BLM5110 Makine Öğrenmesi Projesi Raporu
Alican Tunç
235B7055

Alican Tunç
Veri Bilimi ve Büyük Veri Yüksek Lisans Programı
Yıldız Teknik Üniversitesi
Istanbul, TÜRKİYE:
can.tunc1@std.yildiz.edu.tr

***Abstract*—Bu çalışmada, şarap kalitesini tahmin etmek için farklı makine öğrenimi yöntemlerinin performansları karşılaştırılmıştır. Kullanılan veri kümesi, kimyasal özellikleri ve 0 ile 10 arasında değişen kalite skorlarını içeren, kırmızı ve beyaz şaraplardan oluşmaktadır. Sınıflandırma modelleri olarak Destek Vektör Makineleri (DVM), Rastgele Orman (RF), Lojistik Regresyon (LR), Yapay Sinir Ağları (YSA) ve En Yakın Komşu (KNN) algoritmaları seçilmiştir. DVM'de farklı çekirdek fonksiyonları (lineer, polinomsal ve RBF), YSA'da ise farklı katman yapılarına sahip ağlar değerlendirilmiştir. Modeller, doğruluk, precision, recall ve F1-score metrikleri kullanılarak karşılaştırılmıştır. Elde edilen sonuçlar, kırmızı ve beyaz şarapların ayrı ayrı ele alınmasının model performansını artırabileceğini ve algoritma seçiminde veri özelliklerinin kritik bir rol oynadığını göstermektedir.**

Keywords—Yapay sinir ağları, destek vektör makineleri, sınıflandırma, doğruluk analizi,

I. GIRIŞ

Son yıllarda makine öğrenimi, çeşitli endüstrilerde karmaşık problemlerin çözümünde yaygın olarak kullanılmaktadır. Gıda ve içecek sektöründe de bu teknolojilerin uygulama alanı genişlemiş ve özellikle şarap kalitesinin tahmin edilmesi gibi spesifik alanlarda önemli başarılar elde edilmiştir. Şarap kalitesini etkileyen faktörler arasında kimyasal bileşikler ve üretim süreçleri gibi çok sayıda değişken bulunmaktadır. Bu faktörlerin değerlendirilmesi ve sınıflandırılması, uzmanlar tarafından manuel olarak yapılabildiği gibi, veri odaklı algoritmalarla da otomatik hale getirebilmektedir.

Bu çalışmada, kırmızı ve beyaz şarapların kimyasal özelliklerini ve kalite skorlarını içeren bir veri seti üzerinde farklı makine öğrenimi yöntemlerinin performansları karşılaştırılmıştır. Şarapların sınıflandırılması için kullanılan yöntemler arasında Destek Vektör Makineleri (DVM), Rastgele Orman (RF), Lojistik Regresyon (LR), Yapay Sinir Ağları (YSA) ve En Yakın Komşu (KNN) algoritmaları yer almaktadır. Literatürde, bu algoritmaların şarap kalitesinin tahmin edilmesinde yaygın olarak kullanıldığı ve her birinin veri özelliklerine bağlı olarak farklı başarı oranları gösterdiği belirtilmiştir.

Bu bağlamda, çalışmanın amacı, veri setinin yapısını ve algoritmaların sınıflandırma performanslarını ayrıntılı bir şekilde inceleyerek, şarap kalitesini tahmin etmek için en uygun yaklaşımı belirlemektir. Çalışmada, kırmızı ve beyaz şarapların ayrı olarak değerlendirilmesinin model performansı üzerindeki etkileri de ele alınmıştır. Model performansları doğruluk, precision, recall ve F1-score gibi metriklerle ölçülmüş ve sonuçlar kapsamlı bir şekilde analiz edilmiştir.

II. Veri kümesi ve Önceki çalışmalar

A. Veri Seti Bilgilendirmesi

Bu veri seti, Portekiz'e özgü kırmızı ve beyaz "Vinho Verde" şaraplarına ait fizikokimyasal özellikler ve duyusal kalite puanlarını içermektedir. Sınıflandırma ve regresyon görevleri için kullanılabilir, ancak sınıflar arasında dengesizlik bulunmaktadır.

İçerik:

- ***Girdi Değişkenleri:*** *Şarabın fizikokimyasal özellikleri (örneğin, asitlik, şeker, yoğunluk).*
- ***Çıktı Değişkeni:*** *Duyusal verilerden elde edilen 0-10 arası kalite skoru.*

Kullanım Önerileri:

- *Kalite sınıflandırması için keyfi bir eşik belirlenebilir (örneğin, skor ≥ 7 olanlar "iyi" kabul edilebilir).*
- *Veriyi işlemeye daha uygun olması için normalizasyon veya standardizasyon uygulanabilir*
- *Karar ağacı algoritmaları için hiper parametre optimizasyonu yapılabilir ve AUC ile ROC eğrileri kullanılarak değerlendirme yapılabilir.*

Teşekkür:
Bu veri seti, UCI Makine Öğrenimi Deposunda halka açıktır ve kullanıldığında uygun şekilde atıfta bulunulmalıdır.

Yayın:
Cortez, P., Cerdeira, A., Almeida, F., Matos, T., Reis, J., Decision Support Systems, 47(4):547-553, 2009.

Features Table

Feature No.	Feature Name	Description
1	Fixed Acidity	Acidity level that remains after fermentation
2	Volatile Acidity	Acetic acid amount affecting taste
3	Citric Acid	Enhances flavor, adds freshness
4	Residual Sugar	Sugar left after fermentation
5	Chlorides	Salt content
6	Free Sulfur Dioxide	SO_2 not bound and acts as an antimicrobial
7	Total Sulfur Dioxide	Total amount of SO_2 (bound + free)
8	Density	Wine density, related to alcohol and sugar content
9	pH	Acidity level
10	Sulphates	Contributes to microbial stability
11	Alcohol	Alcohol percentage
12	Quality	Sensory quality score (0 to 10)

Şekil 1: Veri kümesine ait özellikler

Yukarıda Şekil 1'de ilgili veri setine ait özellikler bulunmaktadır.

Table 1: The physicochemical data (input variables) and its corresponding statistics

Features	Red wine			White wine		
	Min	Max	Mean	Min	Max	Mean
Fixed acidity *g(tartaric acid)/dm*	4.6	15.9	8.32	3.8	14.2	6.855
Volatile acidity *g(acetic acid)/dm*	0.12	1.58	0.528	0.08	1.1	0.278
Citric acid *g/dm*	0.0	1.0	0.271	0.0	1.66	0.334
Residual sugar *g/dm*	0.9	15.5	2.539	0.6	65.8	6.391
Chlorides *g(sodium chloride)/dm*	0.12	0.611	0.087	0.009	0.346	0.046
Free sulfur dioxide *mg/dm*	1.0	72.0	15.875	2.0	289.0	35.308
Total sulfur dioxide *mg/dm*	6.0	289.0	46.478	9.0	440.0	138.361
Density *g/dm*	0.99	1.00369	0.997	0.98711	1.03898	0.994
pH	2.74	4.01	3.311	2.72	3.82	3.188
Sulphates *g(potassium sulphate)/dm*	0.33	2.0	0.658	0.22	1.08	0.489
Alcohol *%vol*	8.4	14.9	10.423	3.0	14.2	10.514

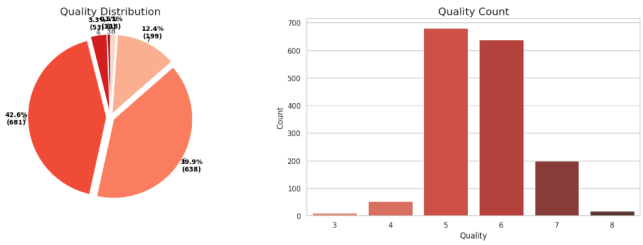
Şekil 2: Özelliklerin dağılımına ait tablo

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.4	0.700	0.00	1.9	0.076	11.0	34.0	0.99780	3.51	0.56	9.4	5
1	7.8	0.880	0.00	2.6	0.098	25.0	67.0	0.99680	3.20	0.68	9.8	5
2	7.8	0.760	0.04	2.3	0.092	15.0	54.0	0.99700	3.26	0.65	9.8	5
3	11.2	0.280	0.56	1.9	0.075	17.0	60.0	0.99800	3.16	0.58	9.8	6
4	7.4	0.700	0.00	1.9	0.076	11.0	34.0	0.99780	3.51	0.56	9.4	5
...	...	...	...	...	...	...	...	...	...	...	...	...
1594	6.2	0.600	0.08	2.0	0.090	32.0	44.0	0.99490	3.45	0.58	10.5	5
1595	5.9	0.550	0.10	2.2	0.062	39.0	51.0	0.99512	3.52	0.76	11.2	6
1596	6.3	0.510	0.13	2.3	0.076	29.0	40.0	0.99574	3.42	0.75	11.0	6
1597	5.9	0.645	0.12	2.0	0.075	32.0	44.0	0.99547	3.57	0.71	10.2	5
1598	6.0	0.310	0.47	3.6	0.067	18.0	42.0	0.99549	3.39	0.66	11.0	6

1599 rows × 12 columns

Şekil 3: Kırmızı şarap datası örneği

Sonrasında Şekil 2 ve 3 de ilgili datamızın hangi değerler arasında dağıldığı ve min-max değerleri göre bilmekteyiz. Aşağıdaki Şekil 4 ve 5'te ise bu özelliklerin nasıl dağıldığına bakıp datayı beraber yorumlayacağız.



Şekil 4: Kalite özelliğinin dağılımı

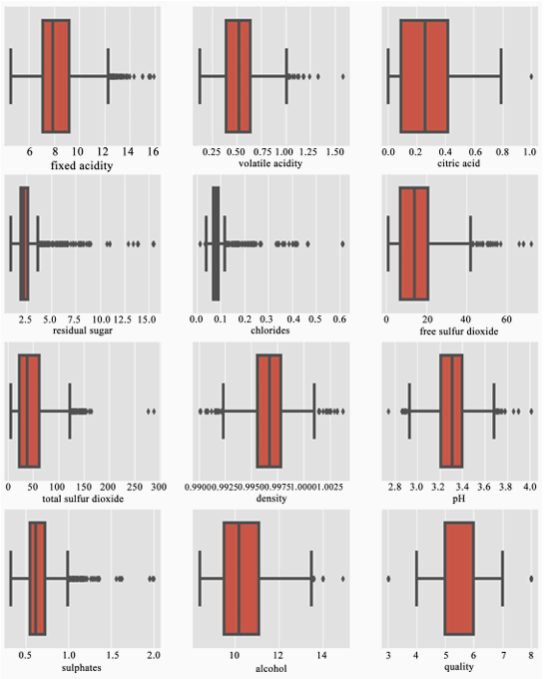
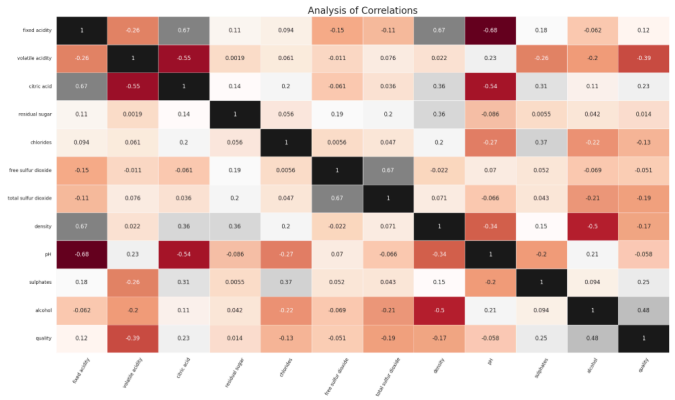


Figure 1. Box plot of the variables of the redwine data.

Şekil 5: Özelliklerin box-plot ile dağılımı



Şekil 6: Dataya ait kolerasyon matrisi

B. *Veri Setine yapılan müdahaleler ve önceki çalışmalar*

İki veya daha fazla değişken, birinin değeri arttığında veya azaldığında diğerinin de aynı ya da ters yönde değişmesi durumunda ilişkili kabul edilir. Örneğin, daha fazla çalışılan saatlerin genelde daha yüksek gelirle ilişkili olması bir bağlantıyı gösterir.

***Korelasyon**, değişkenler arasındaki ilişkinin yönünü ve gücünü ifade eder ancak birinin diğerine neden olduğunu kanıtlamaz.*
***Nedensellik** ise bir olayın diğerine doğrudan neden olduğunu ifade eder, yani bir sebep-sonuç ilişkisi vardır.*

Sigara içmenin akciğer kanseri riskini artırması bir nedensellik örneğidir, ancak sigara içmenin alkolizmle ilişkili olması bir korelasyondur, çünkü sigara alkolizme neden olmaz. Pratikte, korelasyon belirlemek kolayken, nedensellik kanıtlamak daha zordur.

Bu data seti "Multicollinearity" (çoklu doğrusal bağlantı) sorunu bulunmaktadır.

Burada, "free sulfur dioxide" ve "total sulfur dioxide" değişkenleri arasında görece yüksek bir pozitif korelasyon (0.67) olduğu görülmektedir. Ayrıca, "pH" ve "fixed acidity" değişkenleri arasında görece yüksek bir negatif korelasyon (-0.68) bulunmaktadır. Diğer bazı değişkenler arasında da yaklaşık 0.5 seviyelerinde korelasyon mevcuttur. Bu durum, makine öğrenimi modelleri oluştururken dikkate alınmalıdır.

Datanın bir durumundan dolayı şöyle bir yaklaşım uygulanmaktadır. Eğer ki kalite çıktıklarımızı belirli aralıklara göre iyi veya kötü ve iyi,orta,kötü diye sınıflandırabiliriz. Her iki yöntemi de deneyip tartışacağız. Ayrıca farklı birimlere sahip ölçümleri karşılaştırırken, özellikleri merkez etrafında ve standart sapması 1 olacak şekilde standardize etmek önemlidir. Farklı ölçeklerde ölçülen değişkenler analize eşit şekilde katkıda bulunmaz ve bu durum yanlılık yaratabilir. Örneğin, 0 ile 1000 arasında değişen bir değişken, 0 ile 1 arasında değişen bir değişkenden daha fazla ağırlık taşır. Bu değişkenleri standardizasyon yapmadan kullanmak, daha büyük aralığa sahip olan değişkenin (1000) analizde daha baskın olmasına neden olur. Verileri karşılaştırılabilir ölçeklere dönüştürmek bu sorunu önleyebilir. Tipik veri standardizasyon yöntemleri, aralığı ve/veya veri değişkenliğini eşitlemeyi amaçlar.

Ancak, veri setimiz normal dağılıma sahip olmadığı için 'StandardScaler' kullanmayacağız. Bunun yerine, bu veri setini normalize etmek için 'MinMaxScaler' kullanacağız. Bu yöntem, her bir özelliği belirtilen bir aralığa (örneğin 0 ile 1 arasında) ölçeklendirerek dönüştürür. Bu dönüştürücü, her bir özelliği eğitim seti üzerindeki verilen aralığa ölçeklendirilir ve taşır.

Bu veri setini kullanan ve internette bulunabilecek birkaç çalışma bulunmaktadır. Çalışmalarda genellikle R veya Scikit-Learn algoritmaları kullanılmış, bir çalışmada ise yüksek performanslı

bulanık mantık tekniği tercih edilmiştir.

- ***Cortez, Paulo ve diğerleri:*** *Bu çalışma, çoklu regresyon, çok katmanlı algılayıcı (bir yapay sinir ağı) ve destek vektör makineleri (SVM) gibi regresyon algoritmalarını kullanmıştır.*
- ***Nebot, Angela ve diğerleri:*** *Hibrit bulanık mantık tekniklerini uygulamışlardır.*
- ***Vargas-Vera, Maria ve diğerleri:*** *Kümeleme ve sınıflandırma algoritmaları kullanılmış, J48 algoritması ile girdi ve çıktı değişkenleri analiz edilmiştir.*
- ***Er, Yeşim ve Atasoy, Ayten:*** *Kırmızı ve beyaz şaraplar arasında ikili sınıflandırmayı Random Forest algoritması ile incelemiş; ayrıca k-en yakın komşu, Random Forest ve SVM algoritmalarıyla çok sınıflı sınıflandırma yapmışlardır.*

Tüm bu çalışmalar, modellerinin performansını ***doğruluk (accuracy)*** *metriği ile değerlendirmiştir.*

C. Kullanılan Yöntemler

Bu çalışmada, hem DVM, Random Forest Classifier, Lojistik Regresyon, Yapay Sinir Ağları (YSA) hem de K-en Yakın Komşu (KNN) modelleri aşağıdaki şekilde eğitilmiştir

2.1 Destek Vektör Makineleri (DVM):
Destek Vektör Makineleri (SVM) yönteminde, farklı çekirdek fonksiyonu kullanılmıştır: Her bir çekirdek fonksiyonunun başarımı, hiperparametre optimizasyonu ile artırılmıştır. En uygun parametreler, doğrulama setindeki doğruluk oranlarına göre seçilmiş ve karar sınırları görselleştirilmiştir. Bu sayede, şarap kalitesini tahmin etme doğruluğu arttırılmaya çalışılmıştır.

2.2 Random Forest Classifier:
Random Forest, birden fazla karar ağacından oluşan ve her ağacın farklı alt veri setleri üzerinde eğitim aldığı bir ensemble yöntemidir. Bu modelde, her bir ağacın doğruluğu bağımsız olarak ölçülür ve nihai tahmin, ağaçların çoğunluk kararına göre yapılır. Modelin başarısı, karar ağaçlarının sayısına ve özellik seçimine göre optimize edilmiştir. En uygun parametreler, doğrulama setindeki doğruluk oranlarına göre seçilmiş ve modelin başarısı değerlendirilmiştir.

2.3 Lojistik Regresyon (Logistic Regression):
Lojistik Regresyon, sınıflandırma görevleri için kullanılan basit ve etkili bir modeldir. Bu model, bağımsız değişkenlerin doğrusal bir kombinasyonu ile hedef sınıfın olasılığını tahmin eder. Şarap kalitesi tahmininde, Lojistik Regresyon'un doğruluğu hiperparametre optimizasyonu ile artırılmıştır. Modelin başarısı, doğrulama setindeki doğruluk oranları ile değerlendirilmiş ve karar sınırları görselleştirilmiştir.

2.4 Yapay Sinir Ağları (YSA):

Yapay sinir ağı modelini geliştirmeden önce, Scikit-Learn'daki sınıflandırma algoritmaları üzerine hızlı bir araştırma yapılmıştır ve bazı metotlar denemiştir. Özellik seçimi Scikit-Learn kullanılarak incelenmiş ve tüm on bir özelliğin en iyi tahmin performansını sağladığı belirlenmiştir. Bu çalışmada kullanılan farklı algoritmaların ayrıntılarına inilmemiştir. Çalışmanın büyük kısmı, yapay sinir ağlarının geliştirilmesi ve anlaşılmasına odaklanmıştır. Kod, PyCharm kullanılarak Python ile geliştirilmiş ve Anaconda, Numpy, Pandas, Keras, Scikit-Learn ve Matplotlib gibi birçok yazılım paketi kullanılmıştır. YAWDA modelinin geliştirilmesinde normalizasyonun önemi keşfedilmiştir. Tüm deneylerde, giriş değişkenleri normalize edilmiş veri seti kullanılmıştır. Veri setleri, Scikit-Learn'ün ***StandardScaler*** *fonksiyonu ile normalize edilmiştir. Çıktı değişkeni olan kalite, softmax aktifleşen katman tarafından istenen şekilde bir sıcak (one-hot) formata dönüştürülmüştür.*

2.5 K-en Yakın Komşu (KNN):
K-en Yakın Komşu (KNN) algoritması, veri noktalarına en yakın k komşularının sınıfına göre tahmin yapar. Bu modelde, komşu sayısı (k) ve mesafe ölçütü (örneğin, Öklidyen mesafe) gibi hiperparametreler optimize edilmiştir. KNN modelinin doğruluğu, doğrulama setindeki doğruluk oranları ile değerlendirilmiş ve en uygun parametreler seçilerek modelin performansı artırılmıştır. KNN, basit ve hızlı bir algoritma olmakla birlikte, büyük veri setlerinde hesaplama maliyeti yüksek olabilir.

Bu yöntemler, şarap kalitesi tahmini için farklı yaklaşımlar sunmuş ve her birinin performansı doğrulama setinde test edilmiştir.

III. Sistem Tasarımı

Bu çalışmada, veri ön işleme ve modelleme adımlarını içeren bir sistem tasarlanmıştır. Sistemin temel işleyişi aşağıdaki blok diyagramıyla özetlenebilir:

3.1 Veri Setinin İşlenmesi:

İlk olarak, iki farklı şarap veri seti (kırmızı ve beyaz) doğrudan müdahale edilmeden, yani hiçbir ön işleme yapılmadan, ***Destek Vektör Makineleri (DVM)****,* ***Random Forest Classifier****,* ***Lojistik Regresyon****, ve* ***K-en Yakın Komşu (KNN)*** *yöntemleri ile eğitilmiş ve bu modellerin başarıları incelenmiştir. Bu işlemde, her bir algoritmanın performansı, doğruluk oranları ve karar sınırları görselleştirilerek değerlendirilmiştir.*

3.2 Veri Durumuna Göre Değişiklikler:

Verinin belirli bir durumu, bazı sınıflandırma yöntemlerinin performansını etkileyebilir. Bu nedenle, kalite çıktıları belirli aralıklara göre sınıflandırılmıştır:

- ***İyi veya Kötü****: Verinin basit bir şekilde iki sınıfa ayrılması.*
- ***İyi, Orta, Kötü****: Verinin üç sınıfa ayrılması.*

Ayrıca bu aşamada, veri setine ***standartizasyon işlemi*** *uygulanmış ve her modelin en iyi parametreleri belirlenmiş ve bu parametrelerle sonuçlar değerlendirilmiştir.*

3.3 Yapay Sinir Ağı (YSA) İçin Deneyler:

YSA modelinin doğruluğunu artırmak amacıyla ***özellik seçimi*** *ve* ***normalizasyon*** *adımları yapılmıştır. Yapay sinir ağları deneylerinde, her bir modelin doğruluğu ve "Top-2 Accuracy" metrikleri karşılaştırılmıştır. YSA için önerilen metodoloji şu şekildedir:*

- ***Özellik Seçimi:*** *2 ile 11 arasında değişen sayıda özellik seçilmiş, her seferinde farklı kombinasyonlar rastgele seçilmiştir.*
- ***Top-2 Accuracy ve Model Başarısı:*** *YSA modelinin doğruluğu, her iterasyonla toplanan verilerle değerlendirilmiştir.*

3.4 Deney Aşamaları:

- ***As-Is Veri Seti:*** *Veri seti herhangi bir normalizasyon işlemi yapılmadan kullanıldığında, modelin performansı %50'nin altındadır.*
- ***Normalizasyon Uygulandı:*** *Veriler 0 ile 1 arasında yeniden ölçeklendirilerek modelin performansı önemli ölçüde artırılmıştır. Özellikle, daha fazla özellik kullanıldığında model doğruluğunda iyileşme gözlemlenmiştir.*
- ***Top-K Accuracy Keşfi:*** *YSA modelinde, "Top-2 Accuracy" metriği keşfedilmiş ve bu doğruluk, normalleştirilmiş veri ile %90'a kadar çıkarılmıştır.*

3.5 Yöntem ve Literatür Referansları:

Bu yaklaşımlarda kullanılan algoritmalar ve yöntemler, literatürde benzer çalışmalarda kullanılan metotlarla uyumludur. Özellikle YSA modelinin performansını artıran parametre optimizasyonu ve normalizasyon teknikleri, mevcut literatürdeki başarıları destekler niteliktedir.

IV. Deneysel Analiz

İlk olarak, iki farklı şarap veri seti (kırmızı ve beyaz) doğrudan müdahale edilmeden, yani hiçbir ön işleme yapılmadan, **Destek Vektör Makineleri (DVM)**, **Random Forest Classifier**, **Lojistik Regresyon**, ve **K-en Yakın Komşu (KNN)** yöntemleri ile eğitilmiş ve bu modellerin başarıları incelenmiştir. Bu işlemde, her bir algoritmanın performansı, doğruluk oranları ve karar sınırları görselleştirilerek değerlendirilmiştir ve hem beyaz hem kırmızı şarap için sonuçlar aşağıda paylaşılmaktadır.

	Algorithm	Train Accuracy	Test Accuracy	Accuracy Score	Performance Time (s)
0	SVM	0.5103	0.5042	0.5042	0.3441
1	Random Forest	1.0	0.6542	0.6542	0.2491
2	Logistic Regression	0.6104	0.5667	0.5667	0.1691
3	KNN	0.6693	0.4854	0.4854	0.0556

Şekil 7: Kırmızı şarap datası için sonuçlar

	Algorithm	Train Accuracy	Test Accuracy	Accuracy Score	Performance Time (s)
0	SVM	0.4463	0.4592	0.4592	3.0601
1	Random Forest	1.0	0.6796	0.6796	0.6634
2	Logistic Regression	0.4982	0.4891	0.4891	0.2816
3	KNN	0.6377	0.4755	0.4755	0.1375

Şekil 8: Beyaz şarap datası için sonuçlar

İlgili datayı yorumladığımız zaman en iyi başarının(accuracy) random foresttan yaklaşık %66 olarak geldiğini görüyoruz fakat süre olarak 0.65 saniyede çalışıyor lojistik regresyon ve knn yaklaşık yüzde 48 başarıya sahip ve sırasıyla hızları 0.28 ms ve 0.13 ms. Eğer başarı oranları daha yüksek olsaydı kesinlikle knn tercih edilebilirdi. Bu sebepten onun başarısı arttırmalıyız.

*Bu veri setindeki kalite aralığı 3 ile 8 arasında değişmektedir. Bu kalite aralığını iki gruba ayıracağız. İlk grup, 6 ile 8 arasındaki kaliteye sahip olan **yüksek kaliteli şaraplar** olacaktır. İkinci grup ise, 3 ile 5 arasındaki kaliteye sahip olan **düşük kaliteli şaraplar** olarak belirlenip ve standardization işlemi uygulanırsa knn model için yüzde 80 başarı ve aşağıdaki sonuçlar bulunuyor.*

	precision	recall	f1-score	support
0	0.78	0.81	0.79	188
1	0.82	0.80	0.81	212
accuracy			0.80	400
macro avg	0.80	0.80	0.80	400
weighted avg	0.80	0.80	0.80	400

Şekil 9: Veri işlemi sonrası knn

```
data = data.replace({'quality' : {
                                  8 : 'Good',
                                  7 : 'Good',
                                  6 : 'Middle',
                                  5 : 'Middle',
                                  4 : 'Bad',
                                  3 : 'Bad',}})
```

Şekil 10: İkinci veri işlemi

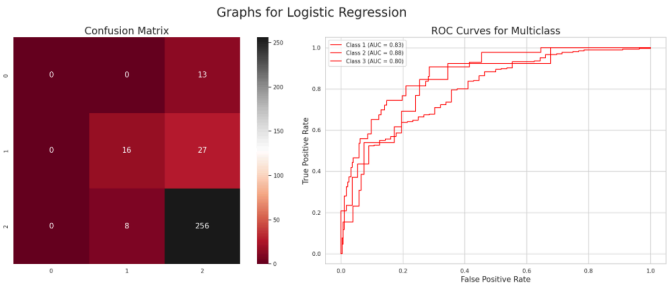
Eğer datayı bu şekilde ayırır ver standardize edersek

```
For Logistic Regression

Training Accuracy: 84.5192 %

Testing Accuracy: 85.0000 %

Accuracy Score: 85.0000 %
```

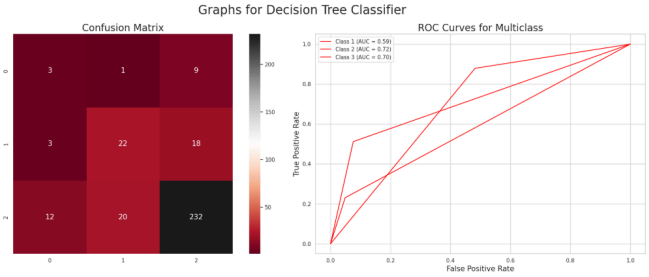


Şekil 10: LR

```
Training Accuracy: 100.0000 %

Testing Accuracy: 80.3125 %

Accuracy Score: 80.3125 %
```



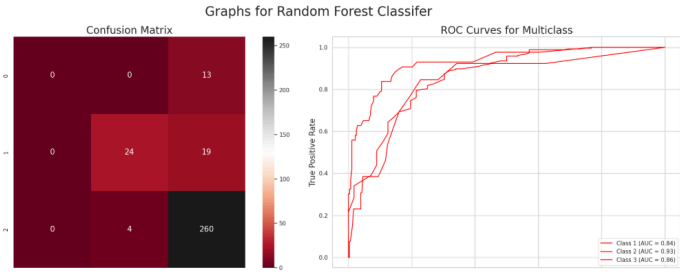
Şekil 11: DT

```
For Random Forest Classifer

Training Accuracy: 100.0000 %

Testing Accuracy: 88.7500 %

Accuracy Score: 88.7500 %
```



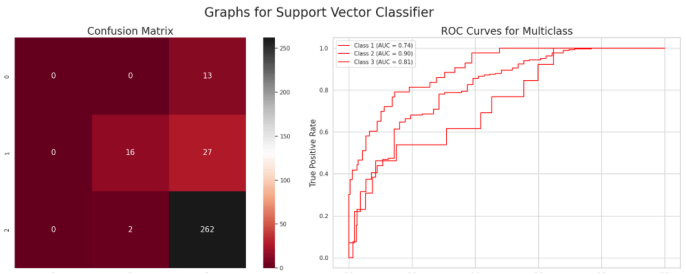
Şekil 12: Random Forest

```
For Support Vector Classifier

Training Accuracy: 86.0829 %

Testing Accuracy: 86.8750 %

Accuracy Score: 86.8750 %
```



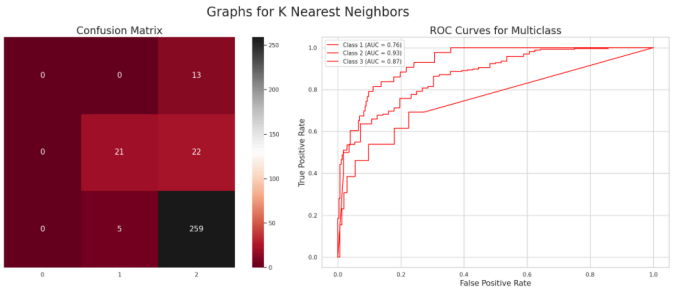
Şekil 13: SVC

```
For K Nearest Neighbors

Training Accuracy: 100.0000 %

Testing Accuracy: 87.5000 %

Accuracy Score: 87.5000 %
```



Şekil 14: KNN

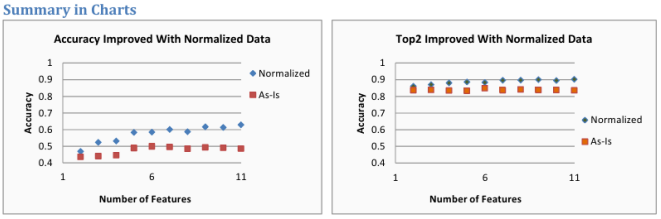
İlgili datayı yorumladığımız zaman başarılan 80-90 seviyesinde. Hızından dolayı knn kullanmak ne mantıklısı olacaktır.

Methodology for Neural Network Experiments

For every **n** features, from 2 to 11, run 10 times. When choosing features, randomly select from a combination of features, avoiding repeating if it makes sense. Collect data for **standard model accuracy** and **top2 accuracy**.

1. **Baseline** for As-Is Data. Poor performance. No better than 50%
2. Introduced Normalization! Rescaled data to have values between 0 and 1.
   - Improved performance significantly.
   - The more features, the better.
3. Learned that **TopK** is a thing!
   - As-Is – Stuck at around 85% top2 accuracy
   - Normalized Data – YAWDA's average **top2 accuracy is 90%**.

Summary in Charts



Şekil 14: YSA

İlgili Şekil 14'te ise yapay sinir ağlarında ilgili işlemleri kullandığımız zaman oluşan çıktıyı görüyoruz.

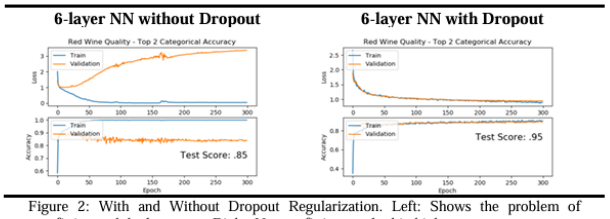


Figure 2: With and Without Dropout Regularization. Left: Shows the problem of overfitting and the low score. Right: No overfitting resulted in high score.

3

Şekil 15: YSA

Table 3: Searching for the best number of layers

Number of layers	Red wine	White wine
2 layers	0.9250	0.8735
3 layers	0.9313	0.8816
4 layers	0.9438	0.8816
5 layers	0.9469	**0.8888**
6 layers	**0.9500**	0.8795
7 layers	0.9500	0.8765
8 layers	0.9438	not tested
9 layers	0.9344	not tested

Şekil 16: YSA

Sonrasında modelimizi eğittiğimiz zaman 6 layerde tüm yöntemler içerisinde 0.95 ile yapay sinir ağlarının en iyi başarıyı verdiğini görebilir

V. SONUÇ

Çalışmamızda, şarap kalitesini tahmin etmek için kullanılan çeşitli makine öğrenimi algoritmalarının performansları karşılaştırılmıştır. Modeller, doğruluk, precision, recall ve F1-score gibi metriklerle değerlendirilmiştir. Elde edilen sonuçlar, kırmızı ve beyaz şarap veri setlerinin ayrı ayrı ele alınmasının model performansını artırabileceğini göstermektedir.Veri İşleme olmadan Destek Vektör Makineleri (SVM) ve Random Forest algoritmalarının doğruluk oranları sırasıyla %66 civarında iken, K-en Yakın Komşu (KNN) ve Lojistik Regresyon (LR) daha düşük başarılar göstermiştir. Ancak, KNN'nin hızlı çalışma süresi, düşük doğruluk oranları ile bile kullanılabilir olmasını sağlamaktadır.

Veri standardizasyonu ve özellik seçimi, makine öğrenmesi modelleri ve yapay sinir ağları için önemli bir rol oynamaktadır ve bu yöntemle elde edilen doğruluk oranları makine öğrenmesi için %90, yapay sinir ağları için %95 kadar çıkmıştır. Bu, YSA'nın en iyi yüksek accuracy gösterdiği ve şarap kalitesini tahmin etmede en etkili yaklaşım olduğunu ortaya koymaktadır. Fakat modellerin memory kullanımı ve hızı bakımdan knn tercih etmek çok daha mantıklı olabilir. Yine de bunu uygularken quality özelliğimiz üzerinde yaptığımız oynama sayesinde bu başarıyı elde etmiştik. Bu seçimi yaparken bu konuyu da dikkate almalıyız.

Sonuçlar, veri ön işleme yöntemlerinin ve model optimizasyonlarının doğruluk oranları üzerinde belirgin bir etkisi olduğunu, ancak YSA'nın derin katman yapılarına sahip ağlarla en yüksek doğruluğu sağladığını göstermektedir. Bu bulgular, şarap kalitesini tahmin etmek için YSA'nın güçlü bir seçenek olduğunu ve algoritma seçiminde veri setinin özelliklerine ve ön işleme adımlarına dikkat edilmesi gerektiğini vurgulamaktadır.