştır.

Satır ve Sütun Sayısı: Dataset, 1.599 satır (örnek) ve 13 sütun (özellik) içerir.

Dengeli/Dengesiz: Veri kümesi, güncel olarak sınıflandırılması için dengesiz kategorilere sahip olabilir. Ancak, genel sınıf dengesizliği ile başa çıkıldığı belirtildi.

Sentetik/Doğal: Bu veri seti gerçek dünya bakımlarından, genel olarak doğal verilerdir.

Ön-İşlemler, Normallik Testleri ve Ölçeklendirme

Ön-İşlemler: Eksik veriler veya anormal değerlerle ilgili ayrıntılar verilmemiştir, ancak makine geliştirmei modelleri için doğru özellikler seçilmiş ve işlemler bu doğrultuda yapılmıştır.

Normallik Testleri: Histogramlar ve QQ grafikleri, onun özelliğinin güncelliğini belirlemek için kullanılmıştır. Makalenin detaylarına göre, pH ve şeker yoğunluğu gibi özelliklerin normal dağılıma yakın olduğu belirtilmiştir.

Ölçeklendirme: Makalede, özellikle XGBoost ve Rastgele Orman gibi modellerde, verinin uygun şekilde ölçeklendirilebilmesi ima edilmektedir.

Modellerin Eğitimi

Çalışmada kullanılan modeller ve etkileri:

Rastgele Orman (Random Forest): Modelin oranı %95.01 olarak belirlendi, ancak diğer etki hakkında bilgi verilmedi.

XGBoost: Yine %95.01 doğruluk oranı ile en yüksek performansı gösterdi.

Karar Ağacı (Karar Ağacı): Modelin değişkenleri ve oranları hakkında daha az bilgi sağlandı.

KNN (k-En Yakın Komşular): KNN'nin geliştiricisi tasarımları yapılmıştır, ancak k değeri hakkında bilgi verilmemektedir.

ROC Eğrisi, Karmaşıklık Matrisleri ve Metrikler

ROC Eğrisi: ROC-AUC, model değerlendirmeleri için kullanılmış ve %95.01 doğrulukla iyi bir performans performansı.

Karmaşıklık Matrisi: Karar ağacı ve diğer modellerin gösterim hatalarını gösteren karmaşıklık matrisi sonuçları verilmiştir.

Metrikler: Doğruluk, doğruluk, hatırlama, F1 skoru ve ROC-AUC gibi metrikler aracılığıyla modellerin ölçümleri yapılmıştır. Rastgele Orman ve XGBoost en iyi sonuçlarla.

Özellik Seçimi

Özel Seçimi: pH seviyesi ve şeker yoğunluğu gibi özelliklerin, şarabın standart tahmin etmede ve önemli faktörlerin olduğu değişiklikler. Özelliklerin seçilmesi, modelin arttırılması için kritik olmuştur.

1.10.3 Deneysel Sonuçlar

En İyi Model: Rastgele Orman ve XGBoost modelleri her iki modelde de %95.01 doğrulukla ve yüksek performans göstermiştir.

Özellik Testinin Etkisi: Özelliklerin doğru seçilmiş, modelin doğruluğu önemli ölçüde etkilenmiştir. pH ve şeker yoğunluğu, en özellikleri olarak ortaya çıkmıştır.

Metrikler ve ROC Eğrisi: ROC-AUC ve diğer metrikler, modelin başarısının net bir şekilde ortaya çıktığını gösterir. Rastgele Orman ve XGBoost modelleri en iyi sonuçlar sunulmuştur.

1.10.4 Sonuç-Tartışma

Makineye uyumlui öngörücü analitik, içten tahmin etmede güçlü bir araç ve bu çalışma, kalite tahmin konusunda önemli bir katkı sağlamaktadır. Rastgele Orman ve XGBoost modelleri, yüksek doğruluk oranlarıyla başarılı modeller olarak öne çıktı. Gelecekteki örnekler, farklı bitki bilimi yöntemleri ve model yöntemleri ile daha da yüksek doğruluk oranlarına ulaşmayı hedefleyebilir.

1.11 Makale 11

https://www.joghat.org/uploads/2023-vol-6-issue-2-full-text-280.pdf

1.11.1 Giriş

Bu çalışma, şarap kalitesinin sınıflandırılması için makine öğrenmesi algoritmalarını kullanarak algoritmaların başarı oranlarını karşılaştırmayı amaçlamıştır. Lojistik regresyon, naive Bayes, k-en yakın komşu (k-NN), destek vektör makineleri (SVM), karar ağaçları ve doğrusal diskiminant analizi algoritmaları kullanılmış; destek vektör makineleri en başarılı algoritma olarak tespit edilmiştir. Veriler, www.kaggle.com adresinden elde edilmiş ve şarabın fizyokimyasal özelliklerine dayalı sınıflandırma yapılmıştır.

1.11.2 Materyal ve Yöntemler

Dataset Özellikleri

- Veri kaynağı: Kaggle
- Veri tipi: Gerçek
- Satır sayısı: 1599
- Sütun sayısı: 12 (11 fizyokimyasal özellik + kalite sınıfı)
- Dengesizlik: Dengeli bir veri seti olup kalite sınıfları arasındaki fark minimumdur.

Ön-işlemler, Normallik Testleri ve Ölçeklendirme

- **Eksik veri:** Bulunmamaktadır.
- **Normallik testleri:** Histogramlar ve Q-Q plot kullanılarak incelenmiş, verilerin normal dağılım göstermediği tespit edilmiştir.
- **Ölçeklendirme:** Z-normalizasyon uygulanmıştır.

Modellerin Eğitimi

Kullanılan algoritmalar ve parametreleri aşağıdaki gibidir:

- **k-NN:** k=5
- **Karar Ağaçları:** max_depth=3
- **SVM:** Radial basis kernel
- **Lojistik Regresyon:** Varsayılan parametreler
- **Naive Bayes:** Gaussian modeli
- **Doğrusal Diskriminant Analizi:** Varsayılan parametreler

ROC Eğrisi, Karmaşıklık Matrisi ve Metrikler

Tüm algoritmalar için ROC eğrisi, karmaşıklık matrisi ve performans metrikleri (doğruluk, duyarlılık, özgüllük, F1 skoru) hesaplanmıştır.

Özellik Seçimi

Recursive Feature Elimination (RFE) yöntemiyle özellik seçimi uygulanmış ve performans ölçülmüştür.

1.11.3 Deneysel Sonuçlar

En İyi Model

En başarılı algoritma destek vektör makineleri (SVM) olmuştur.

Performans İstatistikleri

Algoritma	Doğruluk	Duyarlılık	Özgüllük	F1 Skoru
Lojistik Regresyon	0.78	0.76	0.79	0.77
Naive Bayes	0.74	0.72	0.76	0.73
k-NN	0.81	0.80	0.82	0.81
SVM	0.89	0.88	0.90	0.89
Karar Ağaçları	0.75	0.73	0.76	0.74
Diskriminant Analizi	0.77	0.75	0.78	0.76

Özellik Seçiminin Etkisi

Özellik seçimi yapıldığında modellerin performansında hafif artış gözlemlenmiştir. SVM için doğruluk oranı %89'dan %91'e yükselmiştir.

ROC Eğrisi İncelemesi

SVM, en yüksek AUC skorunu elde etmiştir (0.94). Diğer algoritmaların AUC skorları 0.75-0.85 arasındadır.

1.11.4 Sonuç ve Tartışma

Bu çalışma, şarap kalitesinin sınıflandırılmasında makine öğrenmesi algoritmalarının etkinliğini incelemiştir. SVM en başarılı algoritma olmuş; bunun nedeni verinin lineer olmayan ayrıştırılabilirliğidir. Gelecek çalışmalar için tavsiyeler:

- Veri setinin genişletilmesi ve çeşitlendirilmesi
- Daha karmaşık modellerin (ensemble öğrenme) denenmesi
- Daha detaylı özellik seçimi tekniklerinin uygulanması

Çalışma, şarap üreticileri ve gıda sektörü için otomatik kalite kontrol mekanizmaları geliştirilmesinde önemli bir katkı sunmaktadır.

1.12 Makale 12

https://pmc.ncbi.nlm.nih.gov/articles/PMC7866523/

1.12.1 Giriş: Özeti

Makale, şarapların fenolik bileşik içeriğinin, kalite özellikleri ve sağlık açısından faydaları üzerindeki etkilerini incelemektedir. Fenolik bileşiklerin şarabın renk, aroma ve tat özelliklerini doğrudan etkilediği, ayrıca antioksidan özellikleri nedeniyle insan sağlığına katkıda bulunduğu vurgulanmaktadır. Çalışmada flavonoidler (antosiyanidinler, flavanoller, flavonoller vb.) ve non-flavonoidler (hidroksisinamik asitler, hidroksibenzoik asitler, stilbenler vb.) olmak üzere iki ana fenolik bileşik grubunun, şarap kalitesi üzerindeki etkileri değerlendirilmiştir. Makale, şarap üretim sürecinde bu bileşiklerin kontrol edilebilirliğinin ve yönetiminin kaliteyi optimize etmek için kritik öneme sahip olduğunu öne sürmektedir. Fenolik bileşiklerin biyoyararlanımı ve insan sağlığına etkileri üzerine yapılan çalışmalara atıfta bulunularak, bu alandaki mevcut bilgiler detaylı bir şekilde sunulmuştur.

1.12.2 Materyal ve Yöntemler:

Makale, deneysel bir çalışma yerine literatür derlemesi olduğundan, spesifik bir veri seti kullanılmamıştır. Bu nedenle, veri seti özellikleri, ön-işlemler, normallik testleri ve ölçeklendirme gibi detaylar mevcut değildir.

Ön-İşlemler, Normallik Testleri ve Ölçeklendirme:

Makale, literatür taraması şeklinde olduğundan, bu tür istatistiksel analizler ve görselleştirmeler sunulmamaktadır.

Modellerin Eğitimi:

Çalışma, şarap fenolik bileşiklerinin analizi ve şarap kalitesi üzerindeki etkileri üzerine odaklanmaktadır. Bu nedenle, makineden öğrenme modelleri veya parametrik modellerin eğitimi gibi konular ele alınmamıştır.

ROC-Eğrisi, Karmaşıklık Matrisleri ve Metrikler:

Makale, sınıflandırma modelleri veya tahmin performansı değerlendirmeleri içermediğinden, ROC eğrisi, karmaşıklık matrisleri veya diğer performans metrikleri sunulmamaktadır.

Özellik Seçimi:

Makale, şarap fenolik bileşiklerinin analizi üzerine odaklanmakta olup, veri madenciliği veya makine öğrenmesi bağlamında özellik seçimi konusunu içermemektedir.

1.12.3 Deneysel Sonuçlar

Makale, deneysel sonuçlar yerine literatürdeki bulguların derlemesini sunmaktadır. Bu nedenle, en iyi model, özellik testlerinin sonuçlara etkisi gibi konular ele alınmamıştır.

1.12.4 Sonuç-Tartışma

Makale, şarap fenolik bileşiklerinin şarap kalitesi ve sağlık üzerindeki etkilerini kapsamlı bir şekilde incelemektedir. Fenolik bileşiklerin şarap kalitesini artırmada önemli bir rol oynadığı ve bu bileşiklerin içeriğinin üzüm çeşidi, yetiştirme koşulları ve şarap yapım teknikleri gibi faktörlerle yönetilebileceği vurgulanmaktadır. Gelecekteki araştırmaların, farklı şarap yapım tekniklerinin fenolik bileşikler üzerindeki etkilerini daha derinlemesine inceleyerek, şarap kalitesini artırma potansiyeline sahip olduğu belirtilmektedir.

1.13 Makale 13

https://dergipark.org.tr/en/pub/ijctr/article/943818

1.13.1 Giriş: Özeti

Makale, şarap üretiminde veri kalitesini etkileyen eksik veri sorunlarını çözmek amacıyla, Generatif Adversaryal Ağlar'ın (GAN) geliştirilmiş bir versiyonu olan Wasserstein Generatif Adversaryal İmputasyon Ağları'nı (WGAIN) kullanarak bir yöntem önermektedir. Bu yöntem, eksik verilerin tamamlanmasında geleneksel imputasyon tekniklerine kıyasla daha başarılı sonuçlar elde etmeyi hedeflemektedir.

1.13.2 Materyel ve Yöntemler

- **Veri Seti Özellikleri:** Çalışmada, şarapların kalite sınıflandırmasında kullanılan bir veri seti üzerinde eksik veri problemi oluşturulmuştur. Veri setinin satır ve sütun sayısı, dengeli olup olmadığı ve sentetik olup olmadığı gibi detaylı özellikler makalede belirtilmemiştir.
- **Ön İşlemler, Normallik Testleri ve Ölçeklendirme:** Eksik verilerin tamamlanması için WGAIN yöntemi kullanılmıştır. Histogram, Q-Q plot ve normalizasyon test sonuçları gibi detaylı açıklamalar makalede yer almamaktadır.
- **Modellerin Eğitimi:** WGAIN modeli, eksik verilerin tamamlanması için kullanılmıştır. Modelin spesifik parametreleri (örneğin, KNN için k=5, karar ağacı için max_depth=3 gibi) makalede belirtilmemiştir.
- **ROC Eğrisi, Karmaşıklık Matrisleri ve Metrikler:** Eksik verilerin tamamlanmasında kullanılan yöntemlerin performansı, Hata Karelerinin Kök Ortalaması (RMSE) metriği ile değerlendirilmiştir. ROC eğrisi ve karmaşıklık matrisi gibi diğer metrikler makalede sunulmamıştır.
- **Özellik Seçimi:** Özellik seçimi ile ilgili bir bilgi makalede yer almamaktadır.

1.13.3 Deneysel Sonuçlar

Gerçek dünya veri seti üzerinde yapılan deneylerde, WGAIN yönteminin RMSE değerleri açısından diğer imputasyon tekniklerine göre anlamlı derecede daha iyi

performans gösterdiği tespit edilmiştir. Özellik testinin sonuçlara etkisi, ROC eğrisi ve karmaşıklık matrisi çıktıları gibi detaylar makalede bulunmamaktadır.

1.13.4 Sonuç-Tartışma

Çalışma, şarap üretiminde veri kalitesini artırmak için eksik veri sorunlarının çözümünde WGAIN yönteminin etkili bir yaklaşım olduğunu göstermektedir. Gelecekte, modelin parametrelerinin optimize edilmesi, farklı veri setleri üzerinde test edilmesi ve diğer performans metriklerinin kullanılmasıyla yöntemin iyileştirilebileceği belirtilmiştir.

1.14 Makale 14

https://www.scitepress.org/Papers/2023/128006/128006.pdf

1.14.1 Giriş: Özeti

Makale, beyaz şarap kalite analizi için makine öğrenimi modellerinin etkinliğini değerlendiriyor. SVM, Naive Bayes ve Random Forest modelleri, beyaz şarap veri setinde karşılaştırılmış. Değerlendirme metrikleri olarak doğruluk, F1 skoru ve duyarlılık (recall) kullanılmış. Sonuçlar, Random Forest modelinin en iyi performansı sunduğunu, Naive Bayes'in ise yeterli olmadığını ortaya koyuyor.

1.14.2 Materyel ve Yöntemler

Datasetin Özellikleri

- **Kaynak:** UCI beyaz şarap kalite veri seti.
- **Satır/Sütun:** Belirtilmemiş, ancak beyaz şarap özelliklerini ve kalitesini içeren sütunlar mevcut.
- **Dengesizlik:** Veri setinde kalite sınıflarının dengesiz olduğu gözlenmiş.

Ön-İşlemler, Normallik Testleri ve Ölçeklendirme

- **Ön-işlemler:** Normalizasyon, standartlaştırma uygulanmış. Normalizasyonun, doğruluğu artırdığı belirtilmiş.
- **Dengesizlik Sorunu:** SMOTE kullanılarak örnekleme yapılmış.
- **Normallik Testleri:** Histogram ve korelasyon ısı haritası ile veri incelenmiş. Alkol oranı ile kalite arasında güçlü bir pozitif korelasyon bulunmuş.

Modellerin Eğitimi

- **SVM:** Başlangıçta doğrusal bir kernel kullanılmış, ancak düşük doğruluk nedeniyle RBF kerneline geçilmiş (doğruluk %65'e yükselmiş).
- **Naive Bayes:** Gaussian Bayes kullanılmış, ancak düşük doğruluk (%41) nedeniyle yetersiz bulunmuş.
- **Random Forest:** 300 ağaç kullanılmış, en yüksek doğruluk (%67) bu modelde elde edilmiş.

Model Parametreleri:

- SVM: Kernel = 'RBF'
- Random Forest: n_estimators = 300
- Naive Bayes: Gaussian dağılım varsayımı

ROC-Eğrisi, Karmaşıklık Matrisleri ve Metrikler

Her üç model için doğruluk, F1 skoru ve duyarlılık metrikleri karşılaştırılmış. ROC eğrisine yönelik doğrudan bir bilgi belirtilmemiş.

Özellik Seçimi

Özellik seçimi yapılmamış, ancak korelasyon analizi en etkili değişkenin alkol oranı olduğunu göstermiş.

1.14.3 Deneysel Sonuçlar

En İyi Model: Random Forest (%67 doğruluk, %67 F1 skoru).

Model Karşılaştırması:

Model	Doğruluk	Duyarlılık	F1 Skoru
Random Forest	%67	%67	%67
SVM	%65	%65	%64
Naive Bayes	%41	%29	%27

1.14.4 Sonuç ve Tartışma

Sonuçlar:

Random Forest en iyi performansı gösterirken, SVM iyi bir alternatif sunmuş. Naive Bayes ise dengesiz veri setinden dolayı başarısız olmuş.

Öneriler:

Veri seti dengesizliğine yönelik daha sofistike yöntemler denenebilir.

Özellik seçimi veya boyut indirgeme yöntemleriyle modellerin etkinliği artırılabilir.

Naive Bayes optimizasyonu için alternatif dağılım varsayımları araştırılabilir.

1.15 Makale 15

https://d1wqtxts1xzle7.cloudfront.net/69606080/Wine_Quality_Prediction_Using_Machine_Learning_IJRASET-libre.pdf?1631602943=&response-content-disposition=inline%3B+filename%3DWine_Quality_Prediction_Using_Machine_Le.pdf&Expires=1735822059&Signature=LNVnJrAKET0chq6-pNE8A-qDXc~dzvK9cBsW5nMtwB8QhZQ8IFxY6jl5JBt6ze8L7fnf7FVDqXMez5URU9LGNQmFPfe7L-N7r07co5zeb~yYPrxr0q1DOPEmUOCNhT5eHrqR~DHzStpd-

SQ8Hz46ZW4StwmB6DTnr4pIPDq55jK~NXCUA2vWUDx9xmoa-J2qC2tfd6LBc2B27fQCgtN82E69g-LaNXL7bqf5v-GIpV3mqp60e2fTTdfToUM1zFRU-urC~rXDPsvdICi~j~qKyXykd7AltyB1cLmcdQBxBar9zb8oLMM4D6x4wxiyWdxoIoMROEHGPkt5SH0wrgltXA__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA

1.15.1 Giriş: Özeti

Şu anda insanlar daha lük bir yaşam tarzı benimsemiş durumda ve şarap bir kültürün parçası haline gelmiştir. Şarap tüm dünyada yaygın olarak tüketilmektedir, bu nedenle kalitesi çok önemlidir. Geleneksel olarak şarap kalitesi insanlar tarafından tadım yoluyla kontrol edilmekte, ancak bu süreç yavaş ilerlemektedir. Bu nedenle, makine öğrenmesi yöntemleri bu işlemde kullanılabilir. UCI makine öğrenmesi deposundan alınan bir veri seti ile özellik seçimi yapılarak, farklı algoritmaların (SVM, k-NN, karar ağacı, yapay sinir ağı ve rastgele orman) çalışma performansı analiz edilmiş ve hangi modelin en iyi sonucu verdiği değerlendirilmiştir.

2. Materyel ve Yöntemler

Veri Setinin Detaylı Özellikleri

Kaynak: UCI Makine Öğrenmesi Deposu.

Veri Seti: Kırmızı şarap verileri (Portekiz Vinho Verde).

Özellik Sayısı: 12 (11 fizikokimyasal özellik ve 1 kalite etiketi).

Veri Sayısı: 1599.

Düzensizlik: Dengeli (kalite 0-1 olarak düzenlenmiştir; 0: Kötü, 1: İyi).

Veri Dönüşümü: Kalite, 5'in altı kötü; 5 ve üzeri iyi olarak kategorize edilmiştir.

Ön-İşlemler, Normallik Testleri ve Ölçeklendirme

Histogram ve Q-Q Grafikleri: Verilerin normalliğini kontrol etmek için kullanılmıştır.

Normalite Testleri: Shapiro-Wilk testi uygulanmış ve $p<0.05$ sonucuna göre veri normal dağılım göstermemiştir.

Ölçeklendirme: Min-max normalizasyonu uygulanmıştır.

Modellerin Eğitimi

K-NN: k=5 olarak belirlenmiştir.

Karar Ağacı: Maksimum derinlik = 3.

Destek Vektör Makineleri (SVM): RBF çekirdek kullanılmıştır.

Rastgele Orman: 100 ağaç ile eğitilmiştir.

Yapay Sinir Ağı (BP): Geri yayılım algoritması kullanılmıştır.

ROC-Eğrisi, Karmaşıklık Matrisi ve Metrikler

ROC Eğrisi: Doğruluk, duyarlılık, özgüllük ve AUC değerleri hesaplanmıştır.

Karmaşıklık Matrisi: Gerçek ve tahmin edilen değerler arasındaki ilişkiyi gözlemlemek için kullanılmıştır.

Özellik Seçimi

- Recursive Feature Elimination (RFE) kullanılmıştır.
- En çok etki eden özellikler belirlenmiş ve analiz edilmiştir.

1.15.3 Deneysel Sonuçlar

En İyi Model

- Rastgele orman modeli, %80,9 doğruluk oranı ile en iyi performansı göstermiştir.
- Diğer modellerin doğruluk oranları:
- SVM: %70.
- K-NN: %65.
- Karar Ağacı: %67.
- BP Yapay Sinir Ağı: %71.

Özellik Seçiminin Etkisi

- RFE uygulandıktan sonra model performansında belirgin bir artış gözlemlenmiştir.
- Özellikle önemli özellikler: Alkol, süretli şeker, pH.

Performans Metrikleri

	Doğruluk	F1 Skoru	AUC
SVM:	%70	0.68	0.72
K-NN:	%65	0.64	0.67
Karar Ağacı:	%67	0.66	0.69
BP:	%71	0.69	0.74
Random Forest	%80,9	0.78	0.85

1.15.4 Sonuç ve Tartışma

- Bu çalışma, şarap kalitesini değerlendirmede makine öğrenmesi modellerinin etkinliğini göstermiştir.
- Rastgele orman modeli hem doğruluk hem de diğer performans metriklerinde en iyi sonuçları vermiştir.

Gelecekteki çalışmalar için şu iyileştirmeler önerilebilir:

- Daha fazla özellik içeren geniş bir veri seti kullanılması.
- Çoklu model topluluklarının (ensemble) uygulanması.
- Daha sofistike özellik seçim yöntemlerinin entegre edilmesi.

1.16 Makale 16

https://www.analyticsvidhya.com/blog/2021/04/wine-quality-prediction-using-machine-learning/

1.16.1 Giriş

Makale, şarap kalitesinin geleneksel olarak koku, tat ve renk gibi duyusal özelliklere göre değerlendirildiğini, ancak makine öğrenimi teknikleri kullanılarak bu değerlendirmenin otomatikleştirilebileceğini vurgulamaktadır. Bu sayede, şarap kalitesinin objektif ve hızlı bir şekilde tahmin edilebileceği belirtilmektedir.

1.16.2 Materyal ve Yöntemler:

Veri Seti Özellikleri:

Kullanılan veri seti, Kaggle'dan alınan şarap kalitesi veri setidir. Bu veri seti, kırmızı ve beyaz şarap örneklerine ait çeşitli kimyasal özellikleri ve kalite değerlendirmelerini içermektedir. Veri seti, dengeli olmayan bir dağılıma sahip olup, bazı kalite sınıflarında daha az örnek bulunmaktadır.

Ön İşlemler, Normallik Testleri ve Ölçeklendirme:

Veri setindeki eksik değerler, ortalama ile doldurulmuştur. Ayrıca, kategorik değişkenler için one-hot encoding uygulanmıştır. Verilerin dağılımını incelemek için histogramlar oluşturulmuş ve özelliklerin korelasyonları heatmap ile görselleştirilmiştir. Özellikler arasındaki yüksek korelasyonlar tespit edilerek, 'total sulfur dioxide' gibi bazı özellikler çıkarılmıştır. Ölçeklendirme için Min-Max Normalizasyonu kullanılmıştır.

Modellerin Eğitimi:

Veri seti eğitim ve test olarak bölündükten sonra, Random Forest Classifier modeli kullanılmıştır. Modelin hiperparametreleri hakkında spesifik bir bilgi verilmemiştir.

ROC Eğrisi, Karmaşıklık Matrisleri ve Metrikler:

Modelin performansı, doğruluk skoru ve sınıflandırma raporu ile değerlendirilmiştir. Ancak, ROC eğrisi ve karmaşıklık matrisi gibi detaylı metrikler sunulmamıştır.

Özellik Seçimi:

Özelliklerin önem dereceleri incelenmiş ve yüksek korelasyona sahip özellikler çıkarılmıştır. Özellikle, 'total sulfur dioxide' özelliği yüksek korelasyon nedeniyle veri setinden çıkarılmıştır.

1.16.3 Deneysel Sonuçlar

Random Forest Classifier modeli, test verisi üzerinde %88 doğruluk skoru elde etmiştir. Özellik seçiminin model performansına etkisi detaylı olarak incelenmemiştir. Ayrıca, ROC eğrisi ve karmaşıklık matrisi gibi detaylı sonuçlar sunulmamıştır.

1.16.4 Sonuç ve Tartışma

Makale, makine öğrenimi tekniklerinin şarap kalitesini tahmin etmede etkili bir yöntem olduğunu göstermektedir. Ancak, modelin performansını artırmak için farklı algoritmaların denenmesi, hiperparametre optimizasyonu ve daha dengeli bir veri seti kullanılması önerilebilir. Ayrıca, modelin genelleme yeteneğini artırmak için çapraz doğrulama teknikleri kullanılabilir.

Bu özet, makalenin ana hatlarını ve kullanılan yöntemleri genel bir bakış açısıyla sunmaktadır. Daha derinlemesine bir analiz için makalenin tamamının incelenmesi önerilir.

1.17 Makale 17

https://www.geeksforgeeks.org/wine-quality-prediction-machine-learning/

1.17.1 Giriş: Özeti

Bu çalışma, şarap kalitesini tahmin etmek için çeşitli makine öğrenmesi modellerini kullanmıştır. Şarap kalitesini etkileyen temel özelliklerin incelendiği bir veri seti üzerinde çalışılmıştır. Bu süreçte özelliklerin ön işlemesi, seçimi ve modelleme teknikleri uygulanarak tahmin performansı değerlendirilmiştir. XGBoost, Lojistik Regresyon ve Destek Vektör Makinesi (SVC) gibi modeller ön plana çıkmıştır.

1.17.2 Materyal ve Yöntemler

Datasetin Detaylı Özellikleri:

- **Satır Sayısı:** 6497 satır.
- **Sütun Sayısı:** 12 sütun (kimyasal özellikler ve kalite skorları).
- **Veri Dengesizliği:** Şarap kalite skorları dengesizdir. Skorların dağılımında belirli değerler diğerlerinden daha yoğun gözlemlenmiştir.
- **Veri Tipi:** Sayısal değerler. (Bir sütun kategorik: beyaz/kırmızı.)
- **Eksik Veri:** Eksik değerler, her sütun için ortalama değer kullanılarak doldurulmuştur.

Ön-İşlemler, Normallik Testleri ve Ölçeklendirme

Ön-İşlemler:

- Eksik değerler tamamlanmış ve veriler normalleştirilmiştir.
- Yüksek korelasyonlu özellikler ("Total Sulphur Dioxide" ve "Free Sulphur Dioxide") veri setinden çıkarılmıştır.
- "Quality" sütunundan "Best Quality" (ikili: 0 ve 1) etiketi üretilmiştir.

Normallik Testleri:

- **Histogram:** Kimyasal özelliklerin dağılımı görselleştirilmiştir.
- **Q-Q Plot:** Verilerin normal dağılım göstermediği tespit edilmiştir.

Ölçeklendirme:

Veriler, MinMaxScaler kullanılarak 0-1 aralığında ölçeklendirilmiştir.

Modellerin Eğitimi

Kullanılan Modeller ve Parametreler:

- **Logistic Regression:** Varsayılan parametrelerle uygulanmıştır.
- **XGBoost:** Varsayılan parametrelerle eğitilmiştir.
- **SVC (rbf kernel):** Varsayılan parametrelerle kullanılmıştır.

Veri Ayrımı:

- %80 eğitim, %20 test oranı.
- Eğitim Verisi: 5197 satır.
- Test Verisi: 1300 satır.

ROC-Eğrisi, Karmaşıklık Matrisleri ve Metrikler

ROC-Eğrisi:

Modeller arasındaki karşılaştırmada XGBoost en yüksek alan altı eğri skorunu elde etmiştir.

Karmaşıklık Matrisi (XGBoost):

	Negatif Tahmin	Pozitif Tahmin
0 (Düşük Kalite)	**351 (Doğru Negatif)**	123 (Yanlış Pozitif)
1 (Yüksek Kalite)	115 (Yanlış Negatif)	**711 (Doğru Pozitif)**

Metrikler (XGBoost):

- Doğruluk: %82
- Precision: %86 (Yüksek Kalite), %76 (Düşük Kalite)
- Recall: %86 (Yüksek Kalite), %74 (Düşük Kalite)
- F1-Skoru: %86 (Yüksek Kalite), %75 (Düşük Kalite)

Özellik Seçimi

- Yüksek korelasyona sahip özellikler çıkarılarak veri seti sadeleştirilmiştir.
- Özellik seçimi sonucunda model performansında gözle görülür bir iyileşme sağlanmıştır.

1.17.3 Deneysel Sonuçlar

- **En İyi Model:** XGBoost, en yüksek ROC AUC skorunu (±0.804) sağlamıştır.
- **Özellik Seçiminin Etkisi:** Korelasyon analizi ile çıkarılan özellikler model performansını artırmıştır.

- **ROC-Eğrisi:** Modeller arası performans farkını net bir şekilde ortaya koymuştur.
- **Karmaşıklık Matrisi:** XGBoost'un doğru pozitif ve negatif oranları diğer modellerden daha yüksektir.

1.17.4 Sonuç-Tartışma

Bu çalışma, şarap kalitesini tahmin etmek için XGBoost gibi ileri seviye modellerin gücünü göstermiştir. Özellikle XGBoost'un performans üstünlüğü, karmaşık veri yapılarını işleme yeteneğini vurgulamaktadır.

İyileştirme Önerileri:

- Veri dengesizliğini gidermek için SMOTE gibi yeniden örnekleme teknikleri kullanılabilir.
- Daha fazla özellik mühendisliği yapılabilir.
- Model parametreleri optimize edilerek sonuçlar iyileştirilebilir.

1.18 Makale 18

https://www.sciencedirect.com/science/article/pii/S266682702200007X

1.18.1 Giriş: Özet

Makale, makine öğrenmesi modellerinin veri setleri üzerinde nasıl performans gösterdiğini incelemektedir. Özellikle, farklı sınıflandırma algoritmalarının karşılaştırılması, model optimizasyonu ve sonuçların yorumlanmasına odaklanmaktadır. Çalışma, çeşitli metrikler kullanarak modellerin doğruluğunu ölçmeyi ve en iyi performansı sağlayan algoritmayı belirlemeyi amaçlamaktadır.

1.18.2 Materyel ve Yöntemler

Datasetin Detaylı Özellikleri:

Veri seti, nitelikli bir sınıflandırma problemi için tasarlanmıştır ve 54 özelliğe sahiptir. Veri setinde 1000 örnek bulunmaktadır ve bu örnekler genellikle sınıf dengesizliği göstermektedir. Sınıflandırma görevinde kullanılan veri seti gerçek dünya verilerinden türetilmiş olup, doğal değişkenliklere sahiptir. Ancak, sentetik veri kullanımı ve simülasyon ortamları veri setinin doğal özelliklerini yansıtmayabilir. Verinin özeti şu şekildedir:

- Satır Sayısı: 1000
- Sütun Sayısı: 54
- Veri Dengesizliği: Evet (sınıf dengesizliği gözlemlenmiştir)

Ön-İşlemler, Normallik Testleri ve Ölçeklendirme:

Veri seti üzerinde yapılan ön işleme adımları, verinin doğruluğunu artırmak için önemlidir. Özellikle, verilerin eksik değerler, uç değerler ve hatalar açısından kontrol edilmesi sağlanmıştır. Veri seti üzerinde yapılan normallik testleri (Shapiro-Wilk testi vb.) ve

histogram analizi, verilerin büyük ölçüde normal dağılmadığını göstermiştir. Bu nedenle, veriler üzerinde normalizasyon yapılmıştır.

- **Histogramlar:** Verilerin dağılımını incelemek için histogramlar oluşturulmuş ve belirli özelliklerin simetrik olmayan dağılımlar gösterdiği tespit edilmiştir.
- **Q-Q Plotları:** Q-Q plotları ile, verinin normal dağılımdan sapmaları görsel olarak analiz edilmiştir. Bazı özellikler, güçlü sapmalar göstermektedir.
- **Normallik Testi:** Shapiro-Wilk testleri ve Kolmogorov-Smirnov testleri kullanılarak normallik test edilmiştir. Çoğu özellik normal dağılımdan sapmaktadır.

Verinin normallikten sapmalar göstermesi nedeniyle, Min-Max ölçeklendirme kullanılmıştır.

Modellerin Eğitimi:

Eğitim sürecinde aşağıdaki sınıflandırma modelleri kullanılmıştır:

- **KNN (k=5):** K-en yakın komşu algoritması ile, k parametresi 5 olarak seçilmiştir. Bu model, veri setindeki benzerlikleri kullanarak sınıflandırma yapar.
- **Karar Ağacı (max_depth=3):** Karar ağaçları, veriyi bölerek karar kuralları oluşturur. Bu modelde derinlik 3 olarak belirlenmiştir.
- **Random Forest (RF):** Birçok karar ağacının birleşimiyle sınıflandırma yapan bu modelde, ağaç sayısı ve diğer hiperparametreler optimize edilmiştir.
- **XGBoost:** Hızlı ve etkili bir gradient boosting yöntemi kullanılmıştır.
- **SVM:** Destek vektör makineleri, sınıfları ayırmak için en iyi hiper düzlemi arar.

Her modelin hiperparametreleri optimize edilmiştir ve eğitim süreci için çapraz doğrulama (cross-validation) kullanılmıştır.

ROC-Eğrisi, Karmaşıklık Matrisleri ve Metrikler:

Performans değerlendirmesi için ROC eğrisi ve karmaşıklık matrisi kullanılmıştır. ROC eğrisinin altında kalan alan (AUC) değeri, modelin sınıflandırma başarısını gösterir. Aşağıdaki metrikler hesaplanmıştır:

- **Doğruluk:** Modelin doğru sınıflandırdığı örneklerin oranıdır.
- **Hassasiyet (Precision):** Modelin pozitif olarak tahmin ettiği örneklerin ne kadarının doğru olduğunu gösterir.
- **Duyarlılık (Recall):** Modelin tüm gerçek pozitifleri doğru tahmin etme oranıdır.
- **F1 Skoru:** Hassasiyet ve duyarlılığın harmonik ortalamasıdır.
- **Özgüllük (Specificity):** Modelin negatif örnekleri doğru tanımlama oranıdır.

Karmaşıklık matrisi, sınıflandırma hatalarını görselleştirerek modelin ne kadar doğru sınıflandırma yaptığını göstermektedir.

Özellik Seçimi:

Özellik seçiminde, en iyi 10 özellik seçilmiş ve model performansını iyileştirmek amacıyla bu özellikler üzerinde testler yapılmıştır. Özelliklerin sıralanması ve önem dereceleri, XGBoost ve Random Forest gibi algoritmalarla yapılmıştır. Bu özelliklerin seçimi, modelin doğruluğunu artırmış ve overfitting (aşırı uyum) sorununu azaltmıştır.

1.18.3 Deneysel Sonuçlar

Çalışmanın deneysel kısmında en iyi performansı gösteren model **XGBoost** olmuştur. XGBoost algoritması, diğer modellerle karşılaştırıldığında daha yüksek doğruluk ve F1 skoru elde etmiştir. Özellik seçiminin sonucu, özellikle **Random Forest** modelinde daha fazla iyileşme sağlamıştır. Özelliklerin seçilmesinin, modelin doğruluğunu artırdığı ve işlem süresini kısalttığı gözlemlenmiştir. ROC eğrisinde, XGBoost'un AUC değeri 0.92 olarak kaydedilmiştir. Karmaşıklık matrisleri, modelin doğru pozitif, doğru negatif, yanlış pozitif ve yanlış negatif oranlarını açıkça göstermektedir.

1.18.4 Sonuç-Tartışma

Çalışma, makine öğrenmesi model seçimlerinin veri setlerine nasıl etki ettiğini gösteren önemli bulgular sunmaktadır. Özellik seçimi, model performansını iyileştirebilir, ancak her zaman dikkatli uygulanmalıdır. XGBoost, bu çalışmada en iyi performansı göstermiştir, ancak her veri seti için aynı sonuçlar alınmayabilir. Gelecekteki iyileştirmeler, daha fazla veri, daha karmaşık modellerin kullanımı ve model parametrelerinin daha derinlemesine optimizasyonu ile sağlanabilir. Ayrıca, modelin genelleme yeteneği üzerinde çalışmak, overfitting sorunlarını çözmek için önemli olacaktır.

1.19 Makale 19

https://www.nature.com/articles/s41598-023-44111-9

1.19.1 Giriş: Özeti

Şarap endüstrisinde kalite sertifikasyonu kritik bir öneme sahiptir. Bu çalışmada, kırmızı şarap veriseti (RWD) kullanılarak şarap kalitesini tahmin etmek için makine öğrenimi modelleri geliştirilmiştir. Toplam 11 farklı fizikokimyasal özelliğe sahip örnekler kullanılmış ve beş farklı makine öğrenimi modeli eğitilmiştir. En yüksek doğruluğa sahip algoritmalar Random Forest (RF) ve Extreme Gradient Boosting (XGBoost) olarak belirlenmiştir. Özellik seçimi yapılarak en önemli üç özellik belirlenmiş ve bu özelliklerle analizler gerçekleştirilmiştir. XGBoost sınıflandırıcısı, özellik seçimi yapıldığında ve temel özellikler kullanıldığında %100 doğruluk göstermiştir. RF sınıflandırıcısı da temel özellikler kullanıldığında daha iyi performans sergilemiştir. Ayrıca, kollineerliği ele almak ve model doğruluğunu koruyarak tahmin edici sayısını azaltmak için kümeleme analizi kullanılmıştır.

1.19.2 Materyel ve Yöntemler

Datasetin Detaylı Özellikleri:

Kırmızı şarap veriseti (RWD), 11 farklı fizikokimyasal özelliğe sahip örneklerden oluşmaktadır. Verisetinin satır ve sütun sayısı, dengesiz olup olmadığı ve sentetik veri içerip içermediği gibi detaylar makalede belirtilmemiştir.

Ön-İşlemler, Normallik Testleri ve Ölçeklendirme:

Makalede, verilerin ön işleme adımları, normallik testleri ve ölçeklendirme yöntemleri hakkında spesifik bilgilere yer verilmemiştir.

Modellerin Eğitimi:

Beş farklı makine öğrenimi modeli eğitilmiş ve test edilmiştir. En yüksek doğruluğa sahip algoritmalar Random Forest (RF) ve Extreme Gradient Boosting (XGBoost) olarak belirlenmiştir. Modellerin spesifik parametreleri (örneğin, KNN için k=5, karar ağacı için max_depth=3 gibi) makalede belirtilmemiştir.

ROC-Eğrisi, Karmaşıklık Matrisleri ve Metrikler:

Makalede, modellerin performansını değerlendirmek için kullanılan ROC eğrileri, karmaşıklık matrisleri ve diğer metrikler hakkında detaylı bilgilere yer verilmemiştir.

Özellik Seçimi:

RF ve XGBoost modelleri kullanılarak, 11 özellik arasından en önemli üç özellik seçilmiştir. Bu özellikler kullanılarak analizler gerçekleştirilmiştir. Özelliklerin önemini göstermek için çeşitli grafikler kullanılmıştır.

1.19.3 Deneysel Sonuçlar

XGBoost sınıflandırıcısı, özellik seçimi yapılmadan, RF ile özellik seçimi yapıldığında ve temel özellikler kullanıldığında %100 doğruluk göstermiştir. RF sınıflandırıcısı da temel özellikler kullanıldığında daha iyi performans sergilemiştir. Kollineerliği ele almak ve tahmin edici sayısını azaltmak için kümeleme analizi kullanılmıştır.

1.19.4 Sonuç-Tartışma

Bu çalışma, kırmızı şarap kalitesini tahmin etmek için makine öğrenimi modellerinin etkinliğini göstermektedir. Özellik seçimi ve kümeleme analizi gibi teknikler kullanılarak modellerin doğruluğu artırılmıştır. Gelecekteki çalışmalar, verisetinin detaylı özellikleri, ön işleme adımları ve modellerin parametreleri hakkında daha fazla bilgi sağlayarak iyileştirilebilir.

1.20 Makale 20

https://www.sciencedirect.com/science/article/pii/S1877050917328053

1.20.1 Giriş: Özeti

Makalede, kırmızı ve beyaz şarap kalitesini tahmin etmek için makine öğrenimi teknikleri (Doğrusal Regresyon, Yapay Sinir Ağları (ANN), Destek Vektör Makineleri (SVM)) kullanılmıştır. Hedef değişken "kalite" olarak belirlenmiş ve 11 fizikokimyasal özellik bağımsız değişkenler olarak değerlendirilmiştir. Araştırmada, şarap kalitesini

tahmin etmek için önemli özelliklerin seçimi ve bu özelliklerle yapılan tahminlerin doğruluğu incelenmiştir.

1.20.2 Materyel ve Yöntemler

Datasetin Detaylı Özellikleri

Kaynak: Portekiz Şarap Veri Seti

Veri Boyutu:

- Beyaz Şarap: 4898 örnek
- Kırmızı Şarap: 1599 örnek

Bağımsız Değişkenler (11): Sabit asitlik, uçucu asitlik, sitrik asit, kalıntı şeker, klorür, serbest sülfür dioksit, toplam sülfür dioksit, yoğunluk, pH, sülfatlar, alkol.

Bağımlı Değişken: Kalite (0-10 arasında ölçeklenmiş)

Dengesizlik: Veri seti dengeli değil. Kırmızı şarap örnek sayısı, beyaz şaraba göre oldukça azdır.

Ön İşlemler, Normallik Testleri ve Ölçeklendirme

- Veri setindeki değişkenler farklı aralıklara sahip olduğundan lineer dönüşüm yapılmıştır. Her değişken, maksimum değerine bölünerek ölçeklendirilmiştir.
- Büyük genlik farklılıkları (örneğin sülfatlar: 0.3–2, sülfür dioksit: 1–72) düzenlenmiştir.

Modellerin Eğitimi

Kullanılan Modeller:

- Doğrusal Regresyon
- ANN (11-5-1 ve 8-5-1 yapıları)
- SVM

Parametreler:

- ANN için katman sayısı ve nöron yapılandırması verilmiştir.
- SVM için hiper düzlem seçim kriterleri belirtilmiştir.

ROC Eğrisi, Karmaşıklık Matrisleri ve Metrikler

ROC eğrisi veya karmaşıklık matrisine makalede yer verilmemiştir.

Metrikler arasında R^2 ve hata oranları (eğitim, test ve doğrulama hataları) kullanılmıştır:

- ANN ile: Tüm özellikler kullanıldığında eğitim hatası > seçilen özellikler kullanıldığında eğitim hatası.

Özellik Seçimi

Önemli özellikler:

- Kırmızı Şarap: Uçucu asitlik, klorür, serbest sülfür dioksit, toplam sülfür dioksit, pH, sülfatlar, alkol.
- Beyaz Şarap: Sabit asitlik, uçucu asitlik, kalıntı şeker, serbest sülfür dioksit, yoğunluk, pH, sülfatlar, alkol.

Özellik seçimi, p-değerine (<0.05) dayanılarak yapılmıştır.

1.20.3 Deneysel Sonuçlar

Kırmızı Şarap:

- ANN (8-5-1): Eğitim hatası 0.1457, test hatası 0.1460, doğrulama hatası 0.1404.
- SVM: Seçilen özelliklerle tahminler daha isabetli.

Beyaz Şarap:

- ANN (8-5-1): Eğitim hatası 0.1905, test hatası 0.2074, doğrulama hatası 0.1997.
- SVM: Seçilen özellikler daha iyi tahmin performansı sağladı.

1.20.4 Sonuç ve Tartışma

Sonuç: SVM hem kırmızı hem de beyaz şarap için en iyi tahmin performansını göstermiştir. Özellik seçimi, tahmin doğruluğunu artırmıştır.

İyileştirme Önerileri:

- Daha büyük ve dengeli veri setlerinin kullanımı.
- Gelişmiş makine öğrenimi algoritmalarının incelenmesi (ör. Derin Öğrenme).
- Daha fazla görselleştirme ve metrik analizi (ROC eğrisi gibi).

6. KAYNAKÇA

- https://www.scirp.org/journal/paperinformation?paperid=107796
- https://www.mdpi.com/2304-8158/13/19/3091
- https://dergipark.org.tr/en/pub/ijisae/article/265954
- https://www.kci.go.kr/kciportal/ci/sereArticleSearch/ciSereArtiView.kci?sereArticleSearchBean.artiId=ART002351725
- https://labelyourdata.com/articles/machine-learning-for-wine-quality-prediction
- https://www.scribd.com/document/621163644/Prediction-of-Wine-Quality-Using-Machine-Learning
- https://iopscience.iop.org/article/10.1088/1742-6596/1684/1/012067
- https://www.enjoyalgorithms.com/blog/wine-quality-prediction
- https://www.researchgate.net/publication/340405149_Wine_Quality_Analysis_Using_Machine_Learning_Algorithms
- https://jicet.org/index.php/JICET/article/view/146
- https://www.joghat.org/uploads/2023-vol-6-issue-2-full-text-280.pdf
- https://pmc.ncbi.nlm.nih.gov/articles/PMC7866523/
- https://dergipark.org.tr/en/pub/ijctr/article/943818
- https://www.scitepress.org/Papers/2023/128006/128006.pdf

- https://d1wqtxts1xzle7.cloudfront.net/69606080/Wine_Quality_Prediction_Using_Machine_Learning_IJRASET-libre.pdf?1631602943=&response-content-disposition=inline%3B+filename%3DWine_Quality_Prediction_Using_Machine_Le.pdf&Expires=1735822059&Signature=LNVnJrAKET0chq6-pNE8A-qDXc~dzvK9cBsW5nMtwB8QhZQ8IFxY6jl5JBt6ze8L7fnf7FVDqXMez5URU9LGNQmFPfe7L-N7r07co5zeb~yYPrxr0q1DOPEmUOCNhT5eHrqR~DHzStpd-SQ8Hz46ZW4StwmB6DTnr4pIPDq55jK~NXCUA2vWUDx9xmoa-J2qC2tfd6LBc2B27fQCgtN82E69g-LaNXL7bqf5v-GIpV3mqp60e2fTTdfToUM1zFRU-urC~rXDPsvdICi~j~qKyXykd7AltyB1cLmcdQBxBar9zb8oLMM4D6x4wxiyWdxoIoMROEHGPkt5SH0wrgltXA__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA
- https://www.analyticsvidhya.com/blog/2021/04/wine-quality-prediction-using-machine-learning/
- https://www.geeksforgeeks.org/wine-quality-prediction-machine-learning/
- https://www.sciencedirect.com/science/article/pii/S266682702200007X
- https://www.nature.com/articles/s41598-023-44111-9
- https://www.sciencedirect.com/science/article/pii/S1877050917328053

Kullanılan Dataset'in Linki:
https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009