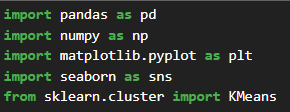
**Customer Segmentation Using K-Means Clustering**

**Nisa Çetin**

**Özet (Abstract)** Bu çalışmada,veri madenciliği yöntemlerinden biri olan K-means kümeleme yöntemi kullanılarak müşteri segmentasyonu gerçekleştirilmiştir. Analizlerde, Kaggle platformundan sağlanan veri seti ile veri ön işleme adımı gerçekleştirilmiş.Ardından, Elbow Yöntemi kullanılarak küme sayısı belirlenmiş. Makine öğrenmesi yöntemlerinden olan K-Means algoritması kullanılarak veriler analiz edilmiştir.Referans olarak alınan akademik bir çalışmadaki yöntemler, aynı veri setiyle tekrar uygulanmıştır. Elde edilen sonuçlar,benzer davranış gösteren müşterileri gruplandırarak kişiselleştirilmiş kampanya oluşturmak gibi çeşitli pazarlama stratejileri geliştirmeyi bu sayede müşteri memnuniyetini ve satışları arttırmaya mümkün kılmaktadır.

1. **Giriş (Introduction)**Veri madenciliği, büyük hacimli veri kümelerinden anlamlı bilgilerin ve gizli desenlerin çıkarılması sürecidir.K-Means Kümeleme,denetimsiz bir makine öğrenmesi yöntemi tercih edilmiştir.Literatürde akedemik çalışmalardan faydalanılmış ve özellikle müşteri verileri üzerinde K-Means kümeleme uygulandığı bir çalışma referans alınarak aynı yöntemler aynı veri seti üzerinde uygulanmıştır.Bu analizde uygulanan yöntem, verileri daha önceden tanımlanmış etiket olmadan benzerliklerine göre gruplandırır.Bu çalışmada, bir alışveriş merkezi müşterilerinin yıllık gelir ve harcama skorları analiz edilerek benzer davranışlara sahip müşteri grupları belirlenmiştir. Bu analiz sonucunda, firmaların hedef odaklı özel kampanyalar geliştirebilmesini ile pazarlama stratejileri optimize etmesi için analiz süreçlerinde kullanılabilecek somut bir referans sağlamaktadır.
2. **Yöntem (Methodology)**
3. Uygulamayı gerçekleştirmek için gerekli Python kütüphane ve modülleri içeri aktarılmıştır.Bunları şu şekilde sıralayabiliriz: pandas ,numpy, matpoltlib.pyplot, seaborn, sklearn.cluster.
4. Daha sonra veri setimizi projeye aktarılmıştır.Veri ön işleme adımında null(eksik) değerleri kontrol edilmektedir.Eğer null değer varsa o veriyi silinebilir,mod değeri ile veya ortalama değer ile doldurulabilirdi.Bu çalışmada kullanılan veri setinde null değer yoktur.
5. Elbow methodu kullanarak uygun küme sayısını belirlenmiştir.Belirlenen küme sayısıyla K-Means kümeleme yöntemini kullanarak analizimi gerçekleştirilmiştir. Sonuç olarak yapılan analizi görselleştirmek için seaborn,matpoltlib.pyplot kütüphanelerini kullanmıştır.
   1. **Veri Seti** Çalışmada Kaggle platformundan sağlanan bir alışveriş merkezine ait veri seti kullanılmıştır.

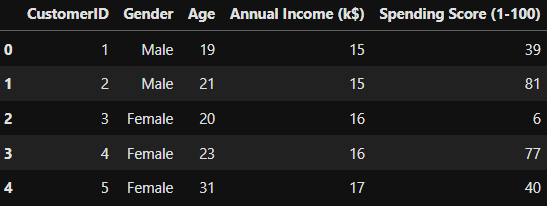
İlk olarak gerekli kütüphane ve modülleri içeri aktarılmaktadır.



Daha sonra veri seti projeye aktarılmaktadır.

2

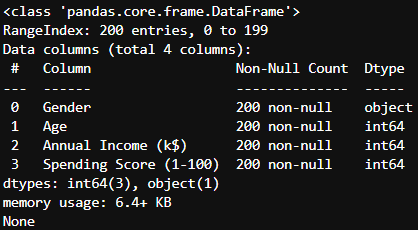
Veri setinin ilk 5 satırı aşağıda görülmektedir.



* 1. **Veri Önişleme**

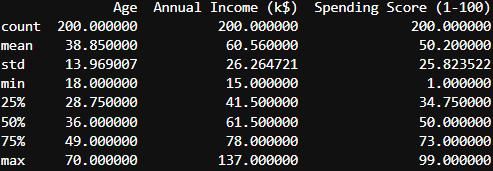
Veri setinin kaç satır ve sütundan oluşmakta olduğu,kolon adları,veri tipleri ve null veri sayısı aşağıda görünmektedir.Görselde görüldüğü üzere veri setinde null değer bulunmamaktadır.

5



Veri setinin ortalaması,standart sapması,minimum.maximum ve çeyreklik bilgileri aşağıdaki görselde yer almaktadır.

8

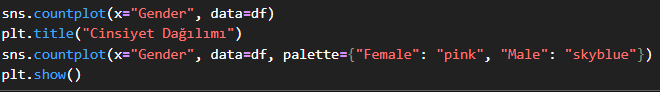


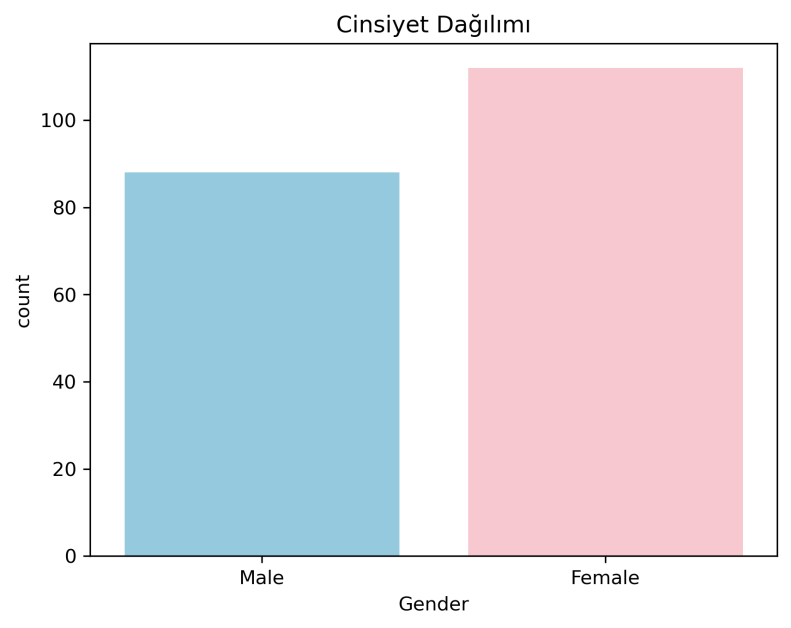
K-Means kümeleme için anlam ifade etmeyen CustomerID sütunu silinmektedir.

7

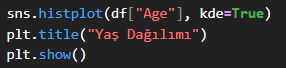
* 1. **Kullanılan Algoritmalar**

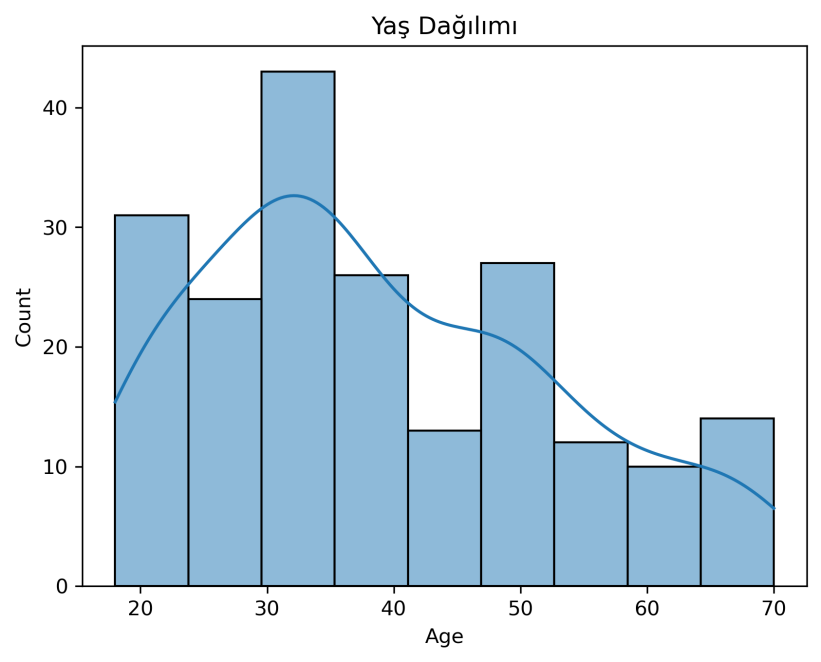
Aşağıdaki grafik,müşterilerin cinsiyet dağılımını göstermektedir.X ekseninde cinsiyet bilgisi(kadın,erkek) , y ekseninde her cinsiyete ait kişi sayıları yer almaktadır. Dağılımı incelendiğinde ,kadınların sayısı 100’den fazla olup erkeklerinki 80’e yakın olduğu gözlemlenmektedir.Buna göre alışveriş merkezindeki müşteri kitlesi daha çok kadınlardan oluşmaktadır.



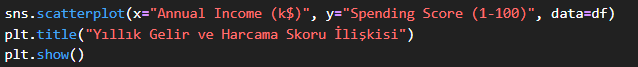


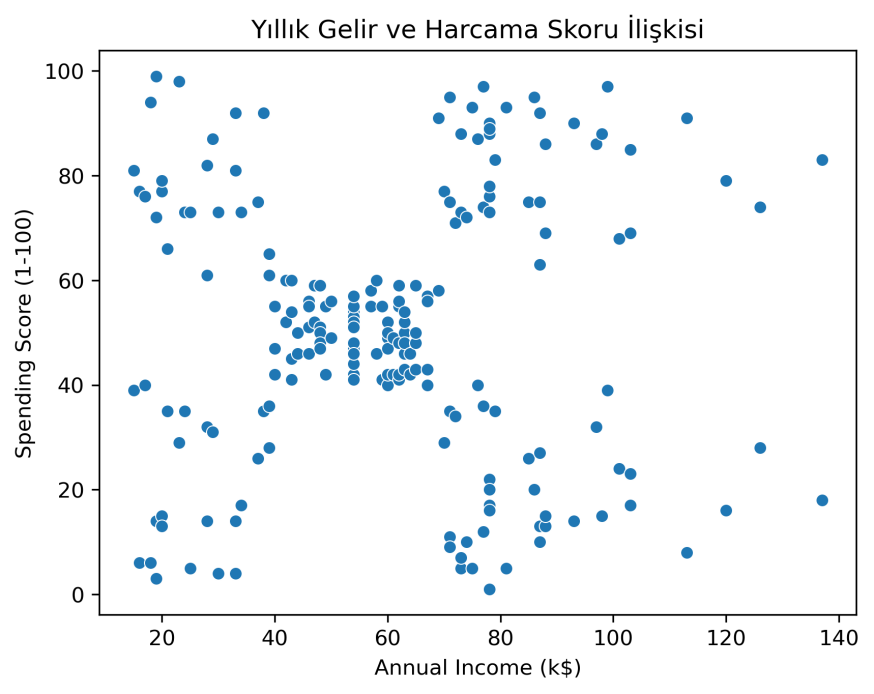
Aşağıdaki grafik,müşterilerin yaş dağılımını göstermektedir.X ekseninde yaş aralıkları, y ekseninde her yaş grubuna ait müşteri sayıları yer almaktadır.Dağılım incelendiğinde,alışveriş merkezinin müşteri kitlesi yoğunluklu olarak 30-35 yaş aralığındaki kişilerden oluşmaktadır.Bunu 20-30 yaş grubundaki müşteriler takip etmektedir.





Aşağıdaki grafik,yıllık gelir ile harcama skoru arasındaki ilişkiyi göstermektedir.X ekseninde yıllık gelir(1000 dolar cinsinden) ,y ekseninde ise harcama skoru(1-100 arası bir ölçekte) yer almaktadır.Dağılım incelendiğinde,müşteriler belirli kümeler oluşturduğu görülmektedir.Bu kümeler, benzer davranış gösteren müşterileri ifade etmektedir.Grafik incelendiğinde gelir düzeyinin harcama miktarını belirlemede tek başına yeterli bir ölçüt olmadığı görünüyor.Bu durumun kanıtı olarak yüksek gelire rağmen düşük harcama yapan ve düşük gelire rağmen yüksek harcama yapan müşteri kümelerini görülebilmektir.Pazarlama açısından en kıymetli müşteriler yüksek harcamaya sahip yüksek gelirli müşterilerdir.

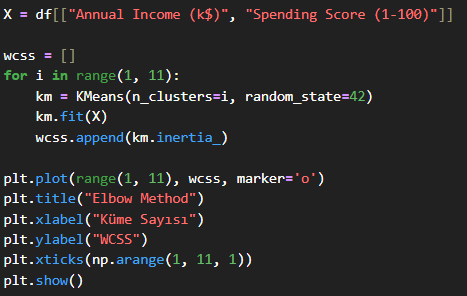


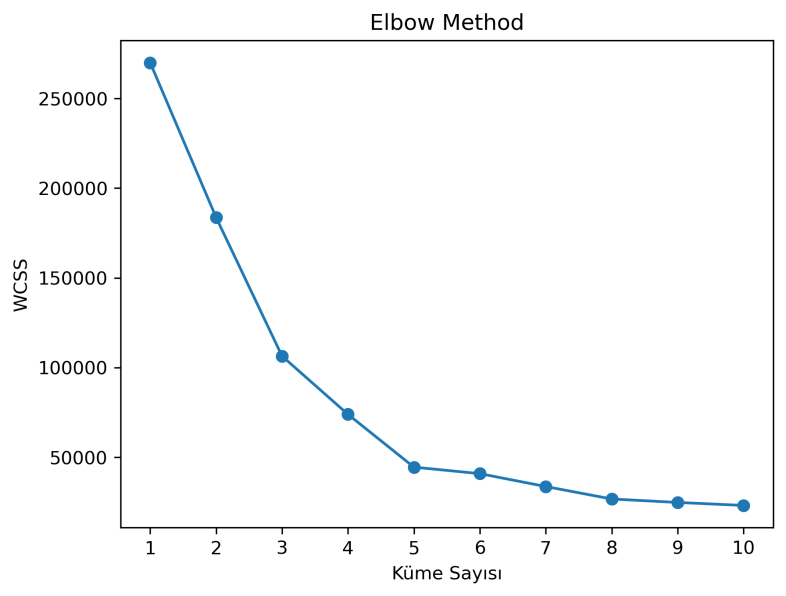


1. Means algoritmasına girdi olarak kullanmak için yıllık gelir(1000 dolar cinsinden) ve harcama skoru(1-100 arasında bir ölçek) iki boyutlu bir X değişkenine atanmaktadır.

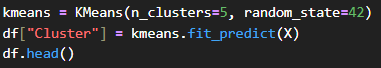
16

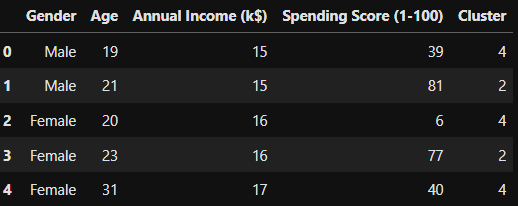
Aşağıdaki grafikte kullanılan elbow method(dirsek yöntemi), gözetimsiz öğrenme algoritmlarında optimum küme sayısını bulmak için kullanılmaktadır.Yöntem kapsamında, küme sayısı 1’den başlayarak 10’a kadar arttıralarak bozulmanın minimum olduğu sayı küme sayısı olarak belirlenmiştir.Elbow yöntemine göre, bozulmanın minimum olduğu grafiğin dirsek yaptığı küme sayısı 5 olduğundan küme sayısı 5 olarak belirlenmiştir.



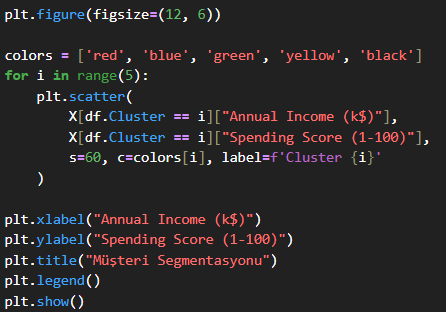


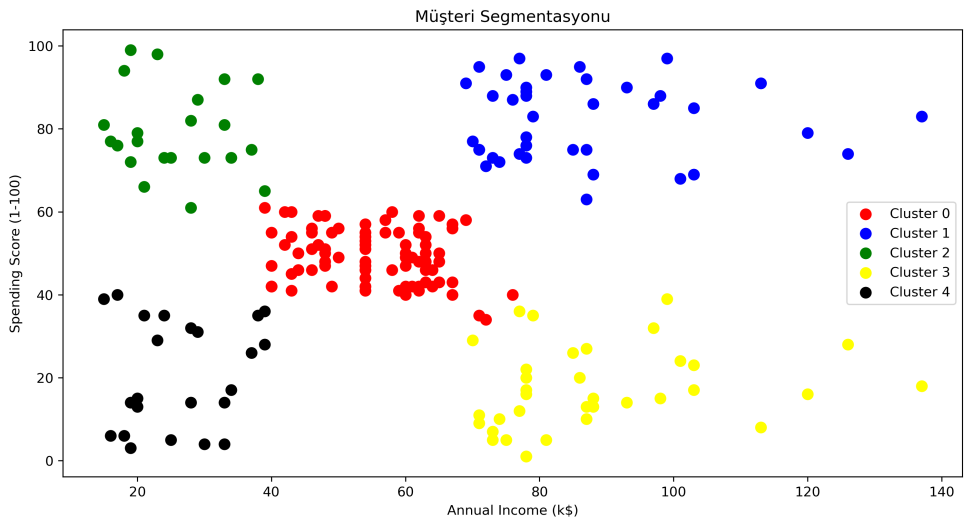
Aşağıdaki kod bloğunda, Elbow Method(Dirsek Yöntemi) ile belirlenen optimum küme sayısı olan 5’e göre K-Means algoritması uygulanmıştır.KMeans(n\_clusters=5, random\_state=42) ifadesiyle her çalıştırmada sabit rastgelelik oluşturularak model tanımlanmıştır.fit\_predict fonksiyonu kullanıralarak model eğitimi gerçekleştirilmiştir ve her bir gözlem için küme atanmıştır. Oluşan kümeler Cluster adlı yeni bir sütunda veri çerçevisine eklenmiştir.Aşağıdaki tablo, veri çerçevesinin ilk 5 satırını göstermektedir.Tablo Gender(cinsiyet), Age(yaş), Annual Income(Yıllık gelir),Spending Score(Harcama Skoru),ait olunan Cluster(küme) bilgilerini içermektedir.





Aşağıdaki kod bloğunda, K-Means algoritması ile oluşturulan kümeler görselleştirilmiştir.X ekseninde yıllık gelir(1000 dolar cinsinden) ,y ekseninde ise harcama skoru(1-100 arası bir ölçek) yer almaktadır.Her küme(Cluster 0-4) farklı bir renk ile ifade edilmiştir.Grafikteki dağılım incelendiğinde, harcama davranışlarının yalnızca gelir düzeyine bağlı olmadığı anlaşılmaktadır. Nitekim, yüksek gelir grubunda yer almasına rağmen düşük harcama yapan bireylerin yanı sıra, düşük gelire sahip olmasına rağmen yüksek harcama gerçekleştiren müşteriler de mevcuttur. Bu durum, müşteri segmentasyonunda yalnızca gelir kriterine odaklanmanın yetersiz kalabileceğini göstermektedir. Alışveriş merkezi açısından en değerli müşteri grubu ise, hem yüksek gelir elde eden hem de yüksek harcama eğiliminde olan bireylerden oluşmaktadır.





**3. Sonuçlar (Results)** Bu çalışmada,bir alışveriş merkezi verileri üzerinde K-Means algoritması uygulanıp müşteriler 5 segmente ayrılmıştır.Elbow Method’una göre optimum küme sayısı 5 olarak belirlenmiştir.Belirlenen küme sayısına göre model eğitilmiştir.Segmentasyon işlemi sonucunda kümeler net bir şekilde ayrışmaktadır.En yoğun gözlemlenen küme,ortalama gelire sahip ortalama harcama yapan müşterilerin bulunduğu kümedir.Bununla birlikte, yüksek geliri olmasına rağmen düşük harcama yapan müşterilerin olduğu ve bunun tam aksi davranış gösteren düşük gelire sahip yüksek harcama yapan müşterilerde bulunmaktadır.Bu durum,harcama skorunun yalnızca gelir düzeyiyle ilişkili olmadığını göstermektedir.

* Cluster 0:Ortalama harcamaya sahip ortalama gelir düzeyine sahip müşteriler
* Cluster 1:Yüksek harcama düzeyine sahip yüksek gelirli müşteriler
* Cluster 2:Yüksek harcama düzeyine sahip düşük gelirli müşteriler
* Cluster 3:Düşük harcama düzeyine sahip yüksek gelirli müşteriler
* Cluster 4:Düşük harcama düzeyine sahip düşük gelirli müşteriler

Referans olarak incelenen çalışmada da Elbow yöntemi kullanılarak optimum küme sayısı 5 olarak belirlenmiş ve segmentasyon işlemi aynı değişkenler olan yıllık gelir ve harcama skoru üzerinden gerçekleştirilmiştir. Her iki çalışmada da, yüksek gelirli ve yüksek harcama düzeyine sahip müşterilerin (Cluster 1) pazarlama açısından en değerli müşteri grubu olduğu sonucuna ulaşılmıştır. Ancak bu çalışmada, harcama davranışlarının yalnızca gelir düzeyiyle açıklanamayacağı, bireysel tercihler ve alışkanlıkların da önemli rol oynadığı vurgulanırken; referans alınan çalışmada bu yönde bir değerlendirmeye yer verilmemiştir.

**4. Tartışma ve Öneriler (Discussion and Recommendations)**

K-Means Kümeleme ile yapılan analiz sonucunda, müşterilerin gelir ve harcama düzeylerine göre anlamlı gruplara ayrıldığı ve bu segmentlerin pazarlama stratejileri açısından değerli bilgiler sunduğu belirlenmiştir.K**üme 1**’deki yüksek gelirli ve çok harcayan müşteriler, özel ayrıcalıklar ve sadakat ödülleriyle desteklenebilir. **Küme 3** için ürün tanıtımı ve fırsat bildirimleri öne çıkarılabilir. **Küme 2**'de yer alan düşük gelirli ama yüksek harcayan kitleye taksitli ödeme seçenekleri sunulabilir.**Küme 4**, düşük potansiyeli nedeniyle düşük bütçeli kampanyalarla hedeflenmelidir.Her müşteri segmenti için özel olarak geliştirilecek stratejiler, pazarlama çalışmalarının verimliliğini önemli ölçüde artırabilir. Analizler, yalnızca gelir ve harcama skorlarının müşteri davranışlarını tam olarak açıklamakta yetersiz kaldığını göstermektedir. Bazı gruplarda, yüksek gelire sahip bireylerin beklenenden düşük harcama eğilimi sergilediği görülmüştür. Bu durum, harcama alışkanlıklarının tek bir değişkene indirgenemeyeceğini ortaya koymaktadır. Müşteri segmentasyonunun daha sağlıklı yapılabilmesi için veri setine **yaş, meslek, medeni durum, eğitim seviyesi** gibi demografik değişkenlerin eklenmesi önerilmektedir. Aynı zamanda, **alışveriş sıklığı** gibi davranışsal veriler de gruplamayı daha isabetli hâle getirebilir. Gelecekteki çalışmalarda zaman içinde değişen müşteri davranışlarını modellemek adına zaman serisi analizlerinden veya esnek yapıdaki dinamik kümeleme algoritmalarından faydalanılabilir.

**5. Kaynakça (References)**

1. Mallik, S. (2019). Customer segmentation using K-means clustering. Kaggle. [https://www.kaggle.com/code/sanikamal/customer-segmentation](https://www.kaggle.com/code/sanikamal/customer-segmentation" \t "_new)
2. Choudhary, V. J. (2019). Mall customer segmentation data. Kaggle. [https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial](https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial" \t "_new)
3. Han, J., Kamber, M., & Pei, J. (2011). Data mining: Concepts and techniques (3rd ed.). Morgan Kaufmann Publishers.
4. Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern Recognition Letters, 31(8), 651–666. <https://doi.org/10.1016/j.patrec.2009.09.011>
5. Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., ... & Steinberg, D. (2008). Top 10 algorithms in data mining. Knowledge and Information Systems, 14(1), 1–37. <https://doi.org/10.1007/s10115-007-0114-2>