24.10.2025

DOKUZ EYLÜL ÜNİVERSİTESİ FEN FAKÜLTESİ
BİLGİSAYAR BİLİMLERİ

BİL3013 VERİ MADENCİLİĞİNE GİRİŞ
ÖĞRETİM ÜYESİ: Prof. Dr. Efendi NASİBOĞLU

ÖDEV 1: Wisconsin Breast Cancer Veri Seti ile Kümeleme Uygulaması

HAZIRLAYAN ÖĞRENCİLER:
İremgül ZEYTİNÖZÜ - 2023280135
Salim Taha KAVAS – 2023280117

VERİ SETİNİN VE KODLARIN LİNKİ:
- Veri seti: Wisconsin Breast Cancer Dataset
- Kodlar: GitHub

---

1. Ödevin Amacı

Bu ödevin amacı, Wisconsin Meme Kanseri teşhis veri setini kullanarak denetimsiz öğrenme yöntemlerinden kümeleme algoritmalarını uygulamaktır. Veri ön işleme, K-Means ve DBSCAN algoritmalarıyla model kurma, kurulan modellerin çeşitli metriklerle değerlendirilmesi ve sonuçların görselleştirilmesi adımları gerçekleştirilmiştir. Amacımız, veri madenciliği teknikleriyle etiket bilgisi olmadan veri içindeki kümeleri ne kadar başarılı bir şekilde ortaya çıkarabildiğimizi görmektir.

2. Kullanılan Teknolojiler ve Proje Ortamının Hazırlanması

Proje, Python programlama dili ve Jupyter Notebook ortamında geliştirilmiştir. Analiz sürecinde aşağıdaki temel kütüphanelerden yararlanılmıştır:

- **Pandas & NumPy:** Veri setini yüklemek, işlemek ve temel sayısal operasyonlar için kullanılmıştır.
- **ucimlrepo:** Wisconsin Breast Cancer veri setini UCI Machine Learning Repository üzerinden standart bir şekilde çekmek için kullanılmıştır.
- **Scikit-learn:** Model oluşturma ve değerlendirme sürecinin ana kütüphanesidir.
  - StandardScaler: Özellikleri standartlaştırmak için kullanılmıştır.
  - KMeans ve DBSCAN: Kümeleme modellerini oluşturmak için kullanılmıştır.
  - PCA: Yüksek boyutlu veriyi görselleştirme amacıyla 2 boyuta indirgemek için kullanılmıştır.
  - adjusted_rand_score, silhouette_score vb. metrikler: Modellerin performansını ölçmek için kullanılmıştır.
- **Matplotlib & Seaborn:** Keşifsel veri analizi ve model sonuçlarının görselleştirilmesi için kullanılmıştır.

3. Veri Setinin Anlaşılması ve Keşifsel Veri Analizi (EDA)

3.1 Veri Setine Genel Bakış

- **Veri Seti:** UCI Wisconsin Breast Cancer (Diagnostic)
- **Örnek Sayısı:** 569
- **Özellik Sayısı:** 30 (tümü sayısal)
- **Sınıf Dağılımı:** 357 İyi Huylu (Benign), 212 Kötü Huylu (Malignant)
- **Eksik Veri:** Bulunmuyor.

Yapılan ilk incelemede, veri setinin **569 adet gözlem (hasta kaydı)** ve her bir gözlemi tanımlayan **30 adet sayısal öznitelikten** oluştuğu görülmüştür.

Kümeleme modelinin başarısını değerlendirmek için kullanılacak olan hedef değişken ('Diagnosis'), iki sınıftan oluşmaktadır: **357 adet 'B' (Benign - iyi huylu)** ve **212 adet 'M' (Malignant - kötü huylu)**. Sınıflar arasında hafif bir dengesizlik olsa da, bu durum analizi olumsuz etkileyecek düzeyde değildir. Bu etiketler, denetimsiz bir öğrenme metodu olan kümeleme algoritmaları eğitilirken

kullanılmayacak, yalnızca sonuçların doğruluğunu ölçmek için son aşamada referans olarak alınacaktır.

Veri kalitesi kontrolü sonucunda, veri setinde **hiçbir eksik değer bulunmadığı** tespit edilmiştir. Bu durum, veri ön işleme sürecini basitleştirmekte ve eksik değer tamamlama gibi ek adımlara gerek kalmamasını sağlamaktadır.

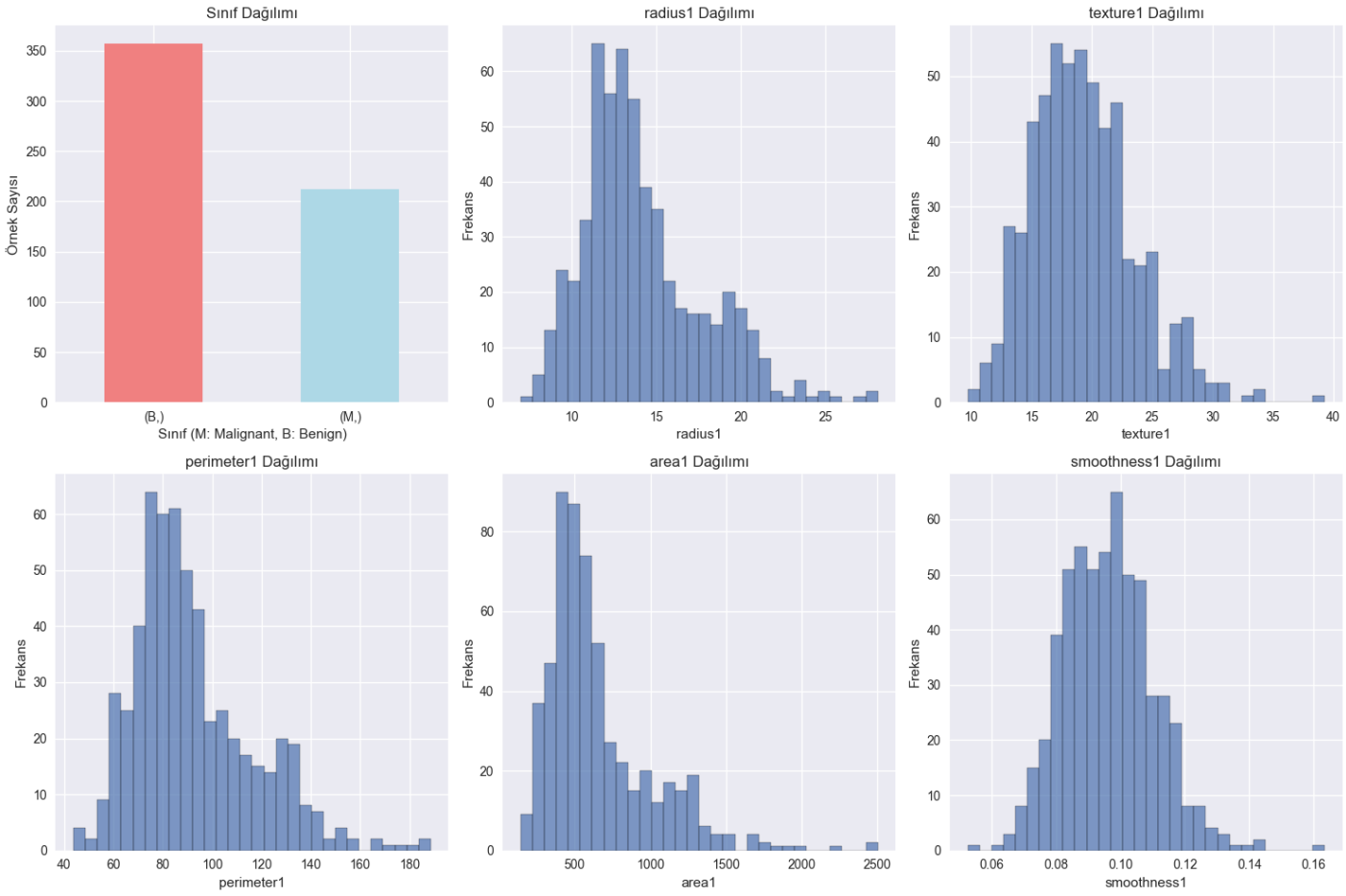
3.2 Özelliklerin İncelenmesi

Veri setinin istatistiksel özeti (describe() metodu ile) incelendiğinde, öznitelikler arasında **belirgin ölçek farklılıkları** olduğu gözlemlenmiştir. Örneğin, **area1** gibi özellikler binli değerlere ulaşırken, **smoothness1** gibi özellikler 0 ile 1 arasında küçük ondalık değerlere sahiptir.

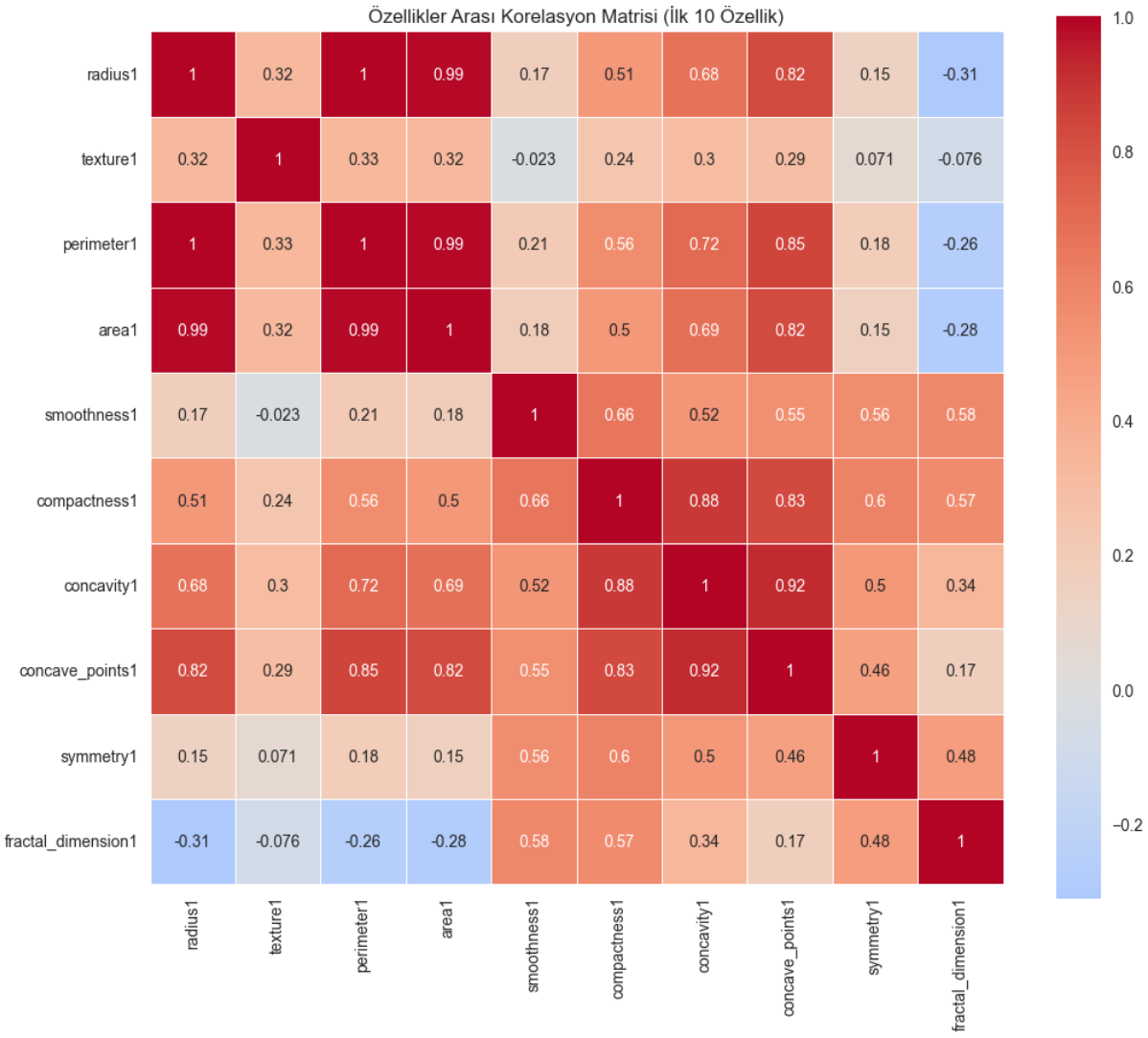
K-Means gibi **uzaklık tabanlı kümeleme algoritmaları**, bu tür ölçek farklılıklarına karşı oldukça hassastır. Eğer ölçeklendirme yapılmazsa, daha büyük sayısal aralığa sahip olan özellikler (area1, perimeter1 vb.) uzaklık hesaplamalarını domine ederek, potansiyel olarak ayırt edici gücü yüksek olan fakat daha küçük aralıktaki özelliklerin (smoothness1 vb.) model üzerindeki etkisini neredeyse sıfırlayacaktır. Bu durum, modelin yanlış veya yanlı sonuçlar üretmesine neden olabilir.

Bu sorunu ortadan kaldırmak ve her özelliğin kümeleme sürecine adil bir şekilde katkıda bulunmasını sağlamak için, modelleme aşamasından önce tüm özelliklerin **standartlaştırılması (Standard Scaling)** zorunlu ve kritik bir ön işleme adımıdır.

3.3 Keşifsel Veri Analizi ve Görselleştirme



Seçilen ilk beş özelliğe ait **histogramlar** ise bu özelliklerin dağılımı hakkında fikir vermektedir. Özellikle radius1, perimeter1 ve area1 gibi tümör boyutuyla ilişkili özelliklerin dağılımlarının tam olarak normal (simetrik) olmadığı, **sağa çarpık** bir yapıya sahip olduğu gözlemlenmiştir. Bu durum, veri setindeki çoğu tümörün belirli bir büyüklük aralığında yoğunlaştığını, ancak az sayıda da olsa çok daha büyük değerlere sahip aykırı (outlier) olabilecek kayıtların bulunduğunu ima etmektedir.



Özellikler arasındaki ilişkileri ortaya çıkarmak amacıyla ilk 10 özelliğe ait **korelasyon matrisi** bir ısı haritası (heatmap) ile görselleştirilmiştir.

Analiz sonucunda, bazı özellikler arasında beklendiği gibi çok güçlü pozitif korelasyonlar olduğu tespit edilmiştir. En çarpıcı örnek, **radius1 (yarıçap), perimeter1 (çevre) ve area1 (alan)** özellikleri arasındaki neredeyse mükemmel ($r > 0.99$) ilişkidir. Bu durum, bu üç özelliğin temelde aynı bilgiyi, yani **tümörün boyutunu** temsil ettiğini ve aralarında yüksek derecede **doğrusal bağımlılık (multicollinearity)** olduğunu göstermektedir.

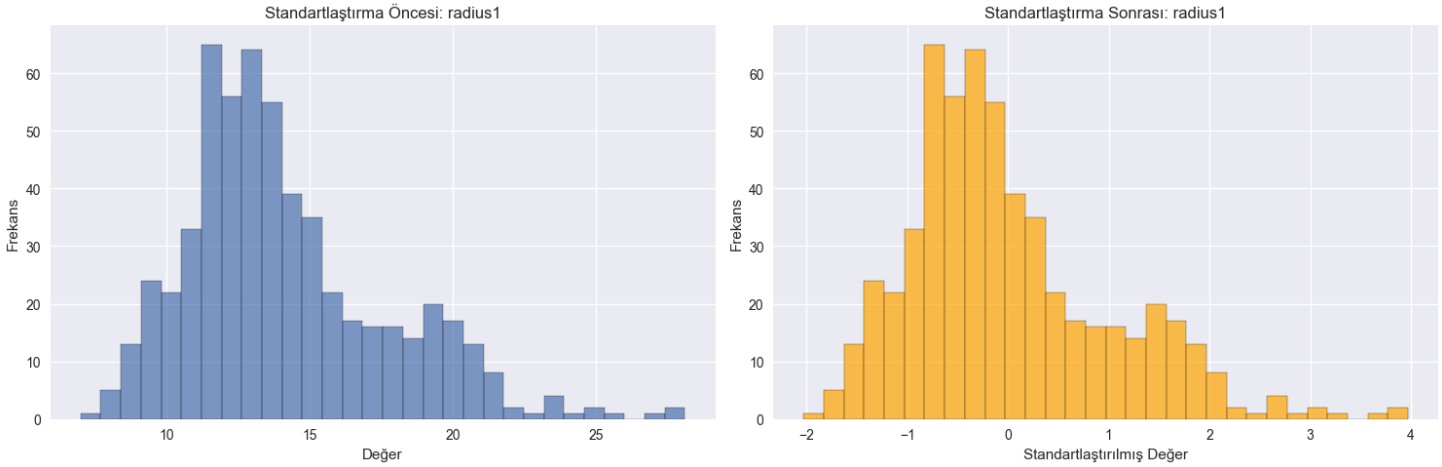
Bu yüksek korelasyon, uzaklık tabanlı algoritmalarda "boyut" bilgisinin etkisinin yapay olarak artmasına neden olabilir. Çünkü algoritma, birbiriyle neredeyse aynı olan üç farklı özellik üzerinden aynı bilgiyi tekrar tekrar hesaba katacaktır. Bu etkinin azaltılması ve kümelerin daha sağlıklı bir şekilde ayrıştırılması için, ilerleyen adımlarda kullanılacak olan **Temel Bileşen Analizi (PCA) ile boyut indirgeme** tekniğinin önemi bir kez daha ortaya çıkmaktadır. PCA, bu tür ilişkili değişkenleri tek bir birleşik bileşende toplayarak bu sorunu etkin bir şekilde çözecektir.

4. Veri Ön İşleme: Özelliklerin Standartlaştırılması

Önceki adımlarda yapılan keşifsel veri analizi, öznitelikler arasında (area1 gibi binli değerlere sahip olanlar ve smoothness1 gibi 0-1 arasında olanlar) ciddi ölçek farklılıkları olduğunu ortaya koymuştu. K-Means gibi uzaklık tabanlı kümeleme algoritmaları bu farklılıklara karşı oldukça hassastır ve ölçeklendirme yapılmadığı takdirde büyük aralıktaki özellikler model sonucunu domine edecektir.

Bu sorunu çözmek ve her özelliğin modele eşit ağırlıkta katkı sağlamasını garantilemek amacıyla, veri seti **StandardScaler** tekniği kullanılarak standartlaştırılmıştır. Bu yöntem, her bir özelliğin ortalamasını 0 ve standart sapmasını 1 olacak şekilde dönüştürür. Matematiksel olarak her veri noktasından özelliğin ortalaması çıkarılır ve sonuç standart sapmasına bölünür.

Kod çıktılarında da görüldüğü gibi, orijinal veri setindeki değer aralığı **0.00 ile 4254.00** gibi çok geniş bir yelpazedeyken, standartlaştırma sonrası bu aralık **-3.11 ile 12.07** arasına çekilmiştir.



Yukarıdaki 'radius1' özelliği için oluşturulan karşılaştırmalı histogramlar, bu dönüşümün etkisini görsel olarak mükemmel bir şekilde özetlemektedir.

- **Sol grafikte (Standartlaştırma Öncesi):** Özelliğin orijinal değer aralığı ve dağılımı görülmektedir.
- **Sağ grafikte (Standartlaştırma Sonrası):** Aynı dağılım yapısının korunduğu, ancak x-ekseninin **ortalama 0 ve standart sapma 1** olacak şekilde yeniden ölçeklendiği net bir şekilde görülmektedir.

Önemle belirtmek gerekir ki, standartlaştırma işlemi bir özelliğin **dağılımının şeklini değiştirmez**, yalnızca ölçeğini değiştirir. Bu sayede, verinin içsel yapısı korunurken, farklı ölçeklerdeki özelliklerin uzaklık hesaplamalarına adil bir şekilde katılması sağlanmış olur. Bu ön işleme adımı, modelin daha doğru ve güvenilir sonuçlar üretmesi için kritik öneme sahiptir.

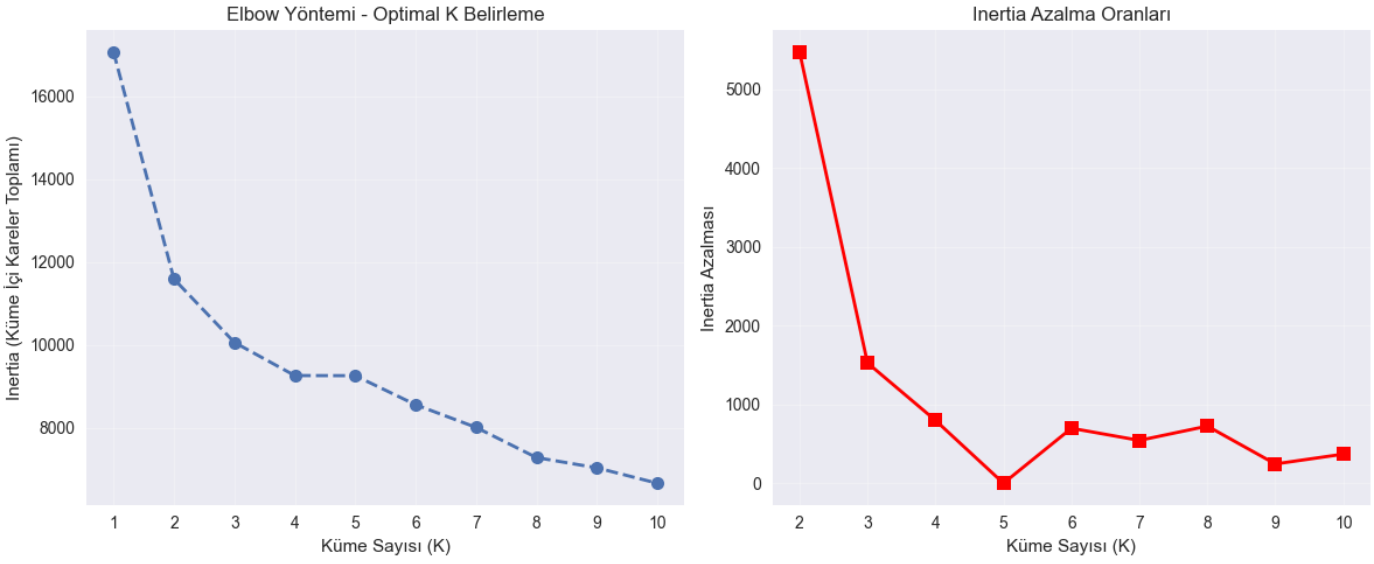
5. Modelleme ve Değerlendirme

Bu bölümde, veri seti üzerinde K-Means ve DBSCAN kümeleme algoritmaları uygulanmış ve performansları çeşitli metrikler kullanılarak karşılaştırılmıştır.

5.1. K-Means için Optimal Küme Sayısının Belirlenmesi

K-Means algoritmasının en önemli parametresi olan küme sayısı (K), modelin başarısını doğrudan etkiler. Optimal K değerini belirlemek amacıyla, yaygın olarak kullanılan iki farklı yöntem uygulanmıştır: Dirsek (Elbow) Yöntemi ve Siluet (Silhouette) Analizi.

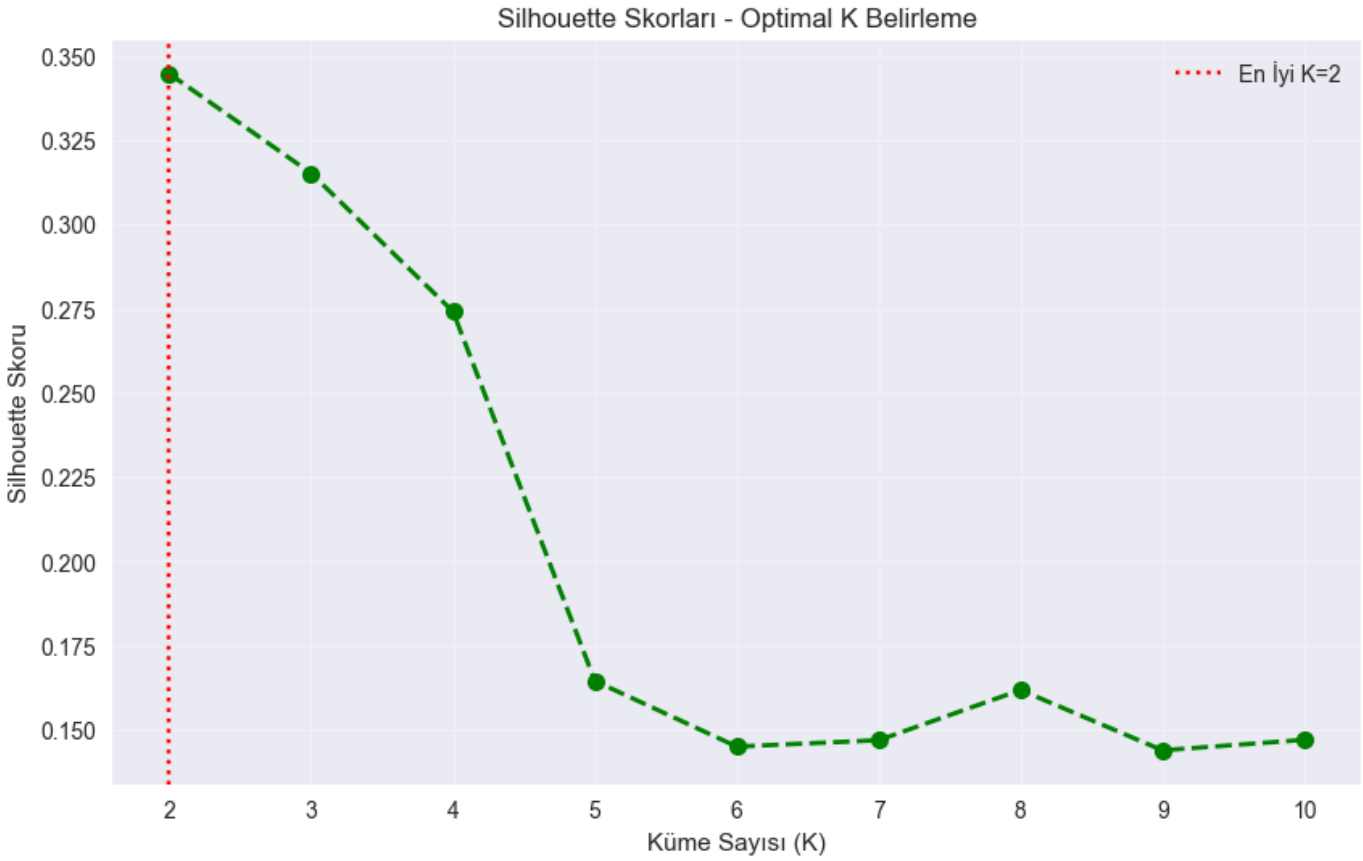
5.1.1. Dirsek (Elbow) Yöntemi



Bu yöntem, farklı K değerleri için modelin **atalet (inertia)** değerini, yani her bir veri noktasının kendi küme merkezine olan uzaklıklarının kareleri toplamını hesaplar. Analiz sonucunda, K değeri arttıkça atalet değerinin beklendiği gibi azaldığı görülmüştür. Grafikteki en belirgin kırılma veya "dirsek",

K=2 noktasında meydana gelmektedir. Bu noktadan sonra eğrinin eğimi belirgin şekilde azalmakta, bu da yeni bir küme eklemenin modeldeki iyileşmeye olan katkısının düştüğünü göstermektedir. Bu analiz, optimal küme sayısının 2 olabileceğini kuvvetle önermektedir.

5.1.2. Siluet (Silhouette) Analizi



Dirsek Yöntemi'nin bulgularını doğrulamak için, kümelerin ne kadar yoğun ve birbirinden ne kadar iyi ayrıldığını ölçen Siluet skoru analizi yapılmıştır. Farklı K değerleri için hesaplanan skorlar incelendiğinde, en yüksek Siluet skorunun **0.3447** ile **K=2** için elde edildiği görülmüştür.

Grafikten de anlaşılacağı üzere, K=2'den sonra skorun sürekli olarak düşmesi, bu veri seti için en anlamlı ve en iyi ayrılmış küme yapısının iki küme ile elde edildiğini göstermektedir. Her iki yöntemin de **K=2** sonucunu vermesi, bu seçimin doğruluğunu pekiştirmekte ve veri setinin doğal yapısı olan "iyi huylu" ve "kötü huylu" olmak üzere iki sınıflı yapısıyla tutarlılık göstermektedir. Bu nedenle, modelleme adımlarında **K=2** değeri kullanılmıştır.

5.2. K-Means Modelinin Uygulanması (K=2)

Optimal küme sayısı 2 olarak belirlendikten sonra, K-Means modeli standartlaştırılmış veri seti üzerine uygulanmıştır.

- **Küme Dağılımı:** Model, veri setini **188** ve **381** örnek içeren iki kümeye ayırmıştır.
- **Ayırt Edici Özellikler:** Küme merkezleri arasındaki farklar incelendiğinde, kümeleri birbirinden ayıran en belirgin özelliklerin concave_points1 (içbükey noktalar), concavity1 (içbükeylik) ve perimeter3 (çevre) gibi tümörün şekli ve boyutuyla ilgili metrikler olduğu tespit edilmiştir. Bu durum, algoritmanın biyolojik olarak anlamlı bir ayrım yaptığını desteklemektedir.

5.3. DBSCAN Modelinin Uygulanması

Yoğunluk tabanlı bir yaklaşım olan DBSCAN algoritması, veri setindeki farklı yoğunluktaki bölgeleri tespit etmek amacıyla uygulanmıştır. Parametre optimizasyonu için çeşitli eps değerleri test edilmiştir.

- **Parametre Seçimi ve Sonuç:** eps=2.5 ve min_samples=5 parametreleri ile yapılan analizde, algoritma 2 adet küme tespit etmiştir. Ancak, veri setindeki 569 örneğin 224 tanesini (%39) gürültü (noise) olarak sınıflandırmıştır. Ayrıca, bulunan kümelerden biri 340 örnek içerirken diğeri yalnızca 5 örnek içermektedir.
- **Yorum:** Gürültü olarak etiketlenen örnek sayısının yüksekliği ve kümeler arasındaki aşırı dengesizlik, DBSCAN'in bu veri setinin küresel yapısına K-Means kadar uygun olmadığını göstermektedir.

5.4. Performans Karşılaştırması ve Değerlendirme

Modellerin başarısı, gerçek sınıf etiketleri referans alınarak **Adjusted Rand Score (ARI), Normalized Mutual Info (NMI)** gibi harici metrikler ve **Silhouette Skoru** gibi dahili metriklerle ölçülmüştür.

Metrik	K-Means (K=2)	DBSCAN (Gürültü Hariç)
Adjusted Rand Score (ARI)	**0.6765**	**0.0775**
Normalized Mutual Info (NMI)	**0.5620**	**0.0774**
Silhouette Score	**0.3447**	**0.1295**

6. Görselleştirme ve Sonuçların Yorumlanması

Modelleme aşamasında elde edilen sayısal sonuçları daha anlaşılır kılmak ve modellerin performansını görsel olarak karşılaştırmak amacıyla, yüksek boyutlu veri seti üzerinde çeşitli görselleştirme teknikleri uygulanmıştır.

6.1. Boyut İndirgeme için Temel Bileşen Analizi (PCA)

30 boyutlu orijinal veri setini 2 boyutlu bir düzlemde görselleştirmek mümkün olmadığından, bu sorunu aşmak için yaygın bir boyut indirgeme tekniği olan **Temel Bileşen Analizi (PCA)** kullanılmıştır. PCA, özellikler arasındaki korelasyonu kullanarak veriyi en çok bilgiyi (varyansı) taşıyan daha az sayıda yeni bileşene dönüştürür.

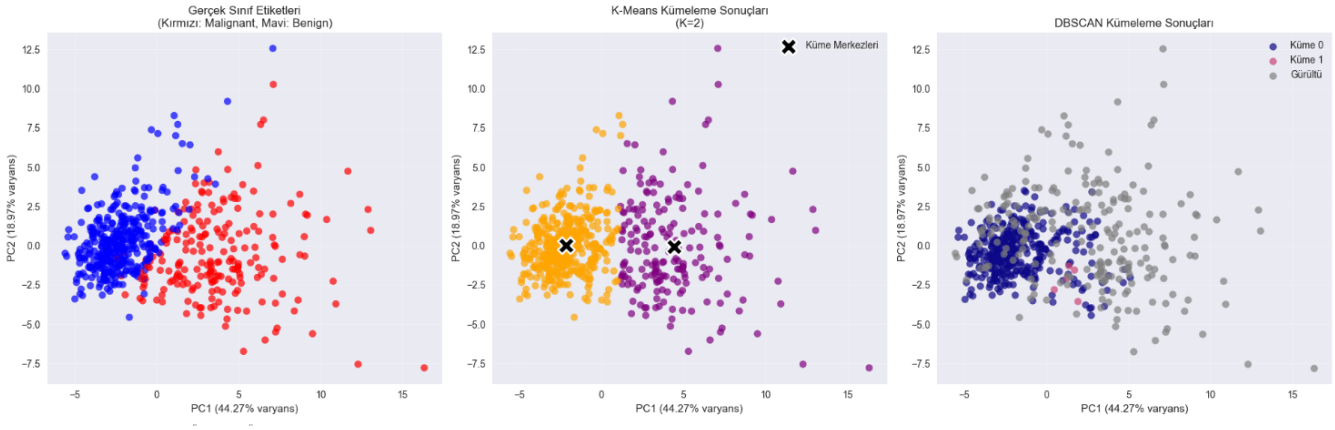
Yapılan analizde, standartlaştırılmış veri seti 2 temel bileşene indirgenmiştir:

1. **Bileşen (PC1):** Toplam varyansın **%44.27'sini** açıklamaktadır.
2. **Bileşen (PC2):** Toplam varyansın **%18.97'sini** açıklamaktadır.

Bu ilk iki bileşen, toplam varyansın **%63.24**'ünü açıklamaktadır. Bu oran, veri setindeki bilginin önemli bir kısmını koruyarak 2 boyutlu uzayda anlamlı bir görselleştirme yapılmasına olanak tanır.

6.2. Kümeleme Sonuçlarının Görsel Karşılaştırması

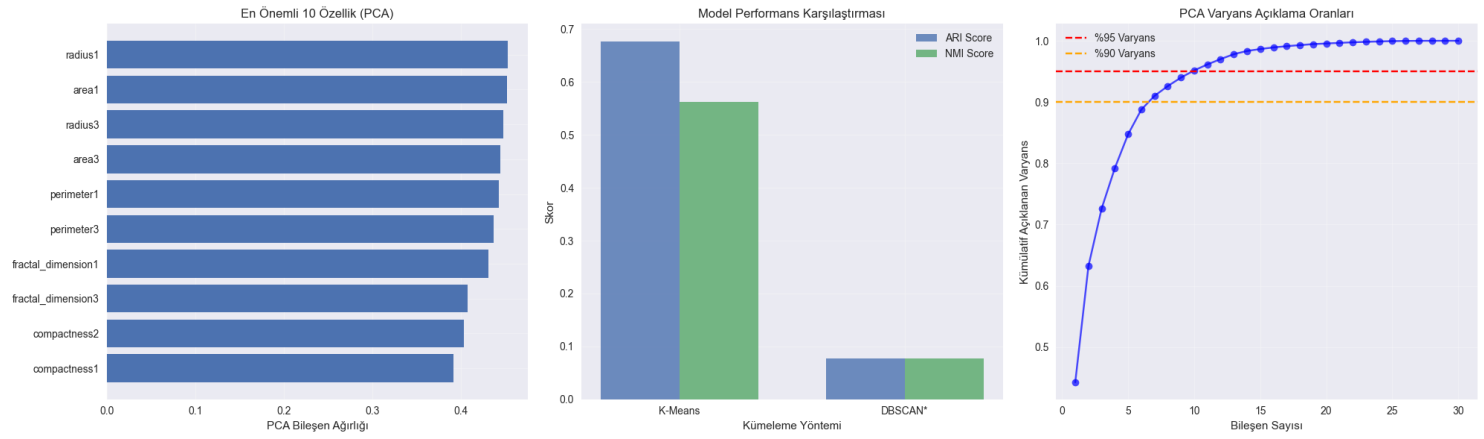
Aşağıdaki saçılım grafikleri (scatter plots), PCA ile elde edilen 2 boyutlu uzayda gerçek sınıf etiketlerini ve modellerin kümeleme sonuçlarını karşılaştırmaktadır.



- **Gerçek Sınıf Etiketleri (Sol Üst Grafik):** Bu grafikte, kırmızı noktalar kötü huylu (Malignant) ve mavi noktalar iyi huylu (Benign) tümörleri temsil etmektedir. Grafikten de görüldüğü gibi, iki sınıf PCA uzayında görsel olarak belirgin bir şekilde ayrışmaktadır. Bu durum, veri setinin kümelemeye uygun bir yapıya sahip olduğunu göstermektedir.
- **K-Means Kümeleme Sonuçları (Orta Üst Grafik):** K-Means modelinin bulduğu kümelerin (turuncu ve mor) gerçek sınıflarla görsel olarak büyük bir uyum içinde olduğu net bir şekilde görülmektedir. Kümeler, gerçekte olduğu gibi iki ayrı ve yoğun grup olarak ayrılmıştır. Bu görsel, K-Means'in sayısal metriklerdeki yüksek başarısını (ARI=0.6765) doğrulamaktadır.
- **DBSCAN Kümeleme Sonuçları (Sağ Üst Grafik):** DBSCAN'in performansı ise görsel olarak da zayıf kalmıştır. Grafikteki gri noktalar, algoritmanın "gürültü" olarak etiketlediği ve herhangi bir kümeye atayamadığı 224 örneği temsil etmektedir. Kalan noktalar anlamlı bir grup yapısı oluşturamamıştır. Bu görsel, DBSCAN'in yoğunluk tabanlı yaklaşımının bu veri setine uygun olmadığını kanıtlamaktadır.

6.3. Analiz Bulgularını Destekleyen Ek Görselleştirmeler

- **Model Performans Karşılaştırması (Orta Alt Grafik):** Bu sütun grafiği, önceki bölümde hesaplanan ARI ve NMI skorlarını görsel olarak özetlemektedir. K-Means'in her iki metrikte de DBSCAN'e göre ezici üstünlüğü açıkça görülmektedir.
- **PCA için En Önemli Özellikler (Sol Alt Grafik):** PCA analizinde, veri setini ayırmada en etkili olan özellikleri belirlemek mümkündür. Bu grafiğe göre, radius (yarıçap), area (alan), perimeter (çevre) ve concave points (içbükey noktalar) gibi özellikler, ilk iki temel bileşenin oluşumuna en çok katkıyı sağlayan metriklerdir. Bu bulgu, tümörün boyutu ve şeklinin, iyi ve kötü huylu vakaları ayırt etmede en önemli faktörler olduğu biyolojik beklentiyle de tutarlıdır.
- **Kümülatif Açıklanan Varyans (Sağ Alt Grafik):** Bu grafik, veri setindeki toplam bilginin (varyansın) ne kadarının kaç temel bileşen ile korunabileceğini göstermektedir. Grafiğe göre, toplam varyansın %90'ını açıklamak için yaklaşık 10 temel bileşene ihtiyaç duyulmaktadır. Bu, gelecekteki olası denetimli öğrenme modelleri için boyut indirgemenin faydalı bir ön işleme adımı olabileceğini göstermektedir.



7. Sonuçların Yorumlanması:

- K-Means Algoritması: K-Means algoritması, tüm metriklerde DBSCAN'e kıyasla çok daha üstün bir performans sergilemiştir. **0.6765**'lik ARI skoru, modelin bulduğu kümelerin gerçek sınıflarla güçlü bir şekilde örtüştüğünü göstermektedir. Kümelerin gerçek etiketlerle karşılaştırıldığı aşağıdaki tablo, bu başarıyı detaylandırmaktadır. Model, kötü huylu tümörlerin (sınıf 0) %82.5'ini (175/212) ve iyi huylu tümörlerin (sınıf 1) %96.3'ünü (344/357) doğru kümelere atayarak oldukça başarılı bir ayrım yapmıştır.

```
K-Means vs Gerçek Etiketler:
Gerçek   0    1   Toplam
-----------------------
0 (M)   175   37    212
1 (B)    13  344    357
Toplam  188  381    569
```

- DBSCAN Performansı: DBSCAN'in ARI ve NMI skorlarının 0'a çok yakın olması, bulduğu kümelerin gerçek sınıflarla neredeyse hiç ilişkili olmadığını göstermektedir. Bu durum, modelin yoğunluk tabanlı yaklaşımının bu veri setinin yapısına uygun olmadığını doğrulamaktadır.

Sonuç olarak, bu veri seti için **K-Means algoritmasının çok daha uygun ve başarılı bir kümeleme modeli olduğu** kanıtlanmıştır.