MAKINE ÖĞRENMESI ÖDEVI

Semih Yazıcı

ÖZETÇE

Müşterilerin davranışlarına göre kümelenmesi birçok işletme için kritik bir konudur. Benzer etnik köken, aile yapısı, ekonomik duruma sahip müşterilerin benzer davranışlar sergiledikleri uzun zamandır bilinmektedir. Bira ile bebek bezi alan insanların hepsinin ortak olarak çocuğa sahip olması örnek verilebilir. Bu alanda yapılmış çalışmalar Kansal et al. (2018); Arul et al. (2021); Rao et al. (2022); Pavithra et al. (2022) incelendiğinde bir işletmenin müşterilerinin davranışlarına göre aldığı aksiyonlar kısa ve uzun vadede işletmeye fayda sağlamıştır.

1 YÖNTEM

Bu projede veri kümesinde bulunan müşterilerin kümelendirilmesi istenmiştir. Projede kullanılan veri kümesi bir tanesi müşteri numarası olmak üzere beş adet özellik barındırmaktadır.

- Cinsiyet: Bu özellik kadın veya erkek olmak üzere iki değer alabilmektedir.
- Yaş: Bu özellik müşterilerin yaş bilgisini içermektedir. Müşterilerin yaşları 18-70 arasında değişmektedir.
- Yıllık Gelir: Bu özellik müşterilerin yıllık gelirlerini içermektedir. Müşterilerin yıllık gelirleri 15.000 dolar ile 137.000 dolar arasında değişmektedir.
- Harcama Puanı: Bu özellik müşterlerin yaptığı harcamlara göre verilen bir puandır. Bu özellik 1-100 arasında değişmektedir.

1.1 ÖN İŞLEMLER

Veri kümesi bir modele verilmeden önce ilk olarak bir ön işleme adımından geçirilir. Bu adımda her özellik için eksik veya hatalı örnek olmadığı gözlemlenmiştir. Bu sebepten ötürü veri kümesinde herhangi bir eksiltme işlemi yapılmaz. Cinsiyet verisinin yazı formatında olmasından ötürü model işleyemez dolayısıyla bu özellik kadın 1 erkek 0 olacak şekilde güncellenir. Verilerin farklı aralıklarda olmasının modele zararı olmasından ötürü normalizasyon işlemi yapılmıştır. Normalizasyon işlemi min-max normalizasyon ile gerçekleştirilmiştir. Bu işlem sonucunda bütün veriler 0-1 aralığına yerleşmiştir.

1.2 ÖZELLIK SEÇIMI

Ön işlem uygulandıktan sonra 0-1 aralığına yerleştirilen veri kümesi üzerinde korelasyon incelenir.

Korelasyon matrisinde de görülebileceği üzere yaş ve harcama puanı özellikleri arasında 0.3 puanlık bir negatif korelasyon bulunmaktadır. Ancak bu iki özellik dışında herhangi bir ikili arasında belirgin pozitif veya negatif bir korelasyon olmadığından ötürü en düşük korelasyona sahip iki özellik seçilmiştir. Bu özellikler gelir ve harcama puanıdır. Seçilen özellikler için farklı miktarda merkeze sahip k ortalamalar kümeleme yöntemi eğitilmiştir. K sayısı 1-20 arasında farklı değerler içermektedir ve her farklı k değeri için küme içi hata hesaplanmaktadır. Bu hataya göre grafik oluşturulmuştur. Üretilen grafik ile elbow yöntemi kullanılarak küme merkez miktarı seçilmiştir. Seçilen küme merkez miktarları ile oluşturulan modeller görselleştirilip en uygun model seçilmiştir.

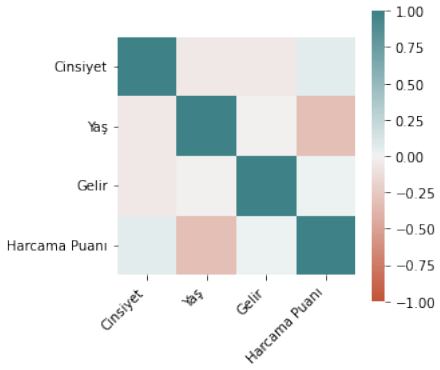

Figure 1: Korelasyon Matrisi

2 UYGULAMA

Bu bölümde k değerleri 1-20 arasında değişen değerler olacak şekilde seçilmiştir. Küme merkezlerinin ilklendirilmesi için KMeans++ Arthur & Vassilvitskii (2007) algoritması kullanılmıştır. Bu algoritma küme merkezlerinin başlangıçta birbirine en uzak olacak şekilde yerleştirilmesini sağlayarak daha başarılı kümeleme sonucu üretmektedir.

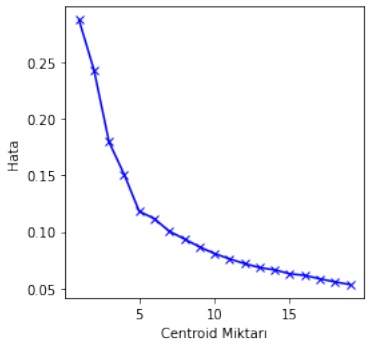

Figure 2: Merkez Miktarı ve Hata İlişkisi

Yukarıdaki grafik incelendiğinde kırılım miktarı k=4 noktasından sonra azalmaya başlıyor. k=6 noktasından itibaren grafiğin eğim miktarı çok azalmakta. Bu durum göz önünde bulundurulduğunda merkez miktarı olarak 5 ve 6 seçilmiştir. Aşağıdaki şekillerde k=5 ve 6 için kümeleme sonuçları her bir küme farklı renkte gösterilmiştir.

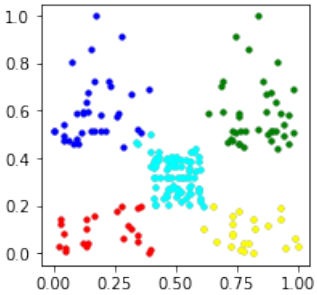

Figure 3: K = 5 iken Kümeleme Sonucu

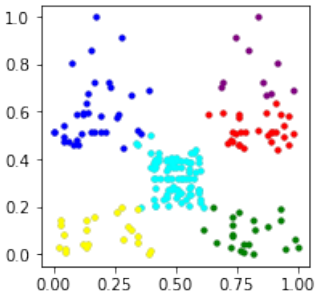

Figure 4: K = 6 iken Kümeleme Sonucu

3 SONUÇ

Veri kümesinin genel yerleşimine incelendiğinde veri kümesi 5 ana kümeden oluşmaktadır. Şekil 3 ve 4'te gözlemlenebildiği gibi en iyi sonuç k = 5 iken elde edilmiştir.

REFERENCES

David Arthur and Sergei Vassilvitskii. k-means++: the advantages of careful seeding. In *ACM-SIAM Symposium on Discrete Algorithms*, 2007.

V. Arul, Ashutosh Kumar, and Aman Agarwal. Segmenting mall customers data to improve business into higher target using k-means clustering. In *2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*, pp. 1602–1604, 2021. doi: 10.1109/ICAC3N53548.2021.9725630.

Tushar Kansal, Suraj Bahuguna, Vishal Singh, and Tanupriya Choudhury. Customer segmentation using k-means clustering. In *2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)*, pp. 135–139, 2018. doi: 10.1109/CTEMS.2018.8769171.

M Pavithra, Ayushman Prashar, and Abirami. Maximizing strategy in customer segmentation using different clustering techniques. In *2022 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES)*, volume 1, pp. 481–485, 2022. doi: 10.1109/SPICES52834.2022.9774200.

V. Chandra Shekhar Rao, Ishwarya Modika, and Niranjan Polala. Customer segmentation with k-means++ as initialization algorithm. In *2022 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI)*, pp. 1–7, 2022. doi: 10.1109/ACCAI53970.2022.9752591.