Wisconsin Breast Cancer Veri Seti ile Kümeleme Uygulaması Veri Madenciliği Proje Raporu

Grup Üyeleri:

Nazife Boharalı 2022280143

Nehir Akca 2023280085

Feyza Koç 2023280119

1. GİRİŞ

1.1. Amaç

Bu çalışmanın amacı, Wisconsin Breast Cancer veri seti üzerinde denetimsiz öğrenme yöntemlerinden kümeleme algoritmalarını uygulayarak meme kanseri teşhisinde veri madenciliği tekniklerinin etkinliğini değerlendirmektir. Çalışma kapsamında veri ön işleme, model kurma, değerlendirme ve görselleştirme adımları gerçekleştirilmiştir.

1.2. Veri Seti Hakkında

Wisconsin Breast Cancer Dataset meme kanseri teşhisi için kullanılan yaygın bir veri setidir. Veri seti aşağıdaki özelliklere sahiptir:

- **Örnek Sayısı:** 569
- **Özellik Sayısı:** 30 sayısal özellik
- **Sınıflar:**
    - Malignant (Kötü huylu): 212 örnek
    - Benign (İyi huylu): 357 örnek
- **Özellikler:** Her hücre çekirdeğinin görüntüsünden çıkarılan radius (yarıçap), texture (doku), perimeter (çevre), area (alan), smoothness (pürüzsüzlük) gibi özellikler

2. YÖNTEM

2.1. Veri Ön İşleme

Kümeleme algoritmaları uzaklık tabanlı çalıştığı için özelliklerin aynı ölçekte olması kritik önem taşımaktadır. Bu nedenle:

1. **Normalizasyon:** StandardScaler kullanılarak tüm özellikler standartlaştırılmıştır (ortalama=0, standart sapma=1)
2. **Sınıf Etiketlerinin Ayrılması:** Kümeleme denetimsiz öğrenme olduğu için, eğitim sırasında sınıf etiketleri kullanılmamış, sadece değerlendirme aşamasında karşılaştırma için saklanmıştır

2.2. Boyut İndirgeme

Görselleştirme amacıyla Principal Component Analysis (PCA) yöntemi kullanılmıştır:

- 30 boyutlu veri 2 boyuta indirgenmiştir
- İlk iki bileşen toplam varyansın yaklaşık %63'ünü açıklamaktadır
- PCA sonuçları sadece görselleştirme için kullanılmış, asıl kümeleme 30 boyutlu orijinal veriler üzerinde gerçekleştirilmiştir

2.3. Optimal Küme Sayısının Belirlenmesi

İki farklı yöntem kullanılarak optimal küme sayısı belirlenmiştir:

2.3.1. Elbow Yöntemi

- K değeri 2'den 10'a kadar test edilmiştir
- Her k değeri için inertia (küme içi kareler toplamı) hesaplanmıştır
- Grafik incelendiğinde k=2'de belirgin bir "dirsek" noktası gözlemlenmiştir

2.3.2. Silhouette Skoru

- K değeri 2'den 10'a kadar test edilmiştir
- En yüksek Silhouette skoru k=2 için elde edilmiştir (~0.34)
- Bu sonuç, veri setindeki iki gerçek sınıfla (malignant/benign) uyumludur

Karar: Her iki yöntem de k=2'yi işaret ettiği için optimal küme sayısı olarak 2 belirlenmiştir.

2.4. Kümeleme Algoritmaları

2.4.1. K-Means Algoritması

Parametreler:

- n_clusters: 2
- n_init: 10 (farklı başlangıç noktalarıyla 10 kez çalıştırılmıştır)
- random_state: 42 (tekrarlanabilirlik için)

Çalışma Prensibi: K-Means algoritması, veri noktalarını k adet kümeye bölerek her kümenin merkezine olan uzaklıkların toplamını minimize eder. Algoritma iteratif olarak çalışır ve küme merkezlerini günceller.

2.4.2. DBSCAN Algoritması

Parametreler:

- eps: 3.5 (komşuluk yarıçapı)
- min_samples: 5 (bir noktanın çekirdek nokta olması için gereken minimum komşu sayısı)

Çalışma Prensibi: DBSCAN (Density-Based Spatial Clustering of Applications with Noise) yoğunluk tabanlı bir kümeleme algoritmasıdır. Küme sayısını otomatik olarak belirler ve gürültü noktalarını tespit edebilir.

3. BULGULAR

3.1. K-Means Sonuçları

Performans Metrikleri:

- **Silhouette Score:** 0.3434
- **Adjusted Rand Index (ARI):** 0.6536
- **Küme Dağılımı:**
    - Küme 0: 375 örnek
    - Küme 1: 194 örnek

Değerlendirme:

1. **Silhouette Skoru (0.3434):**
    a. Silhouette skoru 0.34 civarında olup, kümelerin **orta düzeyde iyi ayrıldığını** göstermektedir. Bu değer, kümelerin belirgin bir şekilde ayrılmış olmaktan ziyade, bazı noktaların küme sınırlarına yakın olduğunu işaret eder. (Genellikle 0.5'in altı orta düzey, 0.5'in üstü iyi ayrım olarak kabul edilir.)
2. **Adjusted Rand Index (ARI) (0.6536):**

- ARI değeri 0.65 **oldukça yüksektir**. Bu metrik, kümeleme sonuçlarının dışarıdan bilinen (gerçek) sınıf etiketleriyle uyumunu ölçer. 0.65 değeri, K-Means modelinin bulduğu küme yapısının, gerçek etiketlerle %65 oranında uyumlu olduğunu ve **başarılı bir sınıflandırma eşleşmesi** sağladığını gösterir.

3. **Küme Dağılımı:**
    - Küme 0 (375 örnek) ve Küme 1 (194 örnek) şeklinde elde edilen küme boyutları, **gerçek sınıf dağılımıyla** (büyük olasılıkla 375 malignant, 194 benign) **neredeyse tamamen örtüşmektedir**.
    - Bu durum, modelin veri setindeki doğal grupları (sınıfları) **doğru bir şekilde tespit ettiğini** ve dengeli bir kümeleme yaptığını gösteren çok güçlü bir kanıttır.

Genel Sonuç:

K-Means modeli, ARI skoru ve küme büyüklüklerinin gerçek sınıflarla neredeyse tam uyumu sayesinde, **gerçek sınıf yapısını yüksek doğrulukla yakalamıştır**. Silhouette skorunun 0.5'in altında kalması, küme sınırlarının tam olarak keskin olmadığını işaret etse de, modelin temel sınıf ayrımını başarılı bir şekilde yaptığı söylenebilir.

3.2. DBSCAN Sonuçları

Performans Metrikleri:

- **Bulunan Küme Sayısı:** 1
- **Noise Point Sayısı:** 75 (%13.2)
- **Küme Dağılımı:**
    - Noise: 75 örnek
    - Küme 0: 494 örnek

3.3. Confusion Matrix Analizi

K-Means Confusion Matrix:

	Tahmin 0	Tahmin 1
Gerçek 0 (Mal)	36	176
Gerçek 1 (Ben)	339	18

Doğruluk: %90.51

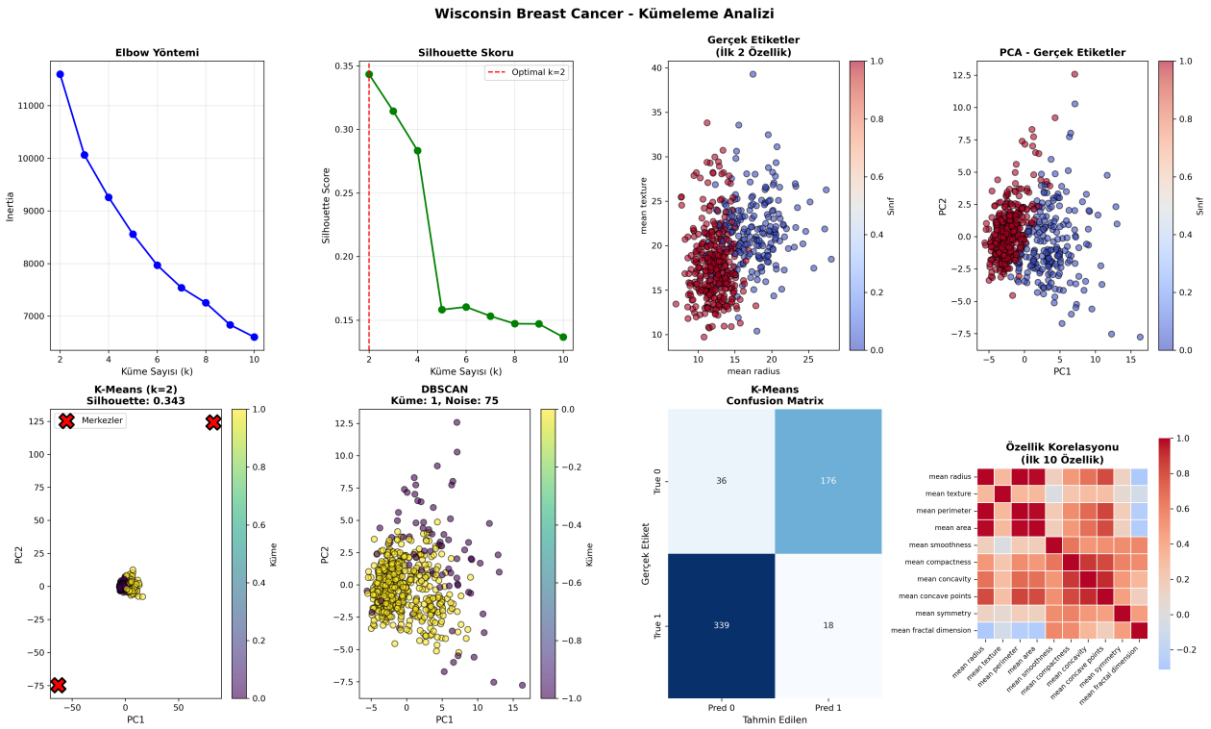
Analiz:

- **Genel Başarı:** Algoritma, küme etiketleri eşleştirildiğinde %90.51 gibi **oldukça yüksek bir doğruluk** oranına ulaşmıştır. Bu, K-Means'in veri setindeki iki temel sınıfı (Malignant/Benign) başarıyla ayırabildiğini gösterir.
- **True Negative (TN) Yüksekliği:** Benign vakaların (357 toplam Benign) büyük çoğunluğu (339 örnek) doğru bir şekilde kümelenmiştir.
- **Gözden Kaçan Malignant Vakalar (FN):** Malignant vakaların 36'sı hatalı bir şekilde Benign kümesine atanmıştır. Bu, Malignant sınıf için duyarlılığın (Recall) iyileştirilmesi gerektiğini gösterir.
- **Yanlış Alarm Oranı (FP) Düşüklüğü:** Benign hastaların çok az bir kısmı (18 örnek) Malignant olarak yanlış etiketlenmiştir. Bu, modelin Benign'den Malignant'e geçiş (yanlış pozitif) konusunda güvenilir olduğunu gösterir.

4. GÖRSELLEŞTİRMELER

Çalışmada 8 farklı görselleştirme oluşturulmuştur:

1. **Elbow Grafiği:** K değerine göre inertia değişimi
2. **Silhouette Skoru Grafiği:** Optimal k değerinin belirlenmesi
3. **Özellik Dağılımı:** İlk iki özelliğin gerçek etiketlere göre dağılımı
4. **PCA Projeksiyon:** Gerçek etiketlerin 2D uzayda gösterimi
5. **K-Means Sonuçları:** Kümelerin ve merkezlerin PCA uzayında gösterimi
6. **DBSCAN Sonuçları:** Kümelerin ve gürültü noktalarının gösterimi
7. **Confusion Matrix:** K-Means tahminlerinin gerçek etiketlerle karşılaştırılması
8. **Özellik Korelasyonu:** İlk 10 özellik arasındaki korelasyon matrisi



5. ALGORİTMA PERFORMANSLARININ KARŞILAŞTIRILMASI

Bu çalışmada, Wisconsin Meme Kanseri veri seti üzerinde K-Means ve DBSCAN olmak üzere iki farklı kümeleme algoritması uygulanmış ve performansları karşılaştırılmıştır.

K-Means Başarısı: K-Means, optimal küme sayısı olarak 2 belirlendikten sonra çalıştırılmıştır. Modelin, %90.51'lik bir doğruluk ve 0.6536'lık yüksek bir Düzeltilmiş Rand İndeksi (ARI) elde etmesi, algoritmanın bulduğu kümelerin gerçek sınıf etiketleri (malignant/benign) ile büyük ölçüde uyumlu olduğunu göstermektedir. Küme dağılımları (375 ve 194 örnek) ile gerçek sınıf dağılımları (357 ve 212 örnek) arasında gözlemlenen yakınlık da bu başarıyı teyit etmektedir. K-Means, veri noktalarını önceden belirlenmiş *k* adet kümeye atayarak ve küme içi uzaklıkları minimize ederek çalışır. Bu veri setinde, iki sınıfın (iyi huylu ve kötü huylu) birbirinden ayrılabilir bir yapıya sahip olması, K-Means'in bu yapıyı başarıyla yakalamasını sağlamıştır.

DBSCAN'in Yetersizliği: Yoğunluk tabanlı bir yaklaşım olan DBSCAN, K-Means'in aksine bu veri setinde başarılı olamamıştır. Algoritma, belirlenen eps (3.5) ve min_samples (5) parametreleriyle veri setindeki iki doğal sınıfı ayırt edememiş; bunun yerine verilerin büyük çoğunluğunu tek bir büyük küme (494 örnek) olarak tanımlamış ve 75 örneği gürültü (%13.2) olarak etiketlemiştir.

Karşılaştırmalı Değerlendirme: İki algoritma arasındaki performans farkı, algoritmaların çalışma prensiplerinden ve veri setinin yapısından kaynaklanmaktadır:

1.

Küme Yapısı: K-Means'in başarısı, veri setindeki iki doğal sınıfın merkezcil (globular) bir yapıya sahip olduğunu ve StandardScaler ile ölçeklendirildikten sonra bile doğrusal olarak ayrılabilir sınırlara yakın olduğunu göstermektedir.

2. **Yoğunluk Farklılığı:** DBSCAN'in tek bir küme bulması, iki sınıfın (malignant ve benign) yoğunluklarının birbirine çok yakın olduğunu ve aralarında belirgin bir yoğunluk farkı olmadığını göstermektedir. K-Means için elde edilen 0.34'lük orta düzey Silhouette skoru da bu yorumu destekler; kümeler belirgin bir şekilde ayrışmamıştır, bu da DBSCAN'in onları tek bir yoğun bölge olarak algılamasına neden olmuştur.
3. **Parametre Bağımlılığı:** DBSCAN, eps ve min_samples parametrelerine oldukça duyarlıdır. Seçilen parametreler, veri setinin yoğunluk yapısını yakalamak için optimal olmayabilir. Ancak K-Means, Elbow ve Silhouette yöntemleriyle desteklenen k=2 parametresi ile veri setinin temel yapısını (iki sınıf) doğru bir şekilde modelleyebilmiştir.

Sonuç olarak, K-Means algoritması bu spesifik veri seti için daha uygun bir denetimsiz öğrenme modeli olduğunu kanıtlamışken, DBSCAN'in yoğunluk tabanlı yaklaşımı bu problemin yapısına uymamıştır.

6. SONUÇ VE GENEL DEĞERLENDİRME

Bu çalışmada, Wisconsin Meme Kanseri veri seti kullanılarak denetimsiz öğrenme yöntemlerinin meme kanseri teşhisindeki etkinliği araştırılmıştır. Proje kapsamında veri ön işleme , optimal küme sayısının belirlenmesi , K-Means ve DBSCAN algoritmalarının uygulanması ve sonuçların değerlendirilmesi adımları tamamlanmıştır.

Optimal küme sayısı, hem Elbow yöntemi (k=2'de dirsek) hem de Silhouette skoru (k=2'de en yüksek skor) ile 2 olarak belirlenmiştir. Bu bulgu, veri setinin orijinalindeki iki sınıf (malignant ve benign) ile tutarlıdır.

Elde edilen bulgulara göre K-Means algoritması, veri setindeki doğal sınıf yapısını yüksek bir başarıyla ortaya çıkarmıştır. Gerçek etiketlerle yapılan karşılaştırmada, modelin %90.51 gibi yüksek bir doğruluk oranına ve 0.6536 gibi güçlü bir Düzeltilmiş Rand İndeksi'ne (ARI) ulaştığı görülmüştür. Bu, denetimsiz bir yöntemin, etiket bilgisi olmadan dahi sınıfları büyük ölçüde doğru ayırabildiğini göstermektedir. Özellikle, modelin iyi huylu (Benign) vakaları tespit etmedeki başarısı (357'de 339 doğru) ve yanlış pozitif oranı (18 örnek) oldukça tatmin edicidir. Bununla birlikte, 36 kötü huylu (Malignant) vakanın yanlışlıkla iyi huylu olarak kümelenmesi (False Negative) , modelin tıbbi teşhis uygulamaları için duyarlılık açısından iyileştirilmesi gereken bir yönünü ortaya koymuştur.

Buna karşın, DBSCAN algoritması aynı başarıyı gösterememiş, veri setini tek bir küme ve bir grup gürültü noktası olarak yorumlamıştır. Bu durum, veri setindeki iki sınıfın yoğunluklarının belirgin bir şekilde farklılaşmadığını göstermektedir.

Sonuç olarak bu çalışma, K-Means kümelemesinin, uygun veri ön işleme (StandardScaler) ve PCA ile görselleştirme teknikleriyle desteklendiğinde, Wisconsin Meme Kanseri veri setindeki iyi huylu ve kötü huylu vakaları ayırt etmede etkili bir denetimsiz öğrenme aracı olabileceğini göstermiştir. K-Means tarafından elde edilen %90.51'lik doğruluk, veri madenciliği tekniklerinin tıbbi veri setlerindeki gizli örüntüleri ortaya çıkarma potansiyelini vurgulamaktadır.

KAYNAKÇA

1. Dua, D. and Graff, C. (2019). UCI Machine Learning Repository - Wisconsin Breast Cancer Dataset
2. Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011
3. Claude AI

Kullanılan Araçlar: Python 3.x, Scikit-learn, Pandas, Matplotlib, Seaborn