BURSA TEKNİK ÜNİVERSİTESİ

ŞARAP VERİ SETİ ANALİZİ

VERİ MADENCİLİĞİ PROJESİ

Taner Solak
18360859034

09,06,022

İÇİNDEKİLER

Sayfa

ŞEKİL LİSTESİ

Sayfa

TABLO LİSTESİ

WINE DATA SET MODEL EĞİTİMİ

ÖZET

Projede UCI Machine Learning Repository sitesinden alınan Wine Data Set veri setinin analizi yapıldı. Bu veriler, İtalya'da aynı bölgede yetişen ancak üç farklı çeşitten elde edilen şarapların kimyasal analizinin sonuçlarıdır. Analiz, üç şarap türünün her birinde bulunan 13 bileşen miktarını belirledi.

1. GİRİŞ

Projede UCI Machine Learning Repository sitesinden alınan Wine Data Set veri setinin analizi yapıldı. Bu veriler, İtalya'da aynı bölgede yetişen ancak üç farklı çeşitten elde edilen şarapların kimyasal analizidir. Bu analiz, üç şarap türünün her birinde bulunan 13 bileşen miktarının belirlenmesi ile sonuçlandı. [5]

2. METODOLOJİ

Wine Data Set kullanılarak analiz edilen verilerin açıklanması, görselleştirilmesi, eksik verilerin doldurulması daha sonrasında ise karar ağacı tekniği kullanılarak veri sınıflandırma modellemesi yapılmıştır. Bu proje veri analizi, ön işleme ve modelleme aşamalarından geçer. Veri analizi ve ön işleme adımları, modelleme adımları için hazılığı amaçlar. Modelleme aşaması, şarap sınıflandırma sorunu için karar ağacı tekniği kullanılarak yüksek doğrulukta bir tahmin modeli oluşturmayı içerir.

2.1 Veri Analizi

UCI Machine Learning Repository sitesinden aldığımız Wine Data Set 13 özniteliğe ve 534 örnekten oluşmaktadır. Bu özniteliklerin açıklaması Tablo 1'de gösterilmiştir.

Tablo 1 Özniteliklerin açıklaması.

Öznitelikler	Açıklama
Malic acid	Sayısal
Ash	Sayısal
Alcalinity of ash	Sayısal
Total phenols	Sayısal
Proanthocyanins	Sayısal
Color intensity	Sayısal
Proline	Sayısal
Flavanoids	Sayısal
OD280/OD315 of diluted wines	Sayısal
Hue	Sayısal

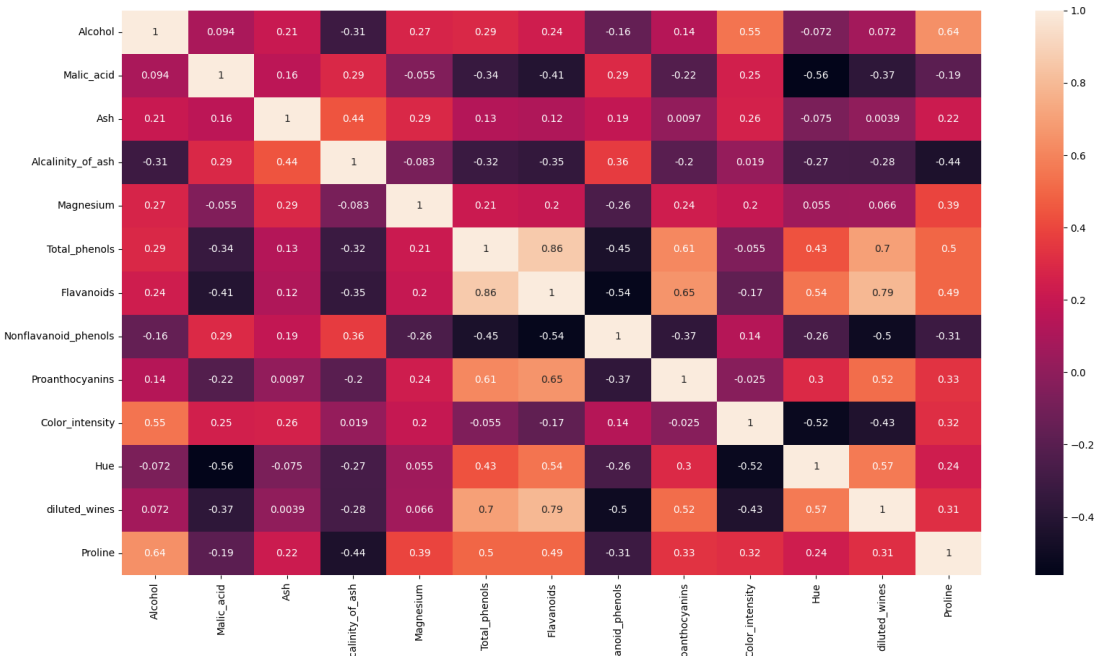
Magnesium	Sayısal
Alcohol	Sayısal
Nonflavanoid phenols	Sayısal

```
Data columns (total 15 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   id                            534 non-null    int64
 1   alcohol                       534 non-null    float64
 2   malic_acid                    534 non-null    float64
 3   ash                           534 non-null    float64
 4   alcalinity_of_ash             534 non-null    float64
 5   magnesium                     534 non-null    float64
 6   total_phenols                 534 non-null    float64
 7   flavanoids                    534 non-null    float64
 8   nonflavanoid_phenols          534 non-null    float64
 9   proanthocyanins               534 non-null    float64
 10  color_intensity               534 non-null    float64
 11  hue                           534 non-null    float64
 12  od280/od315_of_diluted_wines  534 non-null    float64
 13  proline                       534 non-null    float64
 14  target                        534 non-null    int64
dtypes: float64(13), int64(2)
memory usage: 62.7 KB
```

Şekil 2.1 Değişkenlerin bilgileri.

	count	mean	std	min	25%	50%	75%	max
id	534.0	266.500000	154.296792	0.00	133.2500	266.500	399.7500	533.00
alcohol	534.0	12.596404	0.839504	10.40	11.9200	12.605	13.2400	14.69
malic_acid	534.0	1.803371	1.129380	-0.27	0.9525	1.450	2.5875	5.57
ash	534.0	2.233783	0.285172	1.15	2.0500	2.230	2.4300	3.20
alcalinity_of_ash	534.0	17.876180	3.376112	7.99	15.8000	17.760	19.8775	28.74
magnesium	534.0	92.548633	14.859507	57.02	81.3225	90.840	100.6275	159.46
total_phenols	534.0	1.996723	0.652526	0.63	1.4700	1.995	2.5100	3.84
flavanoids	534.0	1.534307	1.047399	-0.49	0.5550	1.590	2.3875	4.35
nonflavanoid_phenols	534.0	0.302210	0.127285	0.02	0.2100	0.280	0.3975	0.62
proanthocyanins	534.0	1.305693	0.614343	-0.15	0.8925	1.260	1.6400	3.44
color_intensity	534.0	3.889401	2.349626	-0.37	2.0625	3.680	5.0800	12.80
hue	534.0	0.852569	0.243022	0.35	0.6800	0.870	1.0300	1.69
od280/od315_of_diluted_wines	534.0	2.266685	0.740055	0.64	1.6125	2.405	2.8300	3.90
proline	534.0	592.646629	333.830211	9.80	352.7250	521.920	807.5650	1654.83
target	534.0	0.947566	0.775482	0.00	0.0000	1.000	2.0000	2.00

Şekil 2.2 Değişkenlerin istatistikleri.



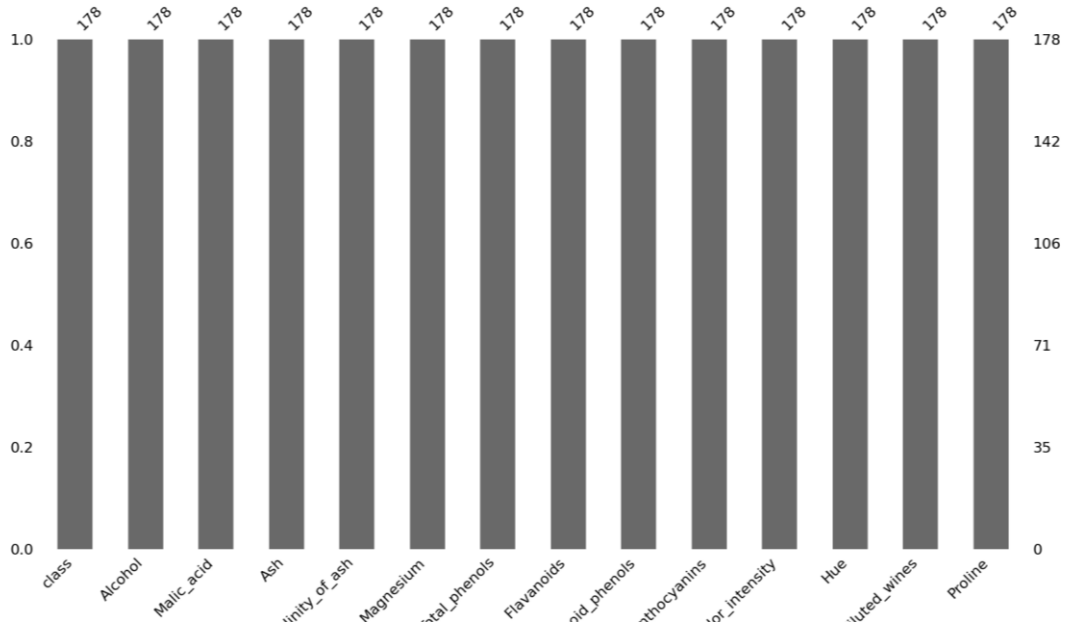
Şekil 2.3 Değişkenler arasındaki korelasyonun ısı haritası ile gösterimi.

2.2 Veri Ön İşleme

Bu projede eksik değer analizi, aykırı değerlerin analizi teknikleri veri ön işleme için kullanılmıştır. Bu teknikler veri setini veri madenciliği algoritmalarını uygulayabilecek hale geetirmek için kullanılır. Veri setinde eksik veri bulunamadığından herhangi bir doldurma veya silme işlemi yapılmamıştır fakat aykırı değer baskılama yöntemi kullanılmak durumunda kalınmıştır.

2.2.1 Eksik Değer Analizi

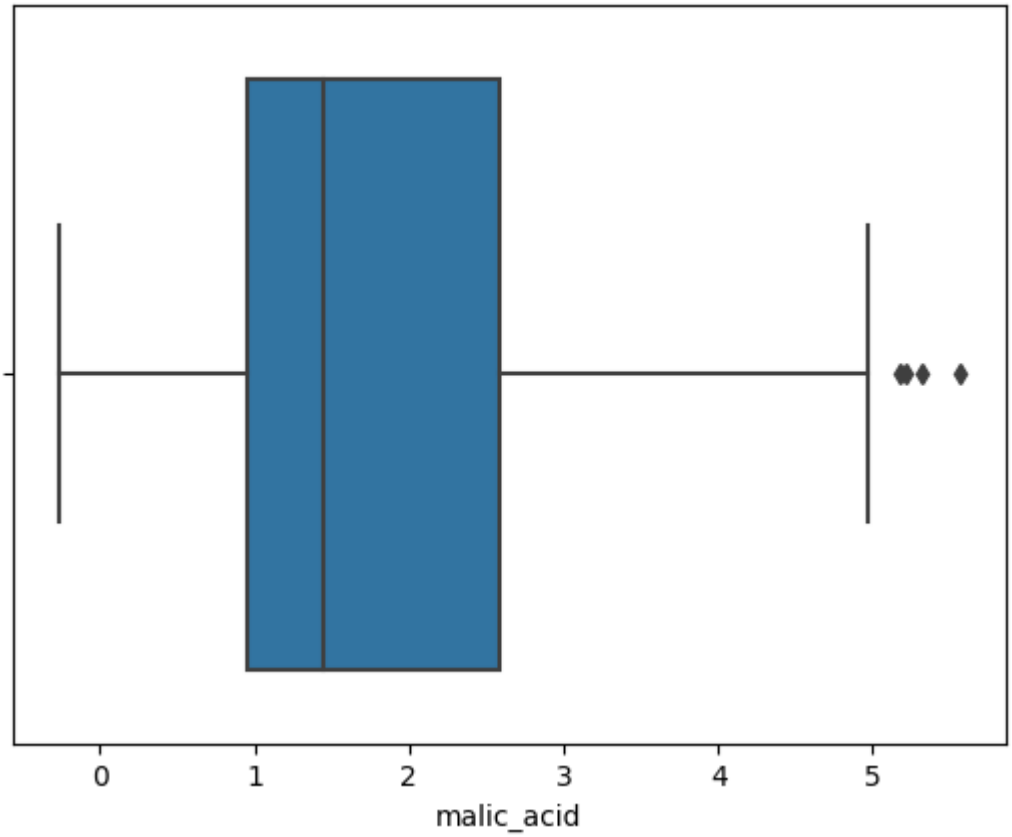
Veri setinde eksik değer bulunmamıştır.



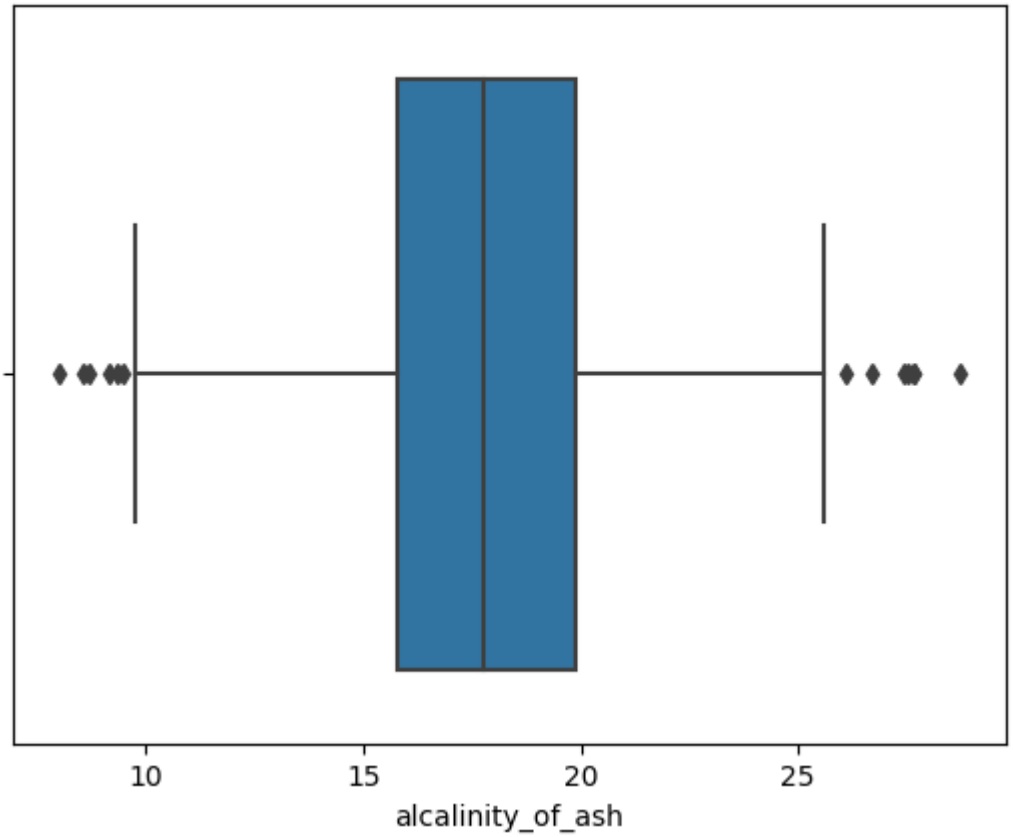
Şekil 2.1 Bu tabloda veri setinde eksik değer bulunmadığı gösterilmiştir.

2.2.2 Aykırı Değer Analizi

İstatistikte aykırı değer, diğer gözlemlerden önemli ölçüde farklı olan bir veri noktasıdır. Aykırı değer, ölçümdeki değişkenlikten kaynaklanabilir veya deneysel hatayı gösterebilir; ikincisi bazen veri setinden hariç tutulur. Bir aykırı değer, istatistiksel analizlerde ciddi sorunlara neden olabilir. Veri setimizde az örnek bulunmasından ve hazırlayan kişinin asıl veri setinden bir kısmını kaybetmesinden dolayı ayrık değerler az çıkmış durumda. Malic acid ve Alcanity of ash(Şekil 2.5, Şekil 2.6) değişkenleri için ayrık değer analizinin sonuçları görselleştirilmiştir. Aykırı değerlerde alt eşik değerinin altında olan değerler alt sınıra üst eşik değerinin üstünde olan değerler üst sınıra eşitlenir. [6]



Şekil 2.5 Malic acid özniteliğinin aykırı değerleri için boxplot gösterimi.



Şekil 2.6 Alcanity of ash özniteliğinin aykırı değerleri için boxplot gösterimi.

3. MODELLEME

Bu bölümde veri setinde yaptığımız veri analizleri sonucunda yaptığımız değişiklerle birlikte modellemenin nasıl yapıldığı açıklanacaktır.

3.1 Karar Ağacı Tekniği(Decision Tree)

Ağaç tabanlı öğrenme algoritmaları, en çok kullanılan gözetimli öğrenme algorimalarındandır. Genel itibariyle ele alınan bütün problemlerin (sınıflandırma ve regression) çözümüne uyarlanabilirler. Karar agaçları, tesadüfi orman, gradyen güçlendirme (gradient boosting) gibi yöntemler, her türlü veri bilimi problemlerinde yaygın bir şekilde kullanılmaktadırlar. [3]

Bir karar ağacı, çok sayıda kayıt içeren bir veri kümesini, bir dizi karar kuralları uygulayarak daha küçük kümelere bölmek için kullanılan bir yapıdır. Yani basit karar verme adımları uygulanarak, büyük miktarlardaki kayıtları, çok küçük kayıt gruplarına bölerek kullanılan bir yapıdır. Anlaması, yorumlaması ve görselleştirilebilmesinin kolaylığı nedeniyle karar ağacı yöntemi seçilmiştir.

3.2 Gini Index

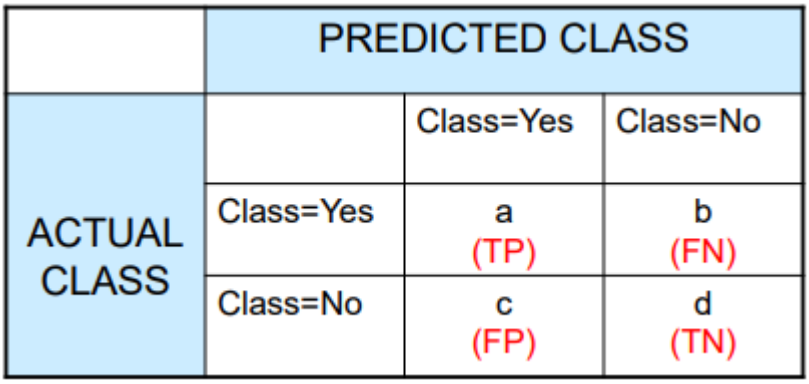
Algoritma seçimi, hedef değişkenin tipine dayanır. Karar ağaçlarında en sık kullanılan algoritmalar; kategorik değişkenler için Entropi, Gini, Sınıflandırma Hatası; sürekli değişkenler için ise En Küçük Karalere yöntemi şeklindedir. Bu veri setinde Gini algoritması kullanılmıştır.

Gini Index

$$GINI(t) = 1 - \sum_j [p(j \mid t)]^2$$

Şekil 3.1 Gini Index formülü.

3.3 Model Başarı Değerlendirme Metrikleri



Şekil 3.2 Performans değerlendirme metrikleri.

- True positive rate (TPR) veya sensitivity (hassasiyet), model tarafındandoğru şekilde tahmin edilen pozitif örneklerin oranı olarak tanımlanır, TPR = TP/(TP + FN).

- True negative rate (TNR) veya specificity, model tarafından doğru bir şekilde tahmin edilen negatif örneklerin oranı olarak tanımlanır, TNR = TN/(TN + FP).

- False positive rate (FPR), pozitif bir sınıf olarak tahmin edilen negatif örneklerin oranıdır, FPR = FP/(TN + FP),

- False negative rate (FNR), negatif bir sınıf olarak tahmin edilen pozitif örneklerin oranıdır, yani, FNR = FN/(TP + FN).

Accuracy, Precision, Recall ve F1 skor değerlerine dayalı olarak sınıflandırıcının performansını değerlendirilecektir. [2]

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1\text{-}score = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Şekil 3.3 Performans metriklerinin formülleri.

- Accuracy, bir modelin başarısını ölçmek için kullanılan bir metriktir. Doğru yaptığımız tahminlerin sayısının bütün yaptığımız tahminlerin sayısına bölümünden bulunur.
- Precision, pozitif olarak tahmin ettiğimiz(TP + FP) örneklerin kaçının doğru tahmin edildiğinin oranıdır.
- Recall, pozitif olarak tahmin edilmesi(TP + FN) gereken örneklerin oransal olarak kaçının doğru tahmin edildiğinin göstergesidir.
- F1 Score, değeri bize Precision ve Recall değerlerinin harmonik ortalamasını göstermektedir.

3.4 Hiperparametre Optimizasyonu

Makine öğrenmesi modelleri tasarlanırken girilmesi gereken parametreler olan hiper parametrelerin optimizasyonu, işletme maliyetini düşürdüğü için makine öğrenmesi modellerinin performansı üzerinde olumlu bir etkiye sahiptir. Bu alanda literatürde sıklıkla kullanılan dört yöntem (grid search, random search, Bayesian, and evolutionary algorithms) vardır ve bu projede grid search yöntemini kullanılmıştır. Çalışma kapsamında bu yöntemler tartışılmış, avantaj ve dezavantajları belirlenmiştir. Çalışmalara baktığımızda hiper parametre değerlerinin belirli aralıklarla seçildiği ve bu değerler arasındaki bağlantının incelendiği bir hiper parametre analizi bölümünün tüm çalışmalarda olması gerektiği belirtilmektedir. [8]

3.4.1 Grid Search

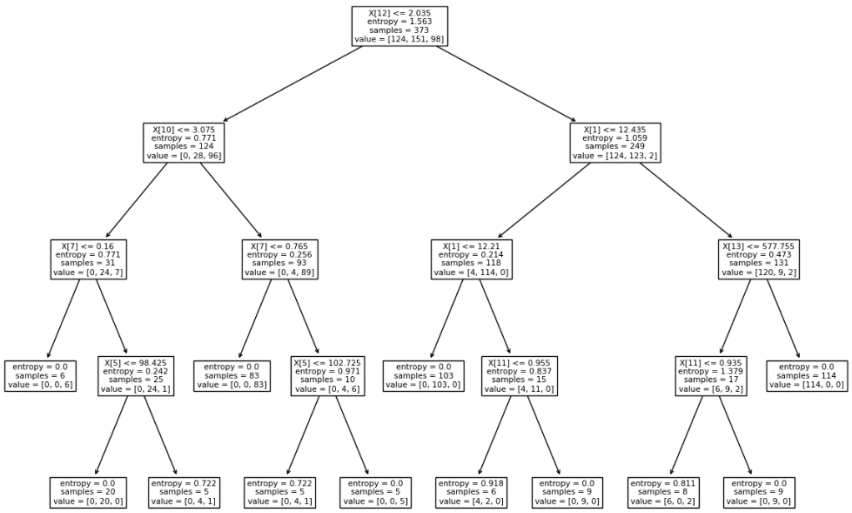
Hiper parametrelerden bazıları sonsuz sayıda değer alabilecek konumdadırlar. Bununla birlikte biz problem hakkında sahip olduğumuz ön bilgileri kullanarak hiper parametrelerin alabilecekleri değerler için aralıklar belirleyebiliriz. Belirlediğimiz bu aralıklardan belirli ana noktalar seçilerek hiper parametreler için değer listeleri oluşturulur. [1]

Grid search ile hiper parametre seçim işleminde; belirlenen aralıkta bulunan tüm değerlerin kombinasyonları için ağ eğitilip sonuçlar gözlenir duruma göre en iyi kombinasyon hiper parametre grubu olarak seçilir.

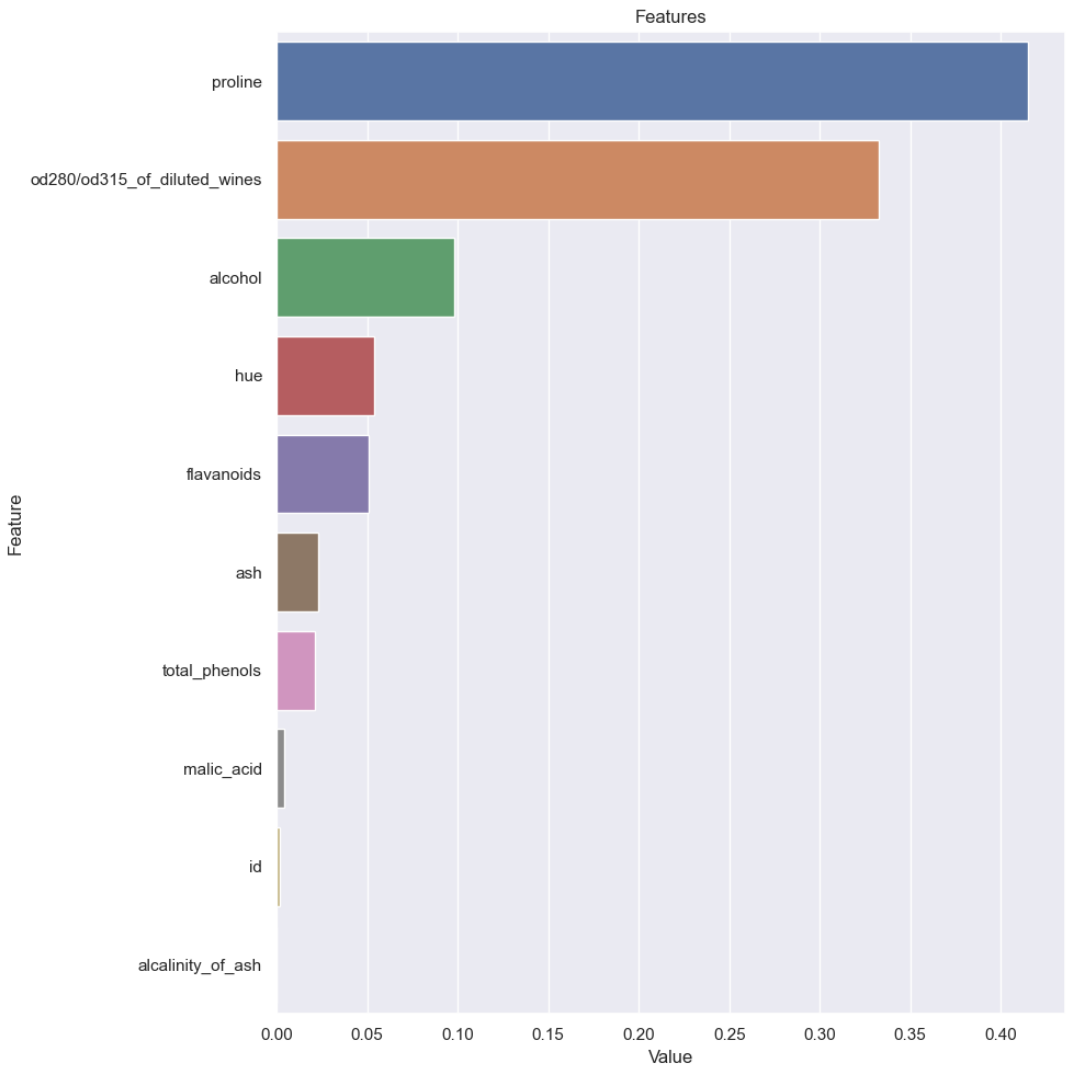
3.4.2 K-fold Cross-validation

Cross-validation, yapılan bir istatistiksel analizin bağımsız bir veri setinde nasıl bir sonuç elde edeceğini sınayan bir model doğrulama tekniğidir. Başlıca kullanım alanı bir öngörü sisteminin pratikte hangi doğrulukla çalışacağını kestirmektir. K-Folds Cross Validation'da verilerimizi k farklı alt kümeye böleriz. Verilerimizi eğitmek ve son alt kümeyi test verisi olarak bırakmak için k-1 adet alt kümeyi kullanırız. k adet deney sonucunda ortaya çıkan ortalama hata değeri modelimizin geçerliliğini belirtir. Bu çalışmada k değerimizi 10 olarak alıyoruz. [7]

4. SONUÇ



Şekil 4.1 Modelimizin oluşturduğu karar ağacının görselleştirilmiş hali.



Şekil 4.2 Özniteliklerin modelimize etki etme oranları.

Tablo 2 Model parametre optimizasyonu yapmadan önce metrik değerler.

METRİK	F1 SCORE	PRECISION	RECALL	ACCURACY
SCORE	0.9313606449444766	0.9339097205238506	0.9316770186335404	0.9316770186335404

Tablo 3 Model parametre optimizsayonu yapıldıktan sonra metrik değerler.

METRİK	F1 SCORE	PRECISION	RECALL	ACCURACY
SCORE	0.912561702 4014224	0.914912164 7782732	0.9130434782 608695	0.91304347826086 95

Bu değerlerden anlaşıldığı gibi parametre optimizasyonu yapıldıktan sonra metrik değerler düşmüştür. Modele herhangi bir cross-validation işlemi uygulanmadığı için rastgelelikten etkilenir ve yanlış sonuç verebilir.

4.1 Sonuçların Karşılaştırılması

	precision	recall	f1-score	support
0	0.88	0.98	0.93	51
1	0.91	0.85	0.88	61
2	0.96	0.92	0.94	49
accuracy			0.91	161
macro avg	0.92	0.92	0.91	161
weighted avg	0.91	0.91	0.91	161

Şekil 4.3 Kendi çalışmamın sonuçları.

	precision	recall	f1-score	support
0	0.88	0.97	0.92	31
1	0.98	0.89	0.93	47
2	0.97	1.00	0.98	29
accuracy			0.94	107
macro avg	0.94	0.95	0.95	107
weighted avg	0.95	0.94	0.94	107

Şekil 4.4 Karşılaştırma yaptığım çalışmanın sonuçları.

Bu görsellerden anlaşılacağı üzere karşılaştırdığım modelin kendi oluşturduğum modelden daha iyi sonuç verdiği anlaşılmıştır. [4]

5. KAYNAKLAR

[1] https://medium.com/deep-learning-turkiye/derin-ogrenme-uygulamalarinda-model-dogrulama-ve-hiper-parametre-secim-yontemleri-823812d95f3

[2] https://devreyakan.com/performans-metrikleri/

[3] https://medium.com/@k.ulgen90/makine-öğrenimi-bölüm-5-karar-ağaçları-c90bd7593010

[4] https://www.kaggle.com/code/iinaam/growing-a-random-forest-from-scratch-acc-96

[5] https://archive.ics.uci.edu/ml/datasets/Wine

[6] https://tr.wikipedia.org/wiki/Aykırı_değer

[7] https://medium.com/@gulcanogundur/model-seçimi-k-fold-cross-validation-4635b61f143c

[8] https://ieeexplore.ieee.org/document/8965609