Makine Öğrenmesi ile Şarap Sınıflandırması

Alper KARACA - Bilgisayar Mühendisliği
25221601001
Github: https://github.com/thealper2/makine-ogrenmesi-vize

1. Proje Konusu Hakkında

Bu çalışmada, scikit-learn kütüphanesinden yüklenen Wine veri seti üzerinde makine öğrenmesi tekniklerinin kapsamlı bir şekilde uygulanması ve değerlendirilmesi amaçlanmıştır. Çalışma kapsamında:

- Veri ön işleme teknikleri uygulanmıştır.
- Çoklu sınıflandırma algoritmaları ile modeller eğitilmiştir.
- Boyut indirgeme (PCA ve LDA) yöntemleri uygulanmıştır.
- Model performansları karşılaştırılmıştır.
- XAI (Explainable AI - Açıklanabilir Yapay Zeka) yöntemleri ile model kararları yorumlanmıştır.

1.1 Kullanılan Kütüphaneler ve Araçlar

Çalışmada aşağıdaki kütüphaneler kullanılmıştır.

- **NumPy ve Pandas**: Veri manipülasyonu, dizi işlemleri ve veri çerçevesi yönetimi.
- **Matplotlib ve Seaborn**: Veri dağılımları ve model sonuçlarının görselleştirilmesi.
- **Scikit-learn**: Makine öğrenmesi modelleri, veri ön işleme ve model değerlendirme.
- **SHAP**: Model tahminlerinin yerel ve global yorumlanabilirliği.
- **PCA ve LDA**: Boyut indirgeme ve öznitelik çıkarımı.

1.2 Çalışma Metodolojisi

Çalışmada aşağıdaki adımlar takip edilmiştir:

1. **Veri Keşfi ve Ön İşleme**: Eksik veri analizi, normalizasyon, veri temizleme.
2. **Temel Modelleme**: Ham ver ile çoklu sınıflandırıcıların eğitimi.
3. **Boyut İndirgeme**: PCA ve LDA uygulamaları.
4. **Gelişmiş Modelleme**: Boyut indirgeme sonrası model performans karşılaştırması.
5. **Model Yorumlama**: XAI yöntemleri ile model kararlarının analizi.
6. **Sonuç ve Değerlendirme**: Tüm bulguların özetlenmesi.

1.3 Wine Veri Seti Hakkında

Wine veri seti, İtalya'nın aynı bölgesinden ancak üç farklı türdeki şaraplardan alınan örneklerin kimyasal analizlerini içermektedir. Veri seti 13 farklı kimyasal bileşen özniteliği ve 3 sınıf etiketinden oluşmaktadır. Toplam 178 örnek içermektedir.

2. Veri Setinin Yüklenmesi

2.1 Scikit-learn'de Veri Seti Yükleme

Bu çalışmada, scikit-learn kütüphanesi içinde bulunan Wine Classification veri seti kullanılmıştır. Veri seti üç farklı İtalya şaraplarının kimyasal analizlerini içermektedir.

2.2 Veri Çerçevesi Oluşturma

Özellikler ve hedef değişken ayrıldıktan sonra, veriler Pandas DataFrame formatına dönüştürülmüştür.

2.3 Çıktılar ve İlk Analiz

Veri setinde toplam 178 adet gözlem bulunmaktadır. Toplam 13 adet özellik bulunmaktadır. Hedef değişken, 3 farklı sınıftan oluşmaktadır. Toplam sütun sayısı 14'tür (13 özellik + 1 hedef). Hedef değişken dağılımı şu şekildedir:

- **Sınıf 0**: 59 örnek.
- **Sınıf 1**: 71 örnek.
- **Sınıf 2**: 48 örnek.

Veri setinin tüm sütunlarının açıklaması şu şekildedir:

- **alcohol (Sayısal)**: Şaraptaki alkol içeriği. Yüzde olarak hacim (% vol). Şarabın gücünü ve vücut yapısını belirler. Değerleri %11-14.83 arasıdır.
- **malic_acid (Sayısal)**: Malik asit miktarı. Şarabın asidik yapısını ve ekşiliğini belirleyen ana organik asittir. Üzümün kendisinden doğal olarak gelir. Yüksek değerler daha keskin, düşük değerler daha yumuşak tat sunar. Değerleri 0.75-5.8 arasındadır.
- **ash (Sayısal)**: Toplam kül miktarı. Şarabın yakılması sonucu geriye kalan mineral içeriğidir. Toprak yapısı ve mineral zenginliği hakkında bilgi verir. Değerleri 1.36-3.23 arasındadır.
- **alcalinity_of_ash (Sayısal)**: Külün alkalinite değeri. Potasyum karbonat eşdeğeri olarak ölçülür. Şaraptaki toplam mineral alkali içeriğidir. Değerleri 10.6-30.0 arasındadır.
- **magnesium (Sayısal)**: Magnezyum içeriği. Hem sağlık açısından hem de şarap stabilitesi için önemi mineraldir. Birimi, miligram (mg). Değerleri 70-162 arasındadır.
- **total_phenols (Sayısal)**: Toplam fenol içeriği. Şarabın antioksidan kapasitesini ve renk stabilitesini belirler. Daha yüksek değerler daha kompleks aroma ve daha uzun raf ömrü sunar. Değerleri 0.98-3.88 arasındadır.
- **flavanoids (Sayısal)**: Flavonoid fenollerinin miktarı. Kateşin, epikateşin, prosiyanidinler. Burukluk (tanin) hissi. Renk stabilitesi. Antioksidan özellikler. Değerleri 0.34-5.08 arasındadır.
- **nonflavanoid_phenols (Sayısal)**: Flavonoid olmayan fenoller. Gallik asit, kafeik asit, kumarik asit. Daha hafif burukluk, farklı aroma profili. Değerleri 0.13-0.66 arasındadır.
- **proanthocyanins (Sayısal)**: Proantosiyaninler (taninler). Ağızda buurukluk hissi. Yapısal bütünlük. Yaşlanma potansiyeli. Değerleri 0.41-3.58 arasındadır.
- **color_indensity (Sayısal)**: Renk yoğunluğu. Absorbans veya optik yoğunluk. Şarabın görsel derinliği ve olgunluk göstergesi. Değeleri 1.28-13.0 arasındadır.
- hue (Sayısal)}: Renk tonu veya gölge. Kırmızı/mor oranı. Daha yüksek değerler daha kırmızı tonlar, daha düşük değerler daha mor/mavi tonlar. Değerleri 0.48-1.71 arasındadır.
- **od280/od315_of_diluted_wines (Sayısal)**: Seyreltilmiş şarapların optik yoğunluk oranı. 280 nm ve 315 nm dalga boylarında absorbans oranı. Protein ve polifenol içeriğinin göstergesi. Şarabın saflık ve kalite kontrolü. Değerleri 1.27-4.0 arasındadır.
- **proline (Sayısal)**: Prolin amino asiti içeri. Maya besini olarak fermantasyonu etkiler. Şarabın stabilitesine katkıda bulunur. Kalite göstergesi olarak kullanılır. Birimi miligram/litre. Değerleri 278-1680 arasındadır.
- **target (Sayısal)**: Şarap türü. Sayıların hangi türe karşılık geldiği gizlenmiştir. 0, 1 ve 2 olmak üzere 3 farklı sınıfa sahiptir.

3. Veri Seti Kalite Kontrolleri

3.1 Eksik Değer Analizi

Veri setindeki eksik değerlerin tespiti için her sütun kontrol edilmiştir. Veri setinde hiç eksik değer bulunmamaktadır. Bu durum, veri setinin temiz olduğunu göstermektedir. Eksik değer doldurma işlemine gerek kalmamıştır.

3.2 Aykırı Değer (Outlier) Analizi

Aykırı değerlerin tespiti için Tukey's IQR (Tukey's Interquartile Range) yöntemi kullanılmıştır. Toplam 17 gözlem aykırı değer içermektedir. Bu, veri setinin %9.55'ine karşılık gelmektedir. Aykırı değer içeren sütunlar:

- malic_acid sütununda 3 adet aykırı değer bulunmuştur.
- ash sütununda 3 adet aykırı değer bulunmuştur.
- alcalinity_of_ash sütununda 4 adet aykırı değer bulunmuştur.
- magnesium sütununda 4 adet aykırı değer bulunmuştur.
- proanthocyanins sütununda 2 adet aykırı değer bulunmuştur.
- color_intensity sütununda 4 adet aykırı değer bulunmuştur.
- hue sütununda 1 adet aykırı değer bulunmuştur.

Aykırı değerlerin etkileri şunlardır:

- **Olumlu Etkiler**: Aykırı değerler doğal varyasyonu temsil edebilir ve modelin genelleme yeteneğini artırabilir.
- **Olumsuz Etkiler**: Doğrusal modelleri (Lojistik Regresyon) olumsuz etkileyebilir. Uzaklık tabanlı algoritmaları (KNN) bozabilir. Model performansını düşürebilir.

Bu çalışmada aykırı değerler çıkarılmamıştır, robust scaler kullanılmıştır.

3.3 Veri Tipi ve Dağılımı İncelenmesi

Toplam 14 adet sütun bulunmaktadır. 14 sütunda sayısal değerlerden oluşmaktadır. Kategorik değişken bulunmamaktadır. Hedef değişken sayısal formatta kodlanmıştır. Özellik sütunları float64 değerlerden, hedef sütun ise int64 değerlerden oluşmaktadır.

4. Keşifsel Veri Analizi (EDA)

4.1 İstatistiksel Özellikler

Veri setinin sütun bazında istatistikleri aşağıdaki tabloda verilmiştir.

Özellik	Mean	Median	Min	Max	Std	Q1	Q2
alcohol	13.00	13.05	11.03	14.83	0.81	12.36	13.67
malic_acid	2.33	1.86	0.74	5.80	1.11	1.60	3.08
ash	2.36	2.36	1.36	3.23	0.27	2.21	2.55
alcalinity_of_ash	19.49	19.50	10.60	30.00	3.33	17.20	21.50
magnesium	99.74	98.00	70.00	162.00	14.28	88.00	107.00
total_phenols	2.29	2.35	0.98	3.88	0.62	1.74	2.80

flavanoids	2.02	2.13	0.34	5.08	0.99	1.20	2.87
nonflavanoid_phenols	0.36	0.34	0.13	0.66	0.12	0.27	0.43
proanthocyanins	1.59	1.55	0.41	3.58	0.57	1.25	1.95
color_intensity	5.05	4.69	1.28	13.00	2.31	3.22	6.20
hue	0.95	0.96	0.48	1.71	0.22	0.78	1.12
od280/od315	2.61	2.78	1.27	4.00	0.70	1.93	3.17
proline	746.89	673.5	278.00	1680.00	314.90	500.50	985.00

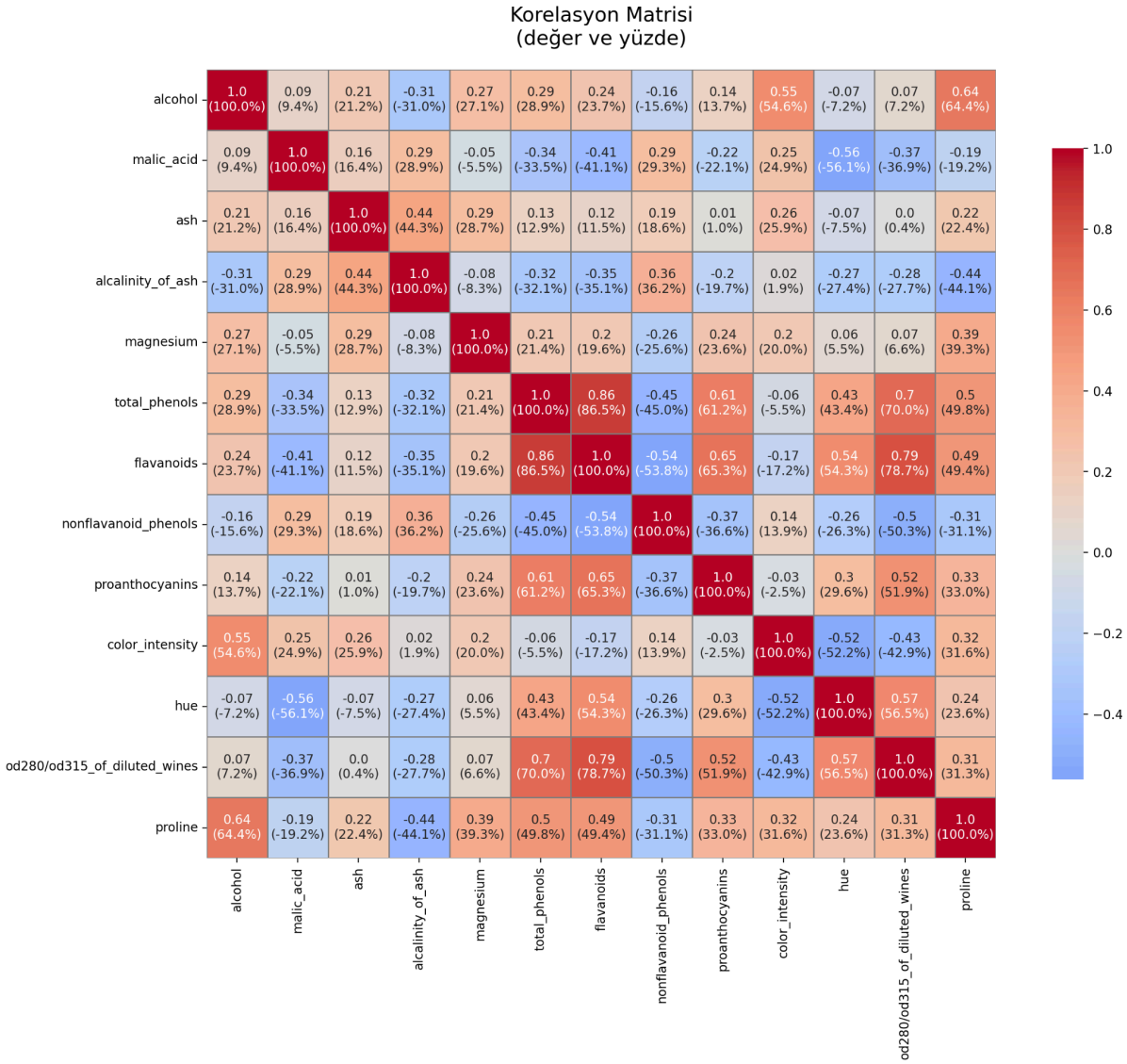
Tabloya göre özellikler arasında belirgin ölçek farkları bulunmaktadır. Bazı satırlar 100 ve 1000 üzeri değerler alırken (örn proline, magnesium) bazı satırlar ise 20'den daha az değerler almıştır (alcohol, ash). Proline sütunu en yüksek standart sapmaya sahipken (314.90), nonflavanoid_phenols en düşük standart sapmaya sahiptir. Bir çok özellikte mean ve max değerleri birbirine yakınken 3.2 adımında tespit edilen aykırı değerlere sahip sütunlar (malic_acid, ash, magnesium..) burada da gözlemlenmiştir.

4.2 Korelasyon Matrisi

Pearson korelasyon matrisi ile özellikler arasındaki lineer ilişkiler incelenmiştir. En yüksek korelasyon çiftleri flavanoids-total_phenols (%86.45), flavanoids-od280/od315_of_diluted_wines (%78.72), total_phenols-od280/od315_of_diluted_wines (%69.91) sütunları arasında gözlemlenmiştir. Buna göre:

- **Yüksek Pozitif Korelasyon (%86.45)**: flavanoids ve total_phenols arasındaki güçlü ilişki beklenen bir durumdur, çünkü flavonoidler toplam fenollerin bir alt grubudur.
- **Orta-Yüksek Pozitif Korelasyon (%78.72)**: flavanoids ve od280/od315_of_diluted_wines arasındaki ilişki, flavonoid içeriğinin optik yoğunluk üzerindeki etkisini göstermektedir.
- **Orta-Yüksek Pozitif Korelasyon (%69.91)**: total_phenols ve od280/od315_of_diluted_wines arasındaki korelasyon, fenolik bileşiklerin şarabın optik özelliklerini etkilediğini doğrulamaktadır.

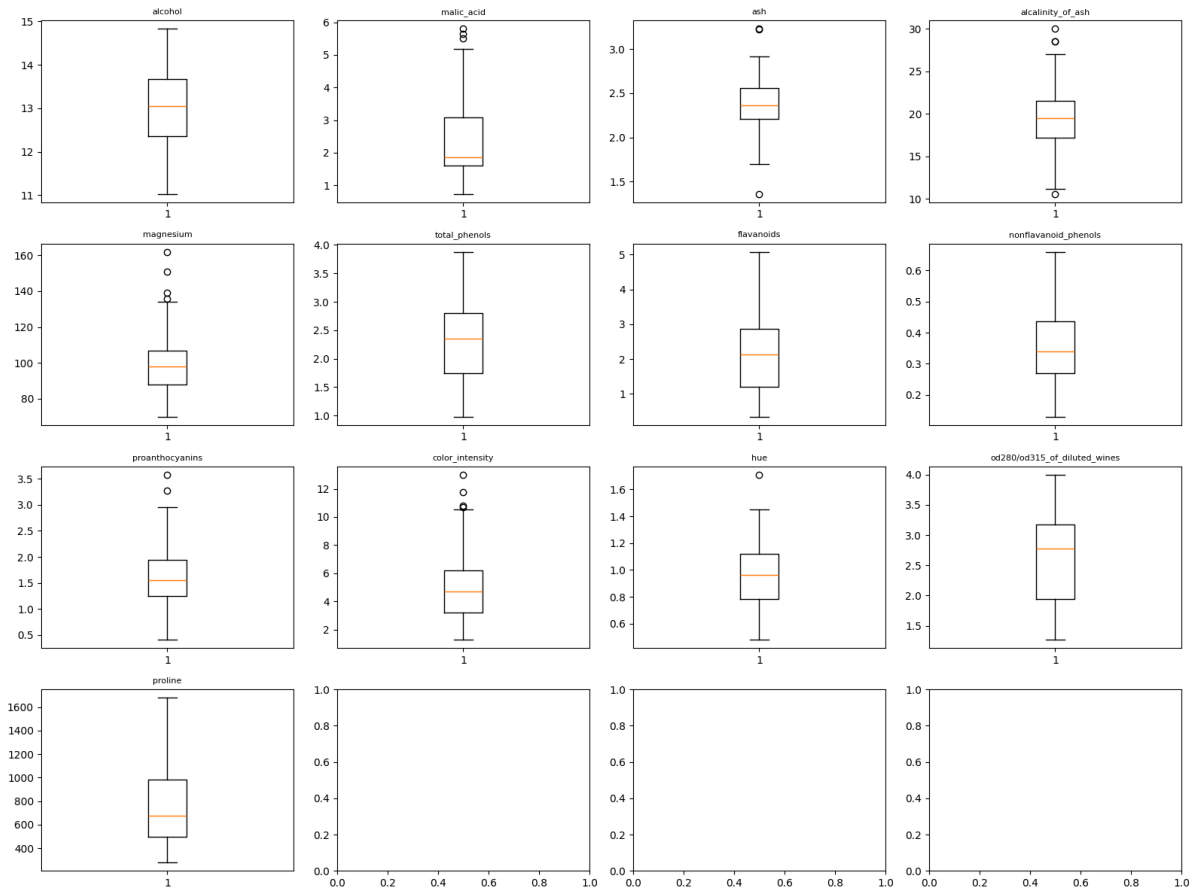
Hesaplanan Pearson korelasyon matrisi aşağıda verilmiştir.



Veri setinde güçlü korelasyonlar bulunmaktadır, bu da özellik seçimi veya boyut indirgeme için uygundur. Ölçek farklılıkları standardizasyon gerektirmektedir.

4.3 Boxplot Analizi

Tüm özelliklerin dağılımı ve aykırı değerleri boxplot grafikleri ile incelenmiştir. Oluşturulan boxplot grafiği aşağıda verilmiştir.



Grafiğe göre belirgin aykırı değerler:

- **malic_acid**: Üst uçta 3 aykırı değer içermektedir.
- **ash**: Üst uçta 1, alt uçta 1 aykırı değer içermektedir.
- **alcalinity_of_ash**: Üst uçta 2, alt uçta 1 aykırı değer içermektedir.
- **magnesium**: Üst uçta 4 aykırı değer içermektedir.
- **proanthocyanins**: Üst uçta 2 aykırı değer içermektedir.
- **color_intensity**: Üst uçta 4 aykırı değer içermektedir.
- **hue**: Üst uçta 1 aykırı değer içermektedir.

Aykırı değerler genellikle üst uçta yoğunlaşmıştır. Bu durum, bazı şarapların belirli kimyasal bileşenlerde olağanüstü yüksek değerlere sahip olduğunu göstermektedir. Aykırı değerler gerçek ölçümler olduğu için çıkarılmayacak, ancak robust ölçeklendirme uygulanacaktır.

5. Veri Ölçeklendirme (Scaling)

Keşifsel veri analizi sonuçlarına dayanarak, veri ölçeklendirme için RobustScaler yöntemi seçilmiştir. Boxplot analizinde tespit edilen aykırı değerler ve özellikler arası belirgin ölçek farkları nedeniyle RobustScaler kullanılmıştır. RobustScaler'in dönüşüm formülü:

$$X_{scaled} = (x - median) / IQR$$

Seçilen özellikler için dönüşüm parametreleri:

Özellik	Median	IQR	Ölçeklendirme Formülü
alcohol	13.05	1.32	(alcohol - 13.05) / 1.32
malic_acid	1.87	1.48	(malic_acid - 1.87) / 1.48
ash	2.36	0.35	(ash - 2.36) / 0.35
proline	673.50	484.50	(proline - 673.50) / 484.50

Tüm özellikler benzer ölçeklere getirilmiştir. Dağılımlar korunmuş, sadece ölçek değiştirilmiştir. Ölçeklendirme işlemi tamamlanmış ve veri 178 gözlem × 13 özellik boyutunda korunmuştur. Ölçeklendirilmiş veri (X_scaled), makine öğrenmesi modelleri için uygun hale getirilmiştir.

6. Veri Setinin Bölünmesi

Veri seti, makine öğrenmesi modelleme sürecinde değerlendirme ve test işlemleri için %70'i eğitim seti, %10'u doğrulama seti ve %20'si test setine bölünmüştür. Her bir bölümdeki veri sayısı aşağıdaki tabloda verilmiştir.

Test	Veri Sayısı	Yüzde (%)
Eğitim Seti	124	69.7
Doğrulama Seti	18	10.1
Test Seti	36	20.2

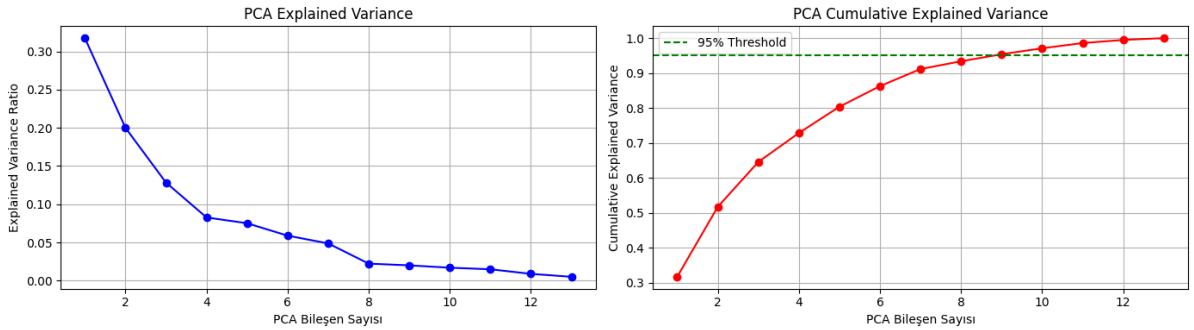
Her bir setteki sınıf dağılımı aşağıdaki tabloda verilmiştir.

Test	Sınıf 0	Sınıf 1	Sınıf 2
Eğitim Seti	41	50	33
Doğrulama Seti	6	7	5
Test Seti	12	14	10

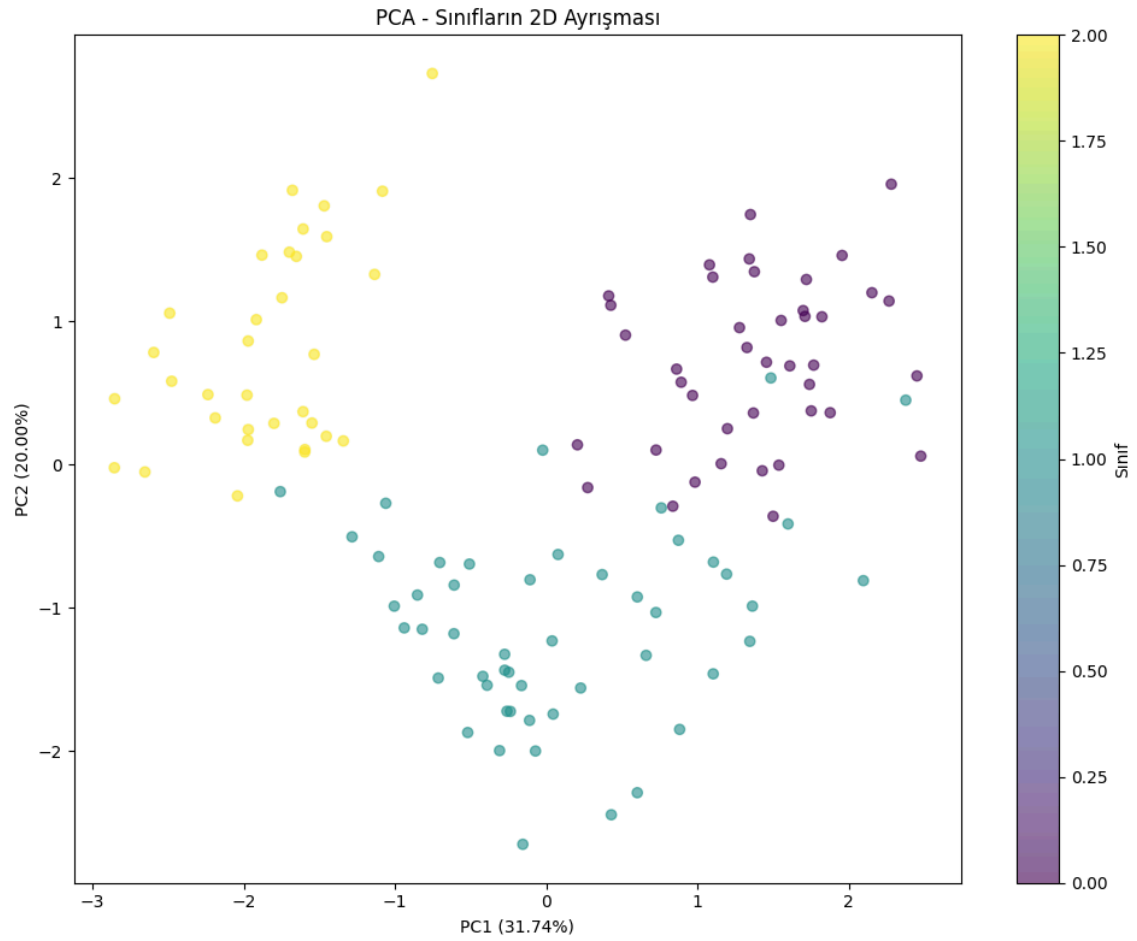
7. Özellik Seçimi ve Boyut İndirgeme

7.1 PCA (Principal Component Analysis)

Toplam bileşen sayısı 13'tür. %95 varyans koruma kriteri ile seçilen bileşen sayısı 9'dur. Seçilen bileşenlere ait grafik aşağıda verilmiştir.



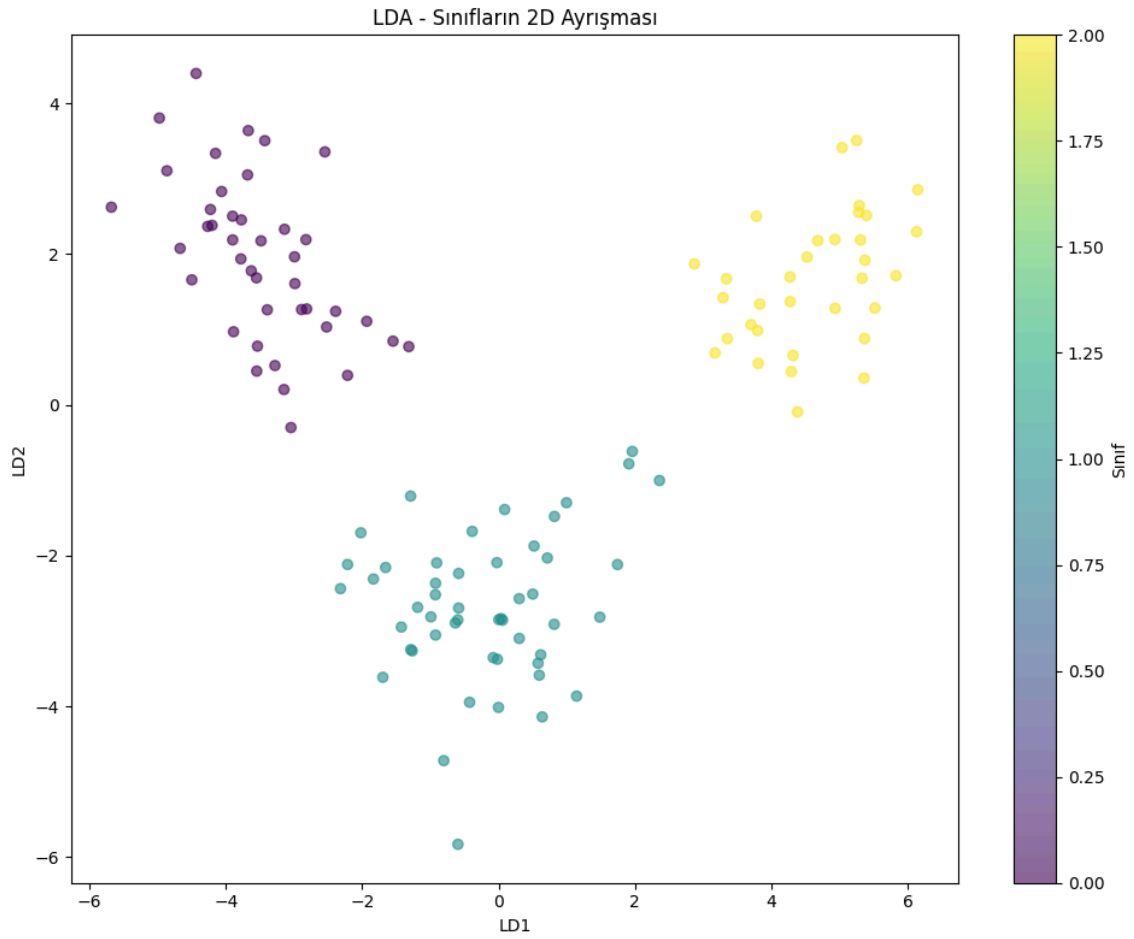
İlk bileşenin varyansı %31.74, ikinci bileşenin varyansı %20.00'dir. Toplam açıklanan varyans %51.74'tür. Seçilen en iyi 2 PCA bileşenin sınıf ayrıştırması aşağıdaki grafikte gösterilmiştir.



Grafiğe göre Sınıf 0 (mor) ve Sınıf 1 (mavi) kısmen örtüşmektedir. Fakat Sınıf 3 diğer sınıflardan belirgin şekilde ayrılmaktadır.

7.2 LDA (Linear Discriminant Analysis)

LDA için n_components parametresi 2 (3 sınıf için maksimum 2 bileşen) seçilmiştir ve eğitilirken sınıf bilgisi kullanılmıştır (supervised learning). LD1, sınıflar arası varyansı maksimize eder. LD2, kalan sınıflar arası varyansı maksimize eder. LDA grafiği aşağıda verilmiştir. Grafiğe göre tüm sınıflar belirgin bir şekilde birbirinden ayrışmıştır.



8. Makine Öğrenmesi Modellerinin Kurulması

Bu bölümde, 5 farklı sınıflandırma algoritması her üç veri temsili üzerinde ayrı ayrı eğitilerek toplam 15 farklı model oluşturulmuştur: 5 algoritma x 3 veri temsili olmak üzere toplam 15 model. Tüm modeller doğrulama seti üzerinde değerlendirilmiştir. Performans metrikleri olarak doğruluk, kesinlik, duyarlılık, f1 skoru ve roc-auc kullanılmıştır. Logistic Regression, Decision Tree, Random Forest, XGBoost ve Gaussian Naive Bayes algoritmaları kullanılmıştır.

9. Validasyon Performanslarının Ölçülmesi

Eğitilen tüm modellere ait sonuçlar aşağıdaki tabloda verilmiştir.

Veri	Algoritma	Doğruluk	Kesinlik	Duyarlılık	F1 Skoru	ROC-AUC
Ham Veri	Logistic Regression	100.00	100.00	100.00	100.00	100.00
Ham Veri	Random Forest	100.00	100.00	100.00	100.00	100.00
Ham Veri	XGBoost	100.00	100.00	100.00	100.00	100.00
Ham Veri	Naive Bayes	100.00	100.00	100.00	100.00	100.00
LDA	Logistic Regression	100.00	100.00	100.00	100.00	100.00

LDA	XGBoost	100.00	100.00	100.00	100.00	100.00
LDA	Decision Tree	94.44	95.24	95.24	94.87	95.83
LDA	Random Forest	94.44	95.24	95.24	94.87	100.00
PCA	Naive Bayes	94.44	95.83	93.33	94.07	98.56
PCA	Logistic Regression	94.44	95.83	93.33	94.07	100.00
LDA	Naive Bayes	94.44	95.83	93.33	94.07	100.00
Ham Veri	Decision Tree	88.89	92.59	87.78	89.10	90.91
PCA	Decision Tree	77.78	78.25	78.25	78.25	82.90
PCA	Random Forest	77.78	76.35	76.35	77.78	95.58
PCA	XGBoost	77.78	76.35	76.35	77.78	93.29

Tabloya göre birden fazla model %100'e yakın performans sergilemiştir. En düşük performansı PCA veri temsili ile XGBoost modeli vermiştir.

10. En İyi Modelin Test Üzerinde Değerlendirilmesi

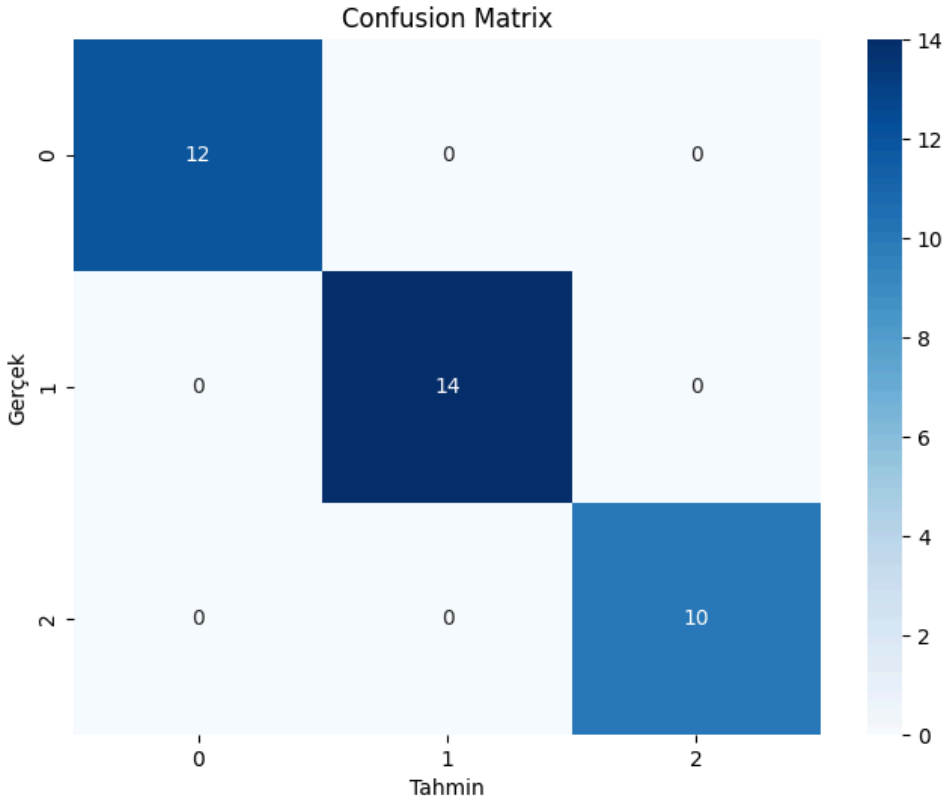
%100'e yakın performans veren modeller arasından Ham Veri ile eğitilen XGBoost modeli seçilmiştir. Bunun nedeni algoritmanın daha güncel ve optimize olmasıdır. Bu aşamada model doğrulama seti ile değerlendirildikten sonra test seti ile değerlendirilecektir.

10.1 Performans Metrikleri

Ham Veri ile eğitilen XGBoost modeli test seti üzerinde %100 doğruluk, %100 kesinlik, %100 duyarlılık, %100 F1 ve %100 ROC-AUC skoru elde etmiştir.

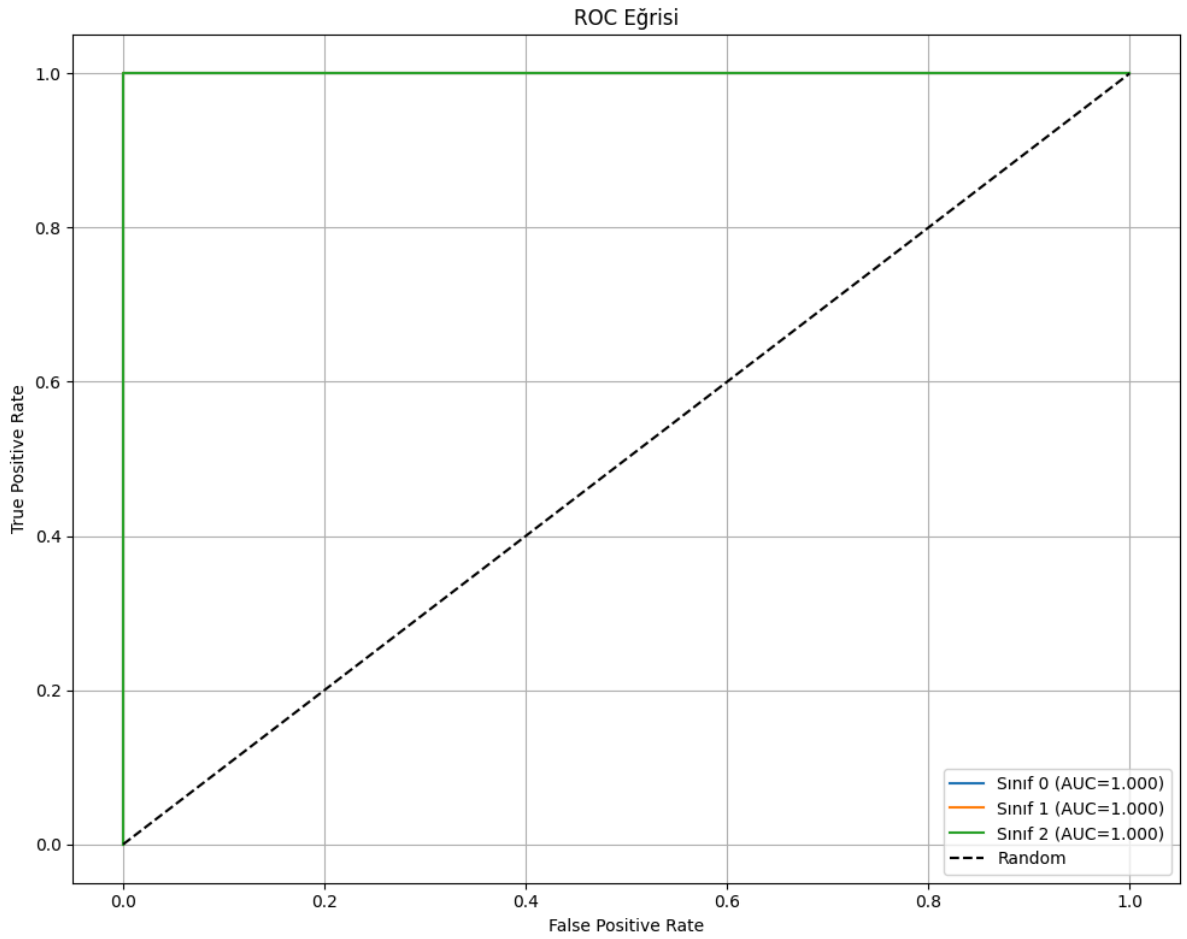
10.2 Confusion Matrix

En iyi modele ait karmaşıklık matrisi aşağıda verilmiştir.

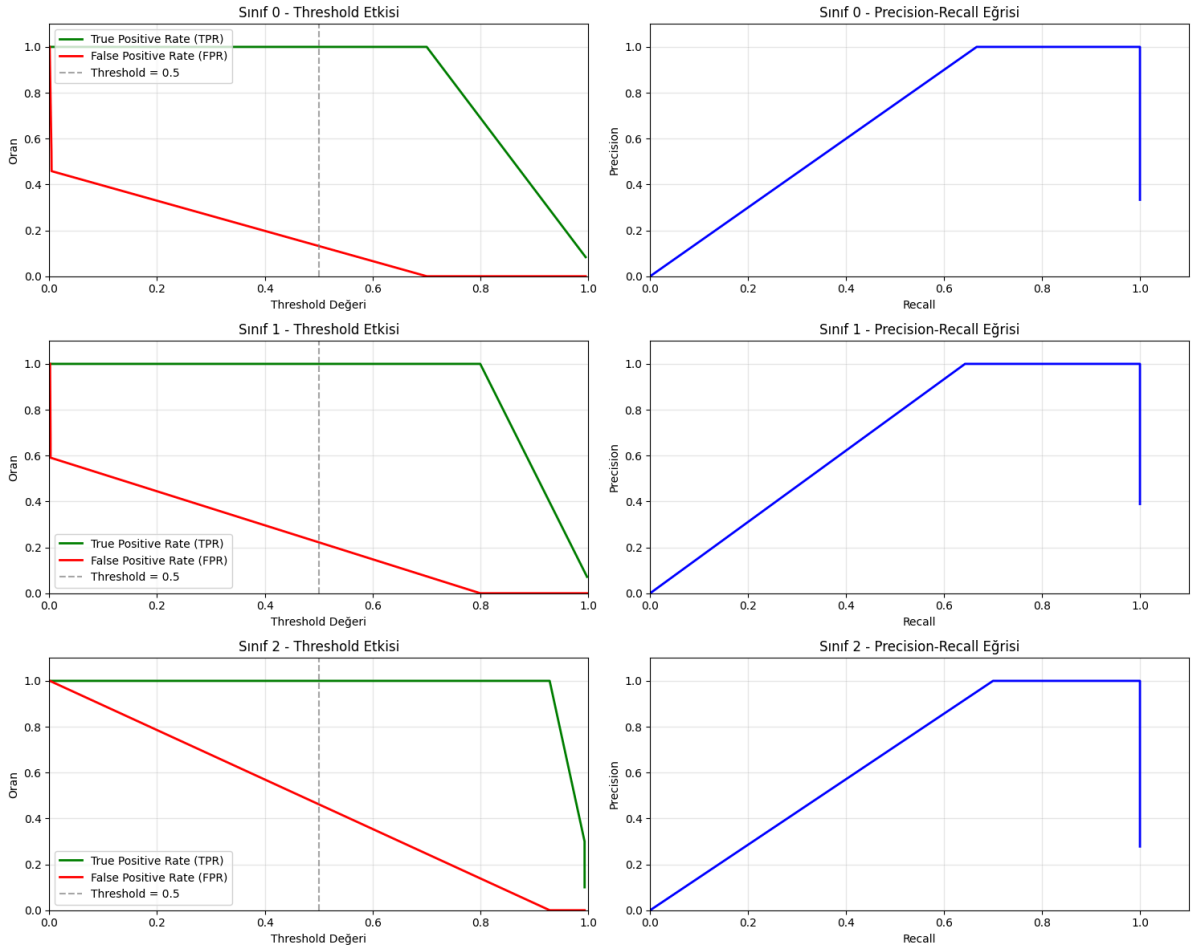


10.3 ROC Eğrisi

En iyi modele ait ROC eğrisi aşağıda verilmiştir. Sınıf bazında hesaplanan AUC skorları da grafikte gösterilmiştir.



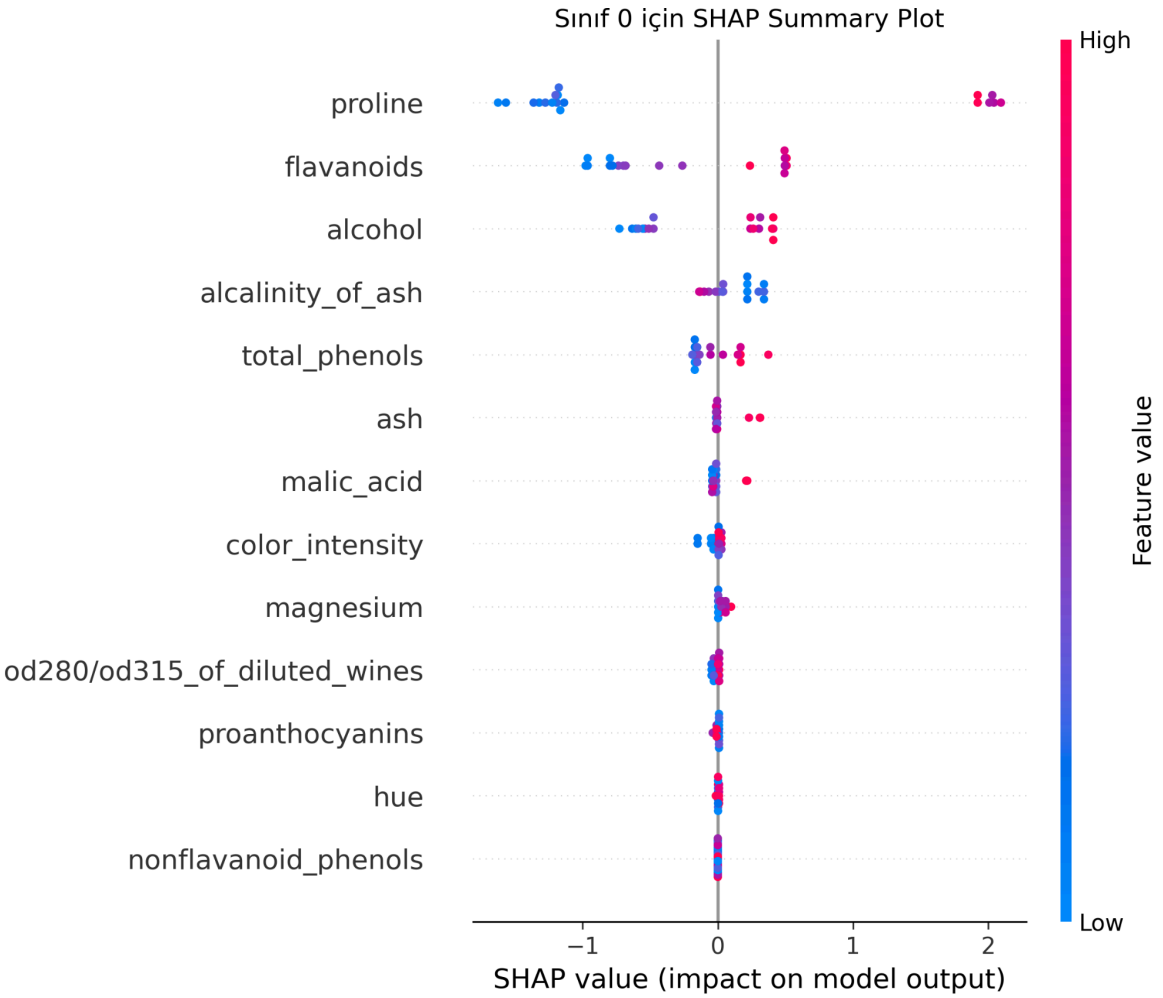
Threshold analizi grafiği aşağıda verilmiştir. Sınıf 0, threshold 0.7, Sınıf 1 için threshold 0.8, Sınıf 2 için threshold 0.9'dur. Threshold arttıkça TPR sabit kalıyor ve FPR düşüyor. Bu, modelin pozitif sınıfı ayırt etme gücünün çok yüksek olduğunu gösterir. Negatifleri yanlış pozitive çevirme eğilimi threshold yükseldikçe tamamen kayboluyor. Daha sıkı eşik bile pozitifleri kaçırmıyor. Negatif yanlışlar azalıyor



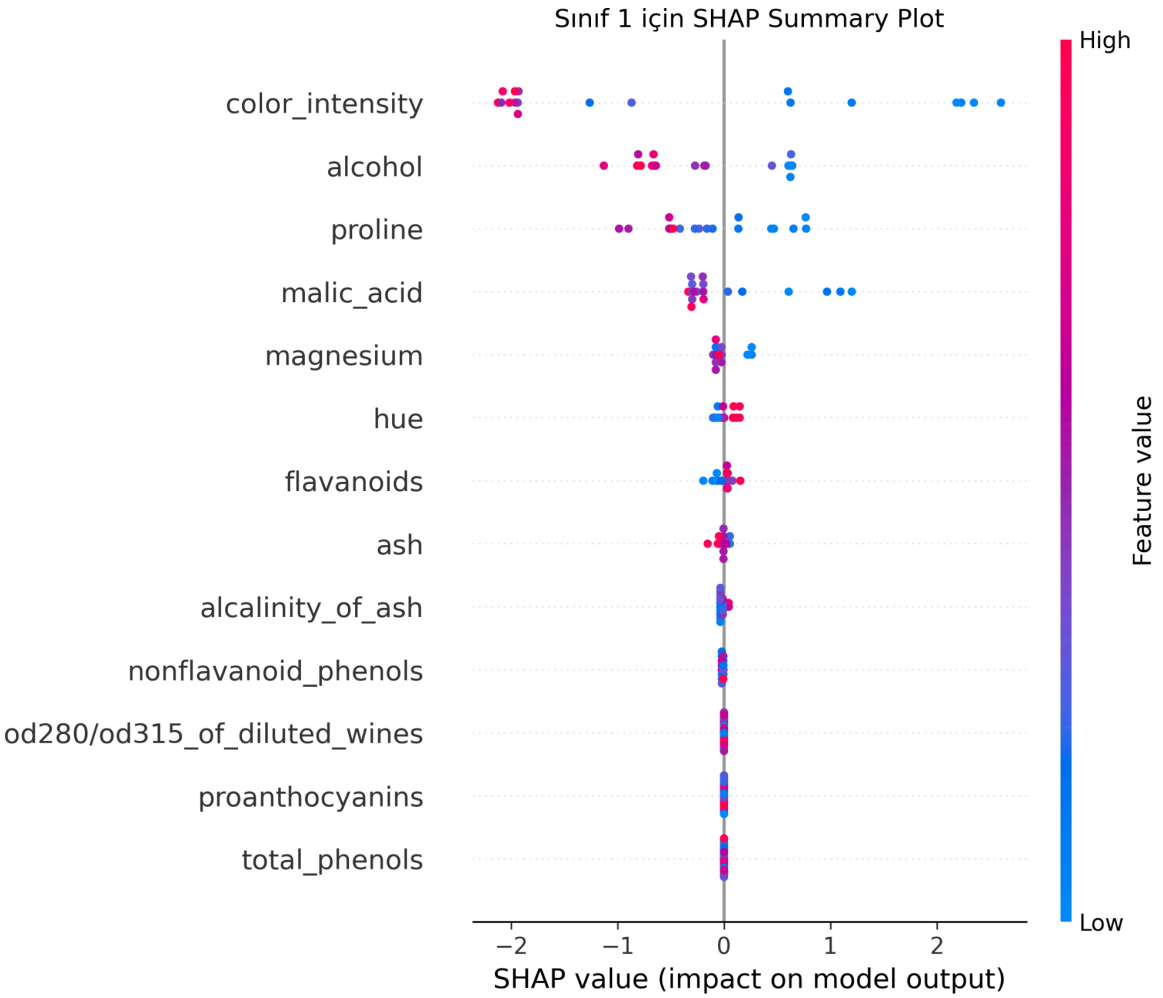
11. XAI - SHAP Açıklanabilirlik Analizi

11.1 En iyi Validasyon Modeli için SHAP Analizi

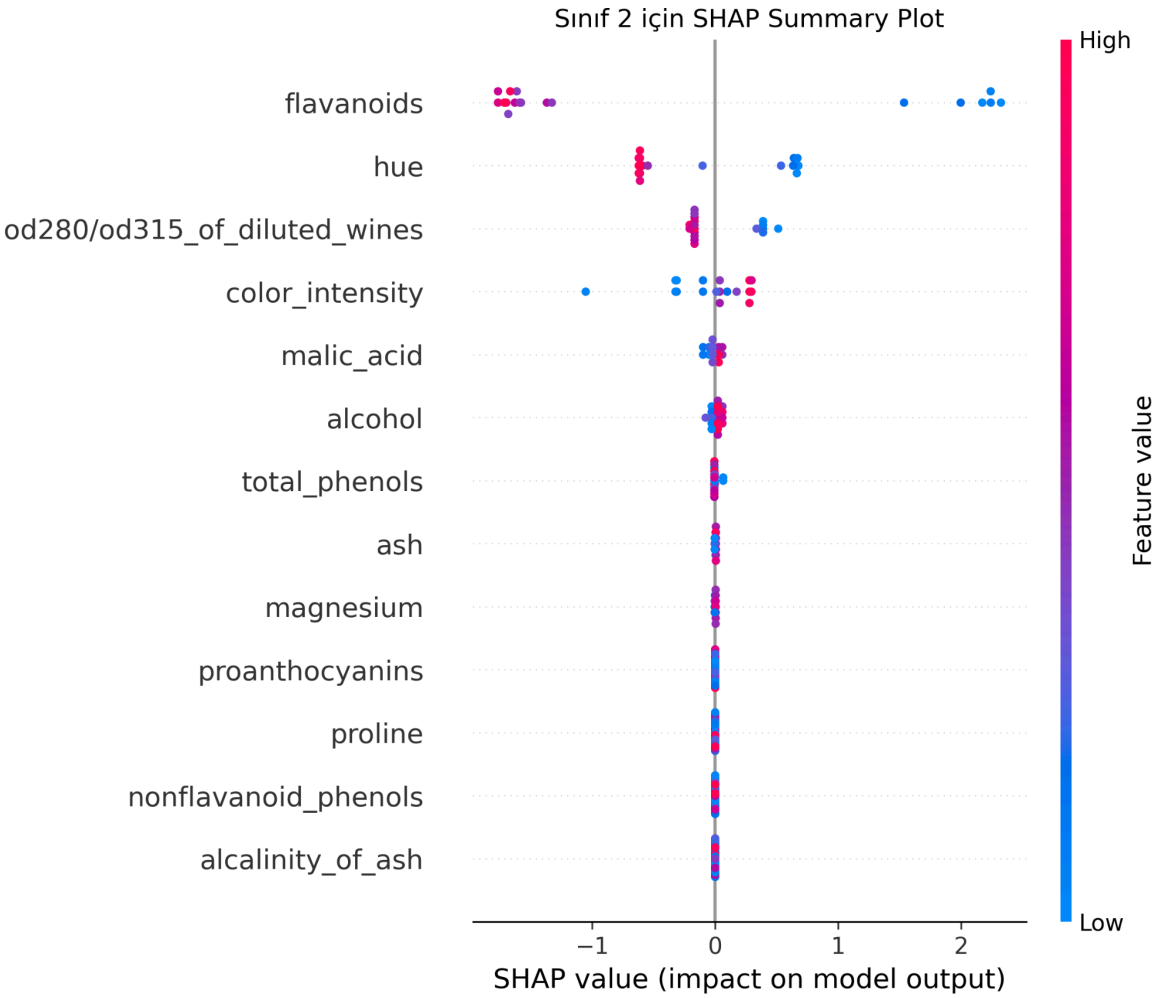
En iyi validasyon modeli için 18 adet shap değeri bulunmaktadır. Bu değerler, 13 özellik ve 3 sınıftan oluşmaktadır. Her bir sınıf için oluşturulan SHAP summary plot aşağıda verilmiştir.



Sınıf 0 için ‘proline’, ‘flavanoids’, ‘alcohol’, ‘alcalinity_of_ash’ ve ‘totel_phenols’ modelin kararını etkileyen en önemli özelliklerdir. Kalan değerler modelin tahminini çok az etkilemektedir. Ayrıca en önemli bu 5 özelliğin pozitif olması, modelin o veriyi ‘Sınıf 0’ olarak değerlendirme ihtimalini artırmaktadır. Yani bir verinin ‘Sınıf 0’ olarak tahmin edilmesi için bu 5 özelliğin 0’dan büyük olması beklenmektedir.

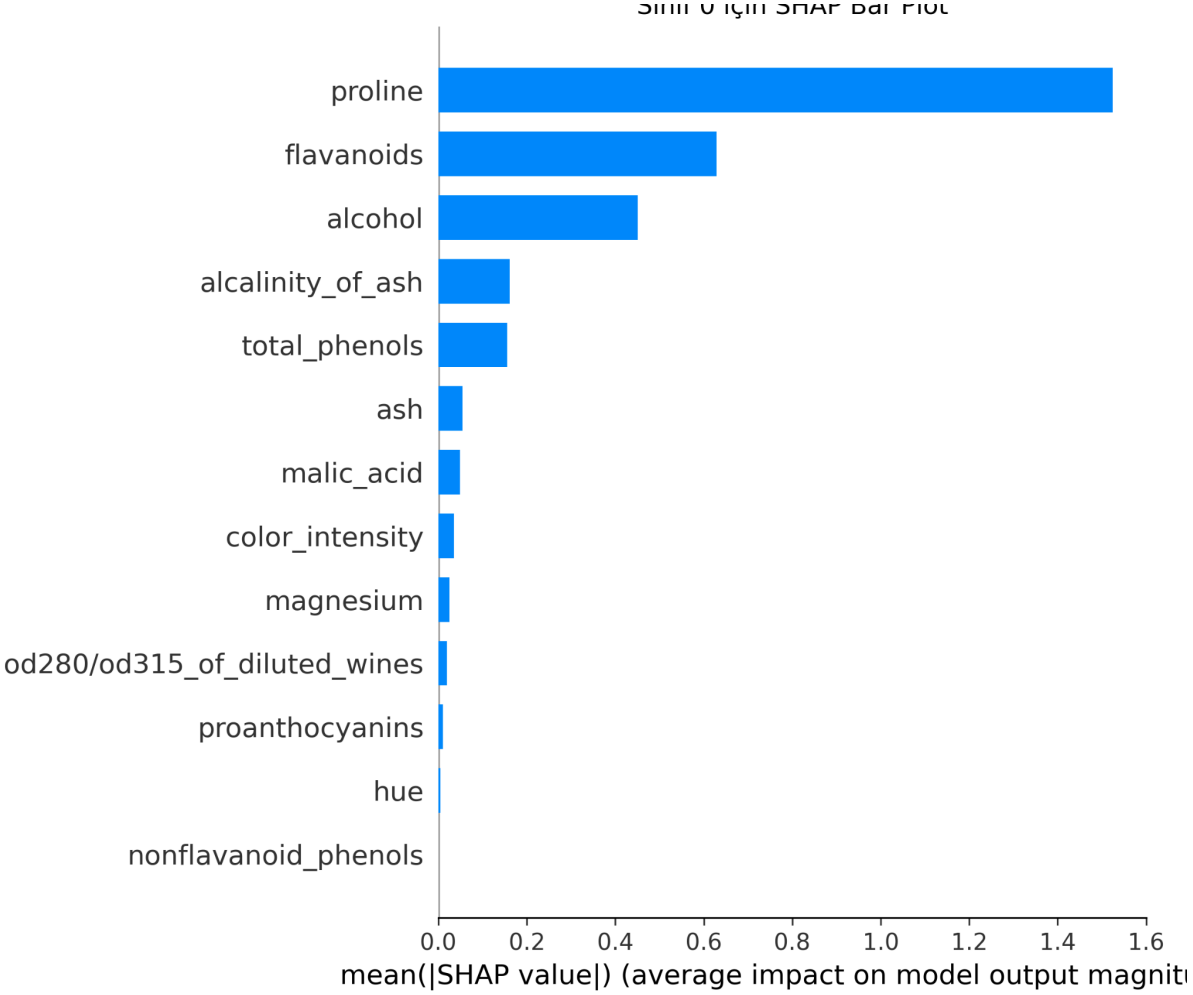


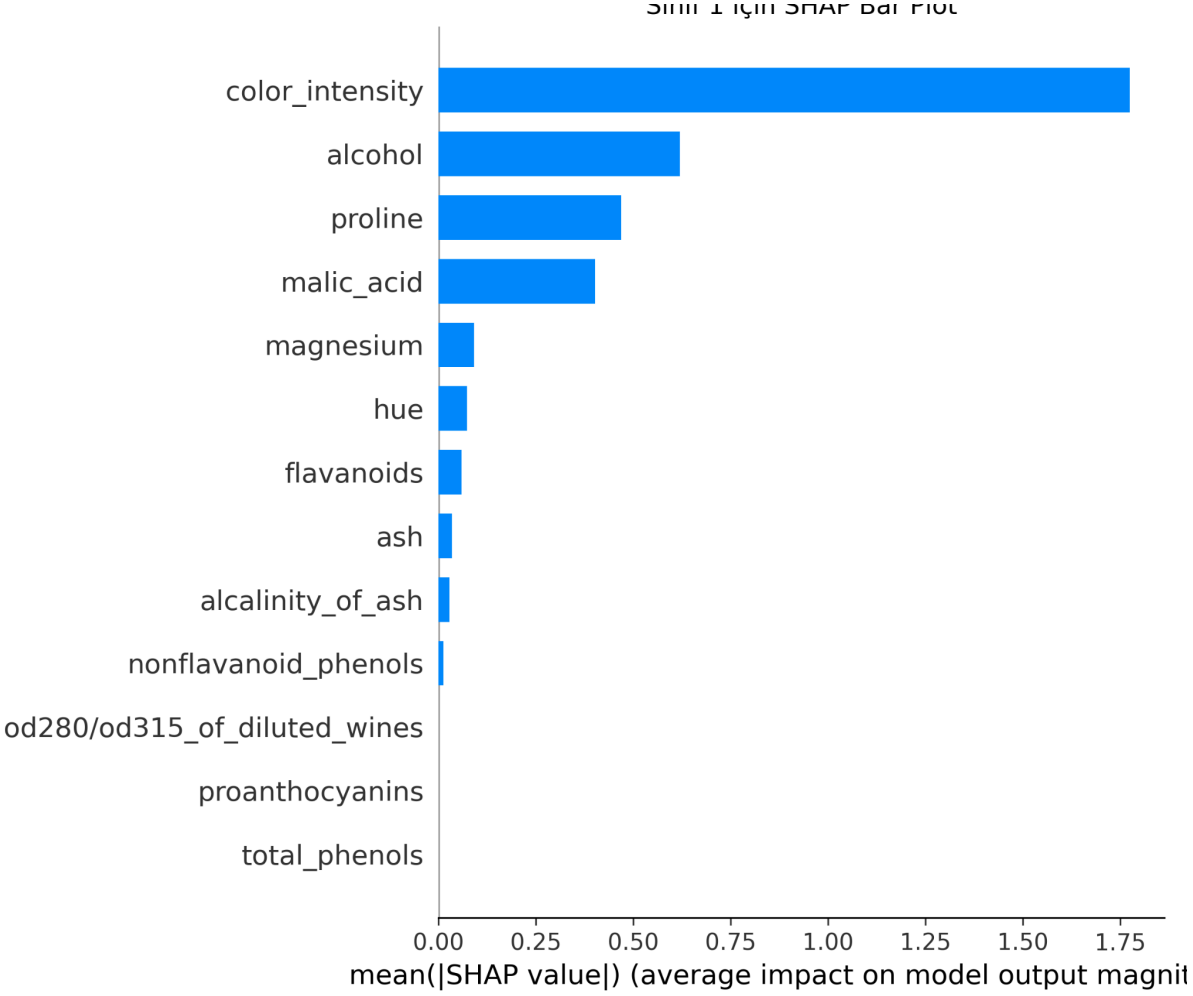
Sınıf 1 için ‘color_intensity’, ‘alcohol’, ‘proline’ ve ‘malic_acid’ modelin kararını etkileyen en önemli özelliklerdir. En önemli bu 4 özelliğin negatif olması, modelin o veriyi ‘Sınıf 1’ olarak değerlendirme ihtimalini artırmaktadır. Sınıf 0’da alcohol ve proline özelliğinin 0’dan büyük olması o sınıfı ‘Sınıf 0’ olarak tahmin etmeyi sağlarken, Sınıf 1’de bu durum tam tersi bir şekildedir.

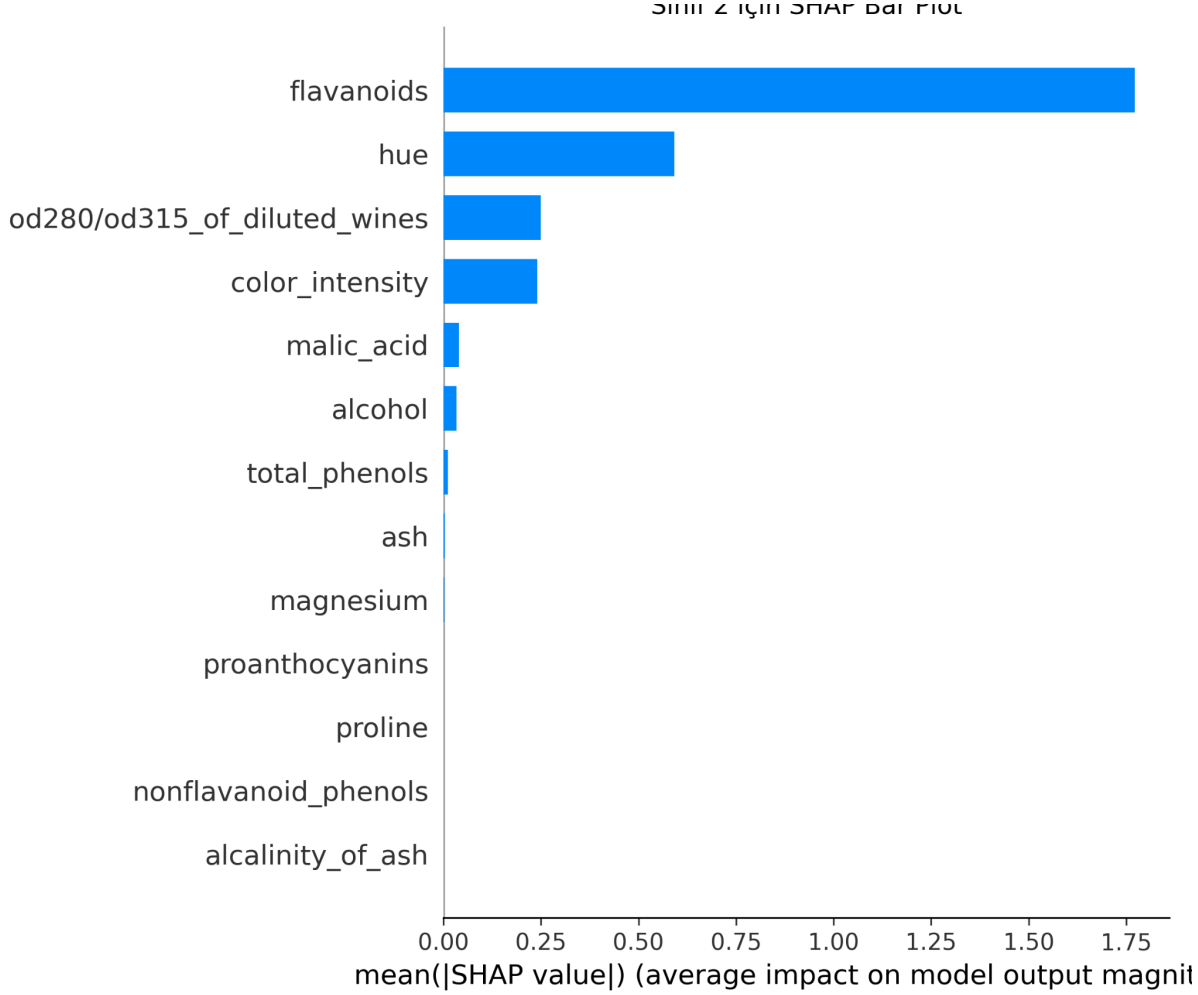


Sınıf 2 için ‘flavanoids’, ‘hue’ ve ‘od280/od315_of_diluted_wines’ modelin kararını etkileyen en önemli özelliklerdir. en önemli bu 3 özelliğin negatif olması, modelin o veriyi ‘Sınıf 2’ olarak değerlendirme ihtimalini azaltmaktadır. Bu 3 özellik Sınıf 0 ve Sınıf 1 için önemli değilken Sınıf 2 için önemlidir.

Her bir sınıf için SHAP Bar Plot aşağıda verilmiştir. Summary plot’da elde edilen en önemli özellikler burada da gösterilmektedir. Sınıf 0 için hue ve nonflavanoid_phenols; Sınıf 1 için od280/od315_of_diluted_wines, proanthocyanins ve total_phenols; Sınıf 2 için ash, magnesium, proanthocyanins, proline, nonflavanoid_phenols, alcalinity_of_ash önemsiz özelliklerdir.


Sınıf 0 için SHAP Bar Plot
proline
flavanoids
alcohol
alcalinity_of_ash
total_phenols
ash
malic_acid
color_intensity
magnesium
od280/od315_of_diluted_wines
proanthocyanins
hue
nonflavanoid_phenols
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
mean(|SHAP value|) (average impact on model output magnitu

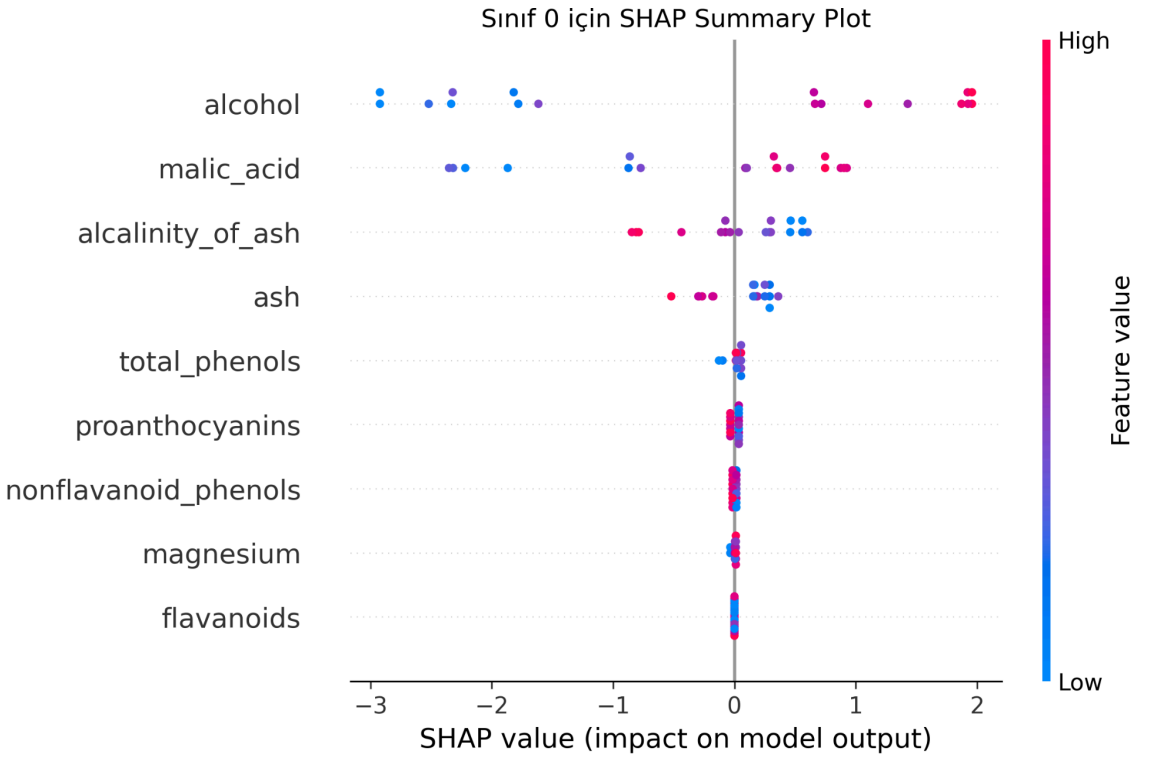

Sınıf 1 için SHAP Bar Plot
color_intensity
alcohol
proline
malic_acid
magnesium
hue
flavanoids
ash
alcalinity_of_ash
nonflavanoid_phenols
od280/od315_of_diluted_wines
proanthocyanins
total_phenols
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
mean(|SHAP value|) (average impact on model output magnit



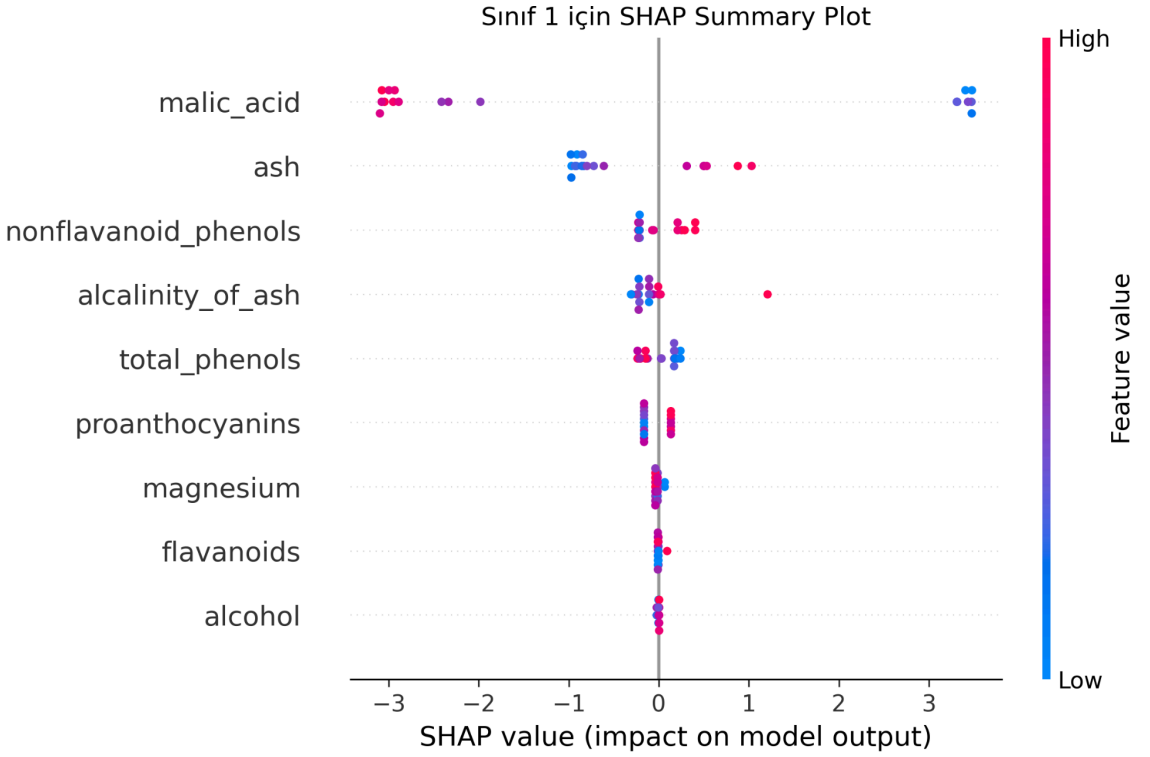
11.2 PCA ve LDA Temsilleri için SHAP Karşılaştırması

11.2.1 PCA için SHAP Karşılaştırması

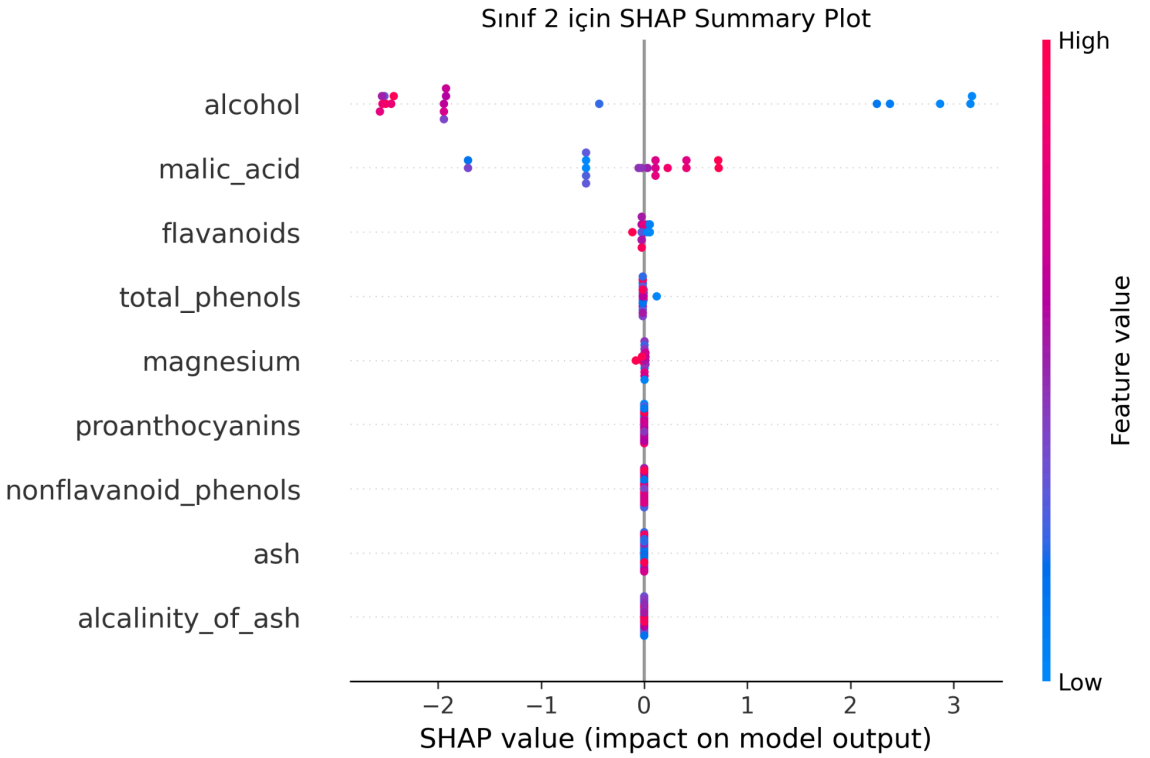
PCA'da en önemli 9 özellik seçilmiştir. PCA ile eğitilmiş XGB modeli için aşağıda SHAP Summary Plot verilmiştir. Sınıf 0 için alcohol, malic_acid, alcalinity_of_ash ve ash en önemli özelliklerdir. alcohol ve malic_acid'in pozitif olması, alcalinity_of_ash ve ash değerinin negatif olması bir verinin Sınıf 0 olarak işaretlenmesini artırmaktadır.



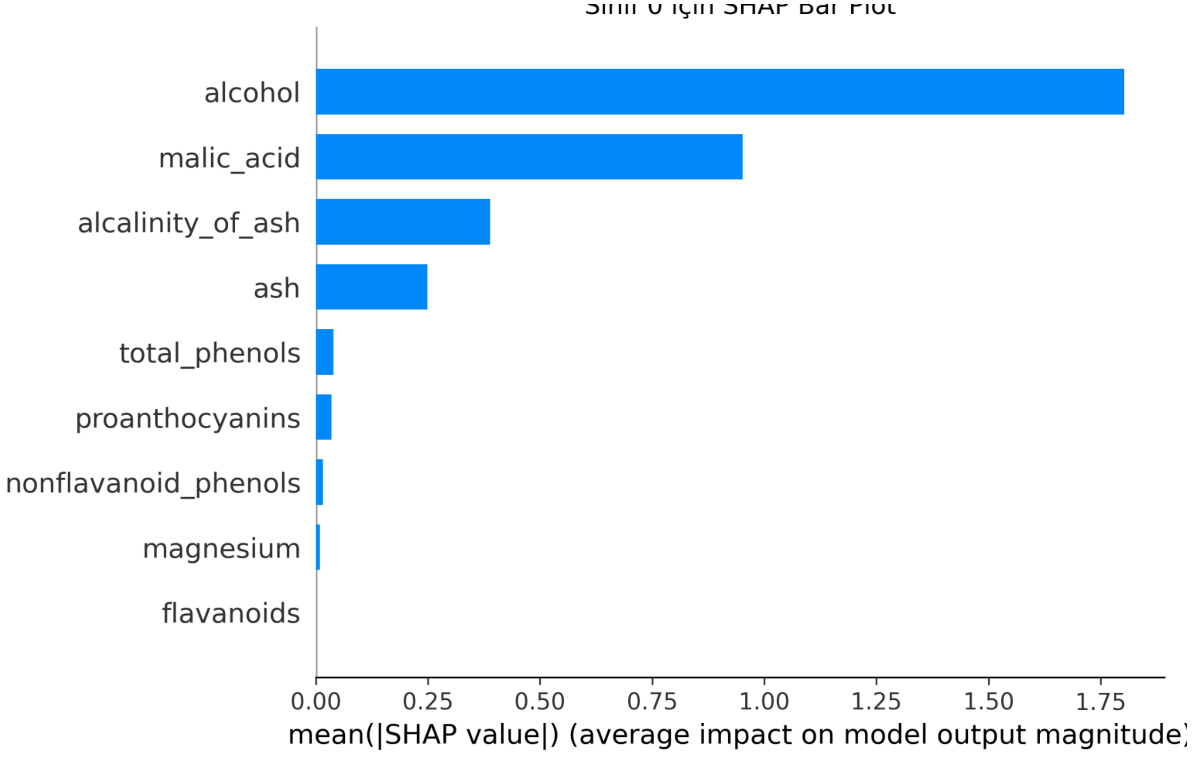
Sınıf 1 için malic_acid, ash ve nonflavanoid_phenols en önemli özelliklerdir. ash ve nonflavanoid_phenols değerinin pozitif olması, malic_acid değerinin negatif olması bir verinin Sınıf 1 olarak işaretlenmesini artırmaktadır.

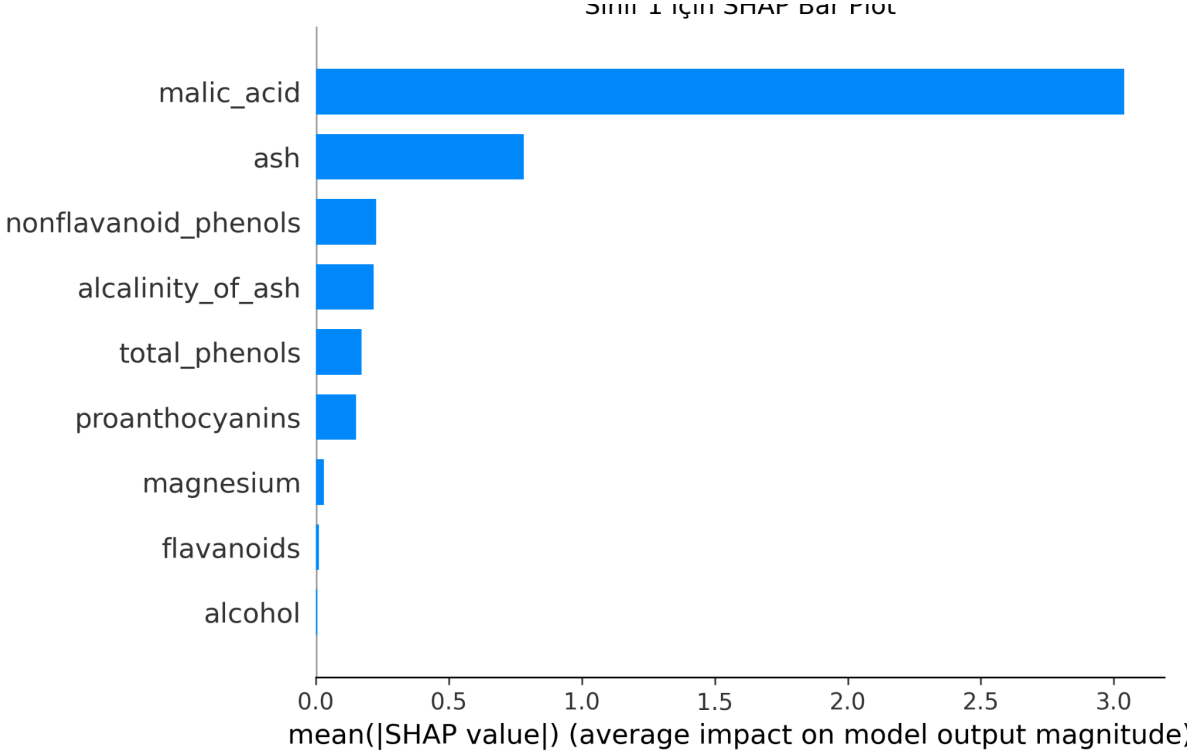


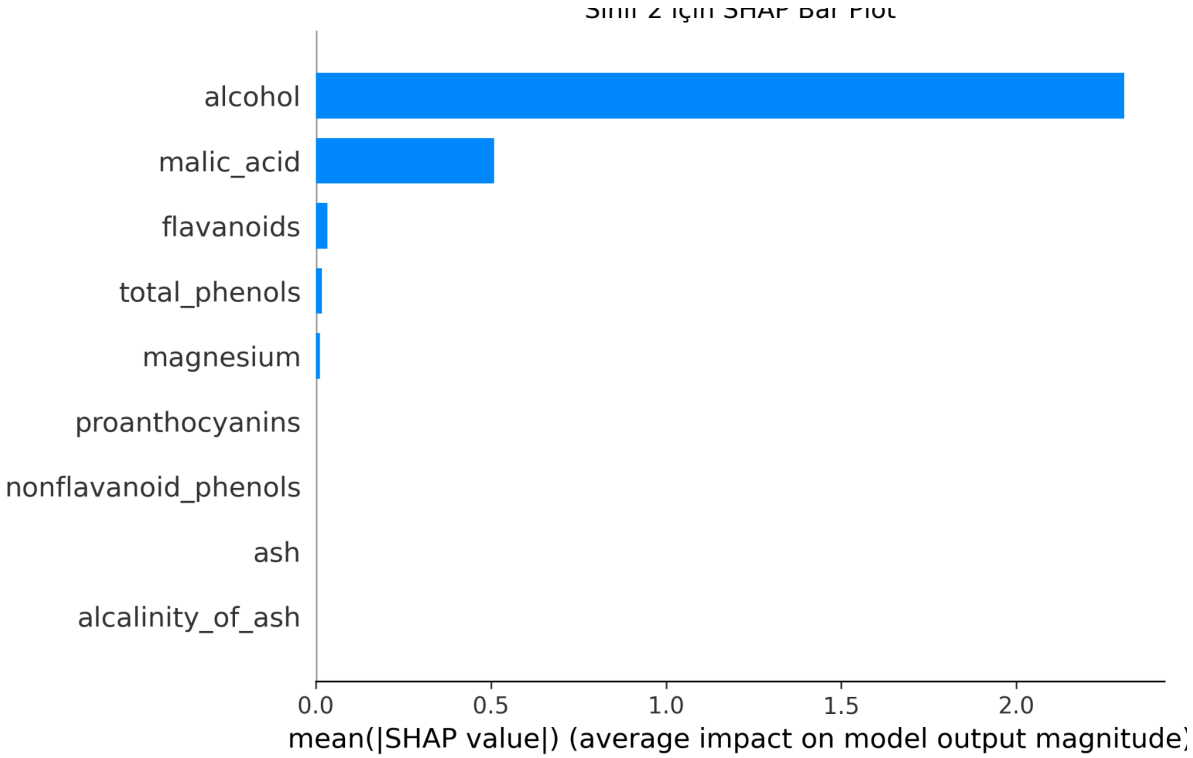
Sınıf 2 için alcohol ve malic_acid en önemli özelliklerdir. alcohol'un negatif ash'in ise pozitif olması bir verinin Sınıf 2 olarak işaretlenmesini artırmaktadır.



Oluşturulan Bar Plot aşağıda verilmiştir.

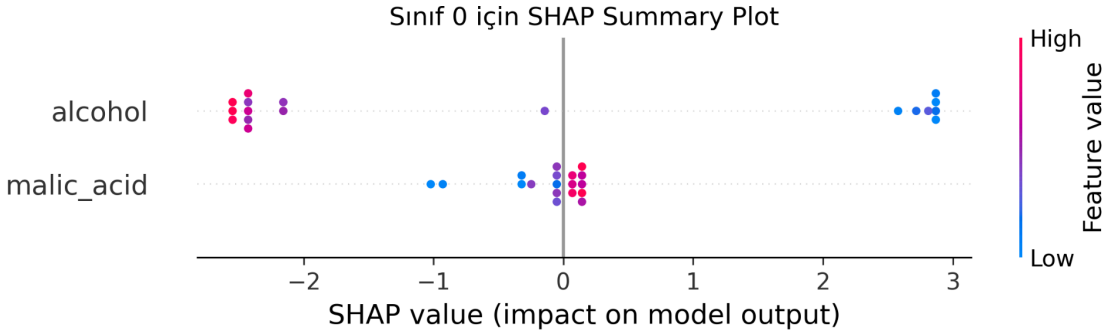


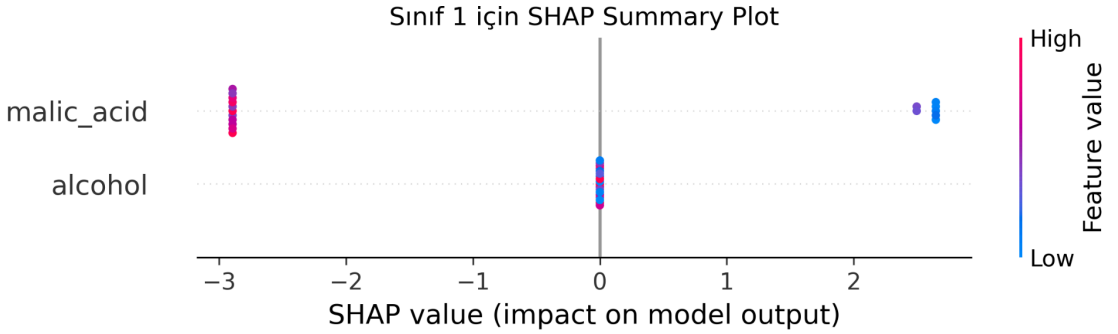


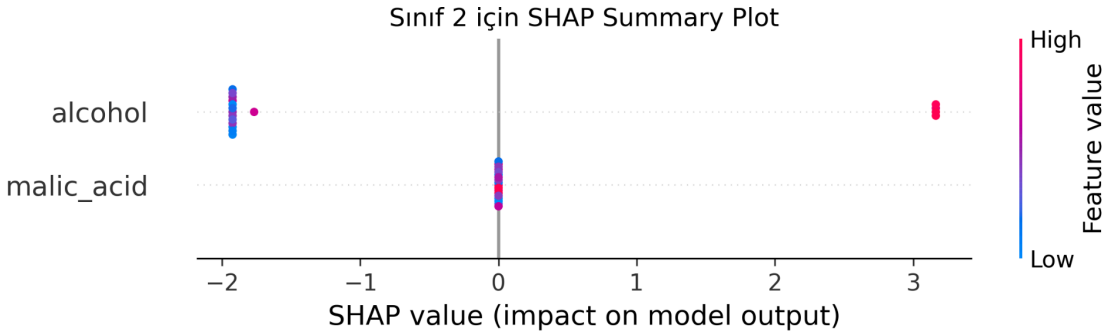


11.2.2 LDA için SHAP Analizi

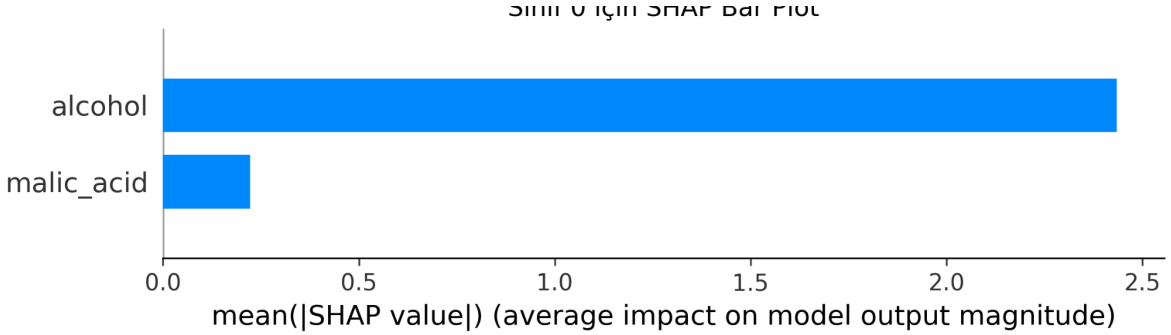
LDA'da en önemli 2 özellik seçilmiştir. Bunlar alcohol ve malic_acid'dir. LDA ile eğitilen modele göre; alcohol'un negatif, malic_acid'in pozitif olması Sınıf 0; alcohol'un 0 malic_acid'in negatif olması Sınıf 1; alcohol'un pozitif, malic_acid'in 0 olması ise Sınıf 2 olarak sınıflandırmaktaıdr. Oluşturulan grafikler aşağıda verilmiştir.

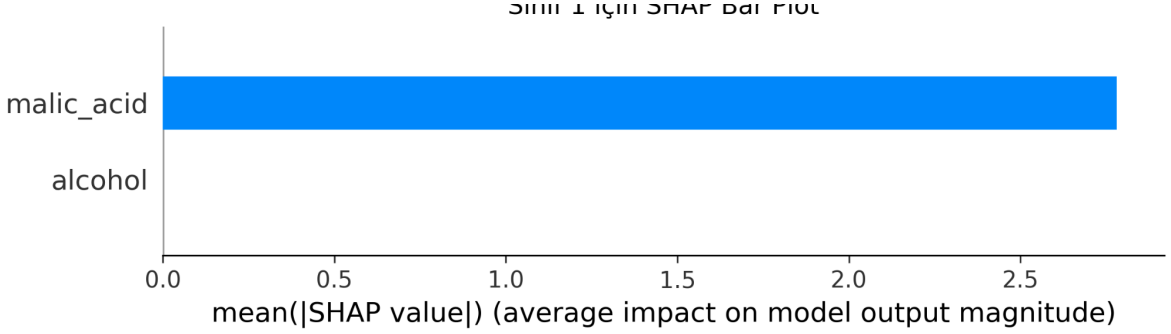


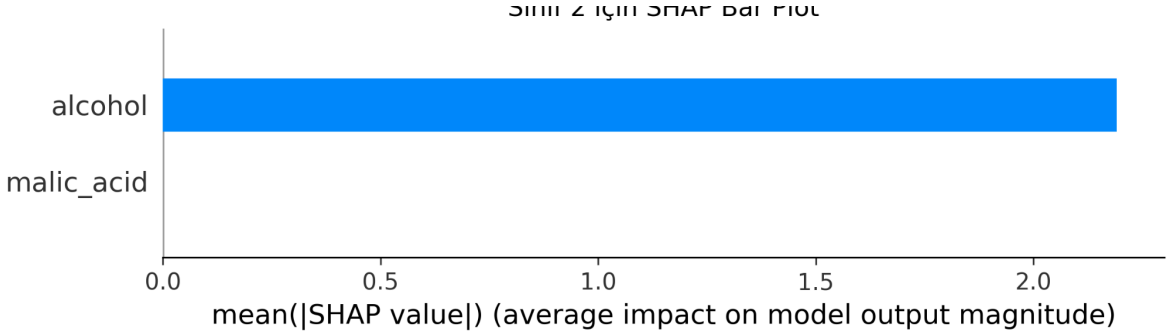




Oluşturulan Bar plotlar aşağıda verilmiştir.







11.2.3 PCA ve LDA Karşılaştırması

Her iki yönteme göre de alcohol ve malic_acid özelliklerinin önemli olduğu görülmektedir. Ancak, LDA en önemli 2 özelliğin alınması modelin karar sınırını çok keskin bir şekilde ayırmaktadır. Buna rağmen LDA ile eğitilen modellerin doğrulama skorları PCA ile eğitilen modellerin doğrulama skorlarından daha yüksektir.