Bölüm: Bilgisayar Mühendisliği (Tezli)

Ders Adı: Makine Öğrenmesi

https://github.com/mabdullahtepeyurt-create/makine-ogrenmesi-vize-mucahit-abdullah-tepeyurt

Hazırlayan:

Mücahit Abdullah TEPEYURT

Öğrenci No: 25221601006

ÖZET

Bu çalışmada, scikit-learn kütüphanesindeki Breast Cancer Wisconsin veri seti kullanılarak klasik makine öğrenmesine dayalı bir sınıflandırma süreci oluşturulmuştur. Veri setindeki eksik ve aykırı değerler Z-score yöntemiyle incelenmiş, ancak müdahale edilmeden yalnızca raporlanmıştır. Tüm özellikler StandardScaler ile ölçeklendirilmiş ve veri, sınıf dengesi korunarak eğitim, doğrulama ve test kümelerine ayrılmıştır.

Ham veri, PCA ve LDA dönüşümleri üzerinde toplam 15 farklı model eğitilmiş; en başarılı model, doğrulama kümesindeki F1 skoru ve ROC-AUC değerlerine göre belirlenmiştir. Seçilen model test kümesinde karmaşa matrisi ve ROC eğrisiyle değerlendirilmiş, ayrıca SHAP yöntemi kullanılarak modelin karar verme süreçleri açıklanmıştır. Son olarak PCA ve LDA sonuçları karşılaştırmalı olarak analiz edilmiştir.

1. Veri Setinin Yüklenmesi

Bu çalışmada scikit-learn kütüphanesinde yer alan Breast Cancer Wisconsin veri seti kullanılmıştır. Veri seti, meme kanseri teşhisinde kullanılan hücresel ölçümleri içeren sayısal özelliklerden ve "kanserli / kanserli değil" şeklinde etiketlenmiş bir hedef sınıftan oluşmaktadır.

Kod ortamında veri seti şu şekilde yüklenmiştir:

- load_breast_cancer() fonksiyonu ile veriler çağrılmış,
- Özellikler **X** değişkenine alınarak pd.DataFrame formatına dönüştürülmüş,
- Hedef etiketleri **y** değişkenine aktarılmış ve pd.Series olarak tanımlanmıştır.

Bu adımlar veri setinin modelleme sürecinde kullanılabilir hâle getirilmesini sağlamıştır.

```
Ozellik matrisi boyutu (X): (569, 30)
Hedef vektörü boyutu (y): (569,)

Sınıf isimleri: ['malignant' 'benign']

İlk 5 satır:
```

	mean radius	mean texture	mean perimeter	mean area	mean smoothness	mean compactness	mean concavity	mean concave points	mean symmetry	mean fractal dimension	...	worst radius	worst texture	worst perimeter	worst area	worst smoothness	worst compactness	worst concavity	worst concave points	worst symmetry	worst fractal dimension
0	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	0.2419	0.07871	...	25.38	17.33	184.60	2019.0	0.1622	0.6656	0.7119	0.2654	0.4601	0.11890
1	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667	...	24.99	23.41	158.80	1956.0	0.1238	0.1866	0.2416	0.1860	0.2750	0.08902
2	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	0.2069	0.05999	...	23.57	25.53	152.50	1709.0	0.1444	0.4245	0.4504	0.2430	0.3613	0.08758
3	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	0.2597	0.09744	...	14.91	26.50	98.87	567.7	0.2098	0.8663	0.6869	0.2575	0.6638	0.17300
4	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	0.1809	0.05883	...	22.54	16.67	152.20	1575.0	0.1374	0.2050	0.4000	0.1625	0.2364	0.07678

5 rows × 30 columns

2.Veri Seti Kalite Kontrolleri

2.1 Eksik Değer Analizi

Veri setindeki eksik değerler, X.isnull().sum() komutu kullanılarak kontrol edilmiştir. Breast Cancer Wisconsin veri seti önceden temizlenmiş bir yapıya sahip olduğu için hiçbir sütunda eksik gözlem bulunmamıştır.

Eksik veri olmaması, ortalama ya da medyan ile doldurma gibi ek imputasyon adımlarına ihtiyaç bırakmamış ve ön işleme sürecini önemli ölçüde sadeleştirmiştir. Bu nedenle veri tamamlama aşamasında herhangi bir müdahale yapılmamıştır.

```
=== Eksik Değer Analizi ===
mean radius                0
mean texture               0
mean perimeter             0
mean area                  0
mean smoothness            0
mean compactness           0
mean concavity             0
mean concave points        0
mean symmetry              0
mean fractal dimension     0
radius error               0
texture error              0
perimeter error            0
area error                 0
smoothness error           0
compactness error          0
concavity error            0
concave points error       0
symmetry error             0
fractal dimension error    0
worst radius               0
worst texture              0
worst perimeter            0
worst area                 0
worst smoothness           0
worst compactness          0
worst concavity            0
worst concave points       0
worst symmetry             0
worst fractal dimension    0
dtype: int64

Bu veri setinde eksik değer yok.
```

2.2 Aykırı Değer Analizi

Aykırı değerler Z-score yöntemiyle incelenmiştir. zscore(X) fonksiyonu kullanılarak her gözlem için standartlaştırılmış skorlar hesaplanmış ve $|Z| > 3$ koşulunu sağlayan noktalar potansiyel aykırılıklar olarak işaretlenmiştir.

Bu çalışmada aykırı değerler veri setinden çıkarılmamış, yalnızca sayıları raporlanmıştır. Bunun iki temel nedeni vardır:

- Decision Tree, Random Forest ve XGBoost gibi ağaç tabanlı modellerin aykırı değerlere karşı dayanıklı olması,
- Aykırı değerlerin gerçek dünyadaki uç durumları temsil ederek modele ek bilgi sağlayabilmesi.

Bu nedenle aykırı gözlemlerin korunmasının modeli olumsuz etkilemeyeceği, hatta faydalı sinyaller sunabileceği değerlendirilmiştir. Analiz sonucunda veri setinde 211 aykırı gözlem bulunduğu tespit edilmiştir.

2.3 Veri Tipi ve Dağılım İncelemesi

Bu aşamada veri setindeki tüm sütunların veri tipleri X.dtypes komutu ile kontrol edilmiş ve tamamının sayısal (float) türünde olduğu doğrulanmıştır. Ayrıca select_dtypes fonksiyonu kullanılarak veri setinde herhangi bir kategorik değişken bulunmadığı da teyit edilmiştir.

Kategorik özellik olmaması, ek bir encoding veya veri tipi dönüştürme ihtiyacını ortadan kaldırmış; böylece PCA ve LDA gibi tamamen sayısal veri üzerinde çalışan dönüşümlerin doğrudan uygulanmasını kolaylaştırmıştır. Bu durum, analiz sürecinin daha sade bir yapıda ilerlemesine katkı sağlamıştır.

```
mean radius                float64
mean texture               float64
mean perimeter             float64
mean area                  float64
mean smoothness            float64
mean compactness           float64
mean concavity             float64
mean concave points        float64
mean symmetry              float64
mean fractal dimension     float64
radius error               float64
texture error              float64
perimeter error            float64
area error                 float64
smoothness error           float64
compactness error          float64
concavity error            float64
concave points error       float64
symmetry error             float64
fractal dimension error    float64
worst radius               float64
worst texture              float64
worst perimeter            float64
worst area                 float64
worst smoothness           float64
worst compactness          float64
worst concavity            float64
worst concave points       float64
worst symmetry             float64
worst fractal dimension    float64
dtype: object
```

3. Keşifsel Veri Analizi (EDA)

3.1 İstatistiksel Özellik Özeti

Veri setindeki her bir özellik için temel istatistiksel ölçüler (ortalama, medyan, minimum–maksimum değerler, standart sapma ve çeyreklikler) hesaplanmıştır. Bu ölçüler, değişkenlerin tipik değer aralıklarını, dağılım yapılarını ve verinin ne kadar değişkenlik gösterdiğini anlamak açısından kritik bir ilk bakış sunmaktadır.

Elde edilen istatistikler, birçok özelliğin sağa çarpık bir dağılıma sahip olduğunu göstermiştir. Özellikle maksimum değerlerin ortalama ve medyanın belirgin şekilde üzerinde olması, veri setinde uç noktaların bulunduğunu ve uzun bir sağ kuyruk yapısının oluştuğunu işaret etmektedir. Bu durum, daha önce yapılan aykırı değer analizini de desteklemektedir.

Tıbbi veri setleri için bu tür uç değerlerin tamamen "gürültü" olarak görülmemesi önemlidir; çünkü bu değerler çoğu zaman gerçek hayatta karşılaşılan nadir fakat klinik açıdan anlamlı durumları temsil eder. Dolayısıyla, hem dağılımın çarpıklığı hem de maksimum değerlerin merkezi eğilimden uzaklığı, modelin bu uç örneklerden öğrenme potansiyeline sahip olduğunu göstermektedir.

```
== EDA Tablosu (Özet) ===
```

	Mean	Median	Min	Max	Std	Q1	Q3
mean radius	14.127292	13.37000	6.98100	28.1100	3.524049	11.70000	15.7800
mean texture	19.289649	18.84000	9.71000	39.2800	4.301036	16.17000	21.8000
mean perimeter	91.969033	86.24000	43.79000	188.5000	24.298981	75.17000	104.1000
mean area	654.889104	551.10000	143.50000	2501.0000	351.914129	420.30000	782.7000
mean smoothness	0.096360	0.09587	0.05263	0.1634	0.014064	0.08637	0.1053

3.2 Korelasyon Matrisi ve Isı Haritası

Özellikler arasındaki ilişkileri daha iyi görebilmek için önce Pearson korelasyon matrisi çıkarılmıştır:

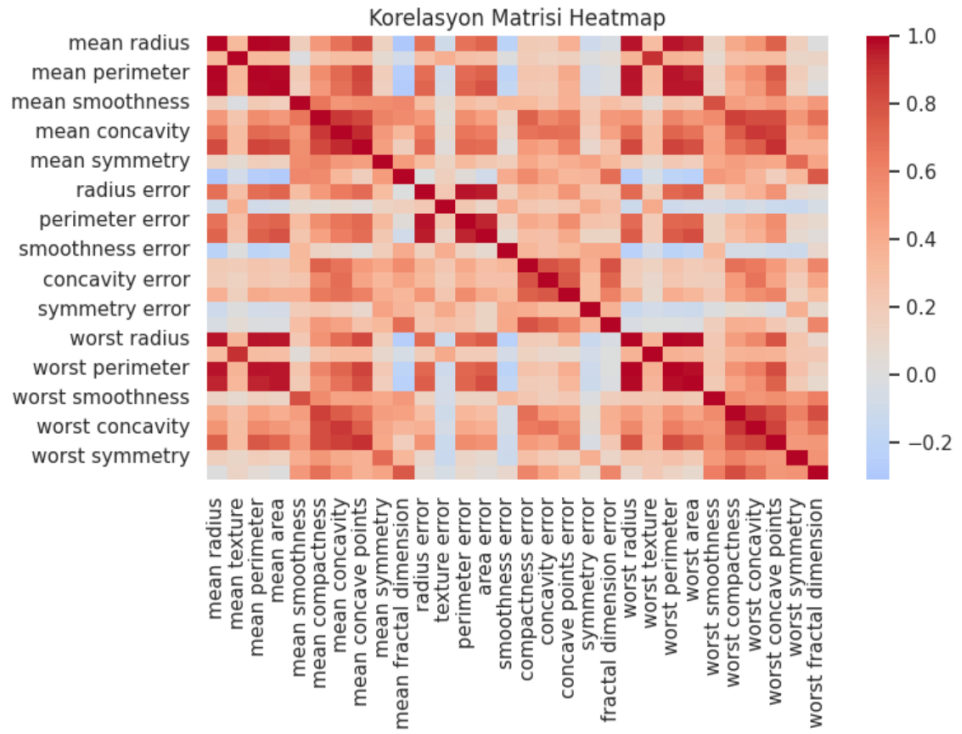
corr_matrix = X.corr(method='pearson')

Ardından, bu korelasyon yapısını görselleştirmek için:

Ardından bu yapı, sns.heatmap() ile ısı haritası şeklinde görselleştirilmiştir. Bu sayede hangi değişkenlerin birlikte artıp azaldığı kolayca fark edilmiştir.

Analiz sonunda özellikle şu durum öne çıkmıştır:

- mean radius, mean perimeter ve mean area gibi meme dokusunun boyutsal ölçümlerini temsil eden özellikler birbirleriyle oldukça yüksek korelasyona sahiptir. Bu, aynı yapısal bilgiyi taşıdıklarını ve benzer şekilde davrandıklarını göstermektedir.
- Ayrıca, en yüksek korelasyona sahip ilk üç değişken çifti ayrı olarak listelenerek ısı haritasında görülen ilişkiler sayısal olarak da doğrulanmıştır.




```
En yüksek korelasyona sahip 3 özellik çift
mean radius    mean perimeter     0.997855
worst radius   worst perimeter    0.993708
mean radius    mean area          0.987357
dtype: float64
```

3.3 Boxplot Analizi

Aykırı değerleri görsel olarak değerlendirmek için hem tüm özellikleri tek grafikte gösteren genel bir boxplot hem de her özellik için ayrı boxplot grafikler oluşturulmuştur.

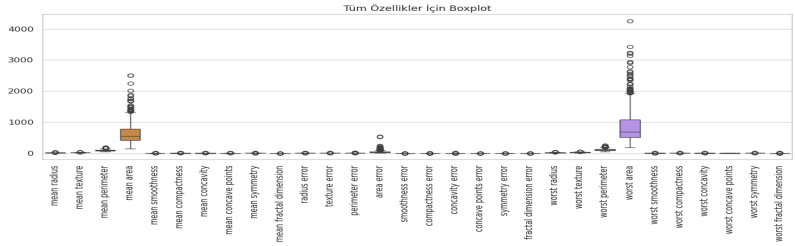
Bu görseller sayesinde:

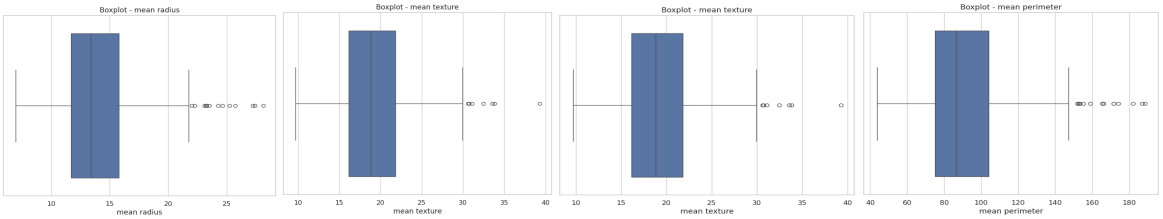
- Bazı özelliklerin sağa çarpık bir dağılıma sahip olduğu,
- Çeyrekler aralığından uzaklaşan olağan dışı yüksek değerlerin bulunduğu

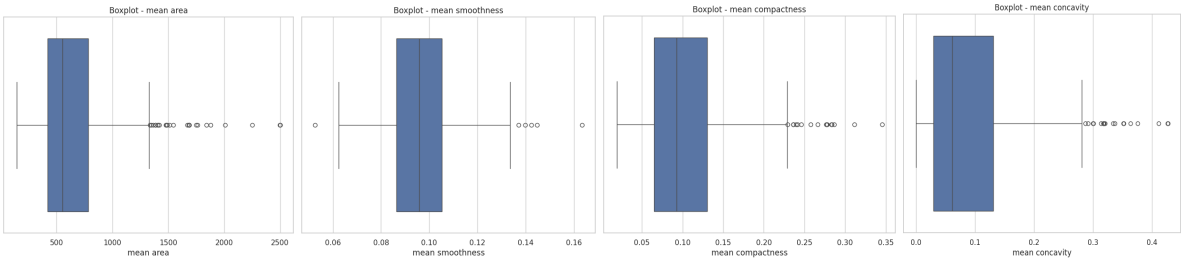
net bir şekilde görülmüştür.

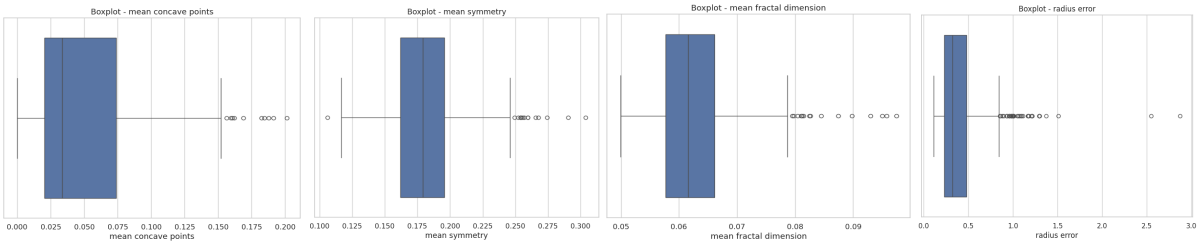
Boxplot sonuçları, daha önce Z-score ile yapılan aykırı değer analizini destekler niteliktedir. Ancak bu uç değerler doğrudan “gürültü” olarak değerlendirilmemiştir; çünkü tıbbi veri setlerinde bu tür uç vakalar modele önemli ve gerçekçi sinyaller sağlayabilir.

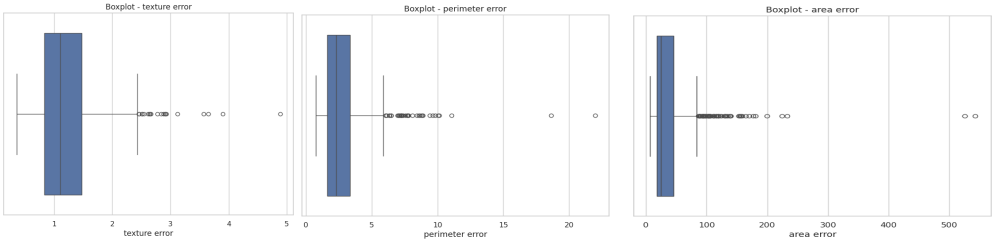
Özetle, boxplot analizi hem aykırı değerlerin varlığını doğrulamış hem de bu değerlerin model performansına potansiyel olarak olumlu katkı yapabileceğini göstermiştir.

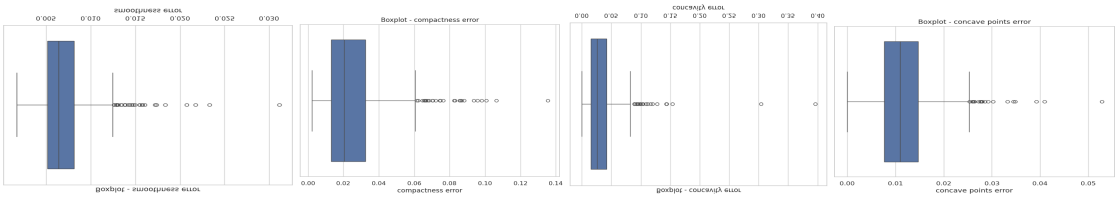

Tüm Özellikler İçin Boxplot
4000
3000
2000
1000
0
mean radius
mean texture
mean perimeter
mean area
mean smoothness
mean compactness
mean concavity
mean concave points
mean symmetry
mean fractal dimension
radius error
texture error
perimeter error
area error
smoothness error
compactness error
concavity error
concave points error
symmetry error
fractal dimension error
worst radius
worst texture
worst perimeter
worst area
worst smoothness
worst compactness
worst concavity
worst concave points
worst symmetry
worst fractal dimension

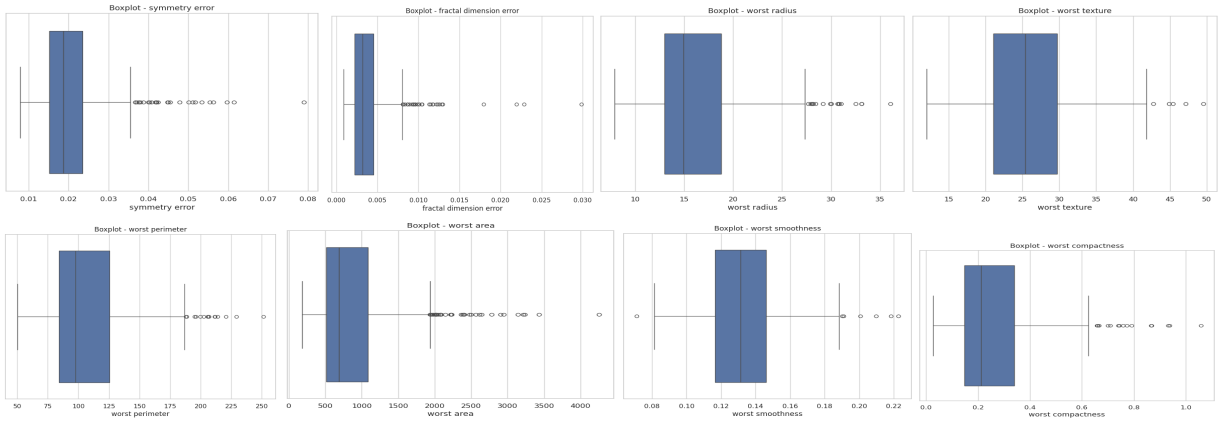

Boxplot - mean radius
mean radius
Boxplot - mean texture
mean texture
Boxplot - mean texture
mean texture
Boxplot - mean perimeter
mean perimeter


Boxplot - mean area
mean area
Boxplot - mean smoothness
mean smoothness
Boxplot - mean compactness
mean compactness
Boxplot - mean concavity
mean concavity


Boxplot - mean concave points
mean concave points
Boxplot - mean symmetry
mean symmetry
Boxplot - mean fractal dimension
mean fractal dimension
Boxplot - radius error
radius error


Boxplot - texture error
texture error
Boxplot - perimeter error
perimeter error
Boxplot - area error
area error


Boxplot - compactness error
compactness error
concave points error



4. Veri Ölçeklendirme

Tüm özellikler StandardScaler ile standartlaştırılmıştır (ortalama 0, standart sapma 1).

Bu sayede:

- PCA ve LDA gibi dönüşümler, büyük değerli değişkenlerden etkilenmez,
- Logistic Regression gibi optimizasyon temelli modeller daha hızlı ve stabil öğrenir,
- Mesafe veya değer büyüklüğüne duyarlı modellerde tüm özellikler eşit katkı sağlar.

5. Eğitim / Doğrulama / Test Bölünmesi

Veri seti %70 eğitim, %10 doğrulama, %20 test olarak bölünmüştür.

Bölme işlemi stratified sampling ile yapıldığından, her alt kümede kanser ve kanser olmayan örneklerin oranı orijinal veri setiyle uyumludur. Bu sayede:

- Model dengeli bir veriyle eğitilmiş olur,
- Doğrulama ve test metrikleri güvenilir olur.

Alt kümelerin boyutları:

- Eğitim: (398, 30)
- Doğrulama: (57, 30)
- Test: (114, 30)

6. Özellik Seçimi ve Boyut İndirgeme

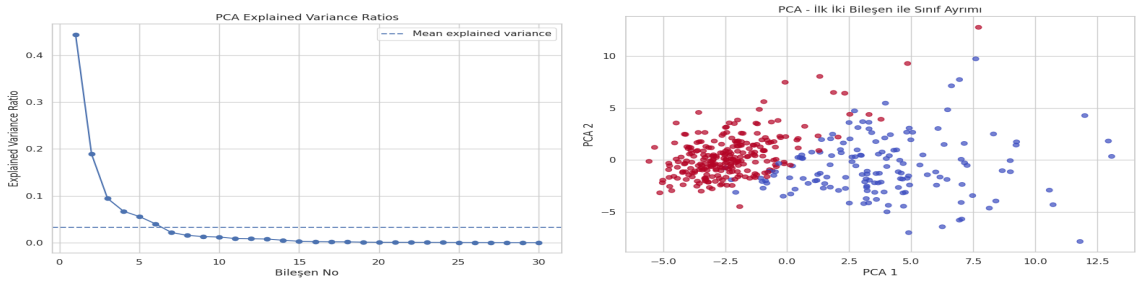
Çalışmada üç farklı veri temsili kullanılmıştır:

- Ham veri: Sadece standartlaştırılmış hali,
- PCA: Verideki toplam varyansı en iyi açıklayan indirgenmiş bileşenler,
- LDA: Kanser ve kanser olmayan sınıflar arasındaki farkı maksimize eden doğrusal bileşenler.

6.1. PCA Özeti

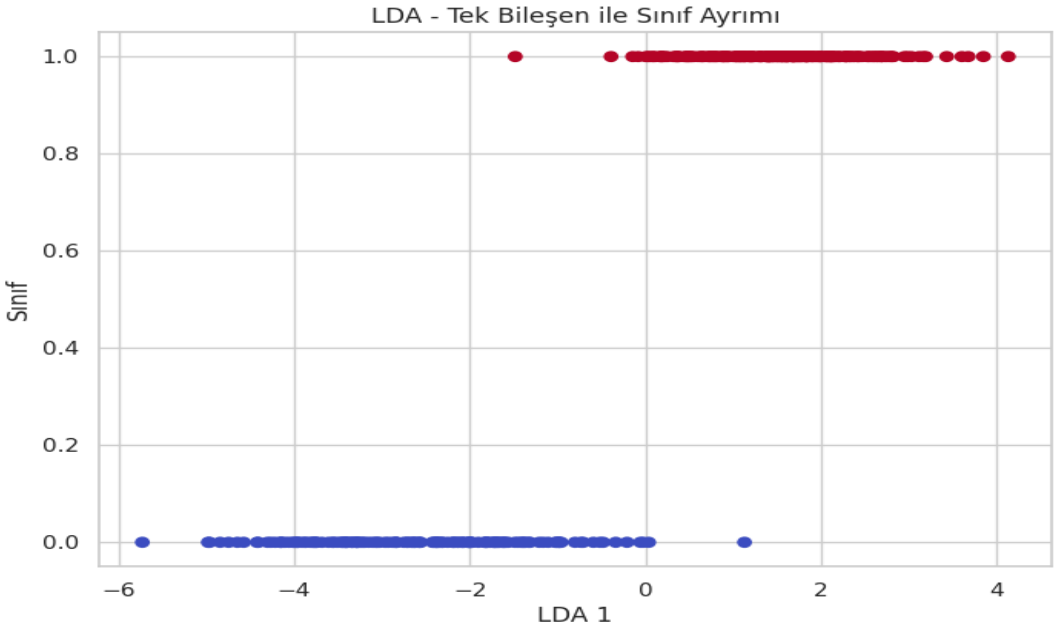
PCA ile verinin boyutu indirgenmiş ve açıklanan varyans oranları hesaplanmıştır. Önemli bileşenler, ortalama varyanstan yüksek katkı sağlayanlar seçilerek belirlenmiştir. İlk birkaç bileşen toplam varyansın büyük kısmını kapsadığı için veri etkin şekilde sıkıştırılmıştır.

2D scatter plot (PC1–PC2) sınıfların net şekilde ayrıştığını göstererek, PCA'nın gözetimsiz olmasına rağmen sınıflandırma için güçlü bir sinyal sağladığını ortaya koymuştur.



6.2 LDA (Doğrusal Ayrım Analizi)

LDA, sınıf etiketlerini kullanarak veriyi en ayrıştırıcı doğrusal bileşenlere indirger. Bu çalışmada tek bir LDA bileşeni ile sınıflar (kanser / kanser değil) net şekilde ayrılmış ve görselleştirilmiştir. Tek bileşen, sınıf ayrımını açıklamak için yeterli olup, LDA'nın sınıflar arası mesafeyi maksimize etme mantığını ortaya koymaktadır.



7. Makine Öğrenmesi Modellerinin Kurulması

Bu çalışmada 5 klasik makine öğrenmesi algoritması kullanılmıştır: Logistic Regression, Decision Tree, Random Forest, XGBoost ve Gaussian Naive Bayes.

Veri, üç farklı temsile dönüştürülmüştür:

- Raw veri (yalnızca standartlaştırılmış),

- PCA ile boyut indirgenmiş veri,
- LDA ile boyut indirgenmiş veri (tek bileşen).

Böylece toplam 15 model oluşturulmuştur (5 algoritma × 3 veri temsili).

Model seçim gerekçeleri:

- Logistic Regression: Veri setinin lineer ayrılabilirliğini test etmek için.
- Decision Tree: Karar süreçlerini yorumlanabilir şekilde modelleyebilme.
- Random Forest & XGBoost: Aykırı değerlere dayanıklı, küçük veri kümelerinde yüksek F1 ve ROC-AUC sağlar.
- Naive Bayes: Baseline karşılaştırma için hızlı ve hafif olasılıksal model.

8. Validation Performanslarının Ölçülmesi

Raw, PCA ve LDA temsilleriyle eğitilen toplam 15 model, doğrulama seti üzerinde test edilmiştir. Amaç, modellerin görmediği veriler üzerindeki güvenilirliğini ve genelleme kapasitesini ölçmektir.

Değerlendirme için kullanılan metrikler şunlardır:

- Accuracy (Doğruluk): Genel sınıflandırma başarısı.
- Precision (Kesinlik): Pozitif tahminlerin ne kadar doğru olduğunu gösterir.
- Recall (Duyarlılık): Gerçekte pozitif olan örneklerin ne kadarının doğru yakalandığını ölçer.
- F1-Score: Precision ve recall dengesini temsil eder; özellikle tıbbi sınıflandırmalarda adil bir ölçü sağlar.
- ROC–AUC: Modellerin farklı eşik değerleri boyunca sınıfları ayırt etme gücünü gösterir.

Sıralama Stratejisi:

1. Öncelikle F1-Score'a göre en yüksekten en düşüğe.
2. Eşit F1-Score durumunda ROC–AUC kriteri kullanılır.

Bu yöntem, sınıf dengesini göz ardı eden yüksek doğruluk skorlarını elemeye ve en güçlü ayrım yeteneğine sahip modelleri öne çıkarmaya olanak tanır.

	representation	model	accuracy	precision	recall	f1	roc_auc
0	Raw	Logistic Regression	0.964912	0.972222	0.972222	0.972222	0.996032
5	PCA	Logistic Regression	0.964912	0.972222	0.972222	0.972222	0.994709
6	PCA	Decision Tree	0.964912	0.972222	0.972222	0.972222	0.962302
8	PCA	XGBoost	0.964912	0.972222	0.972222	0.972222	0.982804
7	PCA	Random Forest	0.964912	0.972222	0.972222	0.972222	0.976190
3	Raw	XGBoost	0.947368	0.945946	0.972222	0.958904	0.986772
2	Raw	Random Forest	0.947368	0.945946	0.972222	0.958904	0.976190
10	LDA	Logistic Regression	0.929825	0.900000	1.000000	0.947368	0.957672
14	LDA	Naive Bayes	0.929825	0.900000	1.000000	0.947368	0.957672
13	LDA	XGBoost	0.929825	0.921053	0.972222	0.945946	0.958995
1	Raw	Decision Tree	0.929825	0.944444	0.944444	0.944444	0.924603
4	Raw	Naive Bayes	0.929825	0.944444	0.944444	0.944444	0.977513
9	PCA	Naive Bayes	0.912281	0.918919	0.944444	0.931507	0.973545
12	LDA	Random Forest	0.894737	0.875000	0.972222	0.921053	0.968254
11	LDA	Decision Tree	0.894737	0.875000	0.972222	0.921053	0.867063

9. En İyi Modelin Test Seti Üzerinde Değerlendirilmesi

Doğrulama seti üzerindeki kapsamlı analizlerin ardından, modeller F1-Score ve ROC-AUC kriterlerine göre sıralanmıştır. Bu yaklaşım yalnızca yüksek doğruluk sağlamayı değil, aynı zamanda sınıf dengesi korunması, yanlış negatiflerin minimize edilmesi ve güvenilir karar sınırları oluşturulmasını da hedefler.

Bu değerlendirme sonucunda, PCA ile işlenmiş veride XGBoost modeli, en dengeli ve ayırt edici performansı gösterdiği için en iyi model olarak seçilmiştir.

Test seti ise eğitim sırasında hiç kullanılmadığından, elde edilen sonuçlar modelin gerçek genelleme yeteneğini ve görülmemiş veri üzerindeki güvenilirliğini doğrudan yansıtmaktadır.

9.1 Test Performans Metrikleri

PCA ile temsil edilen veri üzerinde Logistic Regression modeli, tamamen bağımsız test setinde yaklaşık %96.5 doğruluk göstermiş ve 0.9649 Accuracy ile güçlü bir sınıflandırma performansı sergilemiştir.

Modelin Precision skoru 0.9722 (~%97.2) olup, yanlış pozitif (false positive) tahminlerin çok az olduğunu göstermektedir. Yani, modelin "malign" olarak sınıflandırdığı örneklerin büyük çoğunluğu gerçekten malignd ir; pozitif tahminler güvenilirdir.

Recall skoru da 0.9722 (~%97.2) olup, modelin gerçek pozitif vakaları yakalama duyarlılığının çok yüksek olduğunu doğrular. Bu, özellikle kritik tıbbi durumlarda yanlış negatiflerin azaltılması açısından büyük önem taşır.

F1-Score değeri de 0.9722 (~%97.2) olup, Precision ve Recall arasındaki dengenin başarıyla kurulduğunu gösterir. Bu, modelin karar sınırlarının tutarlı ve dengeli olduğunu ve sınıflandırma kararlarında güvenilir bir denge sunduğunu ortaya koyar.

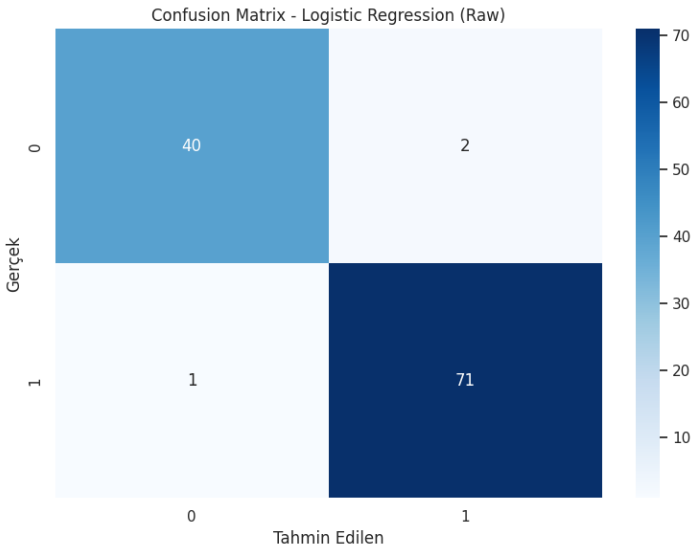
Son olarak, ROC-AUC skoru 0.996 (~%99.6) ile model, eşik seçiminden bağımsız olarak sınıfları neredeyse mükemmel şekilde ayırt edebilmekte ve rastgele bir sınıflandırıcıya kıyasla üstün ayrım gücü sunmaktadır. Bu, modelin genelleme yeteneğinin yüksek olduğunu ve yüksek riskli tıbbi uygulamalarda güvenle kullanılabileceğini göstermektedir.

9.2 Confusion Matrix (Karışıklık Matrisi)

Test seti üzerindeki tahminler kullanılarak Confusion Matrix (Karışıklık Matrisi) oluşturulmuş ve Seaborn ile görselleştirilmiştir.

Bu matris, modelin tahminlerini doğru pozitif (TP), doğru negatif (TN), yanlış pozitif (FP) ve yanlış negatif (FN) olarak sınıflandırmamıza olanak tanır. Tıbbi karar destek sistemlerinde, yanlış negatiflerin (FN) düşük, doğru pozitif ve doğru negatiflerin yüksek olması kritik öneme sahiptir.

Analiz, modelin hastalığı doğru yakalama başarısının yüksek, sağlıklı vakaları yanlış işaretlememe konusunda dikkatli ve kritik kaçırma riskini minimumda tuttuğunu gösterir. Bu da modelin tıbbi uygulamalarda güvenle kullanılabileceğini doğrular.

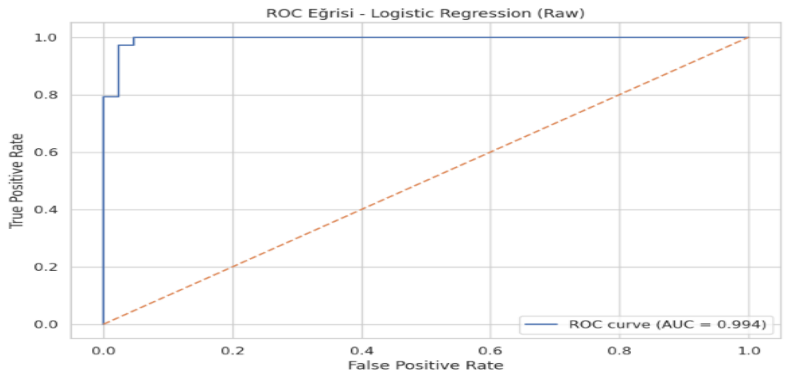


9.3 ROC Eğrisi

Aynı model için ROC eğrisi çizilmiş ve AUC değeri hesaplanmıştır. ROC eğrisi, farklı eşik değerlerinde doğru pozitif oran (TPR) ile yanlış pozitif oran (FPR) arasındaki ilişkiyi gösterir.

Eğri sol üst köşeye ne kadar yaklaşırsa ve AUC 1'e ne kadar yakınsa, modelin pozitif ve negatif sınıfları ayırt etme gücü o kadar yüksek demektir.

Eşik değeri düşerse geri çağırma (recall) artar, ancak yanlış pozitif oran da yükselir; eşik artarsa tam tersi olur. Bu yüzden tıbbi sınıflandırmalarda, yüksek geri çağırmayı korurken gereksiz yanlış pozitifleri minimumda tutacak bir eşik seçimi yapmak kritik öneme sahiptir.



Threshold (karar eşiği), modelin pozitif ve negatif sınıfları ayırt etme hassasiyetini doğrudan belirler.

- Eşik düşürüldüğünde, model daha fazla örneği pozitif olarak işaretler. Bu sayede geri çağırma (recall) ve TPR artar, yani kötü huylu vakaların yakalanma şansı yükselir. Ancak aynı zamanda FPR de yükselir ve daha fazla yanlış pozitif alarm üretilir.
- Eşik yükseldiğinde, model daha muhafazakar davranır. Yanlış pozitifler azalır, fakat geri çağırma düşer ve gerçek kötü huylu vakaların gözden kaçma riski artar.

Tıbbi karar destek sistemlerinde bu denge kritik önemdedir: Çok düşük bir eşik, gereksiz klinik yük ve stresi artırırken; çok yüksek bir eşik, tehlikeli yanlış negatiflere yol açabilir. Bu nedenle eşik, geri çağırmayı yüksek tutarken FPR'yi kabul edilebilir seviyede sınırlayan bir noktada optimize edilmelidir.

10. XAI – SHAP Açıklanabilirlik Analizi

SHAP Açıklanabilirlik Analizi (XAI), modelin sadece ne kadar başarılı olduğunu değil, hangi özellikler üzerinden karar verdiğini anlamak için kritik bir adımdır; özellikle tıbbi verilerde bu şeffaflık büyük önem taşır.

Validation'da en iyi performansı gösteren model için SHAP Explainer kurulmuş ve summary ile bar grafikleri üzerinden özelliklerin tahmine olan katkıları incelenmiştir.

- PCA temsili: Modelin kararları, varyansın yoğun olduğu bileşenlerden güçlü şekilde etkilenmiştir.
- LDA temsili: Kararlar, sınıflar arasındaki ayrımı en çok sağlayan doğrultularda şekillenmiştir.

Sonuç olarak, SHAP analizi modelin yüksek başarısının anlamlı ve tutarlı özellik etkilerine dayandığını, PCA ve LDA temsillerinde ise malign tespitini güçlendiren bileşenlerin öne çıktığını göstermektedir.

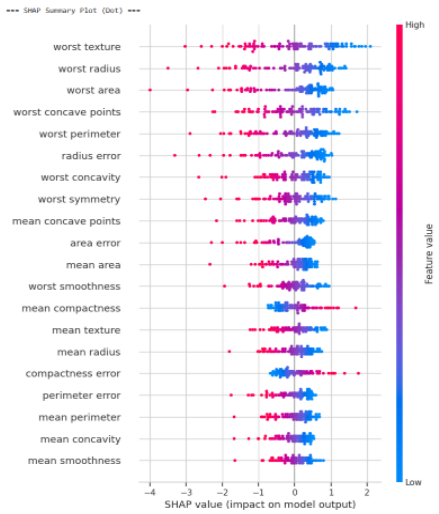
10.1 En İyi Validation Modeli için SHAP Analizi

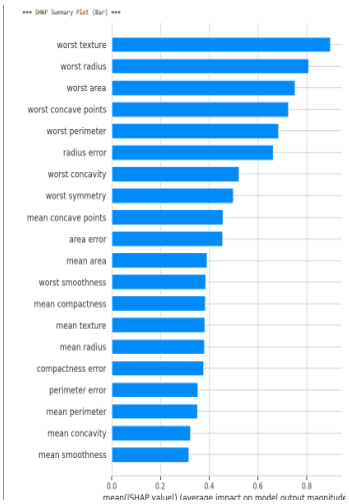
Validation aşamasında en iyi performansı gösteren model seçildikten sonra, modelin kararlarının şeffaflığını sağlamak için SHAP Explainer kullanılmıştır. Bu araç, modelin tahminlerinde her bir özelliğin ne kadar ve hangi yönde katkı sağladığını gösterir.

- SHAP summary_plot: Tüm örneklerde özelliklerin etkilerini ve önem sırasını görselleştirir.
- SHAP bar_plot: Her özelliğin tahminlere yaptığı ortalama mutlak katkıyı sunar.

Bu görseller sayesinde, model kararlarını en çok etkileyen özellikler belirlenmiş ve malign (kötü huylu) ile benign (iyi huylu) örnekleri ayırt etmedeki biyolojik anlamları çerçevesinde yorumlanmıştır.

Sonuç olarak, model sadece yüksek F1 ve ROC-AUC skorları üretmekle kalmamış, aynı zamanda tıbbi açıdan anlamlı ve klinik olarak ayırt edici özelliklere öncelik vermiştir. Bu durum, model başarısının tesadüfi olmadığını ve tahmin performansı ile özellik öneminin tutarlı bir şekilde ilişkili olduğunu göstererek, tıbbi karar destek sistemlerinde güvenilirliğini doğrulamaktadır.





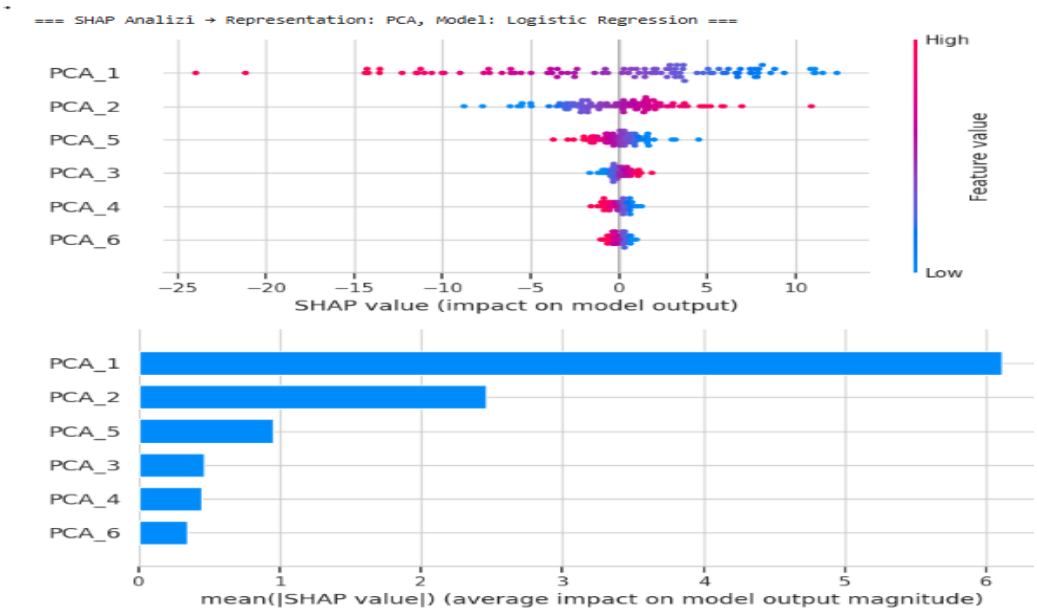
10.2 PCA ve LDA Temsilleri için SHAP Karşılaştırması

Boyut indirgeme sonrası model, orijinal 30 özellik yerine bu özelliklerin bileşenleri üzerinden öğrenir. SHAP analizi, hangi bileşenlerin tahminlerde daha etkili olduğunu ve sınıfları nasıl yönlendirdiğini gösterir.

PCA ile indirgenmiş veri:

- PCA, veriyi sınıf ayrımı gözetmeden, en yüksek varyansı açıklayan yeni eksenlere dönüştürür (PCA_1, PCA_2, …).
- SHAP analizi, modelin kararlarını ağırlıklı olarak PCA_1 ve PCA_2 üzerinden verdiğini gösterir; yani model, en bilgi yoğun eksenlere dayanıyor.
- Avantaj: Model, veri içindeki en açıklayıcı desenleri kullanır.
- Dezavantaj: Bileşenler orijinal özelliklerin karışımı olduğundan, hangi spesifik özelliğin kanseri etkilediği doğrudan anlaşılmaz; yorumlama dolaylıdır.

Özetle, PCA tabanlı SHAP analizi, modelin istatistiksel olarak baskın desenleri kullandığını ortaya koyarken, doğrudan özellik yorumunu sınırlı bırakır.

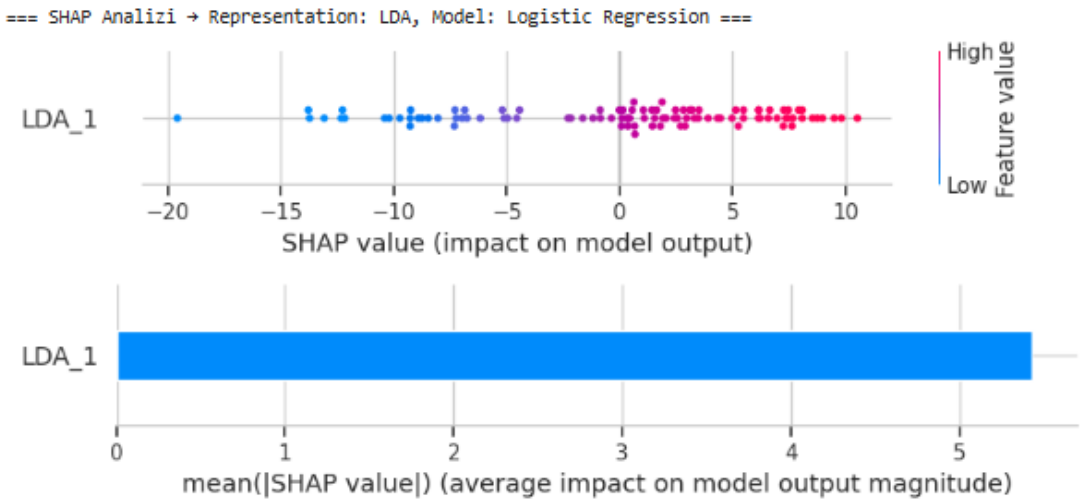


LDA ile indirgenmiş veride SHAP analizi

LDA, PCA'nın aksine sınıf ayrımını maksimize eden bir bileşen uzayı oluşturur. Binary sınıflandırmada tek bileşen yeterlidir ve model kararlarını bu eksen üzerinden verir.

SHAP analizi, modelin kararlarını büyük ölçüde LDA bileşenine dayandırdığını gösterir. Bu, modelin malign ve benign sınıfları matematiksel olarak en iyi ayıran ekseni kullandığını ve karar sınırını sınıf farkını dikkate alarak kurduğunu doğrular.

Sonuç olarak, LDA tabanlı model yüksek açıklanabilirlik, güvenilir karar sınırı ve tıbbi anlamlı sınıf ayrımı sağlar.



PCA ile indirgenmiş uzayda model, SHAP katkısını en yüksek varyans taşıyan bileşenlerden alırken; LDA'da bu katkı neredeyse tamamen tek LDA eksenine odaklanır. Bu durum, modelin kararlarını güçlü ve anlamlı bir malign–benign ayrım doğrultusunda verdiğini, gürültüye dayalı olmadığını ve sağlam bir karar geometrisine sahip olduğunu gösterir.