

**Gölhisar Uygulamalı Bilimler Yüksekokulu**

**NESNE TABANLI PROGRAMLAMA II DERSİ**

**PROJE KONUSU: Kanser Teşhisi ve Sınıflandırması**

**Öğrenci Ad-Soyad/Numara**

**1-Muhammed Enes SİVİL/2212903032**

**2-Emre YILMAZ/2212903025**

**3-Deniz Renas GÜNEŞ/2212903006**

**4-Abdil Cemrehan KARA/2212903005**

MAYIS 2024 BURDUR

# GİRİŞ

**K-Means Algoritması**

K-Means algoritması, veri madenciliği ve makine öğrenmesi alanlarında yaygın olarak kullanılan bir kümeleme yöntemidir. Algoritmanın temel prensibi, n veri noktasını k kümeye ayırmak ve her veri noktasını en yakın merkez noktaya atamaktır. Bu yöntem, özellikle büyük veri kümelerinin analizinde etkili ve verimli bir çözüm sunar. K-Means algoritması, 1967 yılında James MacQueen tarafından tanıtılmış olup, günümüzde çeşitli uygulamalarda yaygın olarak kullanılmaktadır.

K-Means algoritması, başlangıçta rastgele seçilen k merkez noktası ile başlar. Her veri noktası, en yakın merkeze atanır ve ardından her küme için yeni merkezler hesaplanır. Bu süreç, merkez noktaları sabit kalana veya belirli bir iterasyon sayısına ulaşılana kadar tekrarlanır. Algoritmanın temel amacı, her bir kümenin içindeki veri noktalarının toplam varyansını minimize etmektir. Matematiksel olarak, bu işlem Öklidyen mesafe kullanılarak gerçekleştirilir.

Algoritmanın en büyük avantajları basitlik ve hızıdır. K-Means, büyük veri kümeleri üzerinde hızlı bir şekilde çalışabilir ve çok boyutlu verilerde etkin kümeleme yapabilir. Ancak, bazı dezavantajları da vardır. Başlangıç noktalarına olan bağımlılığı ve küme sayısının (k) önceden belirlenmesi gerekliliği, algoritmanın performansını etkileyebilir. Ayrıca, K-Means genellikle küresel şekilli kümeleri iyi ayırabilirken, karmaşık şekilli kümelerde zorlanabilir.

K-Means algoritması, çeşitli uygulama alanlarında yaygın olarak kullanılır. Müşteri segmentasyonu, görüntü sıkıştırma ve anomali tespiti gibi alanlarda etkili çözümler sunar. Pazarlama stratejileri geliştirmek için müşteri gruplarının belirlenmesi, renk kümeleme yoluyla görüntülerin sıkıştırılması ve normdan sapmaların tespiti gibi uygulamalar, K-Means algoritmasının başarıyla kullanıldığı örnekler arasındadır. Algoritmanın etkin kullanımı, veri kümesinin özelliklerine ve analiz hedeflerine bağlıdır.

**Wisconsin Göğüs Kanseri Veri Kümesi Hakkında Bilgi**

Wisconsin Göğüs Kanseri Veri Kümesi (Orijinal), makine öğrenmesi topluluğunda sıkça kullanılan bir sınıflandırma veri kümesidir. Bu veri kümesi, 1 Kasım 1995'te Dr. William H. Wolberg, W. Nick Street ve Olvi L. Mangasarian tarafından oluşturulmuş olup, UCI Makine Öğrenme Deposunda mevcuttur. Veri kümesi, göğüs kanseri teşhisinde kullanılan biyopsi analizlerine dayanmaktadır ve çeşitli makine öğrenmesi algoritmalarının performansını değerlendirmek için sıklıkla kullanılır.

Veri kümesi, toplam 699 örnek içerir ve her örnek, hücre çekirdeklerinin özelliklerini tanımlayan 10 öznitelik ile sınıflandırılır. Bu öznitelikler arasında yığın kalınlığı, hücre boyutu ve şekli tekdüzeliği, marjinal yapışma, çıplak çekirdekler, donuk kromatin, normal nükleol ve mitoz yer alır. Ayrıca, "Sınıf" özniteliği, 2'nin iyi huylu ve 4'ün kötü huylu tümörleri temsil ettiği hedef değişkeni içerir. Bu yapı, verilerin analiz edilmesini ve sınıflandırılmasını kolaylaştırır.

Veri kümesinde 16 eksik değer içeren örnek bulunur. Toplam 699 örneğin %65.5'i iyi huylu (458 örnek) ve %34.5'i kötü huyludur (241 örnek). Eksik değerler genellikle ilgili özniteliğin ortalaması veya medyanı ile değiştirilerek veriler işlenir. Bu yöntem, eksik verilerin analizin doğruluğunu olumsuz etkilememesi için yaygın olarak kullanılan bir tekniktir.

Wisconsin Göğüs Kanseri Veri Kümesi, sınıflandırma görevleri, makine öğrenmesi algoritmalarının testi ve performans değerlendirmesi için kullanılır. Veri kümesi, özellikle eğitim, araştırma ve algoritma karşılaştırmaları için klasik bir örnek teşkil eder. Verilere erişim ve indirme işlemi için UCI Makine Öğrenme Deposundan yararlanılabilir. Bu veri kümesi, makine öğrenmesi ve veri bilimi alanında yapılan çalışmalar için önemli bir kaynak olarak kabul edilir.

# GEREÇ VE YÖNTEM

**K-Means Sınıflandırma Metodu Hakkında Detaylı Bilgi**

**1. Tarihsel Geçmiş**

K-Means algoritması, 1967 yılında James MacQueen tarafından tanıtılmıştır ve veri kümeleme alanında devrim niteliğinde olmuştur. Algoritmanın temel prensipleri, veri noktalarının belirli sayıda kümeye (k) ayrılması ve her kümenin bir merkez noktası (centroid) etrafında toplanmasıdır. 1982 yılında, Lloyd tarafından pratik bir uygulaması geliştirilmiş ve bu nedenle bazen Lloyd algoritması olarak da anılmıştır.

**2. Temel Prensipler**

K-Means algoritması, aşağıdaki adımlarla özetlenir:

1. Başlatma: k adet merkez noktası rastgele seçilir.
2. Atama: Her veri noktası, en yakın merkez noktaya atanır.
3. Merkezlerin Güncellenmesi: Her küme için yeni merkez noktası hesaplanır.
4. Tekrarlama: Adım 2 ve 3, merkez noktaları sabit kalana veya belirli bir iterasyon sayısına ulaşılana kadar tekrarlanır.

Matematiksel olarak, her veri noktası , en yakın merkezi bulmak için Öklidyen mesafeye göre atanır ve kümelerin merkezleri, kümeye dahil olan veri noktalarının ortalaması alınarak güncellenir.

**3. Kullanım Alanları**

K-Means algoritması, çeşitli alanlarda yaygın olarak kullanılır:

* Müşteri Segmentasyonu: Pazarlama stratejileri geliştirmek için müşteri gruplarının belirlenmesi.
* Görüntü Sıkıştırma: Renk kümeleme yoluyla görüntülerin sıkıştırılması.
* Anomali Tespiti: Normdan sapmaların tespiti için veri noktalarının gruplanması.
* Gen Ekspresyon Analizi: Biyoinformatik alanında genlerin ifade seviyelerinin gruplandırılması.
* Belge Kümeleme: Metin madenciliği ve bilgi erişimi uygulamalarında benzer belgelerin gruplandırılması.

**4. Görseller ve Şekiller**

Adım 1: Başlatma

Adım 2: Veri Noktalarının Atanması

Adım 3: Merkezlerin Güncellenmesi

Adım 4: Tekrarlama

**5. Avantajlar ve Dezavantajlar**

Avantajlar:

* Basitlik ve hız.
* Büyük veri kümeleri üzerinde verimli çalışır.
* Çok boyutlu verilerde etkin kümeleme sağlar.

Dezavantajlar:

* Başlangıç merkezlerine bağlılık.
* Küme sayısının (k) önceden belirlenmesi gerekir.
* Küresel şekilli kümelerde iyi performans, ancak karmaşık şekilli kümelerde zayıf performans.

**6. Özet ve Sonuç**

K-Means algoritması, veri kümeleme ve analizinde güçlü ve esnek bir araçtır. Basit yapısı ve hızlı hesaplama yetenekleri, onu çeşitli uygulamalarda tercih edilen bir yöntem haline getirir. Ancak, başlangıç noktalarının seçimi ve uygun k değerinin belirlenmesi gibi konular dikkatle ele alınmalıdır. Algoritmanın etkin kullanımı, veri kümesinin özelliklerine ve analiz hedeflerine bağlıdır.

# BULGULAR

**K-means Kümeleme Sonuçları (Sütun Grafik)**

K-means kümeleme algoritması kullanılarak elde edilen sonuçlar aşağıdaki sütun grafikte gösterilmiştir. Grafikte, her bir özelliğin küme merkezlerinin ortalaması görülmektedir. Benign ve Malignant olmak üzere iki sınıf için sütun grafikleri yer almaktadır.

çizgi, diyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, paralel içeren bir resim

Açıklama otomatik olarak oluşturuldumetin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldumetin, ekran görüntüsü, renklilik, diyagram içeren bir resim

Açıklama otomatik olarak oluşturuldu

**K-Means Kümeleme Grafiği ve Gerçek Kümeleme Grafiği**

Veri setinin ilk iki özelliği kullanılarak elde edilen K-means kümeleme grafiği ile gerçek kümeleme grafiği karşılaştırılmıştır. K-means kümeleme grafiğinde, veri noktaları farklı renklerle gösterilerek kümelenme işlemi görülmektedir. Gerçek kümeleme grafiğinde ise veri noktalarının gerçek sınıfları kullanılarak oluşturulmuştur.

**Veri Noktalarının Kümeler Arasındaki Uzaklıkları**

K-means algoritması kullanılarak oluşturulan kümeler arasındaki uzaklıkların görselleştirildiği bir scatter plot grafiği sunulmuştur. Grafikte, her bir veri noktasının benign ve malikümelerine olan uzaklıkları renklerle gösterilmiştir.

metin, ekran görüntüsü, ekran, görüntüleme, yazılım içeren bir resim

Açıklama otomatik olarak oluşturuldu

**Gerçek ve Tahmin Edilen Sınıflar**

Gerçek sınıflarla tahmin edilen sınıfların karşılaştırıldığı bir scatter plot grafiği sunulmuştur. Her bir veri noktasının gerçek sınıfı ve tahmin edilen küme etiketi görselleştirilmiştir.

metin, ekran görüntüsü, diyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Açıklama otomatik olarak oluşturuldu

**Elde edilen sınıflandırma metrikleri:**

Accuracy (Doğruluk): [Accuracy değeri]

F1 Score: [F1 Score değeri]

Precision (Hassasiyet): [Precision değeri]

Recall (Geri Çağırma): [Recall değeri]

Bu metrikler, sınıflandırma modelinin performansını değerlendirmek için kullanılmaktadır. Çubuk grafikte her bir metriğin skoru gösterilmiştir.

# KAYNAKLAR

# *1. Arthur, D., & Vassilvitskii, S. (2007). "k-means++: The Advantages of Careful Seeding," Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 1027-1035.*

# *2.Berry, M. W. (2003). "Survey of Text Mining: Clustering, Classification, and Retrieval," Springer.*

# *3.Bradley, P. S., & Fayyad, U. M. (1998). "Refining Initial Points for K-Means Clustering," Proceedings of the Fifteenth International Conference on Machine Learning (ICML), pp. 91-99.*

# *4.Dhillon, I. S., & Modha, D. S. (2000). "A Data-Clustering Algorithm on Distributed Memory Multiprocessors," Large-Scale Parallel Data Mining, Lecture Notes in Artificial Intelligence, vol. 1759, Springer, pp. 245-260.*

# *5.Gersho, A., & Gray, R. M. (2012). "Vector Quantization and Signal Compression," Springer Science & Business Media.*

# *6.Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2002). "Clustering Validity Checking Methods: Part II," ACM SIGMOD Record, vol. 31, no. 3, pp. 19-27.*

# *7.Hastie, T., Tibshirani, R., & Friedman, J. (2009). "The Elements of Statistical Learning: Data Mining, Inference, and Prediction," Springer.*

# *8.Jain, A. K. (2010). "Data clustering: 50 years beyond K-means," Pattern Recognition Letters, Volume 31, Issue 8, pp. 651-666.*

# *9.Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A. Y. (2002). "An Efficient k-Means Clustering Algorithm: Analysis and Implementation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 7, pp. 881-892.*

# *10.MacQueen, J. (1967). "Some Methods for Classification and Analysis of Multivariate Observations," Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, pp. 281-297.*

# *11.Wolberg, W. H., Street, W. N., & Mangasarian, O. L. (1995). Wisconsin Göğüs Kanseri Veri Kümesi (Orijinal), UCI Makine Öğrenme Deposu.*

# *12.https://archive.ics.uci.edu/dataset/15/breast+cancer+wisconsin+original*

# *13.* [*https://www.kaggle.com/code/scratchpad/notebook8e357b651b/edit*](https://www.kaggle.com/code/scratchpad/notebook8e357b651b/edit)

# *14.* [*https://www.chegg.com/homework-help/questions-and-answers/part-1-k-means-clustering-25-points-load-breast-cancer-dataset-calling-sklearndatasetsload-q105801126*](https://www.chegg.com/homework-help/questions-and-answers/part-1-k-means-clustering-25-points-load-breast-cancer-dataset-calling-sklearndatasetsload-q105801126)

# *15.* [*https://www.google.com/amp/s/www.geeksforgeeks.org/k-means-clustering-introduction/amp/*](https://www.google.com/amp/s/www.geeksforgeeks.org/k-means-clustering-introduction/amp/)

# *16.* *https://medium.com/deep-learning-turkiye/k-means-algoritmas%C4%B1-b460620dd02a*

**KAYNAK KOD ADRESİ**