T.C.

BEYKENT ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

BİLGİSAYAR MÜHENDİSLİĞİ BİLİM DALI

# (KATEGORİK VE KATEGORİK OLMAYAN VERİLERDEN OLUŞAN VERİ SETLERİ İÇİN K-ORTALAMA TABANLI BİR YAKLAŞIM)

(Yüksek Lisans Tezi)

Tezi Hazırlayan: **Mustafa DEMİRKAN**

İSTANBUL, 2014

T.C.

BEYKENT ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

BİLGİSAYAR MÜHENDİSLİĞİ BİLİM DALI

# (KATEGORİK VE KATEGORİK OLMAYAN VERİLERDEN OLUŞAN VERİ SETLERİ İÇİN K-ORTALAMA TABANLI BİR YAKLAŞIM)

(Yüksek Lisans Tezi)

Tezi Hazırlayan:

**Mustafa DEMİRKAN**

Öğrenci No:

110820033

Danışman:

Doç.Dr. Gökhan SİLAHTAROĞLU

İSTANBUL, 2014

## YEMİN METNİ

Yüksek lisans tezi olarak sunduğum **(Kategorik ve Kategorik Olmayan Verilerden Oluşan Veri Setleri İçin K-Ortalama Tabanlı Bir Yaklaşım)** başlıklı bu çalışmanın, bilimsel ahlak ve geleneklere uygun şekilde tarafımdan yazıldığını, yararlandığım eserlerin tamamının kaynaklarda gösterildiğini ve çalışmamın içinde kullanıldıkları her yerde bunlara atıf yapıldığını belirtir ve bunu onurumla doğrularım. …./…./2014

(imza)

Mustafa DEMİRKAN

T.C.
BEYKENT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

## YÜKSEK LİSANS TEZ SAVUNMA SINAVI SONUÇ TUTANAĞI

**Beykent Üniversitesi**
**Fen Bilimleri Enstitüsü Müdürlüğü'ne,**

Aşağıda tez adı belirtilen yüksek lisans öğrencisi110820033.no'lu Mustafa DEMİRKAN'in 05/03/2014 tarihinde yapılan tez savunma sınavı[1] sonucunda 75. dakika süreyle sunduğu ve savunduğu tezi hakkında[2] oybirliğiyle, KABUL kararı verilmiştir.

Bilgilerinize saygılarımızla arz ederiz.

---

**Anabilim Dalı** : BİLGİSAYAR MÜHENDİSLİĞİ
**Programı** : BİLGİSAYAR MÜHENDİSLİĞİ
**Tez Başlığı**[3] : Kategorik ve Kategorik Olmayan Verilerden Oluşan Veri Setleri için K-Ortalama Tabanlı Bir Yaklaşım

---

| Tez Sınav Jürisi | Öğretim Üyesi | İmza |
|---|---|---|
| **Danışman** | **: Doç. Dr. Gökhan Silahtaroğlu** | |
| Üye | : Yrd. Doç. Dr. Zeynep ALTAN | |
| Üye | : Yrd. Doç. Dr. Ediz ŞAYKOL | |

[1] Jüri üyeleri söz konusu tezin kendilerine teslim edildiği tarihten itibaren en geç bir ay içinde toplanarak öğrenciyi tez savunma sınavına alır. Belirlenen günde yapılamayan jüri toplantısı, katılanların hazırladığı bir tutanakla enstitü yönetimine bildirilir. Bu durumda jüri en geç onbeş gün içinde toplanarak adayı tez savunma sınavına alır. Tez savunma sınav süresi en az 45 dakikadır. Yüksek lisans tez savunma sınavı, tez çalışmasının sunulması ve bunu izleyen soru-yanıt bölümlerinden oluşur ve dinleyiciye açıktır. (Beykent Lisansüstü eğitim ve Öğretim Yönetmeliği-Madde30-3)

[2] Tez sınavının tamamlanmasından sonra jüri, tez hakkında "kabul", "düzeltme" veya "red" kararı verir. Jüri başkanı, jüri üyelerince imzalanmış sınav tutanağını, tez sınavını izleyen üç gün içinde ilgili enstitü yönetimine teslim eder. Tezi başarısız bulunan öğrencinin Enstitü ile ilişiği kesilir. Tezi hakkında düzeltme kararı verilen öğrenci en geç üç ay içinde gerekli düzeltmeleri yaparak ve yönetmelikte belirtilen usullere uygun olarak tezini aynı jüri önünde yeniden savunur. Bu savunma sınavında da tezi kabul edilmeyen öğrencinin enstitü ile ilişiği kesilir. (Beykent Lisansüstü eğitim ve Öğretim Yönetmeliği-Madde30-4)

[3] İleride doğabilecek aksaklıkların engellenmesi için tezin başlığının yazılması gerekmektedir.

# (KATEGORİK VE KATEGORİK OLMAYAN VERİLERDEN OLUŞAN VERİ SETLERİ İÇİN K-ORTALAMA TABANLI BİR YAKLAŞIM)

Tezi Hazırlayan:**Mustafa DEMİRKAN**

## ÖZET

Günümüzde ilerleyen teknoloji ve bilgiye erişebilirliğin artması ile birlikte kurum ve kuruluşlar bu bilgilerden faydalanarak üretim, satış ve kaynakların yönetilmesi gibi konularda daha doğru, düzgün ve yeni kararlar verebilmek için önemli bir yol olarak veri medenciliğinden yararlanmanın faydalı olabileceğini düşünmüşlerdir. Bununla birlikte veri madenciliğinde yeni yöntemler, algoritmalar, düşünceler gelişmiş ve veri madenciliği sektörlerin olmazsa olmazı haline gelmiştir.

Veri madenciliğinin yapı taşlarını veri tabanları, istatistik, görselleştirme, karar verme mekanizması ve makine öğrenmesi alanları oluşturur. Hepimizin çok duyduğu fakat çok yakından bilmediği makine öğrenmesi veri madenciliğinde vazgeçilemeyecek bir unsurdur çünkü makine öğrenmesi verilen bir problemi elde olan verilere göre şekillenen, bilgisayar algoritmalarının hepsini kapsayan bir yöntemdir. Bu tez çalışmasında veri madenciliği bölümlerinde olan makine öğrenmesi içerisindeki K-Means algoritması ve Jaccard benzerlik ölçütünden yaralanarak daha farklı bir çözüm üretmektir. Veri tabanlarında elde ettiğimiz veriler her zaman tam ihtiyaca göre olan veriler olmayabilir. Bu verilerin içerinde gereksiz veriler, eksik veriler, uyumsuz veri ölçekleri, kategorik ve kategorik olmayan veriler mevcuttur. Burada gereksiz veriler veri tabanı içerisinden çıkartılabilir, eksik veriler yok sayılabilir ya da gereksiz veriler gibi veri tanımından çıkartılabilir, verilerin arasındaki ölçeklemeler normalize edilebilir ve sonuç olarak bu sorunlar kolayca düzeltilebilir fakat her zaman kategorik ve kategorik olmayan verilerin bulunduğu veri tabanlarında makine öğrenmesinin algoritmaları kullanıldığı zaman özellikle kümelemelerde sorunlar yaşanmaktadır. Bir algoritma kategorik verilerde başarılı sonuçlar verirken kategorik olmayan verilerde başarısız, bir diğeri de kategorik

olamayan verilerde başarısız sonuç verirken kategorik verilerde başarılı sonuçlar vermektedir. Bunun anlamı normal şartlar altında kategorik veriler kategorik veriler ile kategorik olmayan veriler ise kategorik olmayan veriler ile kümelenebilmektedir.

Fakat bu iki tip veriler birlikte kümelenebilmelidirler. Bu hipotezde kategorik ve kategorik olmayan karışık veri kümeleri alınarak, ne oranda anlamlı, mantıklı ve doğru bir sonuç kümesi oluşturulabileceğine yönelik bir çalışma ele alınmıştır. Tezin içeriğinde de kategorik olmayan veri kümeleri için K-Means algoritması, kategorik veri kümeleri içinse Jaccard benzerlik ölçütünden yararlanılmış ve iki kümeleme yöntemi birleştirilip örnek veri setleri kullanılarak yepyeni anlamlı bir sonuç kümesi, kümeleme yöntemi oluşturulması anlatılmaya çalışılmıştır.

**Anahtar Kelimeler :** K-Means, Jaccard Benzerlik Ölçütü, Kategorik Veri, Kategorik Olmayan Veri

# (K-MEANS BASED APPROACH FOR CATEGORICAL AND NON CATEGORICAL DATA SETS)

Presented by: **Mustafa DEMİRKAN**

## ABSTRACT

Nowadays, corporations and enterprises are used data mining to increase sales and profits through reaching data. Therefore new algorithms and methods are developed in data minig. The machine learning is indispensable component of data mining. In machine learning, there are a lot of algorithms for classification, clustering etc. One of the well known algorithm is K-Means algorithm in machine learning. In K-Means algorithm non categorical data sets are clustering. However in real world categorical and non categorical data sets are nested. The aim of thesis is to develop K-Means algorithm which does clusters categorical and non categorical data sets together. To do this, Jaccard similarity measure is embeded inside K-Means algorithm instead of Euclid for categorical part of data sets then two algorithms are combined each other clustering categorical and non categorical data sets.

**Key Words:** K-Means, Jaccard Similarity Measure, categorical data, non categorical data

Eşim **Fatma DOLMA DEMİRKAN**'a

# İÇİNDEKİLER

**Sayfa No:**

# TABLOLAR LİSTESİ

# ŞEKİLLER LİSTESİ

## KISALTMALAR

**m1** = merkez1

**m2** = merkez2

**m3** = merkez3

**mes** = mesafe fonksiyonu

**ben** = benzerlik

**S1** = Sonuç 1

**S2** = Sonuç 2

**S3** = Sonuç 3

**benj** = Jaccard benzerlik ölçütü benzerliği

# 1 GİRİŞ

Temelde yapay zekanın bir alt dalı olan Makine Öğrenimi robotikten bilgisayar görüsüne veri madenciliğinden bilgisayar oyunlarına kadar bir çok alanda kullanılır. Otomotiv sektöründe kullanılan robotlar, para çektiğimiz ATM makineleri, parmak tanıma ve yüz tanıma sistemleri makine öğrenimine örnek verilebilir.

Makine öğreniminin önemli konularından bir tanesi de kümelemedir. Kümeleme ile ilgili BIRCH, SLINK, CLUCDUH, CHAMELON [14], K-Means gibi bilindik bir çok algoritma bulunmaktadır. Bahsedilen algoritmalardan K-Means algoritması ise en iyi bilinen ve en çok kullanılan algoritmalardan bir tanesidir. K-Means algoritması öklit tabanlı bir algoritmadır ve yaş, boy, hız gibi sayısal verilerde doğru sonuç alınmaktadır. Eğer verilerde renk, cinsiyet, isim gibi parametreler varsa K-Means algoritmasında doğru sonuçlar elde edilmez. Bu tez çalışmasında K-Means algoritmasının sayısal olmayan verilerle de çalışması için bir yöntem önerilmiştir. Ayrıca K-Means algoritmasının karışık veri kümelerinde nasıl çalışacağı anlatılmaya çalışılmıştır.

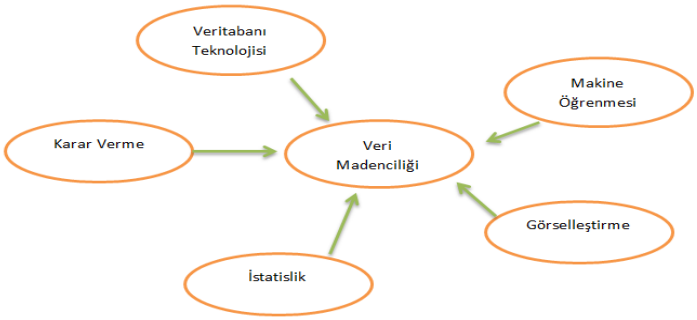
Bölüm 2' de Makine öğrenimi ve makine öğreniminin uygulama alanları ve öğrenme yöntemlerinden bahsedilmiştir.

Bölüm 3 İki ana başlıktan oluşmaktadır.

1. K-Means algoritmasının sayısal olmayan veriler üzerinde nasıl çalışacağına dair yöntem geliştirilmiştir. Bunun için K-Means algoritmasının içine Jaccard benzerlik ölçütü koyulmuştur.
2. K-Means algoritmasının sayısal ve sayısal olmayan karışık veri kümeleri ile çalışmasına yönelik bir yöntem sunulmuştur.

## 2 MAKİNE ÖĞRENİMİ TEMEL BİLGİLERİ

Teknolojinin gelişmesi, bilgisayar sistemlerinin ve internetin en gelişmiş kurumlardan sıradan insanların eline ulaşması ile birlikte veriye ulaşılabilirlikte artmış, yapılan herşey kayıt altına alınmaya başlanmıştır. Yaptığımız sıradan alışverişlerden, bankacılık işlemlerine kadar her şey veri tabanına kaydedilmektedir. Hatta büyük şehirlerde her vatandaşın görüntüsü ortalama otuz sekiz defa mobese kameraları tarafından veri tabanına kaydedilmektedir. Ancak veriler tek başlarına bir anlam ifade etmezler. Verilerin birşey ifade edebilmesi için içindeki mevcut olan bilgiyi su yüzüne çıkarmak gerekir. Bu bilgilerin ise su yüzüne çıkması için veri madenciliğine ve en çokta veri madenciliğiniz bölümü olan makine öğrenmesine ihtiyaç vardır. Makine öğrenimi (Machine Learning) mevcut olan verileri alarak onları çeşitli bilgisayar algoritmaları ile işleyip sonuç bulma işlemidir. Aslında makine öğreniminde nihai amaç makinelerin insanlar gibi düşünmesini sağlamaktır. Bu da bilgi, öğrenme ve tecrübe ile olur. Makine öğreniminde asıl olarak sistemler nasıl otomatik olarak bilgi, tecrübesi ve öğrenmesi artacak şekilde programlanabilir sorusunun cevabı aranmaktadır. Öğrenme sonucunda algoritmalar yeni gelen verilerden faydalanarak kendini geliştirme, yeni gelen verilerin tecrübesi ile aynı algoritmasının çalışması sonucu farklı, adapte olmuş sonuçlar bulabilir. Makine öğrenimi verinin türüne göre ikiye ayrılır. Denetimli (supervised) ve Denetimsiz (unsupervised) öğrenmedir. Makine öğreniminde üç tür yaklaşım vardır. Bu yaklaşımlar kestirim (estimation), tahmin (prediction) ve sınıflandırma (classification)dır. Öğrenme algoritmaları temel olarak dört gruba ayrılır. Bunlar Hebb, Delta, Hopfield ve Kohonen öğrenme algoritmalarıdır.




**Şekil 1. Makine öğrenmesinin veri madenciliği ile ilşkisi**

## 2.1 Makine Öğreniminin Uygulama Alanları

Makine öğrenimi temelde yapay zekanın bir alt dalıdır. Makine öğrenimi robotikten bilgisayar görüsüne veri madenciliğinden bilgisayar oyunlarına kadar bir çok alanda kullanılır. Otomotiv sektöründe kullanılan robotlar, para çektiğimiz ATM makinaları, parmak tanıma ve yüz tanıma sistemleri makine öğrenimine örnek verilebilirler. Makine öğreniminde algoritmalar denetimli ve denetimsiz olarak iki kısıma ayrılabilirler.

### 2.1.1 Denetimli (Supervised) Öğrenme

Denetimli öğrenmenin değişkenleri iki gruba ayrılır. Bunlar açıklayıcı (explanatory) ve bağımlı (dependency) değişkenlerdir. Analizin hedefi açıklayıcı değişkenler ile bağımlı değişkenler arasındaki ilişkiyi belirlemektir. Veri madenciliğindeki teknikleri uygulamak için veri seti içindeki bağımlı değişkenlerin büyük bir kısmının bilinmesi gerekmektedir.

#### 2.1.1.1 Kestirim (Estimation)

Bir veri seti incelenirken o veri seti için bir fonksiyon ya da bir sonuç üretmeyi kestirmektir. Örneğin bir veri seti içerisinde k = 1 değerine karşılık m = 2, k = 2 değerine karşılık m = 4 ve örüntünün sonunda k = 20 değerine karşılık m = 400 değeri geldiği zaman m = $k^2$ eşitliği olduğu kestirilir. İşte eşitliği bulmak bir kestirimdir.

**Tablo 1. Kestirim tablosu**

| **K** | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| **M** | 2 | 4 | 9 | 16 | 25 | 36 |

#### 2.1.1.2 Tahmin (Prediction)

Kestirim sonucu oluşan fonksiyon ya da sonuca göre hareket ederek istenilen problemin ne olduğunu tahmin edebilmektir. Daha önce kestirim sonucu elde edilen fonksiyona göre k = 100 değerine göre m değerinin ne olduğu öğrenilmek istendiğinde m = $k^2$ eşitliğinden yola çıkılırsa m = $100^2$ = 10000 sonucu tahmin edilebilir.

**Tablo 2. Tahmin Tablosu**

| **K** | 1 | 2 | 3 | 4 | ... | 100 |
|---|---|---|---|---|---|---|
| **M** | 2 | 4 | 9 | 16 | ... | 10000 |

Ayrıca tahmin yöntemine en önemli örneklerden biri olarak Birliktelik Kuralı verilebilir.

#### 2.1.1.2.1 Birliktelik Kuralı /Örüntü Tanımlama

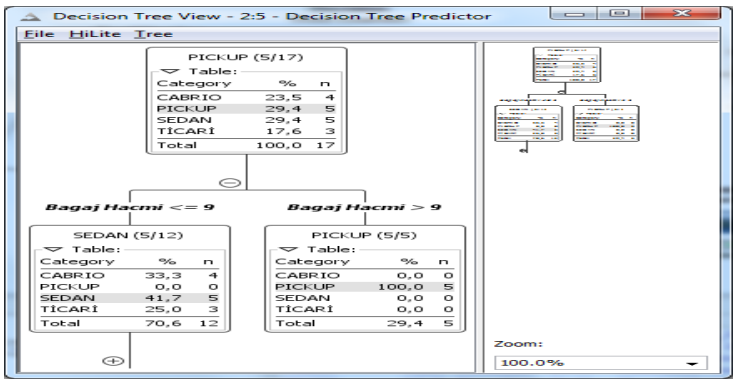
Veri tabanındaki kayıtların birbirleri ile olan ilişkileri belirlenmeye çalışılır. Burada amaç olayları takip ederek neye ihtiyaç duyulabileceğini saptayıp o yönde yönlendirme yapmaktır. Birliktelik kuralında en çok bilinen durum Pazar sepeti alışverişleridir. E-satış sitelerinde gezinti yapan insanların baktığı ürünler tıkladıkları linkler devamlı kayıt altında tutularak buradan çeşitli algoritmalar ile sonuçlar çıkartılır. Bu sonuçlara göre siteyi ziyaret eden ziyaretçinin gerçekten alıcımı yoksa değil mi ya da neye ihtiyaç duyulabileceği ortaya çıkartılıp sadece o siteyi ziyaret eden ziyaretçiye özel kampanyalar, duyurular, yönlendirmeler yapılabilir. Aslında basitçe belirtmek gerekirse birliktelik kuralında alışkanlar öğrenilip bu alışkanlıklara göre hareket edilir.

### 2.1.1.3 Sınıflandırma (Classification)

Sınıflandırma da veri tabanındaki gözle ayırt edilemeyen gizli bilgiler açığa çıkartılır. Bunu yapabilmek için veri setin bir kısmı eğitim için harcanır ve sınıflandırmada izlenecek yol ortaya çıkar bu yolla birlikte kararın ne yönde olacağı belirlenir. Sınıflandırma sonucunda bir iş nasıl yapılırsa optimum sonuç alınır, bir ürün satıldığında bunun yanında başka hangi ürünler müşteriye tükettirebilinir, işe personel alımı yapılacağında hangi özellikteki insanı işe almak daha doğrudur gibi sorulara cevap verilir. Sınıflandırmada karar ağaçları, yapay sinir ağları, mesafe ve istatistiği temel alan algoritmalar kullanılır.

**Tablo 3. Sınıflandırma için örnek veri kümesi**

| Araç | Müşteri Yaşı | Yakıt Tüketimi | Son Sürat | Yolcu Kapasitesi | Bagaj Hacmi |
|---|---|---|---|---|---|
| CABRIO | 20 | 6 | 258 | 5 | 2 |
| CABRIO | 25 | 7 | 308 | 2 | 5 |
| CABRIO | 27 | 10 | 373 | 3 | 6 |
| CABRIO | 28 | 10 | 345 | 3 | 2 |
| CABRIO | 18 | 10 | 252 | 2 | 9 |
| CABRIO | 24 | 8 | 399 | 4 | 1 |
| CABRIO | 30 | 12 | 390 | 4 | 2 |
| CABRIO | 29 | 9 | 282 | 2 | 1 |
| PICKUP | 58 | 7 | 221 | 6 | 14 |
| PICKUP | 62 | 6 | 235 | 7 | 11 |
| PICKUP | 34 | 9 | 209 | 7 | 14 |
| PICKUP | 34 | 12 | 211 | 5 | 13 |
| PICKUP | 19 | 7 | 217 | 6 | 11 |
| PICKUP | 66 | 10 | 204 | 5 | 14 |
| SEDAN | 62 | 9 | 219 | 7 | 6 |
| SEDAN | 18 | 7 | 198 | 5 | 7 |
| SEDAN | 49 | 7 | 232 | 7 | 7 |
| SEDAN | 65 | 11 | 206 | 7 | 6 |
| SEDAN | 22 | 7 | 246 | 5 | 5 |
| SEDAN | 58 | 11 | 225 | 7 | 7 |
| SEDAN | 61 | 9 | 239 | 5 | 6 |
| TİCARİ | 44 | 3 | 178 | 10 | 5 |
| TİCARİ | 62 | 7 | 183 | 14 | 6 |
| TİCARİ | 32 | 4 | 170 | 15 | 6 |
| TİCARİ | 26 | 4 | 197 | 8 | 6 |




**Şekil 2. Karar ağacı**

Tablodan çıkan sonuca göre eğer bagaj hacmi 9 $m^3$ ten büyükse araç kesinlikle pick-up araçtır.

### 2.1.2 Denetimsiz (Unsupervised) Öğrenme

Eldeki verilerin düzenleri, desenleri hakkında bir bilgi olmadan verileri gruplandırma ya da kümelendirme yapma işidir. Kümeleme yöntemleri denetimsiz öğrenmedir. Öğrenme için sadece bir giriş değeri verilir bu giriş değerine göre çıkış değerleri sınıflandırma ya da gruplandırma olarak keşfedilir. Örneğin K-Means algoritmasında eldeki verileri kümelemek için sadece o veri seti kaç kümeye bölüncekse giriş değeri sadece küme sayısı belirtilir ve algoritma tamamen kendi başına kümeleme işlemini üstlenir. Kümeleme yapmak için algoritması dışında başka hiçbir şeye ihtiyaç duymaz.

#### 2.1.2.1 Kümeleme Analizi

Kümeleme bir veri setindeki k adet verinin birbirlerine benzerliğine göre n tane kümeye ayrılarak gruplandırılmasıdır. Kümeleme ile ilgili birçok tanım vardır ve bundan bazıları şu şekildedir.

Kümeleme analizi, araştırma içinde incelenen birimleri aralarındaki benzerliklerine göre belirli gruplar içinde toplayarak sınıflandırma yapmayı, birimlerin ortak özelliklerini ortaya koymayı ve bu sınıflar ile ilgili genel tanımlar yapmayı sağlayan bir yöntemdir [7] .

Kümeleme analizi için başka bir tanım da şu biçimde yapılmaktadır. “ Kümeleme analizi, temel amacı nesneleri (birimleri) sahip oldukları karakteristik özellikleri baz alarak gruplamak olan çok değişkenli teknikler grubudur. Kümeleme analizi, nesneleri küme içerisinde çok benzer biçimde, kümeler arasında farklı olacak biçimde kümeler. Kümeleme işlemi başarılı olursa, bir geometrik çizim yapıldığında nesneler küme içerisinde birbirine çok yakın, kümeler ise birbirinden uzak olacaktır [3] .

Bu gruplama işlemi yapıldığında bir küme içindeki verilerin benzerliği(similarity) en fazla, grup dışındaki verilerin uzaklığı (distance) ise en fazladır. Kümeleme işlemleri hayatımızın her kısmında karşımıza çıkmaktadır. Kümeleme analizi görüntü işleme, finans, bankacılık gibi konularda sıkça kullanılmaktadır.

## 2.2 Kümeleme analizinin aşamaları

- Verilerin hazırlanması
- Benzerlik veya uzaklığın belirlenmesi
- Kümeleme yönteminin belirlenmesi
- Elde edilen çıktıların tefsir edilmesi

### 2.2.1 Verilerin Hazırlanması

Veriler toplanırken verilerin o konuya uygun ve elde edilen verilerin birbirinden ayırt edici özelliklerinin olması gerekmektedir. Konu ile ilgisi olmayan, kümelemeyi saptırabileceği düşünülen bazı verilerin ya da kayıp verilerin ayıklanması veya düzeltilmesi önemlidir.

### 2.2.2 Benzerlik veya Uzaklığın Belirlenmesi

Daha öncede belirtildiği gibi bir küme içindeki verilerin benzerliği(similarity) en fazla, grup dışındaki verilerin uzaklığı (distance) ise en fazladır. Veri tipine bağlı olarak veri setler içinde uzaklık-benzerlik ölçüleri hesaplanır. Uzaklığı ve yakınlığı belirleyebilmek için en az 2 nokta olması gerekir. Benzerlik ve uzaklığı belirlemek için birçok algoritma var fakat bunların içinde en yaygın olanlardan 2 tanesi mesafe için öklit (euclid) ve yakınlık için Jaccard benzerlik ölçütüdür.

**Öklit mesafesi :**

mes$(X_m, X_j)$ = $\sqrt{\sum_{i=1}^{n} \left(X_{mi}, X_{ji}\right)^2}$ formülü,

Jaccard benzerlik ölçütü ise

ben$(X_m, X_j)$ = $X_m \frac{|X_m| \cap |X_j|}{|X_m| \cup |X_j|}$ formülü ile bulunur.

Öklit mesafesi kategorik olamayan verilerde, Jaccard benzerlik ölçütü ise kategorik verilerde büyük önem taşır. Örneğin bir marketler zincirinin veri tabanında meyvelerden elmaya 1, armuta 2, üzüme 3 ve muza 4 değerleri verirsek elma ile üzümün ortalaması armuttur gibi bir sonuca ulaşırız. Bu da öklit mesafesini hipotezini çürütür.

Örnek : $X_m$ = {2,3,4}, $X_j$ = {3,6,9} ve $X_z$ = {3,5,7} kümelereinde $X_m$ kümesinin $X_j$ kümesine mi daha yakın olduğu yoksa $X_z$ kümesine mi daha yakın olduğunu anlamak için öklit bağlantısı kullanılır

$$\text{mes}(X_m, X_j) = \sqrt{\sum_{i=1}^{n}(X_{mi}, X_{ji})}^2 = \sqrt{(2-3)^2 + (3-6)^2 + (4-9)^2} = \sqrt{35} = 5.91$$

$$\text{mes}(X_m, X_z) = \sqrt{\sum_{i=1}^{n}(X_{mi}, X_{ji})}^2 = \sqrt{(2-3)^2 + (3-5)^2 + (4-7)^2} = \sqrt{19} = 4.36$$

Bunun sonucunda $X_m$ noktasının $X_z$ noktasında $X_j$ noktasından daha yakın olduğu söylenebilir. Bu 3 veriyi 2 ayrı küme olarak kümelemek isteseydik $X_m$ ve $X_z$ verileri beraber $X_j$ kümsi de ayrı kümelecekti.

### 2.2.3 Kümeleme yönteminin belirlenmesi

Kümeleme yönteminin belirlenmesi hiyerarşik ne bölümlemeli(partitioning) olarak 2 kısmına ayrılır.

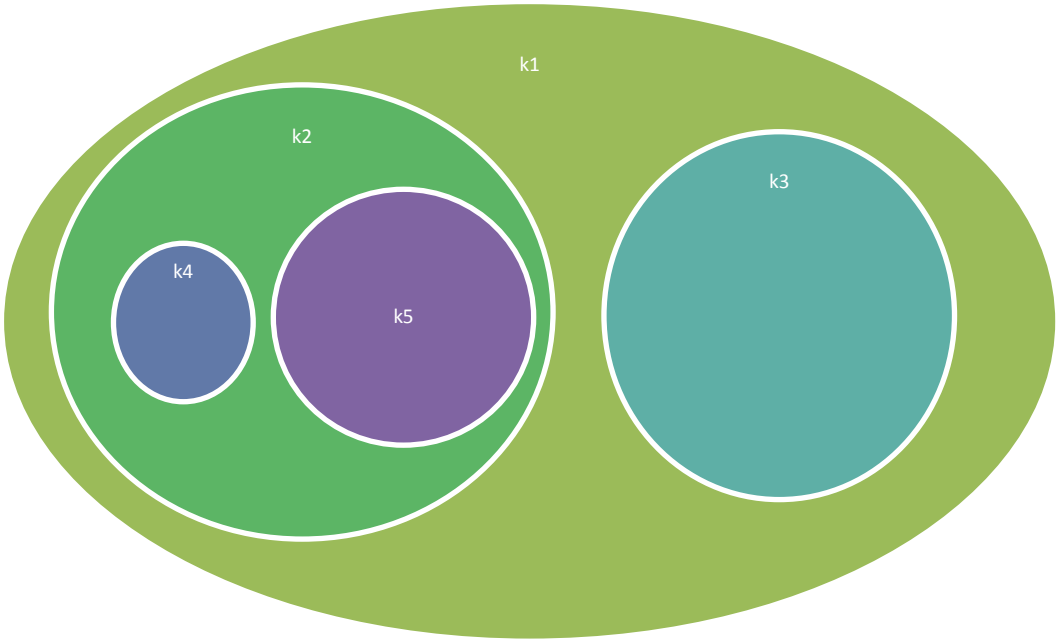
### 2.2.4 Hiyerarşik Yöntemler

Daha çok küçük veri tabanlarında tercih edilir. Hiyerarşik yöntemler karar ağacına sahiptir bu ağaç dallara ve oradan yapraklara ayrılır. Yapraklara ulaşınca ağaç biter. Ağacın dallara ayrıldığı yerlere düğüm adı verilir. Hiyerarşik yöntemler birleştirici ve ayrıştırıcı kümeleme algoritmaları olarak 2 ye ayrılır birleştirici kümeleme algoritmasında en başta her bir veriyi küme olarak görür ve bu kümeleri algoritmalara göre birleştirerek ayrı ayrı kümelere oluşturur. Ayrıştırıcı kümeleme algoritmasındaysa en başta tüm veriler tek küme olarak alınır daha verilerin farklarına başlangıçtaki tek olan küme ayrı ayrı kümelere bölünür. Hiyerarşik yöntemlerde en ünlü algoritmalar BIRCH, SLINK, CLUCDUH, CHAMELON ve CURE algoritmalarıdır. [14]

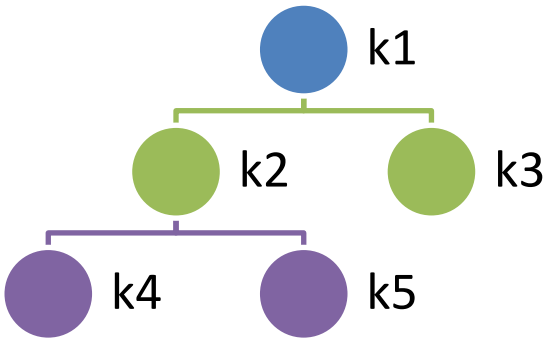
### 2.2.5 Bölümlemeli Yöntemler

Bölümlemeli yöntemlerde veri setindeki veriler daha önceden belirlenmiş kritere göre ayrılır. D veri tabanında m tane veri varsa bu veri tabanı en fazla m-1 kadar bölüme ayrılabilir. Bölümlemeli yöntemlere de hiyerarşik yöntemlerin aksine birbirine bağlı yöntemler yerine her veri bölünen bağımsız bir kümede yer alır. Bu

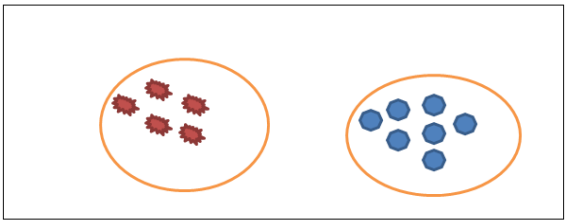
çalışmada bölümlemeli yöntemlerden K-Means algoritması ve Jaccard benzerlik ölçütü üzerinde durulacaktır.




**Şekil 3. Hiyerarşik kümeleme 1**




**Şekil 4. Hiyerarşik kümeleme 2**



**Şekil 5. Bölümlemeli yöntemler**

### 2.3 K-Means Algoritması

1967 yılında temeli J.B. MacQueen tarafından atılan bir kümeleme algoritmasıdır. Algoritmanın temeli n tane olan veri setinin k tane kümeye bölünmesidir. Bu bölme işlemi yapılırken öklit bağlantısı baz alınarak sonuç kümesi bulunana kadar kümelerin devamlı yenilendiği döngüsel bir işlem uygulanır. K-means algoritmasında her veri sadece bir kümeye ait olabilir. Bu algoritmada kümeler arasındaki uzaklık maximum, küme içi elemanlar arasında ise uzaklık minimum olacak şekilde kümeleme işlemi yapılır. Öklit algoritmasının formülü şu şekildedir.

$$d(i,j)=\sqrt{(x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2+\ldots+(x_{ip}-x_{jp})^2} \quad [11]$$

K-Means algoritmasının çalışma mantığı aşağıdaki gibidir.

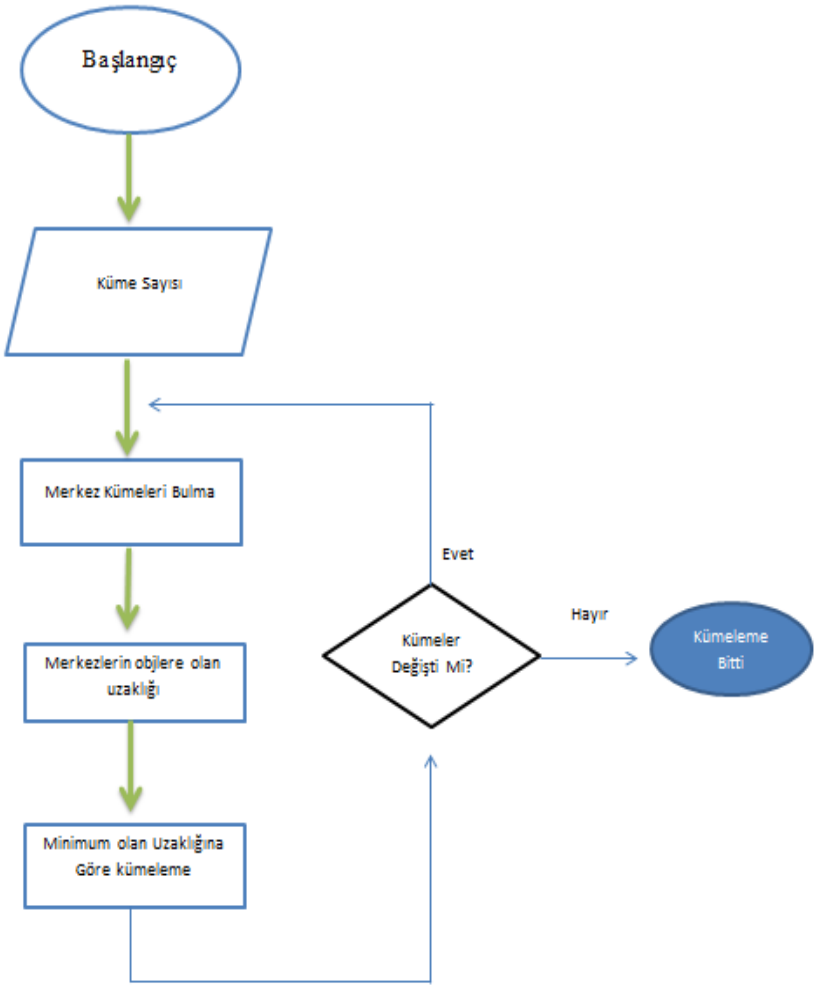
- İlk veri setinin kaç kümeye ayrılacağı belirlenir.
- İlk sefere mahsus eldeki verilerin içinden küme sayısı kadar rastgele merkez kümeler seçilir. Seçilen bu merkezlerin her biri veri setinin ayrılacağı her bir kümeyi temsil etmektedir.
- Veri setinin içindeki her bir elemanın merkez kümelere göre öklit bağlantısı alınır ve öklitten çıkan sonuç en az olan o merkez kümenin kümesine eklenir.
- Veri setindeki her eleman merkez kümelerin temsil ettiği kümelere yerleştirildikten sonra yeni merkez kümeler bulunur. Yeni merkez kümelerin bulunması için kümeye eklenen her veri toplanıp kümedeki eleman sayısına bölünür.

- En sonunda bu işlemler bir döngüde devam ettirilir. Bir önceki döngüdeki kümeler bir sonradaki kümeler ile aynı ise işlem sonlandırılır. Veri seti k-means algoritmasına göre kümelenmiş olur.

K – Means algoritmasının örnek kodu aşağıdaki gibidir.

```
var m = initialCentroids(x, K);

   var N = x.length;
   while (!stoppingCriteria) {
      var w = [][];
      // calculate membership in clusters
      for (var n = 1; n <= N; n++) {
      v = arg min (v0) dist(m[v0], x[n]);
      w[v].push(n);
   }
   // recompute the centroids
   for (var k = 1; k <= K; k++) {
      m[k] = avg(x in w[k]);
   }
}
return m;[13].
```




**Şekil 6. K-Means algoritması akış şeması**

**Tablo 4. K–Means algoritması örnek veri kümesi**

| | | |
|---|---|---|
| 8 | 2 | 1 |
| 3 | 1 | 1 |
| 4 | 2 | 2 |
| 1 | 1 | 1 |
| 6 | 7 | 8 |
| 9 | 90 | 1 |

**Örnek 1:** İlk örnekte veri setin K-Means algoritması ile 2 kümeye bölünmesi

İlk olarak her ayrılmasını istediğimiz kümeler için rastgele bir küme merkezi seçilir. Bu örnek için ilk 2 eleman her küme merkezi seçilmiştir.

M1 = 8, 2, 1

M2 = 3, 1, 1

İlk merkezler bulunduktan sonra merkezler ile veri seti içinde bulunan her eleman arasında öklit bağlantısı kurulur ve değerce büyük olan sonuç kümenin uzak ve değerce küçük olan sonuç ise kümenin yakın olduğunu gösterir. Bunun sonucunda küçük olan değer ökliti alınan merkez kümeye eklenir.

Veri seti elemanı 4, 2, 2 için öklit:

S1 = $\sqrt{(8-4)^2+(2-2)^2+(1-2)^2}$ = $\sqrt{16+0+1}$ = $\sqrt{17}$ = 4.123

S2 = $\sqrt{(3-4)^2+(1-2)^2+(1-2)^2}$ = $\sqrt{1+1+1}$ = $\sqrt{3}$ = 1.73

S2 < S1 bunun için veri seti 4, 2, 2 Küme2 kümesine eklendi.

Veri seti elemanı 1, 1, 1 için öklit:

S1 = $\sqrt{(8-1)^2+(2-1)^2+(1-1)^2}$ = $\sqrt{49+1+0}$ = $\sqrt{50}$ = 7.07,

S2 = $\sqrt{(3-1)^2+(1-1)^2+(1-1)^2}$ = $\sqrt{4+0+0}$ = $\sqrt{4}$ = 2

S2 < S1 bunun için veri seti 1, 1, 1 Küme2 kümesine eklendi

Veri seti elemanı 6, 7, 8 için öklit:

S1 = $\sqrt{(8-6)^2+(2-7)^2+(1-8)^2}$ = $\sqrt{4+25+49}$ = $\sqrt{78}$ = 8.83,

S2 = $\sqrt{(3-6)^2+(1-7)^2+(1-8)^2}$ = $\sqrt{9+36+49}$ = $\sqrt{94}$ =9.69

S2 > S1 bunun için veri seti 6, 7, 8 Küme1 kümesine eklendi

Veri seti elemanı 9, 90, 1 için öklit:

S1 = $\sqrt{(8-9)^2+(2-90)^2+(1-1)^2}$ = $\sqrt{1+7744+0}$ = $\sqrt{7745}$ = 88,

S2 = $\sqrt{(3-9)^2+(1-90)^2+(1-1)^2}$ = $\sqrt{36+7921+0}$ = $\sqrt{7957}$ =89.2

S2 >S1 bunun için veri seti 9, 90, 1 Küme1 kümesine eklendi

Uygulanan öklit işlemleri sonucunda küme1 ve küme2 kümeleri şu şekilde oluşmuştur.

**Tablo 5. K-Means algoritması örnek 1 sonucu oluşan ilk merkez kümeler**

| | | | |
|---|---|---|---|
| **Küme1** | 8 | 2 | 1 |
| | 6 | 7 | 8 |
| | 9 | 90 | 1 |
| **Küme2** | 3 | 1 | 1 |
| | 4 | 2 | 2 |
| | 1 | 1 | 1 |

İlk küme merkezleri bulunduktan sonra yeni küme merkezleri bulunması işlemi yapılır. Yeni küme merkezlerinin bulunması için her merkez küme içerisindeki verilerin sütunları birbirleri ile toplanır ve satır sayısına bölünür.

M1 = (8+6+9)/3 , (2+7+90)/3, (1+8+1)/3 = 7.67, 33.0, 3.33

M2 = (3+4+1)/3, (1+2+1)/3, (1+2+1)/3 = 2.67, 1.33, 1.33

Bu aşamadan sonra sonsuz bir iterasyon başlatılır. Yeni bulunan merkez kümeler veri setindeki her elemana öklit işlemi ile uygulanmaya başlanır. Öklit işlemi her bir veriye uygulandıktan sonra yeniden yeni kümeler ve küme merkezleri elde edilir. Eğer bir önceki merkez kümeler ile yeni bulunan merkez kümeler birbirinin aynısı ise iterasyon biter ve en son bulunan kümeler K-Means algoritması ile gruplanmış kümeler olur şayet merkez kümeler farklı ise aynı öklit işlemi yeni merkez kümeler için tekrarlanır ta ki bir önceki itersyonun küme merkezleri ile en sondaki iterasyonun küme sayıları birbirinin aynısı oluncaya kadar.

**İterasyon 1**

Veri seti elemanı 8, 2, 1 için öklit:

S1 = $\sqrt{(7.67-8)^2+(33-2)^2+(3.33-1)^2}$ = $\sqrt{0.11+961+5.43}$ = 31,

S2 = $\sqrt{(2.66-8)^2+(1.33-2)^2+(1.33-1)^2}$ = $\sqrt{28.51+0.44+0.111}$ = 29

S2 < S1 bunun için veri seti 8, 2, 1 Küme 2 kümesine eklendi.

Veri seti elemanı 3, 1, 1 için öklit:

S1 = $\sqrt{(7.67-3)^2+(33-1)^2+(3.33-1)^2}$ = 32.42 ,

S2 = $\sqrt{(2.66-3)^2+(1.33-1)^2+(1.33-1)^2}$ = 0.574

S2 < S1 bunun için veri seti 3, 1, 1 Küme 2 kümesine eklendi

Veri seti elemanı 4, 2, 2 için öklit:

S1 = $\sqrt{(7.67-4)^2+(33-2)^2+(3.33-2)^2}$ = 31.24 ,

S2 = $\sqrt{(2.67-4)^2+(1.33-2)^2+(1.33-2)^2}$ = 1.63

S2 < S1 bunun için veri seti 4, 2, 2 Küme 2 kümesine eklendi.

Veri seti elemanı 1, 1, 1 için öklit:

S1 = $\sqrt{(7.67-1)^2+(33-1)^2+(3.33-1)^2}$ = 32.77 ,

S2 = $\sqrt{(2.67-1)^2+(1.33-1)^2+(1.33-1)^2}$ = 2.78+0.11+0.11 = 1.73

S2 < S1 bunun için veri seti 1, 1, 1 Küme 2 kümesine eklendi

Veri seti elemanı 6, 7, 8 için öklit:

S1 = $\sqrt{(7.67-6)^2+(33-7)^2+(3.33-8)^2}$ = 26.47,

S2 = $\sqrt{(2.67-6)^2+(1.33-7)^2+(1.33-8)^2}$ = 9.36

S2 <S1 bunun için veri seti 6, 7, 8 Küme 2 kümesine eklendi.

Veri seti elemanı 9, 90, 1 için öklit:

S1 = $\sqrt{(7.67-9)^2+(33-90)^2+(3.33-1)^2}$ = 57.06,

S2 = $\sqrt{(2.67-9)^2+(1.33-90)^2+(1.33-1)^2}$ = 88.89

S2 >S1 bunun için veri seti 9, 90, 1 Küme1 kümesine eklendi.

**Tablo 6. K-Means algoritması örnek 1 iterasyon 1 sonucu oluşan kümeler**

| | | | |
|---|---|---|---|
| **Küme1** | 9 | 90 | 1 |
| **Küme2** | 8 | 2 | 1 |
| | 3 | 1 | 1 |
| | 4 | 2 | 2 |
| | 1 | 1 | 1 |
| | 6 | 7 | 8 |

M1 = (9)/1 , (90)/1, (1)/1 = 9, 90, 1

M2 = (8+3+4+1+6)/5, (2+1+2+1+7)/5, (1+1+2+1+8)/5 = 4.4, 2.6, 2.6

Bir önceki iterasyon da ki merkez kümelerle bu iterasyon da ki merkez kümeler birbirini tutmadığı için yeni bir iterasyon başlatılıyor.

**İterasyon 2**

Veri seti elemanı 8, 2, 1 için öklit:

S1 = $(9-8)^2$+$(90-2)^2$+$(1-1)^2$ = 1+7744+0 = 17 ,

S2 = $(4.44-8)^2$+$(2.6-2)^2$+$(2.6-1)^2$ = 12.96+0.36+2.56 = 3

S2 < S1 bunun için veri seti 8, 2, 1 Küme2 kümesine eklendi.

Veri seti elemanı 3, 1, 1 için öklit:

S1 = $\sqrt{(9-3)^2+(90-1)^2+(1-1)^2}$ = 89.20,

S2 = $\sqrt{(4-3)^2+(2.6-1)^2+(2.6-1)^2}$ = 2.47

S2 < S1 bunun için veri seti 3, 1, 1 Küme2 kümesine eklendi

Veri seti elemanı 4, 2, 2 için öklit:

S1 = $\sqrt{(9-4)^2+(90-2)^2+(1-2)^2}$ = 88.14,

S2 = $\sqrt{(4.44-4)^2+(2.6-2)^2+(2.6-2)^2}$ = 0.94

S2 < S1 bunun için veri seti 4, 2, 2 Küme2 kümesine eklendi.

Veri seti elemanı 1, 1, 1 için öklit:

S1 = $\sqrt{(9-1)^2+(90-1)^2+(1-1)^2}$ = 89.36,

S2 = $(4.4-1)^2$+$(2.6-1)^2$+$(2.6-1)^2$ = 4.084

S2 < S1 bunun için veri seti 1, 1, 1 Küme2 kümesine eklendi

Veri seti elemanı 6, 7, 8 için öklit:

S1 = $\sqrt{(9-6)^2+(90-7)^2+(1-8)^2}$ = 83.34,

S2 = $\sqrt{(4.4-6)^2+(2.6-7)^2+(2.6-8)^2}$ = 7.13

S2 <S1 bunun için veri seti 6, 7, 8 Küme2 kümesine eklendi.

Veri seti elemanı 9, 90, 1 için öklit:

S1 = $\sqrt{(9-9)^2+(90-90)^2+(1-1)^2}$ = 0,

S2 = $\sqrt{(4.4-9)^2+(2.6-90)^2+(2.6-1)^2}$ = 87.53

S2 >S1 bunun için veri seti 9, 90, 1 Küme1 kümesine eklendi.

İterasyon 2 sonucunda oluşan kümeler aşağıdaki gibidir.

**Tablo 7. K-Means algoritması örnek 1 iterasyon 2 sonucu oluşan kümeler**

| | | | |
|---|---|---|---|
| **Küme1** | 9 | 90 | 1 |
| **Küme2** | 8 | 2 | 1 |
| | 3 | 1 | 1 |
| | 4 | 2 | 2 |
| | 1 | 1 | 1 |
| | 6 | 7 | 8 |

M1 = (9)/1 , (90)/1, (1)/1 = 9, 90, 1

M2 = (8+3+4+1+6)/5, (2+1+2+1+7)/5, (1+1+2+1+8)/5 = 4.4, 2.6, 2.6

En son iterasyon ile en sondan bir önceki iterasyonların merkez kümeleri birbirlerinin aynısı olduğu için sonsuz iterasyonlar kırılarak sonlandırıldı ve veri setinde veriler K-Means algoritmasına göre 2 küme olarak kümelenmiş oldu. Algoritma sonucunda tablo 11'deki sonuçlar elde edilmiştir.

**Tablo 8. K-Means algoritması ile örnek 1 için gruplanmış kümeler**

| | | | |
|---|---|---|---|
| **Küme1** | 9 | 90 | 1 |
| **Küme2** | 8 | 2 | 1 |
| | 3 | 1 | 1 |
| | 4 | 2 | 2 |
| | 1 | 1 | 1 |
| | 6 | 7 | 8 |

**Örnek 2:** İlk örnekte veri setinin K-Means algoritması ile 3 kümeye bölünmesi

3 kümeye ayırmak istendiği için veri setinin rastgele 3 satırı ilk merkez kümeler yapıldı.

M1 = 8, 2, 1

M2 = 3, 1, 1

M3 = 4, 2, 2

İlk merkezlerle veri ve veri setinin satırları arasında öklit bağlantısı alınır

Veri seti elemanı 1, 1, 1 için öklit:

S1 = $\sqrt{(8-1)^2 + (2-1)^2 + (1-1)^2}$ = 7.07,

S2 = $\sqrt{(3-1)^2+(1-1)^2+(1-1)^2}$ = 2

S3 = $\sqrt{(4-1)^2+(2-1)^2+(2-1)^2}$ = 3.32

S2 < S3<S1 bunun için veri seti 1, 1, 1 Küme 2 kümesine eklendi

Veri seti elemanı 6, 7, 8 için öklit:

S1 = $\sqrt{(8-6)^2+(2-7)^2+(1-8)^2}$ = 8.83,

S2 = $\sqrt{(3-6)^2+(1-7)^2+(1-8)^2}$ = 9.7

S3 = $\sqrt{(4-6)^2+(2-7)^2+(2-8)^2}$ = 8.06

S3<S1<S2 bunun için veri seti 6, 7, 8 Küme 3 kümesine eklendi

Veri seti elemanı 9, 90, 1 için öklit:

S1 = $\sqrt{(8-9)^2+(2-90)^2+(1-1)^2}$ = 88,

S2 = $\sqrt{(3-9)^2+(1-90)^2+(1-1)^2}$ = 89.2,

S3 = $\sqrt{(4-9)^2+(2-90)^2+(1-1)^2}$ = 88.14

S1<S3 <S2 bunun için veri seti 9, 90, 1 Küme 1 kümesine eklendi

İlk merkez kümeler sonucunda elde edilen kümeler tablo 9'daki gibidir.

**Tablo 9.K-Means algoritması örnek 2 ilk merkez kümeler**

| | | | |
|---|---|---|---|
| **Küme1** | 8 | 2 | 1 |
| | 9 | 90 | 1 |
| **Küme2** | 3 | 1 | 1 |
| | 1 | 1 | 1 |
| **Küme3** | 4 | 2 | 2 |
| | 6 | 7 | 8 |

İlk kümelere göre yeni merkez kümelerin bulunması şu şekildedir;

M1 = (8+9)/2 , (2+90)/2,  (1+1)/2 = 8.5, 46.0, 1.0

M2 = (3+1)/2, (1+1)/2, (1+1)/2 = 2.0, 1.0, 1.0

M3 = (4+6)/2, (2+7)/2, (2+8)/2 = 5.0, 4.5, 5.0

İlk merkez kümeler bulunduğundan iterasyon kısmı başlatılıyor.

**İterasyon 1**

Veri seti elemanı 8, 2, 1 için öklit:

S1 = $\sqrt{(8.5-8)^2+(46-2)^2+(1-1)^2}$ = 44,

S2 = $\sqrt{(2-8)^2+(1-2)^2+(1-1)^2}$ = 6.08,

S3 = $\sqrt{(5-8)^2+(4.5-2)^2+(5-1)^2}$ = 5.59

S3<S2 < S1 bunun için veri seti 8, 2, 1 Küme3 kümesine eklendi.

Veri seti elemanı 3, 1, 1 için öklit:

S1 = $\sqrt{(8.5-3)^2+(46-1)^2+(1-1)^2}$ = 45.33,

S2 = $\sqrt{(2-3)^2+(1-1)^2+(1-1)^2}$ = 1,

S3 = $\sqrt{(5-3)^2+(4.5-1)^2+(5-1)^2}$ = 5.68

S2 < S3<S1 bunun için veri seti 3, 1, 1 Küme2 kümesine eklendi

Veri seti elemanı 4, 2, 2 için öklit:

S1 = $\sqrt{(8.5-4)^2+(46-2)^2+(1-2)^2}$ = 44.24,

S2 = $\sqrt{(2-4)^2+(1-2)^2+(1-4)^2}$ = 3.74,

S3 = $\sqrt{(5-4)^2+(4.5-2)^2+(5-4)^2}$ = 2.87

S2 < S3<S1 bunun için veri seti 4, 2, 2 Küme2 kümesine eklendi.

Veri seti elemanı 1, 1, 1 için öklit:

S1 = $\sqrt{(8.5-1)^2+(46-1)^2+(1-1)^2}$ = 45.62,

S2 = $\sqrt{(2-1)^2+(1-1)^2+(1-1)^2}$ = 1,

S3 = $\sqrt{(5-1)^2+(4.5-1)^2+(5-1)^2}$ = 6.65

S2 < S3<S1 bunun için veri seti 1, 1, 1 Küme2 kümesine eklendi

Veri seti elemanı 6, 7, 8 için öklit:

S1 = $\sqrt{(8.5-6)^2+(46-7)^2+(1-8)^2}$ = 39.70,

S2 = $\sqrt{(2-6)^2+(1-7)^2+(1-8)^2}$ = 16+36+49 = 10,

S3 = $\sqrt{(5-6)^2+(4.5-7)^2+(5-8)^2}$ = 4.03

S3<S2 <S1 bunun için veri seti 6, 7, 8 Küme3 kümesine eklendi.

Veri seti elemanı 9, 90, 1 için öklit:

S1 = $\sqrt{(8.5-9)^2+(46-90)^2+(1-1)^2}$ = 44,

S2 = $\sqrt{(2-9)^2+(1-90)^2+(1-1)^2}$ = 89.27,

S3 = $\sqrt{(5-9)^2+(4.5-90)^2+(5-1)^2}$ = 85.69

S1<S3 <S2 bunun için veri seti 9, 90, 1 Küme1 kümesine eklendi.

İterasyon 1 sonucunda elde edilen kümeler tablo 10'daki gibidir.

**Tablo 10. K-Means algoritması örnek 2 iterasyon 1 sonucunda oluşan kümeler**

| | | | |
|---|---|---|---|
| **Küme1** | 9 | 90 | 1 |
| **Küme2** | 3 | 1 | 1 |
| | 4 | 2 | 2 |
| | 1 | 1 | 1 |
| **Küme3** | 8 | 2 | 1 |
| | 6 | 7 | 8 |

İlk kümelere göre yeni merkez kümelerin bulunması şu şekildedir;

M1 = (9)/1 , (90)/,  (1)/1 = 9, 90, 1

M2 = (3+4+1)/3, (1+2+1)/3, (1+2+1)/3 = 2.67, 1.33, 1.33

M3 = (8+6)/2, (2+7)/2, (1+8)/2 = 7.0, 4.5, 4.5

Bir önceki iterasyondaki ile yeni bulunan merkez kümeler birbirinin aynısı olmadığı için yeni iterasyona geçilir

**İterasyon 2**

Veri seti elemanı 8, 2, 1 için öklit:

S1 = $\sqrt{(9-8)^2+(90-2)^2+(1-1)^2}$ = 88,

S2 = $\sqrt{(2.67-8)^2+(1.33-2)^2+(1.33-1)^2}$ = 5.38,

S3 = $\sqrt{(5-8)^2+(4.5-2)^2+(5-1)^2}$ = 5.6

S3<S2 < S1 bunun için veri seti 8, 2, 1 Küme3 kümesine eklendi.

Veri seti elemanı 3, 1, 1 için öklit:

S1 = $\sqrt{(9-3)^2+(90-1)^2+(1-1)^2}$ = 89.2,

S2 = $\sqrt{(2.66-3)^2+(1.33-1)^2+(1.33-1)^2}$ = 0.57,

S3 = $\sqrt{(75-3)^2+(4.5-1)^2+(4.5-1)^2}$ = 72.17

S2 < S3<S1 bunun için veri seti 3, 1, 1 Küme2 kümesine eklendi

Veri seti elemanı 4, 2, 2 için öklit:

S1 = $\sqrt{(9-4)^2+(9-2)^2+(1-2)^2}$ = 8.66,

S2 = $\sqrt{(2.66-4)^2+(1.33-2)^2+(1.33-4)^2}$ = 3.06,

S3 = $\sqrt{(7-4)^2+(4.5-2)^2+(4.5-4)^2}$ = 3.94

S2 < S3<S1 bunun için veri seti 4, 2, 2 Küme2 kümesine eklendi.

Veri seti elemanı 1, 1, 1 için öklit:

S1 = $\sqrt{(9-1)^2+(90-1)^2+(1-1)^2}$ = 89.36,

S2 = $\sqrt{(2.66-1)^2+(1.33-1)^2+(1.33-1)^2}$ = 1.73

S3 = $\sqrt{(7-1)^2+(4.5-1)^2+(4.5-1)^2}$ = 7.78

S2 < S3<S1 bunun için veri seti 1, 1, 1 Küme2 kümesine eklendi

Veri seti elemanı 6, 7, 8 için öklit:

S1 = $\sqrt{(9-6)^2+(90-7)^2+(1-8)^2}$ = 83.35,

S2 = $\sqrt{(2-6)^2+(1-7)^2+(1-8)^2}$ = 10,

S3 = $\sqrt{(5-6)^2+(4.5-7)^2+(5-8)^2}$ = 4.03

S3<S2 <S1 bunun için veri seti 6, 7, 8 Küme3 kümesine eklendi.

Veri seti elemanı 9, 90, 1 için öklit:

S1 = $\sqrt{(9-9)^2+(90-90)^2+(1-1)^2}$ = 0,

S2 = $\sqrt{(2.67-9)^2+(1.33-90)^2+(1.33-1)^2}$ = 88.9,

S3 = $\sqrt{(7-9)^2+(4.5-90)^2+(4.5-1)^2}$ = 85.69

S1<S3 <S2 bunun için veri seti 9, 90, 1 Küme1 kümesine eklendi.

İterasyon 2 sonucunda elde edilen kümeler tablo 11'deki gibidir.

**Tablo 11. K-Means algoritması örenek 2 iterasyon 2 sonucunda oluşan kümeler**

| | | | |
|---|---|---|---|
| **Küme1** | 9 | 90 | 1 |
| **Küme2** | 3 | 1 | 1 |
| | 4 | 2 | 2 |
| | 1 | 1 | 1 |
| **Küme3** | 8 | 2 | 1 |
| | 6 | 7 | 8 |

İterasyon 2 den elde edilen kümelere göre yeni merkez kümelerin bulunması şu şekildedir;

M1 = (9)/1 , (90)/1, (1)/1 = 9, 90, 1

M2 = (3+4+1)/3, (1+2+1)/3, (1+2+1)/3 = 2.67, 1.33, 1.33

M3 = (8+6)/2, (2+7)/2, (1+8)/2 = 7.0, 4.5, 4.5

En son iterasyon ile en sondan bir önceki iterasyonların merkez kümeleri birbirlerinin aynısı olduğu için sonsuz iterasyonlar kırılarak sonlandırıldı ve veri setinde veriler K-Means algoritmasına göre 3 küme olarak kümelendirilmiş oldu. Algoritma sonucunda tablo 12'deki sonuçlar elde edilmiştir.

**Tablo 12. K-Means algoritması örenek 2 sonucunda oluşan kümeler**

| | | | |
|---|---|---|---|
| **Küme1** | 9 | 90 | 1 |
| **Küme2** | 3 | 1 | 1 |
| | 4 | 2 | 2 |
| | 1 | 1 | 1 |
| **Küme3** | 8 | 2 | 1 |
| | 6 | 7 | 8 |

K- Means algoritmasında veri kümesinin kaç kümeye bölmenin daha optimum sonuç vereceğini anlamak çok önemlidir. Bunun için Dunn Geçerlilik İndeksi, Davies-Bouldin İndeksi, Silhoutte Geçerlilik Yöntemi, C İndexi gibi algoritmalar mevcuttur. [14].

### 2.4 Jaccard Benzerlik Ölçütü

"Bu index değeri [12] benzerlik değerlendirmesine göre küme kalitesi belirler. J(C,K) değeri C ile gösterilen sınıf ve K ile gösterilen küme sonuçlarının karşılaştırılması sonucudur. Diğer indexlerden farklı olarak veri seti üzerinde ayırt

edici bir C sınıfının belirlenmesi ve küme algoritması ile birlikte C sınıflandırmanın da yapılması gerekmektedir. Ayrıca veri setindeki elemanlar çiftler halinde değerlendirmesi yapılmalıdır." [14]. Jaccard benzerlik ölçütü kategorik verilerin benzerliğinin bulunmasında başarılı sonuçlar vermektedir.

benj$(X_m, X_j)$ $= \frac{|X_m| \cap |X_j|}{|X_m| \cup |X_j|}$ formülü ile bulunur.

Bu algoritmada karşılaştırma yapıldığında her zaman 0-1 arası sonuçlar alınmaktadır. Eğer karşılaştırmadan çıkan sonuç 1 ise kümeler tıpa tıp aynı, sonuç 0 ise kümeler birbirinden tamamen bağımsızdır.

**Tablo 13. Jaccard Benzerlik Ölçütü için örnek veri kümesi**

| **Telefon Adı** | **Kamera** | **İşlemci** | **İşletim Sis.** | **Ekran** |
|---|---|---|---|---|
| **Tel1** | Var | Intel | Android | Led |
| **Tel2** | Var | Intel | IOS | Led |
| **Tel3** | Yok | Amd | Android | Amoled |

Tablo 15 içindeki verilere göre Tel2 ve Tel3 telefonlarından hangisinin Tel1 telefonuna daha benzer olduğunu görmek için Jaccard benzerlik ölçütü şu şekilde kullanılır.

benj$(Tel1 \ , Tel2) = \frac{|Var,Intel,Android,Led| \cap |Var,Intel,IOS,Led|}{|Var,Intel,Android,Led| \cup |Var,Intel,IOS,Led|}$

$=3/5 = 0.6$

benj$(Tel1 \ , Tel3) = \frac{|Var,Intel,Android,Led| \cap |Yok,Amd,Android,Amoled|}{|Var,Intel,Android,Led| \cup |Yok,Amd,Android,Amoled|}$

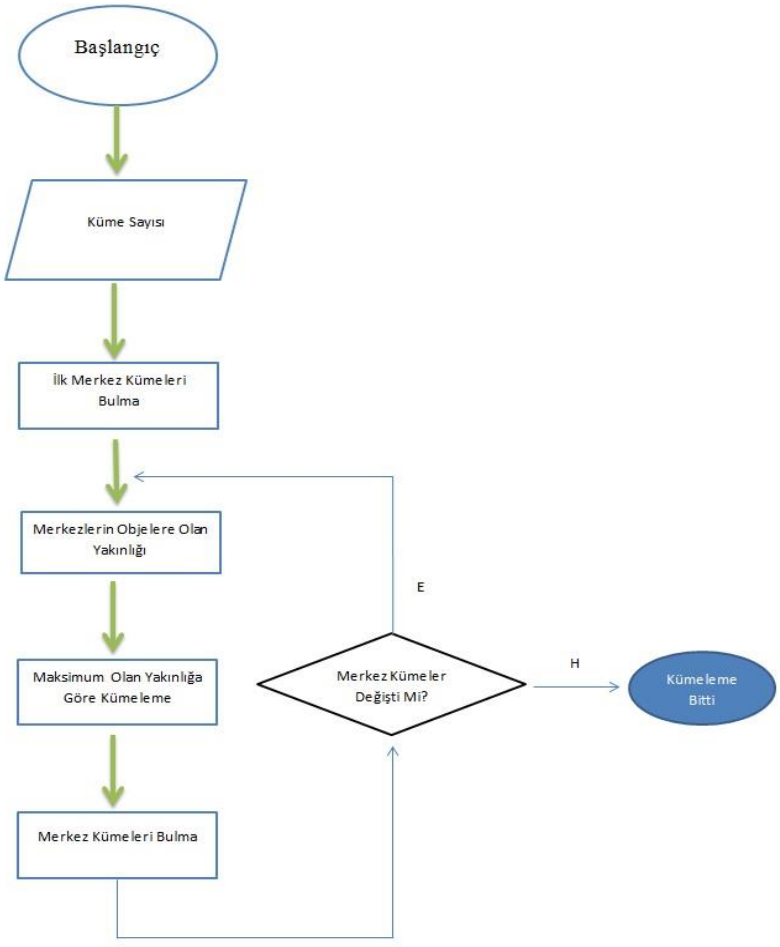
$=1/7 = 0.14$

Tel1 ve Tel2'nin benzerliğinden çıkan sonuç 1 sayısına Tel1 ve Tel3'ten çıkan sonuca göre daha yakın olduğu için Tel2 telefonu Tel1 telefonuna Tel3 telefonundan daha benzer olduğu sonucu ortaya çıkar.

# 3 K-MEANS ALGORİTMASI İLE SADECE KATEGORİK VERİ KÜMELERİ VE KARIŞIK VERİ KÜMELERİ ÜZERİNDE ÇALIŞMAK

## 3.1 K-Means Algoritması İle Kategorik Veri Kümelerinde Çalışmak

K-Means algoritması yapısal olarak yani öklit uzaklık ölçütünü kullandığı için sadece sayısal değerlerle çalışır. Oysa gerçek hayatta hem sayısal hem de sözel verilerin karışık bir biçimde bulunduğu veri tabanları mevcuttur. Bu çalışmada ortaya koyulan şey Jaccard benzerlik ölçütünü K-Means prensibiyle çalıştırmaktır. Başka bir deyişle K-Means algoritmasında öklit yerine Jaccard benzerlik ölçütünü kullanarak kategorik olmayan verilerin yanında kategorik verileri de kümeleyebilmektir. Şekil 4' teki K-Means akış şemasından farklı olarak şekil 5'teki gibi öklit yerine Jaccard benzerlik ölçütünün geldiği anlaşılmaktadır.




**Şekil 7. K-Means Algoritması ve Jaccard Benzerlik Ölçütü ile Bulunan Çözümün Akış Şeması**

**Tablo 14. Örnek Kategorik Veri Kümesi**

| | | |
|---|---|---|
| Ela | Kahve | bakımlı |
| Mavi | Sarı | bakımlı |
| Kahve | Siyah | bakımsız |
| Yeşil | Sarı | bakımlı |
| Yeşil | Sarı | bakımlı |
| Kahve | Kahve | bakımsız |
| Kahve | Siyah | bakımsız |
| Kahve | Sarı | bakımlı |
| Mavi | Sarı | bakımlı |

**Örnek 1:** Jaccard benzerlik ölçütü ile örnek veri kümesinin klasik K-Means algoritması gibi üç kümeye ayırma işlemi.

İlk adım alarak veri kümesinin rastgele seçilen elemanları ilk merkezler kabul edilir. Bu örnek için ilk üç satır ilk merkezler kabul edildi.

M1 = ela, kahve, bakımlı

M2 = mavi, sarı, bakımlı

M3 = kahve, siyah, bakımsız

Merkez kümeler bulunduktan sonra sonsuz iterasyon başlatılır. Bu iterasyonlarda tüm küme elemanları ile merkez kümeler arasında Jaccard benzerlik ölçütü uygulanır. Bu uygulama son iterasyondaki merkez kümeler ile bir önceki iterasyondaki merkez kümeler birbirinin aynısı oluncaya kadar devam ettirilir.

**İterasyon 1:**

M1 ve veri kümesi elemanı yeşil, sarı, bakımlı için Jaccard benzerlik ölçütü,

$$\text{benj}(m_1, X_3) = \frac{|ela,kahve,bakımlı| \cap |yeşil,sarı,bakımlı|}{|ela,kahve,bakımlı| \cup |yeşil,sarı,bakımlı|} = 1/5 = 0.2$$

M2 ve veri kümesi elemanı yeşil, sarı, bakımlı için Jaccard benzerlik ölçütü,

$$\text{benj}(m_2, X_3) = \frac{|mavi,sarı,bakımlı| \cap |yeşil,sarı,bakımlı|}{|mavi,sarı,,bakımlı| \cup |yeşil,sarı,bakımlı|} = 2/4 = 0.5$$

M3 ve veri kümesi elemanı yeşil, sarı, bakımlı için Jaccard benzerlik ölçütü,

$$\text{benj}(m_3, X_3) = \frac{|\text{kahve,siyah,bakımsız}| \cap |yeşil,sarı,bakımlı|}{|\text{kahve,siyah,bakımsız}| \cup |yeşil,sarı,bakımlı|} = 0/6 = 0$$

M2>M1>M3 olduğundan yeşil, sarı, bakımlı verisi ikinci kümeye eklendi.

M1 ve veri kümesi elemanı yeşil, sarı, bakımlı için Jaccard benzerlik ölçütü,

$$\text{benj}(m_1, X_4) = \frac{|ela,kahve,bakımlı| \cap |yeşil,sarı,bakımlı|}{|ela,kahve,bakımlı| \cup |yeşil,sarı,bakımlı|} = 1/5 = 0.2$$

M2 ve veri kümesi elemanı yeşil, sarı, bakımlı için Jaccard benzerlik ölçütü,

$$\text{benj}(m_2, X_4) = \frac{|mavi,sarı,bakımlı| \cap |yeşil,sarı,bakımlı|}{|mavi,sarı,,bakımlı| \cup |yeşil,sarı,bakımlı|} = 2/4 = 0.5$$

M3 ve veri kümesi elemanı yeşil, sarı, bakımlı için Jaccard benzerlik ölçütü,

$$\text{benj}(m_3, X_4) = \frac{|\text{kahve,siyah,bakımsız}| \cap |yeşil,sarı,bakımlı|}{|\text{kahve,siyah,bakımsız}| \cup |yeşil,sarı,bakımlı|} = 0/6 = 0$$

M2>M1>M3 olduğundan yeşil, sarı, bakımlı verisi ikinci kümeye eklendi.

M1 ve veri kümesi elemanı kahve, kahve, bakımsız için Jaccard benzerlik ölçütü ,

$$\text{benj}(m_1, X_6) = \frac{|ela,kahve,bakımlı| \cap |\text{kahve,kahve,bakımsız}|}{|ela,kahve,bakımlı| \cup |\text{kahve,kahve,bakımsız}|} = 2/4 = 0.5$$

M2 ve veri kümesi elemanı kahve, kahve, bakımsız için Jaccard benzerlik ölçütü ,

$$\text{benj}(m_2, X_6) = \frac{|mavi,sarı,bakımlı| \cap |\text{kahve,kahve,bakımsız}|}{|mavi,sarı,,bakımlı| \cup |\text{kahve,kahve,bakımsız}|} = 0/5 = 0$$

M3 ve veri kümesi elemanı kahve, kahve, bakımsız için Jaccard benzerlik ölçütü ,

$$\text{benj}(m_3, X_6) = \frac{|\text{kahve,siyah,bakımsız}| \cap |\text{kahve,kahve,bakımsız}|}{|\text{kahve,siyah,bakımsız}| \cup |\text{kahve,kahve,bakımsız}|} = 2/3 = 0.67$$

M3 > M1 > M2 olduğundan kahve, kahve, bakımsız üçüncü kümeye eklendi.

M1 ve veri kümesi elemanı kahve, siyah, bakımsız için Jaccard benzerlik ölçütü ,

$$\text{benj}(m_1, X_7) = \frac{|ela,kahve,bakımlı| \cap |\text{kahve,siyah,bakımsız}|}{|ela,kahve,bakımlı| \cup |\text{kahve,siyah,bakımsız}|} = 1/5 = 0.2$$

M2 ve veri kümesi elemanı kahve, siyah, bakımsız için Jaccard benzerlik ölçütü ,

$$\text{benj}(m_2, X_7) = \frac{|mavi,sarı,bakımlı| \cap |\text{kahve,siyah,bakımsız}|}{|mavi,sarı,,bakımlı| \cup |\text{kahve,siyah,bakımsız}|} = 0/6 = 0$$

M3 ve veri kümesi elemanı kahve, siyah, bakımsız için Jaccard benzerlik ölçütü ,

$$\text{benj}(m_3, X_7) = \frac{|\text{kahve,siyah,bakımsız}| \cap |\text{kahve,siyah,bakımsız}|}{|\text{kahve,siyah,bakımsız}| \cup |\text{kahve,siyah,bakımsız}|} = 3/3 = 1$$

M3 > M1 > M2 kahve, siyah, bakımsız üçüncü kümeye eklendi.

M1 ve veri kümesi elemanı kahve, sarı, bakımlı için Jaccard benzerlik ölçütü ,

$$\text{benj}(m_1, X_8) = \frac{|ela,kahve,bakımlı| \cap |\text{kahve,sarı,bakımlı}|}{|ela,kahve,bakımlı| \cup |\text{kahve,sarı,bakımlı}|} = 2/4 = 0.5$$

M2 ve veri elemanı kahve, sarı, bakımlı için Jaccard benzerlik ölçütü ,

$$\text{benj}(m_2, X_8) = \frac{|mavi,sarı,bakımlı| \cap |\text{kahve,sarı,bakımlı}|}{|mavi,sarı,,bakımlı| \cup |\text{kahve,sarı,bakımlı}|} = 2/4 = 0.5$$

M3 ve veri kümesi elemanı kahve, sarı, bakımlı için Jaccard benzerlik ölçütü ,

$$\text{benj}(m_3, X_8) = \frac{|\text{kahve,siyah,bakımsız}| \cap |kahve,sarı,bakımlı|}{|\text{kahve,siyah,bakımsız}| \cup |kahve,sarı,bakımlı|} = 1/5 = 0.2$$

M2 = M1 > M3 olduğundan kahve, sarı, bakımlı m1 veya m2 kümelerinden bir tanesine eklenebilir. Bu örnekte birinci kümeye eklendi.

M1 ve veri kümesi elemanı mavi, sarı, bakımlı için Jaccard benzerlik ölçütü ,

$$\text{benj}(m_1, X_9) = \frac{|ela,kahve,bakımlı| \cap |\text{mavi,sarı,bakımlı}|}{|ela,kahve,bakımlı| \cup |\text{mavi,sarı,bakımlı}|} = 1/5 = 0.2$$

M2 ve veri kümesi elemanı mavi, sarı, bakımlı için Jaccard benzerlik ölçütü ,

$$\text{benj}(m_2, X_9) = \frac{|mavi,sarı,bakımlı| \cap |\text{mavi,sarı,bakımlı}|}{|mavi,sarı,,bakımlı| \cup |\text{mavi,sarı,bakımlı}|} = 3/3 = 1$$

M3 ve veri kümesi elemanı mavi, sarı, bakımlı için Jaccard benzerlik ölçütü ,

$$\text{benj}(m_3, X_9) = \frac{|\text{kahve,siyah,bakımsız}| \cap |mavi,sarı,bakımlı|}{|\text{kahve,siyah,bakımsız}| \cup |mavi,sarı,bakımlı|} = 2/4 = 0.5$$

M2>M3>M1 olduğundan mavi, sarı, bakımlı ikinci kümeye eklendi.

İterasyon 1 sonunda oluşan kümeler tablo 15'deki gibidir.

**Tablo 15. Jaccard benzerlik ölçütü ile klasik K-Means gibi üçlü kümelemede iterasyon 1 sonucu oluşan kümeler**

| | | | |
|---|---|---|---|
| **Küme 1** | ela | Kahve | bakımlı |
| | kahve | Sarı | bakımlı |
| **Küme 2** | mavi | Sarı | bakımlı |
| | yeşil | Sarı | bakımlı |
| | yeşil | Sarı | bakımlı |
| | mavi | Sarı | bakımlı |
| **Küme 3** | kahve | Siyah | bakımsız |
| | kahve | Kahve | bakımsız |
| | kahve | Siyah | bakımsız |

Yeni küme merkezlerini bulmak için her kümenin sütununda en çok geçen eleman alınır. Eğer eleman sayıları eşitse rastgele olarak alınır. Yeni kümelerden yola çıkılarak yeni küme merkezleri şu şekilde oluşur.

M1 = ela, kahve, bakımlı

M2 = mavi, sarı, bakımlı

M3 = kahve, siyah, bakımsız

Bir önceki küme merkezleri ile en son iterasyondaki küme merkezleri kıyaslandığında birbirlerinin aynısı olduğundan sonsuz iterasyon kırıldı ve en son belirlenen kümeler ayrılmak istenilen kümeler elde edildi.

**Tablo 16. Jaccard benzerlik ölçütü ile klasik K-Means gibi üçlü kümelemede elde edilen kümeler**

| | | | |
|---|---|---|---|
| **Küme 1** | ela | Kahve | bakımlı |
| | kahve | Sarı | bakımlı |
| **Küme 2** | mavi | Sarı | bakımlı |
| | yeşil | Sarı | bakımlı |
| | yeşil | Sarı | bakımlı |
| | mavi | Sarı | bakımlı |
| **Küme 3** | kahve | Siyah | bakımsız |
| | kahve | Kahve | bakımsız |
| | kahve | Siyah | bakımsız |

Anlatılan kümeleme işleminde rastlantı olarak doğru kümeleme sonucu elde edildi fakat kategorik veriler için klasik K-Means gibi öklit yerine, Jaccard benzerlik ölçütü

koyularak yapılan kümelemelerde Tablo 17'deki gibi bir veri kümesi olursa anlatılan hatalar alınır.

**Tablo 17. Örnek Kategorik Veri Kümesi**

| | | |
|---|---|---|
| Ela | Kahve | bakımlı |
| Mavi | Sarı | bakımlı |
| Yeşil | Sarı | bakımlı |
| Yeşil | Sarı | bakımlı |
| Kahve | Siyah | bakımsız |
| Kahve | Kahve | bakımsız |
| Kahve | Siyah | bakımsız |
| Kahve | Sarı | bakımlı |
| Mavi | Sarı | bakımlı |

**Örnek 1:** Jaccard benzerlik ölçütü ile örnek veri kümesinin klasik K-Means algoritması gibi üç kümeye ayırma işlemi.

İlk adım alarak veri kümesinin rastgele seçilen elemanları ilk merkezler kabul edilir. Bu örnek için ilk üç satır ilk merkezler kabul edildi.

M1 = ela, kahve, bakımlı

M2 = mavi, sarı, bakımlı

M3 = yeşil, sarı, bakımlı

Merkez kümeler bulunduktan sonra sonsuz iterasyon başlatılır. Bu iterasyonlarda tüm küme elemanları ile merkez kümeler arasında Jaccard benzerlik ölçütü uygulanır. Bu uygulama son iterasyondaki merkez kümeler ile bir önceki iterasyondaki merkez kümeler birbirinin aynısı oluncaya kadar devam ettirilir.

**İterasyon 1:**

M1 ve veri kümesi elemanı yeşil, sarı, bakımlı için Jaccard benzerlik ölçütü,

$$\text{benj}(m_1, X_3) = \frac{|ela,kahve,bakımlı| \cap |yeşil,sarı,bakımlı|}{|ela,kahve,bakımlı| \cup |yeşil,sarı,bakımlı|} = 1/5 = 0.2$$

M2 ve veri kümesi elemanı yeşil, sarı, bakımlı için Jaccard benzerlik ölçütü,

$$\text{benj}(m_2, X_3) = \frac{|mavi,sarı,bakımlı| \cap |yeşil,sarı,bakımlı|}{|mavi,sarı,,bakımlı| \cup |yeşil,sarı,bakımlı|} = 2/4 = 0.5$$

M3 ve veri kümesi elemanı yeşil, sarı, bakımlı için Jaccard benzerlik ölçütü,

$$\text{benj}(m_3, X_3) = \frac{|\text{yeşil,sarı,bakımlı}| \cap |yeşil,sarı,bakımlı|}{|\text{yeşil,sarı,bakımlı}| \cup |yeşil,sarı,bakımlı|} = 3/3 = 1$$

M3>M2>M1 olduğundan yeşil, sarı, bakımlı verisi üçüncü kümeye eklendi.

M1 ve veri kümesi elemanı yeşil, sarı, bakımlı için Jaccard benzerlik ölçütü,

$$\text{benj}(m_1, X_4) = \frac{|ela,kahve,bakımlı| \cap |yeşil,sarı,bakımlı|}{|ela,kahve,bakımlı| \cup |yeşil,sarı,bakımlı|} = 1/5 = 0.2$$

M2 ve veri kümesi elemanı yeşil, sarı, bakımlı için Jaccard benzerlik ölçütü,

$$\text{benj}(m_2, X_4) = \frac{|mavi,sarı,bakımlı| \cap |yeşil,sarı,bakımlı|}{|mavi,sarı,,bakımlı| \cup |yeşil,sarı,bakımlı|} = 2/4 = 0.5$$

M3 ve veri kümesi elemanı yeşil, sarı, bakımlı için Jaccard benzerlik ölçütü,

$$\text{benj}(m_3, X_4) = \frac{|\text{yeşil,sarı,bakımlı}| \cap |yeşil,sarı,bakımlı|}{|\text{yeşil,sarı,bakımlı}| \cup |yeşil,sarı,bakımlı|} = 3/3 = 1$$

M3 > M2 > M1olduğundan yeşil, sarı, bakımlı verisi üçüncü kümeye eklendi.

M1 ve veri kümesi elemanı kahve, siyah, bakımsız için Jaccard benzerlik ölçütü,

$$\text{benj}(m_1, X_5) = \frac{|ela,kahve,bakımlı| \cap |\text{kahve,siyah,bakımsız}|}{|ela,kahve,bakımlı| \cup |\text{kahve,siyah,bakımsız}|} = 1/5 = 0.2$$

M2 ve veri kümesi elemanı kahve, siyah, bakımsız için Jaccard benzerlik ölçütü,

$$\text{benj}(m_2, X_5) = \frac{|mavi,sarı,bakımlı| \cap |\text{kahve,siyah,bakımsız}|}{|mavi,sarı,,bakımlı| \cup |\text{kahve,siyah,bakımsız}|} = 0/6 = 0$$

M3 ve veri kümesi elemanı kahve, siyah, bakımsız için Jaccard benzerlik ölçütü ,

$$\text{benj}(m_3, X_5) = \frac{|\text{yeşil,sarı,bakımlı}| \cap |\text{kahve,siyah,bakımsız}|}{|\text{yeşil,sarı,bakımlı}| \cup |\text{kahve,siyah,bakımsız}|} = 0/6 = 0$$

M1 > M2 = M3 kahve, siyah, bakımsız verisi birinci kümeye eklendi.

M1 ve veri kümesi elemanı kahve, kahve, bakımsız için Jaccard benzerlik ölçütü ,

$$\text{benj}(m_1, X_6) = \frac{|ela,kahve,bakımlı| \cap |\text{kahve,kahve,bakımsız}|}{|ela,kahve,bakımlı| \cup |\text{kahve,kahve,bakımsız}|} = 2/4 = 0.5$$

M2 ve veri kümesi elemanı kahve, kahve, bakımsız için Jaccard benzerlik ölçütü ,

$$\text{benj}(m_2, X_6) = \frac{|mavi,sarı,bakımlı| \cap |\text{kahve,kahve,bakımsız}|}{|mavi,sarı,,bakımlı| \cup |\text{kahve,kahve,bakımsız}|} = 0/5 = 0$$

M3 ve veri kümesi elemanı kahve, kahve, bakımsız için Jaccard benzerlik ölçütü ,

$$\text{benj}(m_3, X_6) = \frac{|\text{yeşil,sarı,bakımlı}| \cap |\text{kahve,kahve,bakımsız}|}{|\text{yeşil,sarı,bakımlı}| \cup |\text{kahve,kahve,bakımsız}|} = 0/5 = 0$$

M1 > M2 = M3 olduğundan kahve, kahve, bakımsız birinci kümeye eklendi.

M1 ve veri kümesi elemanı kahve, siyah, bakımsız için Jaccard benzerlik ölçütü ,

$$\text{benj}(m_1, X_7) = \frac{|ela,kahve,bakımlı| \cap |\text{kahve,siyah,bakımsız}|}{|ela,kahve,bakımlı| \cup |\text{kahve,siyah,bakımsız}|} = 1/5 = 0.2$$

M2 ve veri kümesi elemanı kahve, siyah, bakımsız için Jaccard benzerlik ölçütü ,

$$\text{benj}(m_2, X_7) = \frac{|mavi,sarı,bakımlı| \cap |\text{kahve,siyah,bakımsız}|}{|mavi,sarı,,bakımlı| \cup |\text{kahve,siyah,bakımsız}|} = 0/6 = 0$$

M3 ve veri kümesi elemanı kahve, siyah, bakımsız için Jaccard benzerlik ölçütü ,

$$\text{benj}(m_3, X_7) = \frac{|\text{yeşil,sarı,bakımlı}| \cap |\text{kahve,siyah,bakımsız}|}{|\text{yeşil,sarı,bakımlı}| \cup |\text{kahve,siyah,bakımsız}|} = 0/6 = 0$$

M3 > M2 = M1 kahve, siyah, bakımsız üçüncü kümeye eklendi.

M1 ve veri kümesi elemanı kahve, sarı, bakımlı için Jaccard benzerlik ölçütü ,

$$\text{benj}(m_1, X_8) = \frac{|ela,kahve,bakımlı| \cap |\text{kahve,sarı,bakımlı}|}{|ela,kahve,bakımlı| \cup |\text{kahve,sarı,bakımlı}|} = 2/4 = 0.5$$

M2 ve veri elemanı kahve, sarı, bakımlı için Jaccard benzerlik ölçütü ,

$$\text{benj}(m_2, X_8) = \frac{|mavi,sarı,bakımlı| \cap |\text{kahve,sarı,bakımlı}|}{|mavi,sarı,,bakımlı| \cup |\text{kahve,sarı,bakımlı}|} = 2/4 = 0.5$$

M3 ve veri kümesi elemanı kahve, sarı, bakımlı için Jaccard benzerlik ölçütü ,

$$\text{benj}(m_3, X_8) = \frac{|\text{yeşil,sarı,bakımlı}| \cap |kahve,sarı,bakımlı|}{|\text{yeşil,sarı,bakımlı}| \cup |kahve,sarı,bakımlı|} = 2/4 = 0.5$$

M2 = M1 = M3 olduğundan kahve, sarı, bakımlı M1, M2 veya M3 kümelerinden bir tanesine eklenebilir. Bu örnekte ikinci kümeye eklendi.

M1 ve veri kümesi elemanı mavi, sarı, bakımlı için Jaccard benzerlik ölçütü ,

benj($m_1, X_9$) = $\frac{|ela,kahve,bakımlı| \cap |\text{mavi,sarı,bakımlı}|}{|ela,kahve,bakımlı| \cup |\text{mavi,sarı,bakımlı}|}$ = 1/5 = 0.2

M2 ve veri kümesi elemanı mavi, sarı, bakımlı için Jaccard benzerlik ölçütü ,

benj($m_2, X_9$) = $\frac{|mavi,sarı,bakımlı| \cap |\text{mavi,sarı,bakımlı}|}{|mavi,sarı,,bakımlı| \cup |\text{mavi,sarı,bakımlı}|}$ = 3/3= 1

M3 ve veri kümesi elemanı mavi, sarı, bakımlı için Jaccard benzerlik ölçütü ,

benj($m_3, X_9$) = $\frac{|\text{yeşil,sarı,bakımlı}| \cap |mavi,sarı,bakımlı|}{|\text{yeşil,sarı,bakımlı}| \cup |mavi,sarı,bakımlı|}$ = 2/4 = 0.5

M2>M3>M1 olduğundan mavi, sarı, bakımlı birinci kümeye eklendi.

İterasyon 1 sonunda oluşan kümeler tablo 18'deki gibidir.

**Tablo 18. Jaccard benzerlik ölçütü ile klasik K-Means gibi üçlü kümelemede iterasyon 1 sonucu oluşan kümeler**

| | | | |
|---|---|---|---|
| **Küme 1** | Ela | Kahve | Bakımlı |
| | Kahve | Siyah | Bakımsız |
| | Kahve | Kahve | Bakımsız |
| | Kahve | Siyah | Bakımsız |
| **Küme 2** | kahve | Sarı | bakımlı |
| | Mavi | Sarı | bakımlı |
| | Mavi | Sarı | bakımlı |
| **Küme 3** | Yeşil | Sarı | bakımlı |
| | Yeşil | Sarı | bakımlı |

Yeni küme merkezlerini bulmak için her kümenin sütununda en çok geçen eleman alınır. Eğer eleman sayıları eşitse rastgele olarak alınır. Yeni kümelerden yola çıkılarak yeni küme merkezleri şu şekilde oluşur.

M1 = kahve, kahve, bakımsız

M2 = mavi, sarı, bakımlı

M3 = yeşil, sarı, bakımlı

Bir önceki küme merkezleri ile en son iterasyondaki küme merkezleri kıyaslandığında birbirlerinin aynısı olmadığından yeni küme merkezlerine göre başka bir iterasyon başlatılır ve bu iterasyonda tüm küme elemanları ile merkez kümeler arasında Jaccard benzerlik ölçütü tekrardan uygulanır.

**İterasyon 2:**

M1 ve veri kümesi elemanı ela, kahve, bakımlı için Jaccard benzerlik ölçütü ,

$$\text{benj}(m_1, X_1) = \frac{|\text{kahve,kahve,bakımsız}| \cap |ela,kahve,bakımlı|}{|\text{kahve,kahve,bakımsız}| \cup |ela,kahve,bakımlı|} = 1/4 = 0.25$$

M2 ve veri kümesi elemanı ela, kahve, bakımlı için Jaccard benzerlik ölçütü ,

$$\text{benj}(m_2, X_1) = \frac{|mavi,sarı,bakımlı| \cap |ela,kahve,bakımlı|}{|mavi,sarı,,bakımlı| \cup |ela,kahve,bakımlı|} = 1/5 = 0.2$$

M3 ve veri kümesi elemanı ela, kahve, bakımlı için Jaccard benzerlik ölçütü ,

$$\text{benj}(m_3, X_1) = \frac{|\text{yeşil,sarı,bakımlı}| \cap |ela,kahve,bakımlı|}{|\text{yeşil,sarı,bakımlı}| \cup |ela,kahve,bakımlı|} = 1/5 = 0.2$$

M1>M2=M1 olduğundan ela, kahve, bakımlı verisi üçüncü kümeye eklendi.

M1 ve veri kümesi elemanı mavi, sarı, bakımlı için Jaccard benzerlik ölçütü ,

$$\text{benj}(m_1, X_2) = \frac{|\text{kahve,kahve,bakımsız}| \cap |mavi,sarı,bakımlı|}{|\text{kahve,kahve,bakımsız}| \cup |mavi,sarı,bakımlı|} = 0/6 = 0$$

M2 ve veri kümesi elemanı mavi, sarı, bakımlı için Jaccard benzerlik ölçütü ,

$$\text{benj}(m_2, X_2) = \frac{|mavi,sarı,bakımlı| \cap |mavi,sarı,bakımlı|}{|mavi,sarı,,bakımlı| \cup |mavi,sarı,bakımlı|} = 3/3 = 1$$

M3 ve veri kümesi elemanı mavi, sarı, bakımlı için Jaccard benzerlik ölçütü ,

$$\text{benj}(m_3, X_2) = \frac{|\text{yeşil,sarı,bakımlı}| \cap |mavi,sarı,bakımlı|}{|\text{yeşil,sarı,bakımlı}| \cup |mavi,sarı,bakımlı|} = 2/4 = 0.5$$

M2 > M3 > M1olduğundan mavi, sarı, bakımlı verisi ikinci kümeye eklendi.

M1 ve veri kümesi elemanı yeşil, sarı, bakımlı için Jaccard benzerlik ölçütü ,

$$\text{benj}(m_1, X_3) = \frac{|\text{kahve,kahve,bakımsız}| \cap |yeşil,sarı,bakımlı|}{|\text{kahve,kahve,bakımsız}| \cup |yeşil,sarı,bakımlı|} = 1/5 = 0.2$$

M2 ve veri kümesi elemanı yeşil, sarı, bakımlı için Jaccard benzerlik ölçütü ,

$$\text{benj}(m_2, X_3) = \frac{|mavi,sarı,bakımlı| \cap |yeşil,sarı,bakımlı|}{|mavi,sarı,,bakımlı| \cup |yeşil,sarı,bakımlı|} = 2/4 = 0.5$$

M3 ve veri kümesi elemanı yeşil, sarı, bakımlı için Jaccard benzerlik ölçütü ,

$$\text{benj}(m_3, X_3) = \frac{|\text{yeşil,sarı,bakımlı}| \cap |yeşil,sarı,bakımlı|}{|\text{yeşil,sarı,bakımlı}| \cup |yeşil,sarı,bakımlı|} = 3/3 = 1$$

M3>M2>M1 olduğundan yeşil, sarı, bakımlı verisi üçüncü kümeye eklendi.

M1 ve veri kümesi elemanı yeşil, sarı, bakımlı için Jaccard benzerlik ölçütü ,

$$\text{benj}(m_1, X_4) = \frac{|\text{kahve,kahve,bakımsız}| \cap |yeşil,sarı,bakımlı|}{|\text{kahve,kahve,bakımsız}| \cup |yeşil,sarı,bakımlı|} = 1/5 = 0.2$$

M2 ve veri kümesi elemanı yeşil, sarı, bakımlı için Jaccard benzerlik ölçütü ,

$$\text{benj}(m_2, X_4) = \frac{|mavi,sarı,bakımlı| \cap |yeşil,sarı,bakımlı|}{|mavi,sarı,,bakımlı| \cup |yeşil,sarı,bakımlı|} = 2/4 = 0.5$$

M3 ve veri kümesi elemanı yeşil, sarı, bakımlı için Jaccard benzerlik ölçütü ,

$$\text{benj}(m_3, X_4) = \frac{|\text{yeşil,sarı,bakımlı}| \cap |yeşil,sarı,bakımlı|}{|\text{yeşil,sarı,bakımlı}| \cup |yeşil,sarı,bakımlı|} = 3/3 = 1$$

M3 > M2 > M1olduğundan yeşil, sarı, bakımlı verisi üçüncü kümeye eklendi.

M1 ve veri kümesi elemanı kahve, siyah, bakımsız için Jaccard benzerlik ölçütü ,

$$\text{benj}(m_1, X_5) = \frac{|\text{kahve,kahve,bakımsız}| \cap |\text{kahve,siyah,bakımsız}|}{|\text{kahve,kahve,bakımsız}| \cup |\text{kahve,siyah,bakımsız}|} = 2/3 = 0.67$$

M2 ve veri kümesi elemanı kahve, siyah, bakımsız için Jaccard benzerlik ölçütü ,

$$\text{benj}(m_2, X_5) = \frac{|mavi,sarı,bakımlı| \cap |\text{kahve,siyah,bakımsız}|}{|mavi,sarı,,bakımlı| \cup |\text{kahve,siyah,bakımsız}|} = 0/6 = 0$$

M3 ve veri kümesi elemanı kahve, siyah, bakımsız için Jaccard benzerlik ölçütü ,

$$\text{benj}(m_3, X_5) = \frac{|\text{yeşil,sarı,bakımlı}| \cap |\text{kahve,siyah,bakımsız}|}{|\text{yeşil,sarı,bakımlı}| \cup |\text{kahve,siyah,bakımsız}|} = 0/6 = 0$$

M1 > M2 = M3 kahve, siyah, bakımsız verisi birinci kümeye eklendi.

M1 ve veri kümesi elemanı kahve, kahve, bakımsız için Jaccard benzerlik ölçütü ,

$$\text{benj}(m_1, X_6) = \frac{|\text{kahve,kahve,bakımsız}| \cap |\text{kahve,kahve,bakımsız}|}{|\text{kahve,kahve,bakımsız}| \cup |\text{kahve,kahve,bakımsız}|} = 3/3 = 1$$

M2 ve veri kümesi elemanı kahve, kahve, bakımsız için Jaccard benzerlik ölçütü ,

$$\text{benj}(m_2, X_6) = \frac{|mavi,sarı,bakımlı| \cap |\text{kahve,kahve,bakımsız}|}{|mavi,sarı,,bakımlı| \cup |\text{kahve,kahve,bakımsız}|} = 0/5 = 0$$

M3 ve veri kümesi elemanı kahve, kahve, bakımsız için Jaccard benzerlik ölçütü ,

$$\text{benj}(m_3, X_6) = \frac{|\text{yeşil,sarı,bakımlı}| \cap |\text{kahve,kahve,bakımsız}|}{|\text{yeşil,sarı,bakımlı}| \cup |\text{kahve,kahve,bakımsız}|} = 0/5 = 0$$

M1 > M2 = M3 olduğundan kahve, kahve, bakımsız birinci kümeye eklendi.

M1 ve veri kümesi elemanı kahve, siyah, bakımsız için Jaccard benzerlik ölçütü ,

$$\text{benj}(m_1, X_7) = \frac{|\text{kahve,kahve,bakımsız}| \cap |\text{kahve,siyah,bakımsız}|}{|\text{kahve,kahve,bakımsız}| \cup |\text{kahve,siyah,bakımsız}|} = 2/3 = 0.67$$

M2 ve veri kümesi elemanı kahve, siyah, bakımsız için Jaccard benzerlik ölçütü ,

$$\text{benj}(m_2, X_7) = \frac{|mavi,sarı,bakımlı| \cap |\text{kahve,siyah,bakımsız}|}{|mavi,sarı,,bakımlı| \cup |\text{kahve,siyah,bakımsız}|} = 0/6 = 0$$

M3 ve veri kümesi elemanı kahve, siyah, bakımsız için Jaccard benzerlik ölçütü ,

$$\text{benj}(m_3, X_7) = \frac{|\text{yeşil,sarı,bakımlı}| \cap |\text{kahve,siyah,bakımsız}|}{|\text{yeşil,sarı,bakımlı}| \cup |\text{kahve,siyah,bakımsız}|} = 0/6 = 0$$

M3 >M2 = M1 kahve, siyah, bakımsız üçüncü kümeye eklendi.

M1 ve veri kümesi elemanı kahve, sarı, bakımlı için Jaccard benzerlik ölçütü ,

$$\text{benj}(m_1, X_8) = \frac{|\text{kahve,kahve,bakımsız}| \cap |\text{kahve,sarı,bakımlı}|}{|\text{kahve,kahve,bakımsız}| \cup |\text{kahve,sarı,bakımlı}|} = 1/4 = 0.25$$

M2 ve veri elemanı kahve, sarı, bakımlı için Jaccard benzerlik ölçütü ,

$$\text{benj}(m_2, X_8) = \frac{|mavi,sarı,bakımlı| \cap |\text{kahve,sarı,bakımlı}|}{|mavi,sarı,,bakımlı| \cup |\text{kahve,sarı,bakımlı}|} = 2/4 = 0.5$$

M3 ve veri kümesi elemanı kahve, sarı, bakımlı için Jaccard benzerlik ölçütü ,

$$\text{benj}(m_3, X_8) = \frac{|\text{yeşil,sarı,bakımlı}| \cap |kahve,sarı,bakımlı|}{|\text{yeşil,sarı,bakımlı}| \cup |kahve,sarı,bakımlı|} = 2/4 = 0.5$$

M2 = M3 >M1 olduğundan kahve, sarı, bakımlı m2 veya m3 kümelerinden bir tanesine eklenebilir. Bu örnekte birinci m2 kümesine eklendi.

M1 ve veri kümesi elemanı mavi, sarı, bakımlı için Jaccard benzerlik ölçütü ,

$$\text{benj}(m_1, X_9) = \frac{|\text{kahve,kahve,bakımsız}| \cap |\text{mavi,sarı,bakımlı}|}{|\text{kahve,kahve,bakımsız}| \cup |\text{mavi,sarı,bakımlı}|} = 0/6 = 0$$

M2 ve veri kümesi elemanı mavi, sarı, bakımlı için Jaccard benzerlik ölçütü ,

$$\text{benj}(m_2, X_9) = \frac{|mavi,sarı,bakımlı| \cap |\text{mavi,sarı,bakımlı}|}{|mavi,sarı,,bakımlı| \cup |\text{mavi,sarı,bakımlı}|} = 3/3= 1$$

M3 ve veri kümesi elemanı mavi, sarı, bakımlı için Jaccard benzerlik ölçütü ,

$$\text{benj}(m_3, X_9) = \frac{|\text{yeşil,sarı,bakımlı}| \cap |mavi,sarı,bakımlı|}{|\text{yeşil,sarı,bakımlı}| \cup |mavi,sarı,bakımlı|} = 2/4 = 0.5$$

M2>M3>M1 olduğundan mavi, sarı, bakımlı birinci kümeye eklendi.

**Tablo 19. Jaccard benzerlik ölçütü ile klasik K-Means gibi üçlü kümelemede iterasyon 2 sonucu oluşan kümeler**

| | | | |
|---|---|---|---|
| **Küme 1** | Ela | Kahve | Bakımlı |
| | Kahve | Siyah | Bakımsız |
| | Kahve | Kahve | Bakımsız |
| | Kahve | Siyah | Bakımsız |
| **Küme 2** | kahve | Sarı | Bakımlı |
| | Mavi | Sarı | Bakımlı |
| | Mavi | Sarı | Bakımlı |
| **Küme 3** | Yeşil | Sarı | Bakımlı |
| | Yeşil | Sarı | Bakımlı |

İterasyon 2 sonucu oluşan yeni merkezler şu şekildedir.

M1 = kahve, kahve, bakımsız

M2 = mavi, sarı, bakımlı

M3 = yeşil, sarı, bakımlı

Bir önceki iterasyondaki küme merkezleri ile en son iterasyondaki küme merkezleri kıyaslandığında birbirlerinin aynısı olduğundan sonsuz iterasyon kırılarak algoritma sonlandırılır. Buna göre örnek veri kümesi üç kümeye ayrılmak istendiğinde tablo 20'deki sonuçlar alınmaktadır.

**Tablo 20. Kategorik verilerin üç kümeye klasik K-Means algoritması gibi ayrıldığındaki sonuçlar**

| | | | |
|---|---|---|---|
| **Küme 1** | Ela | Kahve | Bakımlı |
| | Kahve | Siyah | Bakımsız |
| | Kahve | Kahve | Bakımsız |
| | Kahve | Siyah | Bakımsız |
| **Küme 2** | kahve | Sarı | Bakımlı |
| | Mavi | Sarı | bakımlı |
| | Mavi | Sarı | bakımlı |
| **Küme 3** | Yeşil | Sarı | bakımlı |
| | Yeşil | Sarı | bakımlı |

Tablo 20 incelendiğinde mavi, sarı, bakımlı verileri ile yeşil, sarı, bakımlı verilerinin birbirlerine çok benzediği ve aslında aynı küme içinde olmaları gerektiği halde farklı kümelerin içinde olduğu açıkça görülüyor. Bunun nedeni ise ilk merkez kümeleri belirlenirken klasik K-Means algoritmasına göre belirlenmesidir. Bu hatayı gidermek için ilk kümeler belirlenirken klasik K-Means algoritması yerine şu şekilde bir yol izlenmelidir. İlk olarak K-Means algoritması gibi veri kümesinin kaç kümeye ayrılacağı belirlenmelidir. Bundan sonraki adımda K-Means algoritmasından farklı olarak ilk merkez kümeler rastgele değil veri kümesi içinden özel hesaplamalarla bulunması gerekmektedir.

İlk merkez kümeleri bulmak için:

- Veri kümesinin ilk elemanı ilk merkez kümesi
- İkinci merkez küme, ilk merkez küme ile diğer veri kümesi elemanlarına Jaccard benzerlik ölçütü uygulandığında 0'a en yakın yani en uzak olan elemandır.

- Eğer küme ikiden fazla kümeye ayrılacaksa bir önce bulunan merkez kümelerle veri kümesindeki veriler arasında teker teker jaccard benzerlik ölçütü alınır. Veri kümesi elemanı kaç merkez küme ile benzerlik ölçütüne sokulduysa tüm çıkan sonuçlar toplanır ve o sayıya bölünür. En küçük olan değer diğer küme merkezi olur.

İlk merkez kümeler bulunduktan sonra K-Means algoritması içerisinde Jaccard benzerlik ölçütü her küme elemanı için uygulanır ve bu uygulama sonucunda elde bölünmek istenilen kadar küme olur. Bu kümelerin içerisinde sütun bazında bakıldığında sayısı en fazla olan kayıtlar oluşan kümelerin yeni merkezlerini oluşturur. Daha sonra bu işlemler sonsuz bir iterasyon içerisinde tekrar edilir taki en son küme merkezleri bir önceki küme merkezlerinin aynısı oluncaya kadar. Bu aşamadan sonra sonsuz iterasyon kırılarak bölünmek istenilen sayı kadar küme ve bu kümelerin içerisinde benzerliği en fazla olan veriler bulunur. Aynı K-Means algoritmasındaki gibi aynı küme içerisindeki elemanların benzerliği en fazla farklı kümelerin içindeki elemanların birbirlerine benzerliği en azdır. Fakat burada dikkat edilmesi gereken husus veri kümesi elemanları kümelenmek istendiğinde optimum olarak kaç kümeye ayrılabileceğinin iyi hesaplanmasıdır. [14].

**Örnek 2:** K-Means algoritması içerisinde Jaccard benzerlik ölçütü ile örnek veri kümesini iki kümeye ayırma işlemi.

İlk adım alarak veri kümesinin ilk elemanı ilk kümenin merkezi seçilir.

M1 = ela, kahve, bakımlı

İkinci adım olarak Jaccard benzerlik ölçütü kullanılarak ilk merkez kümeye en uzak olan veri ikinci merkez küme seçilir.

M1 ve veri kümesi elemanı mavi, sarı, bakımlı için Jaccard benzerlik ölçütü,

$$\text{benj}(X_1, X_2) = \frac{|ela,kahve,bakımlı| \cap |mavi,sarı,bakımlı|}{|ela,kahve,bakımlı| \cup |mavi,sarı,bakımlı|} = 1/5 = 0.2$$

M1 ve veri kümesi elemanı yeşil, sarı, bakımlı için Jaccard benzerlik ölçütü,

$$\text{benj}(X_1, X_3) = \frac{|ela,kahve,bakımlı| \cap |yeşil,sarı,bakımlı|}{|ela,kahve,bakımlı| \cup |yeşil,sarı,bakımlı|} = 1/5 = 0.2$$

M1 ve veri kümesi elemanı yeşil, sarı, bakımlı için Jaccard benzerlik ölçütü,

benj($X_1, X_4$) $= \frac{|ela,kahve,bakımlı| \cap |yeşil,sarı,bakımlı|}{|ela,kahve,bakımlı| \cup |yeşil,sarı,bakımlı|}$ = 1/5 = 0.2

M1 ve veri kümesi elemanı kahve, siyah,bakımsız için Jaccard benzerlik ölçütü,

benj($X_1, X_5$) $= \frac{|ela,kahve,bakımlı| \cap |kahve,siyah,bakımsız|}{|ela,kahve,bakımlı| \cup |kahve,siyah,bakımsız|}$ = 1/5 = 0.2

M1 ve veri kümesi elemanı kahve, kahve, bakımsız için Jaccard benzerlik ölçütü,

benj($X_1, X_6$) $= \frac{|ela,kahve,bakımlı| \cap |kahve,kahve,bakımsız|}{|ela,kahve,bakımlı| \cup |kahve,kahve,bakımsız|}$ = 2/4 = 0.5

M1 ve veri kümesi elemanı kahve, siyah, bakımsız için Jaccard benzerlik ölçütü,

benj($X_1, X_7$) $= \frac{|ela,kahve,bakımlı| \cap |kahve,siyah,bakımsız|}{|ela,kahve,bakımlı| \cup |kahve,siyah,bakımsız|}$ = 1/5 = 0.2

M1 ve veri kümesi elemanı kahve, sarı, bakımlı için Jaccard benzerlik ölçütü,

benj($X_1, X_8$) $= \frac{|ela,kahve,bakımlı| \cap |kahve,sarı,bakımlı|}{|ela,kahve,bakımlı| \cup |kahve,sarı,bakımlı|}$ = 2/4 = 0.5

M1 ve veri kümesi elemanı mavi, sarı, bakımlı için Jaccard benzerlik ölçütü,

benj($X_1, X_9$) $= \frac{|ela,kahve,bakımlı| \cap |mavi,sarı,bakımlı|}{|ela,kahve,bakımlı| \cup |mavi,sarı,bakımlı|}$ = 1/5 = 0.2

İlk merkez kümeye en uzak olan kümeler benj($X_1, X_2$) , benj($X_1, X_3$) , benj($X_1, X_4$) , benj($X_1, X_5$) , benj($X_1, X_7$) ve benj($X_1, X_9$). Birden fazla en uzak nokta olduğu için rastgele bir tanesi ikinci merkez küme seçilir. Bu örnekte benj($X_1, X_2$) ikinci merkez küme seçildi.

M1 = ela, kahve, bakımlı

M2 = mavi, sarı, bakımlı

Merkez küme bulma işlemleri tamamlandıktan sonra sonsuz iterasyon başlatılır. Bu iterasyonların içerisinde bulunan merkezler ile tüm veri kümesi Jaccard benzerlik ölçütü ile kıyaslanıp değerce büyük olan o merkeze yakın olarak kabul edilip o yakın olduğu merkezin kümesine eklenir.

**İterasyon 1:**

M1 ve veri kümesi yeşil, sarı, bakımlı için Jaccard benzerlik ölçütü

benj($m_1, X_3$) = $\frac{|ela,kahve,bakımlı| \cap |yeşil,sarı,bakımlı|}{|ela,kahve,bakımlı| \cup |yeşil,sarı,bakımlı|}$ = 1/5 = 0.2

M2 ve veri kümesi yeşil, sarı, bakımlı için Jaccard benzerlik ölçütü

benj($m_2, X_3$) = $\frac{|mavi,sarı,bakımlı| \cap |yeşil,sarı,bakımlı|}{|mavi,sarı,,bakımlı| \cup |yeşil,sarı,bakımlı|}$ = 2/4 = 0.5

M2>M1 olduğundan yeşil, sarı, bakımlı ikinci kümeye eklendi.

M1 ve veri kümesi yeşil, sarı, bakımlı için Jaccard benzerlik ölçütü

benj($m_1, X_4$) = $\frac{|ela,kahve,bakımlı| \cap |yeşil,sarı,bakımlı|}{|ela,kahve,bakımlı| \cup |yeşil,sarı,bakımlı|}$ = 1/5 = 0.2

M2 ve veri kümesi yeşil, sarı, bakımlı için Jaccard benzerlik ölçütü

benj($m_2, X_4$) = $\frac{|mavi,sarı,bakımlı| \cap |yeşil,sarı,bakımlı|}{|mavi,sarı,,bakımlı| \cup |yeşil,sarı,bakımlı|}$ = 2/4 = 0.5

M2>M1 olduğundan yeşil, sarı, bakımlı ikinci kümeye eklendi.

M1 ve veri kümesi kahve, siyah, bakımsız için Jaccard benzerlik ölçütü

benj($m_1, X_5$) = $\frac{|ela,kahve,bakımlı| \cap |kahve,siyah,bakımsız|}{|ela,kahve,bakımlı| \cup |kahve,siyah,bakımsız|}$ = 1/5 = 0.2

M2 ve veri kümesi kahve, siyah, bakımsız için Jaccard benzerlik ölçütü

benj($m_2, X_5$) = $\frac{|mavi,sarı,bakımlı| \cap |kahve,siyah,bakımsız|}{|mavi,sarı,,bakımlı| \cup |kahve,siyah,bakımsız|}$ = 0/6 = 0

M1>M2 kahve, siyah, bakımsız birinci kümeye eklendi.

M1 ve veri kümesi kahve, kahve, bakımsız için Jaccard benzerlik ölçütü

benj($m_1, X_6$) = $\frac{|ela,kahve,bakımlı| \cap |kahve,kahve,bakımsız|}{|ela,kahve,bakımlı| \cup |kahve,kahve,bakımsız|}$ = 2/4 = 0.5

M2 ve veri kümesi kahve, kahve, bakımsız için Jaccard benzerlik ölçütü benj($m_2, X_6$) = $\frac{|mavi,sarı,bakımlı| \cap |kahve,kahve,bakımsız|}{|mavi,sarı,,bakımlı| \cup |kahve,kahve,bakımsız|}$ = 0/5 = 0

M1>m2 olduğundan kahve, kahve, bakımsız birinci kümeye eklendi.

M1 ve veri kümesi kahve, siyah, bakımsız için Jaccard benzerlik ölçütü

$$\text{benj}(m_1, X_7) = \frac{|ela,kahve,bakımlı| \cap |\text{kahve,siyah,bakımsız}|}{|ela,kahve,bakımlı| \cup |\text{kahve,siyah,bakımsız}|} = 1/5 = 0.2$$

M2 ve veri kümesi kahve, siyah, bakımsız için Jaccard benzerlik ölçütü

$$\text{benj}(m_2, X_7) = \frac{|mavi,sarı,bakımlı| \cap |\text{kahve,siyah,bakımsız}|}{|mavi,sarı,,bakımlı| \cup |\text{kahve,siyah,bakımsız}|} = 0/6 = 0$$

M1>M2 kahve, siyah, bakımsız birinci kümeye eklendi.

M1 ve veri kümesi kahve, sarı, bakımlı için Jaccard benzerlik ölçütü

$$\text{benj}(m_1, X_8) = \frac{|ela,kahve,bakımlı| \cap |\text{kahve,sarı,bakımlı}|}{|ela,kahve,bakımlı| \cup |\text{kahve,sarı,bakımlı}|} = 2/4 = 0.5$$

M2 ve veri kahve, sarı, bakımlı için Jaccard benzerlik ölçütü

$$\text{benj}(m_2, X_8) = \frac{|mavi,sarı,bakımlı| \cap |\text{kahve,sarı,bakımlı}|}{|mavi,sarı,,bakımlı| \cup |\text{kahve,sarı,bakımlı}|} = 2/4 = 0.5$$

M2 = M1 olduğundan kahve, sarı, bakımlı iki kümeye de eklenebilir. Bu örnekte birinci kümeye eklendi.

M1 ve veri kümesi mavi, sarı, bakımlı için Jaccard benzerlik ölçütü

$$\text{benj}(m_1, X_9) = \frac{|ela,kahve,bakımlı| \cap |\text{mavi,sarı,bakımlı}|}{|ela,kahve,bakımlı| \cup |\text{mavi,sarı,bakımlı}|} = 1/5 = 0.2$$

M2 ve veri kümesi yeşil, sarı, bakımlı için Jaccard benzerlik ölçütü

$$\text{benj}(m_2, X_9) = \frac{|mavi,sarı,bakımlı| \cap |\text{mavi,sarı,bakımlı}|}{|mavi,sarı,,bakımlı| \cup |\text{mavi,sarı,bakımlı}|} = 3/3 = 1$$

M2>M1 olduğundan mavi, sarı, bakımlı ikinci kümeye eklendi.

İterasyon 1 sonunda oluşan kümeler şu şekildedir.

**Tablo 21. Kategorik verilerde K-Means algoritması örnek 2 iterasyon 1 sonucu bulunan ilk kümeler**

| | | | |
|---|---|---|---|
| **Küme 1** | Ela | Kahve | Bakımlı |
| | Kahve | Siyah | Bakımsız |
| | Kahve | Kahve | Bakımsız |
| | Kahve | Siyah | Bakımsız |
| | Kahve | Sarı | Bakımlı |
| **Küme 2** | Mavi | Sarı | Bakımlı |
| | Yeşil | Sarı | Bakımlı |
| | Yeşil | Sarı | Bakımlı |
| | Mavi | Sarı | Bakımlı |

İlk kümeler bulunduktan sonra yeni küme merkezlerinin bulunması gerekmektedir. Bu işlemi de her küme için ayrı ayrı olarak kümenin her sütununda en fazla geçen değer o sütün için yeni merkezdir. Örneğin Küme1 için ilk sütunda en fazla kahve, ikinci sütunda en fazla kahve veya siyah ve üçüncü sütunda ise en fazla olarak bakımsız bulunmaktadır. O halde Küme1 için yeni küme merkezi kahve, kahve, bakımsız yada kahve, siyah, bakımsızdır. Küme2 içinse küme merkezleri mavi, sarı, bakımlı ya da yeşil, sarı, bakımlıdır.

Bu sonuçtan yola çıkarak yeni küme merkezleri şu şekilde oluşmaktadır.

M1 = kahve, kahve, bakımsız

M2 = yeşil, sarı, bakımlı

Yeni merkezler bulunduktan sora en son bulunan merkez kümeler ile bir önceki merkez kümelerin aynı olup olmadığı kontrol edilir. Eğer bu örnekteki gibi merkez kümeler yer değiştirmişse yeni iterasyona başlatılır.

**İterasyon 2:**

M1 ve veri kümesi ela, kahve, bakımlı için Jaccard benzerlik ölçütü

$$\text{benj}(m_1, X_1) = \frac{|\text{kahve,kahve,bakımsız}| \cap |\text{ela,kahve,bakımlı}|}{|\text{kahve,kahve,bakımsız}| \cup |\text{ela,kahve,bakımlı}|} = 1/4 = 0.25$$

M2 ve veri kümesi yeşil, sarı, bakımlı için Jaccard benzerlik ölçütü

$$\text{benj}(m_2, X_1) = \frac{|yeşil,sarı,bakımlı| \cap |\text{ela,kahve,bakımlı}|}{|yşil,sarı,,bakımlı| \cup |\text{ela,kahve,bakımlı}|} = 1/5 = 0.2$$

M1>M2 olduğundan ela, kahve, bakımlı birinci kümeye eklendi.

M1 ve veri kümesi mavi, sarı, bakımlı için Jaccard benzerlik ölçütü

$$\text{benj}(m_1, X_2) = \frac{|\text{kahve,kahve,bakımsız}| \cap |mavi,sarı,bakımlı|}{|\text{kahve,kahve,bakımsız}| \cup |mavi,sarı,bakımlı|} = 0/5 = 0$$

M2 ve veri kümesi yeşil, sarı, bakımlı için Jaccard benzerlik ölçütü

$$\text{benj}(m_2, X_2) = \frac{|yeşil,sarı,bakımlı| \cap |mavi,sarı,bakımlı|}{|yeşil,sarı,,bakımlı| \cup |mavi,sarı,bakımlı|} = 2/4 = 0.5$$

M2>m1 olduğundan mavi, sarı, bakımlı ikinci kümeye eklendi.

M1 ve veri kümesi yeşil, sarı, bakımlı için Jaccard benzerlik ölçütü

$$\text{benj}(m_1, X_3) = \frac{|\text{kahve,kahve,bakımsız}| \cap |yeşil,sarı,bakımlı|}{|\text{kahve,kahve,bakımsız}| \cup |yeşil,sarı,bakımlı|} = 0/5 = 0$$

M2 ve veri kümesi yeşil, sarı, bakımlı için Jaccard benzerlik ölçütü

$$\text{benj}(m_2, X_3) = \frac{|yeşil,sarı,bakımlı| \cap |yeşil,sarı,bakımlı|}{|yeşil,sarı,,bakımlı| \cup |yeşil,sarı,bakımlı|} = 3/3 = 1$$

M2>M1 olduğundan yeşil, sarı, bakımlı ikinci kümeye eklendi.

M1 ve veri kümesi yeşil, sarı, bakımlı için Jaccard benzerlik ölçütü

$$\text{benj}(m_1, X_4) = \frac{|\text{kahve,kahve,bakımsız}| \cap |yeşil,sarı,bakımlı|}{|\text{kahve,kahve,bakımsız}| \cup |yeşil,sarı,bakımlı|} = 0/5 = 0$$

M2 ve veri kümesi yeşil, sarı, bakımlı için Jaccard benzerlik ölçütü

$$\text{benj}(m_2, X_4) = \frac{|yeşil,sarı,bakımlı| \cap |yeşil,sarı,bakımlı|}{|yeşil,sarı,,bakımlı| \cup |yeşil,sarı,bakımlı|} = 3/3= 1$$

M2>M1 olduğundan yeşil, sarı, bakımlı ikinci kümeye eklendi.

M1 ve veri kümesi kahve, siyah, bakımsız için Jaccard benzerlik ölçütü

$$\text{benj}(m_1, X_5) = \frac{|\text{kahve,kahve,bakımsız}| \cap |\text{kahve,siyah,bakımsız}|}{|\text{kahve,kahve,bakımsız}| \cup |\text{kahve,siyah,bakımsız}|} = 2/3 = 0.67$$

M2 ve veri kümesi kahve, siyah, bakımsız için Jaccard benzerlik ölçütü

$$\text{benj}(m_2, X_5) = \frac{|yeşil,sarı,bakımlı| \cap |\text{kahve,siyah,bakımsız}|}{|yeşil,sarı,,bakımlı| \cup |\text{kahve,siyah,bakımsız}|} = 0/6 = 0$$

M1>M2 kahve, siyah, bakımsız birinci kümeye eklendi.

M1 ve veri kümesi kahve, kahve, bakımsız için Jaccard benzerlik ölçütü

$$\text{benj}(m_1, X_6) = \frac{|\text{kahve,kahve,bakımsız}| \cap |\text{kahve,kahve,bakımsız}|}{|\text{kahve,kahve,bakımsız}| \cup |\text{kahve,kahve,bakımsız}|} = 3/3 = 0.5$$

M2 ve veri kümesi kahve, kahve, bakımsız için Jaccard benzerlik ölçütü

$$\text{benj}(m_2, X_6) = \frac{|yeşil,sarı,bakımlı| \cap |\text{kahve,kahve,bakımsız}|}{|yeşil,sarı,,bakımlı| \cup |\text{kahve,kahve,bakımsız}|} = 0/5 = 0$$

M1>M2 olduğundan kahve, kahve, bakımsız birinci kümeye eklendi.

M1 ve veri kümesi kahve, siyah, bakımsız için Jaccard benzerlik ölçütü

$$\text{benj}(m_1, X_7) = \frac{|\text{kahve,kahve,bakımsız}| \cap |\text{kahve,siyah,bakımsız}|}{|\text{kahve,kahve,bakımsız}| \cup |\text{kahve,siyah,bakımsız}|} = 2/3 = 0.67$$

M2 ve veri kümesi kahve, siyah, bakımsız için Jaccard benzerlik ölçütü

$$\text{benj}(m_2, X_7) = \frac{|yeşil,sarı,bakımlı| \cap |\text{kahve,siyah,bakımsız}|}{|yeşil,sarı,,bakımlı| \cup |\text{kahve,siyah,bakımsız}|} = 0/6 = 0$$

M1>M2 kahve, siyah, bakımsız birinci kümeye eklendi.

M1 ve veri kümesi kahve, sarı, bakımlı için Jaccard benzerlik ölçütü

$$\text{benj}(m_1, X_8) = \frac{|\text{kahve,kahve,bakımsız}| \cap |\text{kahve,sarı,bakımlı}|}{|\text{kahve,kahve,bakımsız}| \cup |\text{kahve,sarı,bakımlı}|} = 1/4 = 0.25$$

M2 ve veri kahve, sarı, bakımlı için Jaccard benzerlik ölçütü

$$\text{benj}(m_2, X_8) = \frac{|yeşil,sarı,bakımlı| \cap |\text{kahve,sarı,bakımlı}|}{|yeşil,sarı,,bakımlı| \cup |\text{kahve,sarı,bakımlı}|} = 2/4 = 0.5$$

M2 > M1 olduğundan kahve, sarı, bakımlı ikinci kümeye eklendi

M1 ve veri kümesi mavi, sarı, bakımlı için Jaccard benzerlik ölçütü

$$\text{benj}(m_1, X_9) = \frac{|\text{kahve,kahve,bakımsız}| \cap |\text{mavi,sarı,bakımlı}|}{|\text{kahve,kahve,bakımsız}| \cup |\text{mavi,sarı,bakımlı}|} = 0/5 = 0$$

M2 ve veri kümesi yeşil, sarı, bakımlı için Jaccard benzerlik ölçütü

$$\text{benj}(m_2, X_9) = \frac{|\textit{yeşil,sarı,bakımlı}| \cap |\text{mavi,sarı,bakımlı}|}{|\textit{yeşil,sarı,,bakımlı}| \cup |\text{mavi,sarı,bakımlı}|} = 2/4 = 0.5$$

M2>M1 olduğundan mavi, sarı, bakımlı ikinci kümeye eklendi.

İterasyon 2 sonunda oluşan kümeler şu şekildedir.

**Tablo 22. Kategorik verilerde K-Means algoritması örnek 2 iterasyon 2 sonucu bulunan ilk kümeler**

| | | | |
|---|---|---|---|
| **Küme 1** | Ela | Kahve | bakımlı |
| | Kahve | Siyah | bakımsız |
| | Kahve | Kahve | bakımsız |
| | Kahve | Siyah | bakımsız |
| **Küme 2** | Mavi | Sarı | bakımlı |
| | Yeşil | Sarı | bakımlı |
| | Yeşil | Sarı | bakımlı |
| | Kahve | Sarı | bakımlı |
| | Mavi | Sarı | bakımlı |

Yeni kümelerin küme merkezleri

M1 = kahve, kahve, bakımsız

M2 = yeşil, sarı, bakımlı

Bir önceki iterasyonun küme merkezleri ile son iterasyonun küme merkezleri birbirlerinin aynı olduğu için iterasyon sonlandırıldı ve veri kümesi iki kümeye ayrılmış oldu.

**Örnek 3:** K-Means algoritması içerisinde Jaccard benzerlik ölçütü ile örnek veri kümesini üç kümeye ayırma işlemi.

İlk adım olarak veri kümesinin ilk elemanı ilk kümenin merkezi seçilir.

M1 = ela, kahve, bakımlı

İkinci adım olarak Jaccard benzerlik ölçütü algoritması kullanılarak ilk merkez kümeye en uzak olan veri ikinci merkez küme seçilir.

M1 ve veri kümesi elemanı mavi, sarı, bakımlı için Jaccard benzerlik ölçütü,

benj($X_1, X_2$) = $\frac{|ela,kahve,bakımlı| \cap |mavi,sarı,bakımlı|}{|ela,kahve,bakımlı| \cup |mavi,sarı,bakımlı|}$ = 1/5 = 0.2

M1 ve veri kümesi elemanı yeşil, sarı, bakımlı için Jaccard benzerlik ölçütü,

benj($X_1, X_3$) = $\frac{|ela,kahve,bakımlı| \cap |yeşil,sarı,bakımlı|}{|ela,kahve,bakımlı| \cup |yeşil,sarı,bakımlı|}$ = 1/5 = 0.2

M1 ve veri kümesi elemanı yeşil, sarı, bakımlı için Jaccard benzerlik ölçütü,

benj($X_1, X_4$) = $\frac{|ela,kahve,bakımlı| \cap |yeşil,sarı,bakımlı|}{|ela,kahve,bakımlı| \cup |yeşil,sarı,bakımlı|}$ = 1/5 = 0.2

M1 ve veri kümesi elemanı kahve, siyah,bakımsız için Jaccard benzerlik ölçütü,

benj($X_1, X_5$) = $\frac{|ela,kahve,bakımlı| \cap |kahve,siyah,bakımsız|}{|ela,kahve,bakımlı| \cup |kahve,siyah,bakımsız|}$ = 1/5 = 0.2

M1 ve veri kümesi elemanı kahve, kahve, bakımsız için Jaccard benzerlik ölçütü,

benj($X_1, X_6$) = $\frac{|ela,kahve,bakımlı| \cap |kahve,kahve,bakımsız|}{|ela,kahve,bakımlı| \cup |kahve,kahve,bakımsız|}$ = 2/4 = 0.5

M1 ve veri kümesi elemanı kahve, siyah, bakımsız için Jaccard benzerlik ölçütü,

benj($X_1, X_7$) = $\frac{|ela,kahve,bakımlı| \cap |kahve,siyah,bakımsız|}{|ela,kahve,bakımlı| \cup |kahve,siyah,bakımsız|}$ = 1/5 = 0.2

M1 ve veri kümesi elemanı kahve, sarı, bakımlı için Jaccard benzerlik ölçütü,

benj($X_1, X_8$) = $\frac{|ela,kahve,bakımlı| \cap |kahve,sarı,bakımlı|}{|ela,kahve,bakımlı| \cup |kahve,sarı,bakımlı|}$ = 2/4 = 0.5

M1 ve veri kümesi elemanı mavi, sarı, bakımlı için Jaccard benzerlik ölçütü,

benj($X_1, X_9$) = $\frac{|ela,kahve,bakımlı| \cap |mavi,sarı,bakımlı|}{|ela,kahve,bakımlı| \cup |mavi,sarı,bakımlı|}$ = 1/5 = 0.2

İlk merkez kümeye en uzak olan kümeler benj($X_1, X_2$) , benj($X_1, X_3$) , benj($X_1, X_4$) , benj($X_1, X_5$) , benj($X_1, X_7$) ve benj($X_1, X_9$). Birden fazla en uzk nokta olduğu için rastgele bir tanesi ikinci merkez küme olarak seçilir. Bu örnekte benj($X_1, X_2$) ikinci merkez küme olarak seçildi.

M1 = ela, kahve, bakımlı

M2 = mavi, sarı, bakımlı

Üçüncü küme merkezini bulmak için ilk ve ikinci küme merkezleri ile veri kümesinin her elemanına Jaccard benzerlik ölçütü uygulanır. Buradan çıkan iki ayrı sonuç toplanıp ikiye bölündükten sonra elde edilen en küçük sonuç iki küme merkezine en uzak sonuçtur.

ela, kahve, bakımlı verisi ile üçüncü kümeyi bulmak için Jaccard benzerlik ölçütü

$$\text{benj}(m_1, X_1) = \frac{|\text{ela,kahve,bakımlı}| \cap |\text{ela,kahve,bakımlı}|}{|\text{ela,kahve,bakımlı}| \cup |\text{ela,kahve,bakımlı}|} = 3/3 = 1$$

$$\text{benj}(m_2, X_1) = \frac{|\text{mavi,sarı,bakımlı}| \cap |\text{ela,kahve,bakımlı}|}{|\text{mavi,sarı,bakımlı}| \cup |\text{ela,kahve,bakımlı}|} = 1/5 = 0.2$$

(1+0.2)/2 = 0.6

mavi, sarı, bakımlı verisi ile üçüncü kümeyi bulmak için Jaccard benzerlik ölçütü

$$\text{benj}(m_1, X_2) = \frac{|\text{ela,kahve,bakımlı}| \cap |\text{mavi,sarı,bakımlı}|}{|\text{ela,kahve,bakımlı}| \cup |\text{mavi,sarı,bakımlı}|} = 1/5 = 0.2$$

$$\text{benj}(m_2, X_2) = \frac{|\text{mavi,sarı,bakımlı}| \cap |\text{mavi,sarı,bakımlı}|}{|\text{mavi,sarı,bakımlı}| \cup |\text{mavi,sarı,bakımlı}|} = 3/3 = 1$$

(0.2+1)/2 = 0.6

yeşil, sarı, bakımlı verisi ile üçüncü kümeyi bulmak için Jaccard benzerlik ölçütü

$$\text{benj}(m_1, X_3) = \frac{|ela,kahve,bakımlı| \cap |yeşil,sarı,bakımlı|}{|ela,kahve,bakımlı| \cup |yeşil,sarı,bakımlı|} = 1/5 = 0.2$$

$$\text{benj}(m_2, X_3) = \frac{|mavi,sarı,bakımlı| \cap |yeşil,sarı,bakımlı|}{|mavi,sarı,,bakımlı| \cup |yeşil,sarı,bakımlı|} = 2/4 = 0.5$$

(0.2+0.5)/2 = 0.35

yeşil, sarı, bakımlı verisi ile üçüncü kümeyi bulmak için Jaccard benzerlik ölçütü

$$\text{benj}(m_1, X_4) = \frac{|ela,kahve,bakımlı| \cap |yeşil,sarı,bakımlı|}{|ela,kahve,bakımlı| \cup |yeşil,sarı,bakımlı|} = 1/5 = 0.2$$

$$\text{benj}(m_2, X_4) = \frac{|mavi,sarı,bakımlı| \cap |yeşil,sarı,bakımlı|}{|mavi,sarı,,bakımlı| \cup |yeşil,sarı,bakımlı|} = 2/4 = 0.5$$

(0.2+0.5)/2 = 0.35

kahve, siyah, bakımsız verisi ile üçüncü kümeyi bulmak için Jaccard benzerlik ölçütü

$$\text{benj}(m_1, X_5) = \frac{|ela,kahve,bakımlı| \cap |\text{kahve,siyah,bakımsız}|}{|ela,kahve,bakımlı| \cup |\text{kahve,siyah,bakımsız}|} = 1/5 = 0.2$$

$$\text{benj}(m_2, X_5) = \frac{|mavi,sarı,bakımlı| \cap |\text{kahve,siyah,bakımsız}|}{|mavi,sarı,,bakımlı| \cup |\text{kahve,siyah,bakımsız}|} = 0/6 = 0$$

(0.2+0)/2 = 0.1

kahve, kahve, bakımsız verisi ile üçüncü kümeyi bulmak için Jaccard benzerlik ölçütü

$$\text{benj}(m_1, X_6) = \frac{|ela,kahve,bakımlı| \cap |\text{kahve,kahve,bakımsız}|}{|ela,kahve,bakımlı| \cup |\text{kahve,kahve,bakımsız}|} = 2/4 = 0.5$$

$$\text{benj}(m_2, X_6) = \frac{|mavi,sarı,bakımlı| \cap |\text{kahve,kahve,bakımsız}|}{|mavi,sarı,,bakımlı| \cup |\text{kahve,kahve,bakımsız}|} = 0/5 = 0$$

(0.5+0)/2 = 0.25

kahve, siyah, bakımsız verisi ile üçüncü kümeyi bulmak için Jaccard benzerlik ölçütü

$$\text{benj}(m_1, X_7) = \frac{|ela,kahve,bakımlı| \cap |\text{kahve,siyah,bakımsız}|}{|ela,kahve,bakımlı| \cup |\text{kahve,siyah,bakımsız}|} = 1/5 = 0.2$$

$$\text{benj}(m_2, X_7) = \frac{|mavi,sarı,bakımlı| \cap |\text{kahve,siyah,bakımsız}|}{|mavi,sarı,,bakımlı| \cup |\text{kahve,siyah,bakımsız}|} = 0/6 = 0$$

(0.2+0)/2 = 0.1

kahve, sarı, bakımlı verisi ile üçüncü kümeyi bulmak için Jaccard benzerlik ölçütü

$$\text{benj}(m_1, X_8) = \frac{|ela,kahve,bakımlı| \cap |\text{kahve,sarı,bakımlı}|}{|ela,kahve,bakımlı| \cup |\text{kahve,sarı,bakımlı}|} = 2/4 = 0.5$$

$$\text{benj}(m_2, X_8) = \frac{|mavi,sarı,bakımlı| \cap |\text{kahve,sarı,bakımlı}|}{|mavi,sarı,,bakımlı| \cup |\text{kahve,sarı,bakımlı}|} = 2/4 = 0.5$$

(0.5+0.5)/2 = 0.5

mavi, sarı, bakımlı verisi ile üçüncü kümeyi bulmak için Jaccard benzerlik ölçütü

$$\text{benj}(m_1, X_9) = \frac{|ela,kahve,bakımlı| \cap |\text{mavi,sarı,bakımlı}|}{|ela,kahve,bakımlı| \cup |\text{mavi,sarı,bakımlı}|} = 1/5 = 0.2$$

$$\text{benj}(m_2, X_9) = \frac{|mavi,sarı,bakımlı| \cap |\text{mavi,sarı,bakımlı}|}{|mavi,sarı,,bakımlı| \cup |\text{mavi,sarı,bakımlı}|} = 3/3 = 1$$

(0.2+1)/2 = 0.6

Buradan çıkan en uzak mesafe 0.1 olduğu için üçüncü küme merkezi kahve, siyah, bakımsız oldu. Eğer veri kümesi dört kümeye ayrılacak olsaydı bu sefer veri kümesinin her elemanı tüm merkez kümelerinin elemanlarına ayrı ayrı Jaccard benzerlik ölçütü uygulanıp, çıkan sonucu ikiye bölünüp en küçük olan değer diğer küme merkezi olacaktı fakat doğru sonuç alabilmek için veri kümesinin optimum kaç kümeye bölünebileceği önceden tespit edilmelidir. En son durumda üç küme merkezi şu şekilde oluştu.

M1 = ela, kahve, bakımlı

M2 = mavi, sarı, bakımlı

M3 = kahve, siyah, bakımsız

Bu aşamadan sonra sonsuz iterasyon başlatılır.

**İterasyon 1:**

M1 ve veri kümesi elemanı yeşil, sarı, bakımlı için Jaccard benzerlik ölçütü,

$$\text{benj}(m_1, X_3) = \frac{|ela,kahve,bakımlı| \cap |yeşil,sarı,bakımlı|}{|ela,kahve,bakımlı| \cup |yeşil,sarı,bakımlı|} = 1/5 = 0.2$$

M2 ve veri kümesi elemanı yeşil, sarı, bakımlı için Jaccard benzerlik ölçütü,

$$\text{benj}(m_2, X_3) = \frac{|mavi,sarı,bakımlı| \cap |yeşil,sarı,bakımlı|}{|mavi,sarı,,bakımlı| \cup |yeşil,sarı,bakımlı|} = 2/4 = 0.5$$

M3 ve veri kümesi elemanı yeşil, sarı, bakımlı için Jaccard benzerlik ölçütü,

$$\text{benj}(m_3, X_3) = \frac{|\text{kahve,siyah,bakımsız}| \cap |yeşil,sarı,bakımlı|}{|\text{kahve,siyah,bakımsız} \cup |yeşil,sarı,bakımlı|} = 0/6 = 0$$

M2>M1>M3 olduğundan yeşil, sarı, bakımlı ikinci kümeye eklendi.

M1 ve veri kümesi elemanı yeşil, sarı, bakımlı için Jaccard benzerlik ölçütü,

$$\text{benj}(m_1, X_4) = \frac{|ela,kahve,bakımlı| \cap |yeşil,sarı,bakımlı|}{|ela,kahve,bakımlı| \cup |yeşil,sarı,bakımlı|} = 1/5 = 0.2$$

M2 ve veri kümesi elemanı yeşil, sarı, bakımlı için Jaccard benzerlik ölçütü,

$$\text{benj}(m_2, X_4) = \frac{|mavi,sarı,bakımlı| \cap |yeşil,sarı,bakımlı|}{|mavi,sarı,,bakımlı| \cup |yeşil,sarı,bakımlı|} = 2/4 = 0.5$$

M3 ve veri kümesi elemanı yeşil, sarı, bakımlı için Jaccard benzerlik ölçütü,

$$\text{benj}(m_3, X_4) = \frac{|\text{kahve,siyah,bakımsız}| \cap |yeşil,sarı,bakımlı|}{|\text{kahve,siyah,bakımsız} \cup |yeşil,sarı,bakımlı|} = 0/6 = 0$$

M2 > M1 > M3olduğundan yeşil, sarı, bakımlı ikinci kümeye eklendi.

M1 ve veri kümesi elemanı kahve, siyah, bakımsız için Jaccard benzerlik ölçütü,

$$\text{benj}(m_1, X_5) = \frac{|ela,kahve,bakımlı| \cap |\text{kahve,siyah,bakımsız}|}{|ela,kahve,bakımlı| \cup |\text{kahve,siyah,bakımsız}|} = 1/5 = 0.2$$

M2 ve veri kümesi elemanı kahve, siyah, bakımsız için Jaccard benzerlik ölçütü,

$$\text{benj}(m_2, X_5) = \frac{|mavi,sarı,bakımlı| \cap |\text{kahve,siyah,bakımsız}|}{|mavi,sarı,,bakımlı| \cup |\text{kahve,siyah,bakımsız}|} = 0/6 = 0$$

M3 ve veri kümesi elemanı kahve, siyah, bakımsız için Jaccard benzerlik ölçütü,

$$\text{benj}(m_3, X_5) = \frac{|\text{kahve,siyah,bakımsız}| \cap |\text{kahve,siyah,bakımsız}|}{|\text{kahve,siyah,bakımsız} \cup |\text{kahve,siyah,bakımsız}|} = 3/3 = 1$$

M3>M1>M2 kahve, siyah, bakımsız üçüncü kümeye eklendi.

M1 ve veri kümesi elemanı kahve, kahve, bakımsız için Jaccard benzerlik ölçütü,

$$\text{benj}(m_1, X_6) = \frac{|ela,kahve,bakımlı| \cap |\text{kahve,kahve,bakımsız}|}{|ela,kahve,bakımlı| \cup |\text{kahve,kahve,bakımsız}|} = 2/4 = 0.5$$

M2 ve veri kümesi elemanı kahve, kahve, bakımsız için Jaccard benzerlik ölçütü,

$$\text{benj}(m_2, X_6) = \frac{|mavi,sarı,bakımlı| \cap |\text{kahve,kahve,bakımsız}|}{|mavi,sarı,,bakımlı| \cup |\text{kahve,kahve,bakımsız}|} = 0/5 = 0$$

M3 ve veri kümesi elemanı kahve, kahve, bakımsız için Jaccard benzerlik ölçütü,

$$\text{benj}(m_3, X_6) = \frac{|\text{kahve,siyah,bakımsız}| \cap |\text{kahve,kahve,bakımsız}|}{|\text{kahve,siyah,bakımsız} \cup |\text{kahve,kahve,bakımsız}|} = 2/3 = 0.67$$

M3 > M1 > M2 olduğundan kahve, kahve, bakımsız üçüncü kümeye eklendi.

M1 ve veri kümesi elemanı kahve, siyah, bakımsız için Jaccard benzerlik ölçütü,

$$\text{benj}(m_1, X_7) = \frac{|ela,kahve,bakımlı| \cap |\text{kahve,siyah,bakımsız}|}{|ela,kahve,bakımlı| \cup |\text{kahve,siyah,bakımsız}|} = 1/5 = 0.2$$

M2 ve veri kümesi elemanı kahve, siyah, bakımsız için Jaccard benzerlik ölçütü,

$$\text{benj}(m_2, X_7) = \frac{|mavi,sarı,bakımlı| \cap |\text{kahve,siyah,bakımsız}|}{|mavi,sarı,,bakımlı| \cup |\text{kahve,siyah,bakımsız}|} = 0/6 = 0$$

M3 ve veri kümesi elemanı kahve, siyah, bakımsız için Jaccard benzerlik ölçütü,

$$\text{benj}(m_3, X_7) = \frac{|\text{kahve,siyah,bakımsız}| \cap |\text{kahve,siyah,bakımsız}|}{|\text{kahve,siyah,bakımsız} \cup |\text{kahve,siyah,bakımsız}|} = 3/3 = 1$$

M3 > M2 = M1 kahve, siyah, bakımsız üçüncü kümeye eklendi.

M1 ve veri kümesi elemanı kahve, sarı, bakımlı için Jaccard benzerlik ölçütü,

$$\text{benj}(m_1, X_8) = \frac{|ela,kahve,bakımlı| \cap |\text{kahve,sarı,bakımlı}|}{|ela,kahve,bakımlı| \cup |\text{kahve,sarı,bakımlı}|} = 2/4 = 0.5$$

M2 ve veri elemanı kahve, sarı, bakımlı için Jaccard benzerlik ölçütü,

$$\text{benj}(m_2, X_8) = \frac{|mavi,sarı,bakımlı| \cap |\text{kahve,sarı,bakımlı}|}{|mavi,sarı,,bakımlı| \cup |\text{kahve,sarı,bakımlı}|} = 2/4 = 0.5$$

M3 ve veri kümesi elemanı kahve, sarı, bakımlı için Jaccard benzerlik ölçütü,

$$\text{benj}(m_3, X_8) = \frac{|\text{kahve,siyah,bakımsız}| \cap |kahve,sarı,bakımlı|}{|\text{kahve,siyah,bakımsız} \cup |kahve,sarı,bakımlı|} = 1/5 = 0.2$$

M2 = M1 > M3 olduğundan kahve, sarı, bakımlı iki m2 veya m1 kümelerinden bir tanesine eklenebilir. Bu örnekte birinci kümeye eklendi.

M1 ve veri kümesi elemanı mavi, sarı, bakımlı için Jaccard benzerlik ölçütü,

$$\text{benj}(m_1, X_9) = \frac{|ela,kahve,bakımlı| \cap |\text{mavi,sarı,bakımlı}|}{|ela,kahve,bakımlı| \cup |\text{mavi,sarı,bakımlı}|} = 1/5 = 0.2$$

M2 ve veri kümesi elemanı mavi, sarı, bakımlı için Jaccard benzerlik ölçütü,

$$\text{benj}(m_2, X_9) = \frac{|mavi,sarı,bakımlı| \cap |\text{mavi,sarı,bakımlı}|}{|mavi,sarı,,bakımlı| \cup |\text{mavi,sarı,bakımlı}|} = 3/3 = 1$$

M3 ve veri kümesi elemanı mavi, sarı, bakımlı için Jaccard benzerlik ölçütü,

$$\text{benj}(m_3, X_9) = \frac{|\text{kahve,siyah,bakımsız}| \cap |mavi,sarı,bakımlı|}{|\text{kahve,siyah,bakımsız}| \cup |mavi,sarı,bakımlı|} = 0/6 = 0$$

M2>M1>M3 olduğundan mavi, sarı, bakımlı ikinci kümeye eklendi.

İterasyon 1 sonunda oluşan kümeler tablo 23'deki gibidir.

**Tablo 23. Kategorik verilerde K-Means algoritması örnek 3 iterasyon 1 sonucu bulunan ilk kümeler**

| | | | |
|---|---|---|---|
| **Küme 1** | Ela | Kahve | Bakımlı |
| | Kahve | Sarı | Bakımlı |
| **Küme 2** | Mavi | Sarı | Bakımlı |
| | Yeşil | Sarı | Bakımlı |
| | yeşil | Sarı | Bakımlı |
| | Mavi | Sarı | Bakımlı |
| **Küme 3** | Kahve | Siyah | Bakımsız |
| | Kahve | Kahve | Bakımsız |
| | Kahve | Siyah | Bakımsız |

Yeni küme merkezlerini bulmak için her kümenin sütununda en çok geçen eleman alınır. Eğer eleman sayıları eşitse rastgele olarak alınır. Yeni kümelerden yola çıkılarak yeni küme merkezleri şu şekilde olur.

M1 = ela, kahve, bakımlı

M2 = mavi, sarı, bakımlı

M3 = kahve, siyah, bakımsız

Bir önceki iterasyondaki küme merkezleri ile en son iterasyondaki küme merkezleri kıyaslandığında birbirlerinin aynısı olduğu için sonsuz iterasyon kırılarak algoritma sonlandırıldı. Buna göre örnek veri kümesi üç kümeye ayrılmak istendiğinde tablo 24'deki sonuçlar alınmaktadır.

**Tablo 24. Kategorik verilerde K-Means algoritması örnek 3 verilerin üç kümeye ayrıldığındaki sonuçlar**

| | | | |
|---|---|---|---|
| **Küme 1** | Ela | Kahve | Bakımlı |
| | Kahve | Sarı | Bakımlı |
| **Küme 2** | Mavi | Sarı | Bakımlı |
| | Yeşil | Sarı | Bakımlı |
| | yeşil | Sarı | Bakımlı |
| | Mavi | Sarı | Bakımlı |
| **Küme 3** | Kahve | Siyah | Bakımsız |
| | Kahve | Kahve | Bakımsız |
| | Kahve | Siyah | Bakımsız |

### 3.2 K-Means Algoritması İle Karışık Veri Kümelerinde Çalışmak

Makine öğreniminde sayısal verilerin kümelenmesi kendini kanıtlamış K-means algoritması ve sözel verilen kümelenebilmesi bölüm 3.1'de açıklanan yöntemle rahatça yapılabildiği açıklanmıştı fakat bu açıklanan yöntemlerle sayısal veriler sadece sayısal verilerle sözel veriler de sadece sözel verilerle kümelenebiliyor. Hâlbuki gerçek hayatta bu veriler birbirleri ile iç içe karışık bir şekildedir. Makine öğreniminde daha net yargılara varabilmek için sayısal ve sözel verilerin birlikte kümelenebilmesi gerekmektedir. Bu tezin modelleme kısmında satısal ve sözel veriler karışık olarak alınıp doğru bir şekilde kümelenmeye çalışılacaktır. Bu modelleme işlemi temel olarak sözel veriler için bölüm 3.1'de anlatılan yöntem ve sayısal verilen için ise K-Means uzaklık algoritmasından birlikte faydalanılacaktır.

Sayısal ve sözel verileri birlikte kümeleyebilmek için aşağıdaki adımlar yapılmalıdır.

- Veri kümesinin kaç kümeye ayrılacağı,
- Karışık olan veri kümesi alındıktan sonra verilerin sayısal ve sayısal olmayan veri olup olmadığının anlaşılması,
- Sayısal veriler sayısal, sayısal olmayan veriler ise sayısal olmayan verilerle kümelenmesi. Burada önemli olan husus veriler sayısal ve sözel olarak iki kümeye ayrılıken verilerin karıştırılmamalı ve kaydırmalar yapılmamalıdır.

**Tablo 25. Örnek veri tabanı**

| | | | | |
|---|---|---|---|---|
| Elma | Armut | 3 | 2 | 1 |
| Elma | Muz | 3 | 5 | 5 |

Örneğin tablo 25'deki verileri kategorik ve kategorik olmayan verilere göre ayrılması istenirse kategorik olan veri tablosu tablo 26 ve kategorik olmayan veri tablosu tablo 27 gibi olmalıdır.

**Tablo 26. Kategorik verilere göre bölünmüş tablo**

| | |
|---|---|
| Elma | Armut |
| Elma | Muz |

**Tablo 27. Kategorik Olmayan Verilere Göre Bölünmüş Tablo**

| | | |
|---|---|---|
| 3 | 2 | 1 |
| 3 | 5 | 5 |

- İlk merkezlerin belirlenmesi. Daha öncede belirtildiği gibi kategorik verileri kümelemek için bölüm 3.1'de anlatılan yöntem ve kategorik olmayan verileri kümelemek için ise K-Means uzaklık algoritması kullanılıyordu. İlk merkezlerin belirlenmesi bölüm 3.1'de anlatılan algoritma veya K-Means algoritmasına göre yapılmalıdır. K-Means algoritmasında ilk merkezlerin belirlenmesi rastgele olabiliyordu fakat bölüm 3.1'de anlatılan algoritmada ilk küme merkezlerinin seçimini bir önemi vardı ve bu merkezlerin hangi veriler olacağının özel bir yöntemle seçilmesi gerektiği aynı bölümde anlatılmıştı. O halde ilk merkezler bölüm 3.1'e göre belirlenmelidir.

**Tablo 28. Karışık Veri Kümesi Örneği**

| | | | | |
|---|---|---|---|---|
| Elma | Armut | 3 | 2 | 1 |
| Elma | Muz | 3 | 5 | 5 |
| Kivi | Üzüm | 1 | 3 | 2 |

Yukarıdaki veri kümesine göre ilk merkezlerin belirlenmesi şu şekildedir.

İlk adım alarak veri kümesinin ilk elemanı ilk kümenin merkezi seçilir.

M1 = elma, armut

İkinci adım olarak bölüm 3.1'deki yöntem kullanılır ve ilk merkez kümeye en uzak olan veri ikinci merkez küme seçilir.

M1 ve veri kümesi elemanı elma, muz için Jaccard benzerlik ölçütü,

$\text{benj}(X_1, X_2) = \frac{|elma,armut| \cap |elma,muz|}{|elma,armut| \cup |elma,muz|} = 1/3 = 0.33$

M1 ve veri kümesi elemanı kivi, üzüm için Jaccard benzerlik ölçütü,

$\text{benj}(X_1, X_3) = \frac{|elma,armut| \cap |kivi,\text{üzüm}|}{|elma,armut| \cup |elma,\text{üzüm}|} = 0/4 = 0$

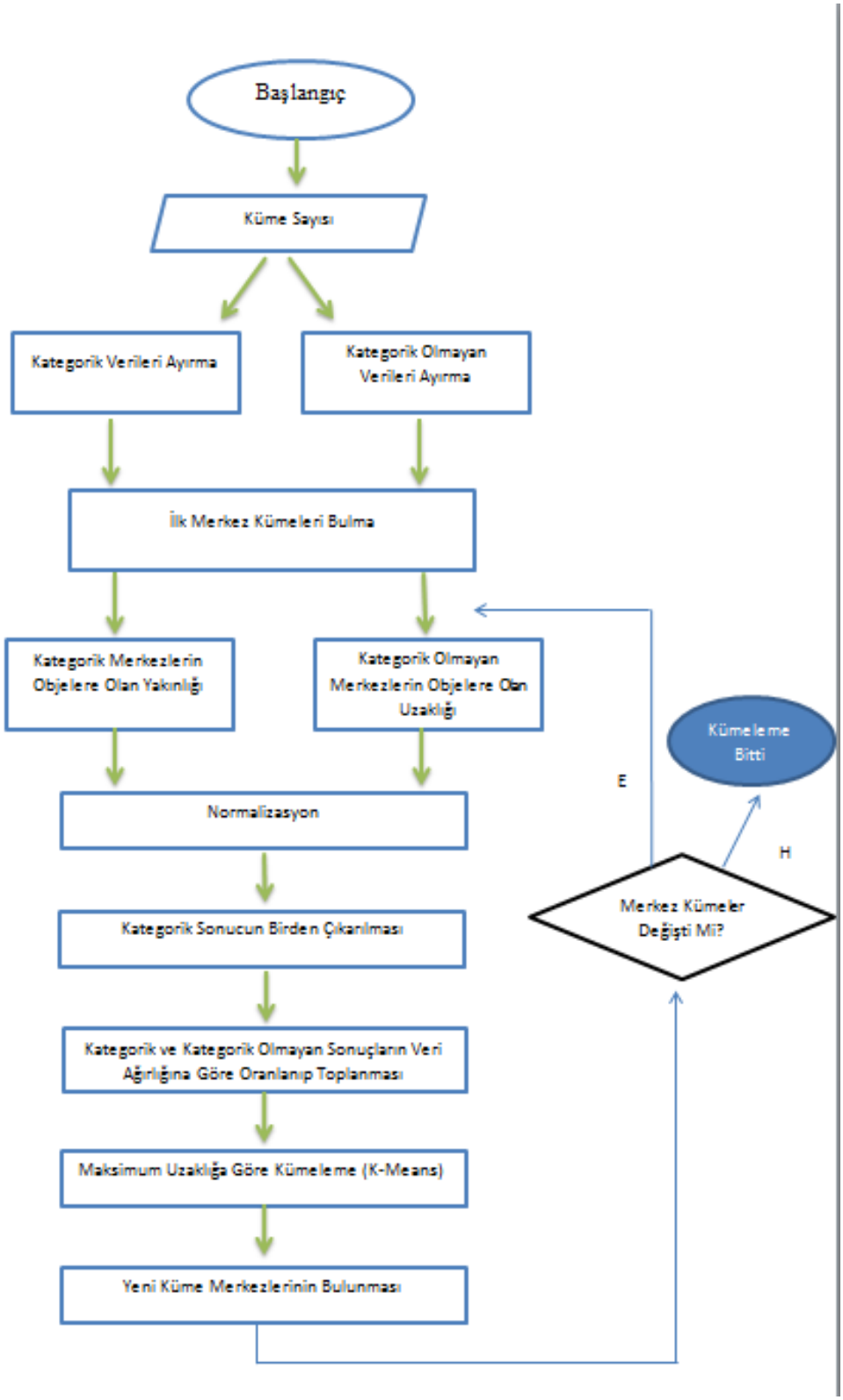
m1'e en uzak olan kivi, üzüm verisi olduğundan ikinci merkez küme kivi, üzüm kümesidir. Bu sonuca göre karışık kümelerin ilk merkezleri aşağıdaki gibi oluşur.

M1 = elma, armut, 3, 2, 1

M2 = kivi, üzüm, 1, 3, 2

- İlk merkez kümeler bulunduktan sonra sözel veriler bölüm 3.1'deki yönteme ve kategorik olmayan veriler ise K-Means algoritmasına sokularak sonsuz bir iterasyon başlatılır.
- K-Means algoritmasından çıkan sonuçlar uzaklık, bölüm 3.'deki algoritmadan gelen değerler yakınlık olduğu için veriler arasında bir uyumsuzluk söz konusu olur. Algoritmalardan çıkan sonuçları tek tip sonuca çevirmek için bölüm 3.1'deki algoritmadan çıkan sonuçlar 1 sayısından çıkartılır. Bu şekilde bölüm 3.1'deki algoritmadan dönen değerler de artık uzaklık olmuş olur.
- Eğer bölüm 3.1'deki algoritmadan ve K-Means algoritmasından gelen değerlerin arasında çok büyük farklar varsa normalizasyon yapılır.
- Buraya kadar yapılan işlemlerin sonucunda iki adet uzaklık sonucu elde edilir. Bu verilerden biri bölüm 3.1'deki algoritmadan bir tanesi ise K-Means algoritmasından gelir. Bu aşamadan çıkan sonuçlar kategorik ve kategorik olmayan verilerin ağırlıkları oranı ile çarpılır.
- Çarpım sonucunda çıkan iki değer toplanır ve o satırdaki verinin merkeze olan uzaklığı hesaplanmış olur.
- Yukarıda belirtilen işlemler her satırdaki veri elemanı ile tüm merkez kümeler arasında gerçekleştirilir. Veri kümesinden bir eleman ile merkez kümeler arasında yapılan tüm işlemler sonucunda veri kümesi elemanı aynı K-Means algoritmasındaki gibi küçük sonucun çıktığı kümeye eklenir.
- Tüm merkezler ile tüm veri kümesi elemanları eşleştirildikten sonra veri kümesi kaç kümeye bölünecekse o kadar küme elde edilir. Kümeler elde edildikten sonra yeni küme merkezlerinin bulunması gerekmektedir. Bu işlemi yapmak için elde edilen her kümenin içerisinde sözel verilerin küme merkezlerini bulmak için bölüm 3.1'deki algoritma ve sayısal verilerin küme merkezlerini belirlemek içim K-Means algoritması ile aynı işlemler tekrardan uygulanır.
- Yeni merkezler bulunduktan sonra en son merkezler ile bir önceki merkezler kıyaslanır eğer merkezler aynıysa sonsuz iterasyon kırılır ve veriler

kümelenmiş olur. Eğer merkezler de farklılık varsa iterasyonun başladığı yere gidilir ve yeni merkezlerle aynı işlemler tekrarlanır.




**Şekil 8. Karışık veri kümelerinin kümelenmesinin akış şeması**

**Tablo 29. Örnek Veri Kümesi**

| | | | | |
|---|---|---|---|---|
| Ela | kahve | Bakımlı | 3 | 2 |
| Mavi | sarı | Bakımlı | 1 | 5 |
| Yeşil | sarı | Bakımlı | 2 | 2 |
| Yeşil | sarı | Bakımlı | 7 | 4 |
| Kahve | siyah | Bakımsız | 1 | 3 |

**Örnek 1:** Tablo 29'daki karışık veri kümesini iki kümeye ayırma işlemi.

İlk adım olarak kategorik olan ve olmayan verileri ayırma işleminin yapılması gerekiyor. Verilen tabloda özellikle sözel olan veriler bir tarafa sayısal veriler ise diğer tarafa toplanmıştır. Bunun sonucunda kategorik olan ve olmayan veriler tablo 30 ve tablo 31'deki gibidir.

**Tablo 30. Kategorik olarak ayrılan veriler**

| | | |
|---|---|---|
| ela | Kahve | Bakımlı |
| mavi | Sarı | Bakımlı |
| yeşil | Sarı | Bakımlı |
| yeşil | sarı | Bakımlı |
| kahve | siyah | bakımsız |

**Tablo 31.Kategorik olmayan olarak ayrılan veriler**

| | |
|---|---|
| 3 | 2 |
| 1 | 5 |
| 2 | 2 |
| 7 | 4 |
| 1 | 3 |

Veri kümesinde kategorik olan ve olmayan veriler belirlendikten sonra ilk merkezleri bulma işlemi başlatılır. İlk merkez bulma işlemi kategorik verilere göre yapılır çünkü kategorik verilerde doğru kümeleme yapabilmek için ilk merkezler büyük önem taşır. İlk merkezleri bulmada birinci adım olarak veri kümesinin ilk elemanı ilk kümenin merkezi seçilir.

M1 = ela, kahve, bakımlı, 3,2

İkinci adım olarak Jaccard benzerlik ölçütü kullanılarak ilk merkez kümeye en uzak olan veri ikinci merkez küme seçilir.

M1 ve veri kümesi elemanı mavi, sarı, bakımlı için Jaccard benzerlik ölçütü ,

benj($X_1, X_2$) = $\frac{|ela,kahve,bakımlı| \cap |mavi,sarı,bakımlı|}{|ela,kahve,bakımlı| \cup |mavi,sarı,bakımlı|}$ = 1/5 = 0.2

M1 ve veri kümesi elemanı yeşil, sarı, bakımlı için Jaccard benzerlik ölçütü ,

benj($X_1, X_3$) = $\frac{|ela,kahve,bakımlı| \cap |yeşil,sarı,bakımlı|}{|ela,kahve,bakımlı| \cup |yeşil,sarı,bakımlı|}$ = 1/5 = 0.2

M1 ve veri kümesi elemanı yeşil, sarı, bakımlı için Jaccard benzerlik ölçütü ,

benj($X_1, X_4$) = $\frac{|ela,kahve,bakımlı| \cap |yeşil,sarı,bakımlı|}{|ela,kahve,bakımlı| \cup |yeşil,sarı,bakımlı|}$ = 1/5 = 0.2

M1 ve veri kümesi elemanı kahve, siyah,bakımsız için Jaccard benzerlik ölçütü ,

benj($X_1, X_5$) = $\frac{|ela,kahve,bakımlı| \cap |kahve,siyah,bakımsız|}{|ela,kahve,bakımlı| \cup |kahve,siyah,bakımsız|}$ = 1/5 = 0.2

İlk merkez kümeye en uzak olan kümeler benj($X_1, X_2$) , benj($X_1, X_3$) , benj($X_1, X_4$) ve benj($X_1, X_5$). Birden fazla en uzak nokta olduğu için rastgele bir tanesi ikinci merkez küme seçilir. Bu örnekte benj($X_1, X_2$) ikinci merkez küme seçildi.

M1 = ela, kahve, bakımlı, 3, 2

M2 = mavi, sarı, bakımlı, 1, 5

Merkez küme bulma işlemleri tamamlandıktan sonra sonsuz iterasyon başlatılır. Bu iterasyonların içerisinde bulunan tüm merkezler ile veri kümesinde buluna tüm elemanlar arasında sözel olan veriler bölüm 3.1'deki algoritma ve sayısal veriler ise K-Means algoritmasına sokulma işlemi yapılır. Daha sonra elde edilen değerlerden bir tanesi uzaklık bir tanesi yakınlık olduğu için yakınlık olan değer bir sayısından çıkartılarak tüm sonuçlar uzaklık cinsine çevrilir. Bu çevirme işlemini takiben çıkan sonuçların normalizasyonu, sayısal ve sözel verilerin veri setindeki ağırlığı oranında çarpılması ve ağırlığa göre sonuç bulunduktan sonra bu değerlerin toplanması işlemleri yapılır. Toplanan sayılarda değerce küçük olan o merkeze yakın olarak kabul edilip o yakın olduğu merkezin kümesine eklenir.

**İterasyon 1:**

M1 ve kategorik veri kümesi elemanı yeşil, sarı, bakımlı için Jaccard benzerlik ölçütü

benj($m_1, X_3$) = $\frac{|ela,kahve,bakımlı| \cap |yeşil,sarı,bakımlı|}{|ela,kahve,bakımlı| \cup |yeşil,sarı,bakımlı|}$ = 1/5 = 0.2

M1 ve sayısal veri kümesi elemanı 2,2 için öklit:

S1 = $\sqrt{(3-2)^2+(2-2)^2}$ = 1,

S2 = 1-benj($m_1, X_3$) = 1-0.2 = 0.8

Normalizasyon işlemi sonuçu:

S1 = 10,

S2 = 8

Ağırlığa göre çarpım ve çıkan sonuçların toplamı:

T1 = (S1*0.4)+(S2*0.6) = (10*0.4) +(8*0.6) = 4+4.8 = 8.8

M2 ve veri kümesi elemanı yeşil, sarı, bakımlı için Jaccard benzerlik ölçütü

benj($m_2, X_3$) = $\frac{|mavi,sarı,bakımlı| \cap |yeşil,sarı,bakımlı|}{|mavi,sarı,,bakımlı| \cup |yeşil,sarı,bakımlı|}$ = 2/4 = 0.5

M2 ve sayısal veri kümesi elemanı 2,2 için öklit:

S1 = $\sqrt{(1-2)^2+(5-2)^2}$ = 3.16,

S2 = 1-benj($m_1, X_3$) = 1-0.5 = 0.5

Normalizasyon işlemi sonuçu:

S1 = 3.16,

S2 = 5

Ağırlığa göre çarpım ve çıkan sonuçların toplamı:

T2 = (S1*0.4)+(S2*0.6) = (3.16*0.4) +(5*0.6) = 1.264+3 = 4.264

T2<T1 olduğundan yeşil, sarı, bakımlı, 2, 2 ikinci kümeye eklendi.

M1 ve sözel veri kümesi elemanı yeşil, sarı, bakımlı için Jaccard benzerlik ölçütü

benj($m_1, X_4$) = $\frac{|ela,kahve,bakımlı| \cap |yeşil,sarı,bakımlı|}{|ela,kahve,bakımlı| \cup |yeşil,sarı,bakımlı|}$ = 1/5 = 0.2

M1 ve sayısal veri kümesi elemanı 7,4 için öklit:

S1 = $\sqrt{(3-7)^2+(2-4)^2}$ = 4.47,

S2 = 1-benj($m_1, X_3$) = 1-0.2 = 0.8

Normalizasyon işlemi sonuçu:

S1 = 44.7,

S2 = 80

Ağırlığa göre çarpım ve çıkan sonuçların toplamı:

T1 = (S1*0.4)+(S2*0.6) = (44.7*0.4) +(80*0.6) = 65.88

M2 ve sözel veri kümesi elemanı yeşil, sarı, bakımlı için Jaccard benzerlik ölçütü

benj($m_2, X_4$) = $\frac{|mavi,sarı,bakımlı| \cap |yeşil,sarı,bakımlı|}{|mavi,sarı,,bakımlı| \cup |yeşil,sarı,bakımlı|}$ = 2/4 = 0.5

M2 ve sayısal veri kümesi elemanı 7,4 için öklit:

S1 = $\sqrt{(1-7)^2+(5-4)^2}$ = 6.08,

S2 = 1-benj($m_1, X_3$) = 1-0.5 = 0.5

Normalizasyon işlemi sonuçu:

S1 = 60.8,

S2 = 50

Ağırlığa göre çarpım ve çıkan sonuçların toplamı:

T2 = (S1*0.4)+(S2*0.6) = (60.8*0.4) +(50*0.6) = 54.32

T2<T1 olduğundan yeşil, sarı, bakımlı ikinci kümeye eklendi.

M1 ve sözel veri kümesi elemanı kahve, siyah,bakımsız için Jaccard benzerlik ölçütü

benj($m_1, X_5$) = $\frac{|ela,kahve,bakımlı \cap |\text{kahve,siyah,bakımsız}|}{|ela,kahve,bakımlı \cup |\text{kahve,siyah,bakımsız}|}$ = 1/5 = 0.2

M1 ve sayısal veri kümesi elemanı 1,3 için öklit:

S1 = $\sqrt{(3-1)^2 + (2-3)^2}$ = 2.24,

S2 = 1-benj($m_1, X_3$) = 1-0.2 = 0.8

Normalizasyon işlemi sonuçu:

S1 = 2.24,

S2 = 8

Ağırlığa göre çarpım ve çıkan sonuçların toplamı:

T1 = (S1*0.4)+(S2*0.6) = (2.24*0.4) +(8*0.6) =0. 9+4.8 = 5.7

M2 ve sözel veri kümesi elemanı kahve, siyah,bakımsız için Jaccard benzerlik ölçütü

benj($m_2, X_5$) = $\frac{|mavi,sarı,bakımlı \cap |\text{kahve,siyah,bakımsız}|}{|mavi,sarı,,bakımlı \cup |\text{kahve,siyah,bakımsız}|}$ = 0/6 = 0

M2 ve sayısal veri kümesi elemanı 1,3 için öklit:

S1 = $\sqrt{(1-1)^2 + (5-3)^2}$ = 2,

S2 = 1-benj($m_1, X_3$) = 1-0 = 1

Normalizasyon işlemi sonuçu:

S1 = 2,

S2 = 10

Ağırlığa göre çarpım ve çıkan sonuçların toplamı:

T2 = (S1*0.4)+(S2*0.6) = (2*0.4) +(10*0.6) = 0.8+6 = 6.8

T1<T2 kahve, siyah, bakımsız birinci kümeye eklendi.

İterasyon 1 sonunda oluşan kümeler şu şekildedir.

**Tablo 32. Karışık verileri kümelemede iterasyon 1 sonucu oluşan yeni kümeler**

| | | | | | |
|---|---|---|---|---|---|
| **Küme 1** | Ela | kahve | bakımlı | 3 | 2 |
| | Kahve | siyah | bakımsız | 1 | 3 |
| **Küme 2** | Mavi | Sarı | bakımlı | 1 | 5 |
| | Yeşil | Sarı | bakımlı | 7 | 4 |
| | Yeşil | Sarı | bakımlı | 1 | 3 |

Bölüm 3.1'deki algoritma için yeni merkezler

M1 = ela, kahve, bakımlı

M2 = yeşil, sarı, bakımlı

K-Means için merkezler

M1 = (3+1+1)/3, (2+3+5)/3 = 1.67, 3.33

M2 = (7+1)/2, (4+3)/2 = 4, 3.5

En son küme merkezleri ile bir önceki küme merkezleri aynı olmadığı için yeni bir iterasyon başlatıldı.

**İterasyon 2:**

M1 ve sözel veri kümesi elemanı ela, kahve, bakımlı için Jaccard benzerlik ölçütü

benj$(m_1, X_1)$ = $\frac{|ela,kahve,bakımlı| \cap |ela,kahve,bakımlı|}{|ela,kahve,bakımlı| \cup |ela,kahve,bakımlı|}$ = 3/3 = 1

M1 ve sayısal veri kümesi elemanı 3,2 için öklit:

S1 = $\sqrt{(1.67-3)^2+(3.33-2)^2}$ = 1.88 ,

S2 = 1-benj$(m_1, X_3)$ = 1-1 = 0

Normalizasyon işlemi sonuçu:

S1 = 1.88,

S2 = 0

Ağırlığa göre çarpım ve çıkan sonuçların toplamı:

T1 = (S1*0.4)+(S2*0.6) = (1.88*0.4) +(0*0.6) = 0.75+0 = 0.75

M2 ve sözel veri kümesi elemanı ela, kahve, bakımlı için Jaccard benzerlik ölçütü

benj($m_2, X_1$) = $\frac{|yeşil,sarı,bakımlı| \cap |ela,kahve,bakımlı|}{|yeşil,sarı,bakımlı| \cup |ela,kahve,bakımlı|}$ = 1/5 = 0.2

M1 ve sayısal veri kümesi elemanı 3,2 için öklit:

S1 = $\sqrt{(4-2)^2 + (3.5-2)^2}$ = 2.5,

S2 = 1-benj($m_1, X_3$) = 1-0.2 = 0.8

Normalizasyon işlemi sonuçu:

S1 = 2.5,

S2 = 8

Ağırlığa göre çarpım ve çıkan sonuçların toplamı:

T2 = (S1*0.4)+(S2*0.6) = (2.5*0.4) +(8*0.6) = 1+4.8 = 5.8

T1<T2 olduğundan ela, kahve, bakımlı, 3, 2 birinci kümeye eklendi.

M1 ve sözel veri kümesi elemanı mavi, sarı, bakımlı için Jaccard benzerlik ölçütü

benj($m_1, X_2$) = $\frac{|ela,kahve,bakımlı| \cap |mavi,sarı,bakımlı|}{|ela,kahve,bakımlı| \cup |mavi,sarı,bakımlı|}$ = 1/5 = 0.2

M1 ve sayısal veri kümesi elemanı 1,5 için öklit:

S1 = $\sqrt{(1.67-1)^2 + (3.33-5)^2}$ = 1.8,

S2 = 1-benj($m_1, X_3$) = 1-0.2 = 0.8

Normalizasyon işlemi sonuçu:

S1 = 1.8,

S2 = 0.8 *2

Ağırlığa göre çarpım ve çıkan sonuçların toplamı:

T1 = (S1*0.4)+(S2*0.6) = (1.8*0.4) +(1.6*0.6) = 0.72+0.96 = 1.68

M2 ve sözel veri kümesi elemanı mavi, sarı, bakımlı için Jaccard benzerlik ölçütü

benj($m_2, X_2$) = $\frac{|yeşil,sarı,bakımlı| \cap |mavi,sarı,bakımlı|}{|yeşil,sarı,,bakımlı| \cup |mavi,sarı,bakımlı|}$ = 2/4 = 0.5

M2 ve sayısal veri kümesi elemanı 1,5 için öklit:

S1 = $\sqrt{(4-2)^2 + (3.5-5)^2}$ = 2.5,

S2 = 1-benj($m_1, X_3$) = 1-0.5 = 0.5

Normalizasyon işlemi sonuçu:

S1 = 2.5,

S2 = 0.5*2

Ağırlığa göre çarpım ve çıkan sonuçların toplamı:

T2 = (S1*0.4)+(S2*0.6) = (2.5*0.4) +(1*0.6) = 1+0.6 = 1.6

T2<T1 olduğundan mavi, sarı, bakımlı, 1, 5 ikinci kümeye eklendi.

M1 ve sözel veri kümesi elemanı yeşil, sarı, bakımlı için Jaccard benzerlik ölçütü

benj($m_1, X_3$) = $\frac{|ela,kahve,bakımlı| \cap |yeşil,sarı,bakımlı|}{|ela,kahve,bakımlı| \cup |yeşil,sarı,bakımlı|}$ = 1/5 = 0.2

M1 ve sayısal veri kümesi elemanı 2,2 için öklit:

S1 = $\sqrt{(1.67-2)^2 + (3.33-2)^2}$ = 1.37,

S2 = 1-benj($m_1, X_3$) = 1-0.2 = 0.8

Normalizasyon işlemi sonuçu:

S1 = 1.37,

S2 = 0.8

Ağırlığa göre çarpım ve çıkan sonuçların toplamı:

T1 = (S1*0.4)+(S2*0.6) = (1.37*0.4) +(0.8*0.6) = 0.55+0.48 = 1.03

M2 ve sözel veri kümesi elemanı yeşil, sarı, bakımlı için Jaccard benzerlik ölçütü

benj($m_2, X_3$) $= \frac{|yeşil,sarı,bakımlı| \cap |yeşil,sarı,bakımlı|}{|yeşil,sarı,,bakımlı| \cup |yeşil,sarı,bakımlı|}$ = 3/3 = 1

M2 ve sayısal veri kümesi elemanı 2,2 için öklit değeri:

S1 = $\sqrt{(4-2)^2 + (3.5-2)^2}$ = 2.5,

S2 = 1-benj($m_1, X_3$) = 1-0 = 0

Normalizasyon işlemi sonuçu:

S1 = 2.5,

S2 = 0

Ağırlığa göre çarpım ve çıkan sonuçların toplamı:

T2 = (S1*0.4)+(S2*0.6) = (2.5*0.4) +(0*0.6) = 1+0 = 1

T2<T1 olduğundan yeşil, sarı, bakımlı, 2, 2 ikinci kümeye eklendi.

M1 ve sözel veri kümesi elemanı yeşil, sarı, bakımlı için Jaccard benzerlik ölçütü

benj($m_1, X_4$) $= \frac{|ela,kahve,bakımlı| \cap |yeşil,sarı,bakımlı|}{|ela,kahve,bakımlı| \cup |yeşil,sarı,bakımlı|}$ = 1/5 = 0.2

M1 ve sayısal veri kümesi elemanı 7,4 için öklit değeri:

S1 = $\sqrt{(1.67-7)^2 + (3.33-4)^2}$ = 5.37,

S2 = 1-benj($m_1, X_3$) = 1-0.2 = 0.8

Normalizasyon işlemi sonuçu:

S1 = 5.37,

S2 = 8

Ağırlığa göre çarpım ve çıkan sonuçların toplamı:

T1 = (S1*0.4)+(S2*0.6) = (5.37*0.4) +(8*0.6) = 2.15+4.8 = 6.95

M2 ve sözel veri kümesi elemanı yeşil, sarı, bakımlı için Jaccard benzerlik ölçütü

benj($m_2, X_4$) = $\frac{|yeşil,sarı,bakımlı| \cap |yeşil,sarı,bakımlı|}{|yeşil,sarı,,bakımlı| \cup |yeşil,sarı,bakımlı|}$ = 3/3 = 1

M2 ve sayısal veri kümesi elemanı 7,4 için öklit değeri:

S1 = $\sqrt{(4-7)^2 + (3.5-4)^2}$ = 3.04 ,

S2 = 1-benj($m_1, X_3$) = 1-1 = 0

Normalizasyon işlemi sonuçu:

S1 = 3.04,

S2 = 0

Ağırlığa göre çarpım ve çıkan sonuçların toplamı:

T2 = (S1*0.4)+(S2*0.6) = (3.04*0.4) +(0*0.6) = 1.22+0 = 1.22

T2<T1 olduğundan yeşil, sarı, bakımlı, 7, 4 ikinci kümeye eklendi.

M1 ve sözel veri kümesi elemanı kahve, siyah,bakımsız için Jaccard benzerlik ölçütü

benj($m_1, X_5$) = $\frac{|ela,kahve,bakımlı| \cap |kahve,siyah,bakımsız|}{|ela,kahve,bakımlı| \cup |kahve,siyah,bakımsız|}$ = 1/5 = 0.2

M1 ve sayısal veri kümesi elemanı 1,3 için öklit değeri:

S1 = $\sqrt{(1.67-1)^2 + (3.33-3)^2}$ = 0.74 ,

S2 = 1-benj($m_1, X_3$) = 1-0.2 = 0.8

Normalizasyon işlemi sonuçu:

S1 = 0.74,

S2 = 0.8

Ağırlığa göre çarpım ve çıkan sonuçların toplamı:

T1 = (S1*0.4)+(S2*0.6) = (0.74*0.4) +(0.8*0.6) =0.97+0.48 = 1.45

M2 ve sözel veri kümesi elemanı kahve, siyah,bakımsız için Jaccard benzerlik ölçütü

benj($m_2, X_5$) $= \frac{|yeşil,sarı,bakımlı\ | \cap |\text{kahve,siyah,bakımsız}|}{|yeşil,sarı,,bakımlı\ | \cup |\text{kahve,siyah,bakımsız}|}$ = 0/6 = 0

M2 ve sayısal veri kümesi elemanı 1,3 için öklit değeri:

S1 = $\sqrt{(4-1)^2 + (3.5-3)^2}$ = 3.04 ,

S2 = 1-benj($m_1, X_3$) = 1-0 = 1

Normalizasyon işlemi sonuçu:

S1 = 3.04,

S2 = 1

Ağırlığa göre çarpım ve çıkan sonuçların toplamı:

T2 = (S1*0.4)+(S2*0.6) = (3.04*0.4) +(1*0.6) = 1.22+0.6 = 1.82

T1<T2 kahve, siyah, bakımsız, 1, 3 birinci kümeye eklendi.

İterasyon 2 sonunda oluşan kümeler şu şekildedir.

**Tablo 33. Karışık verileri kümelemede iterasyon 2 sonucu oluşan yeni kümeler**

| | | | | | |
|---|---|---|---|---|---|
| **Küme 1** | Ela | Kahve | Bakımlı | 3 | 2 |
| | kahve | Siyah | Bakımsız | 1 | 3 |
| **Küme 2** | mavi | Sarı | Bakımlı | 1 | 5 |
| | yeşil | Sarı | Bakımlı | 7 | 4 |
| | yeşil | Sarı | Bakımlı | 1 | 3 |

Bölüm 3.1'deki algoritma için yeni merkezler

M1 = ela, kahve, bakımlı

M2 = yeşil, sarı, bakımlı

K-Means algoritması için merkezler

M1 = (3+1+1)/3, (2+3+5)/3 = 1.67, 3.33

M2 = (7+1)/2, (4+3)/2 = 4, 3.5

En son küme merkezleri ile bir önceki küme merkezleri aynı olduğu için sonsuz iterasyon kırıldı ve algoritma sonlandırıldı.

**Tablo 34. Karışık verileri kümelemede sonucunda oluşan kümeler**

| | | | | | |
|---|---|---|---|---|---|
| **Küme 1** | Ela | Kahve | bakımlı | 3 | 2 |
| | kahve | Siyah | bakımsız | 1 | 3 |
| **Küme 2** | mavi | Sarı | bakımlı | 1 | 5 |
| | yeşil | Sarı | bakımlı | 7 | 4 |
| | yeşil | Sarı | bakımlı | 1 | 3 |

Önerilen bu algoritmada doğru sonuç alabilmek için ilk küme merkezlerinin özenle seçilmesi, eldeki veri kümesinin optimum olarak kaç kümeye ayrılabileceği ve algoritmalardan sonra yapılan normalizasyonların çok düzgün bir şekilde yapılması gerekmektedir. Bu çalışmada normalizasyon için çok uygun bir yöntem geliştirilememiş olup decimal normalizasyona benzer bir normalizasyon yapılmıştır. Eğer bu algoritmaya uygun bir normalizasyon işlemi bulunursa algoritmadan çok yüksek başarılı sonuçlar alınabilir.

# SONUÇ

Daha öncede bahsedildiği gibi alışılmış kümeleme algoritmalarında sayısal veriler sayısal veriler ile kümelenebilmektedir. Eğer veriler sadece sözel veya karışık olduğunda özellikle K-Means algoritmasında mantıksal hatalar oluşmaktadır. Bu tez çalışmasında öncelikle sayısal olmayan verilerin sayısal verilerle çalışan K-Means algoritmasında nasıl çalışabiliceği anlatılmaya çalışılmıştır. Bu işlemi yapabilmek için K-Means algoritmasının içine Jaccard benzerlik ölçütü bölüm 3.1'deki gibi adapte edilmiştir. Sayısal olmayan verilerin K-Means algoritmasında çalışması işlemi bittikten sonra bu yöntemden faydalanarak sayısal ve sayısal olmayan karışık veri kümelerinin K-Means algoritmasında nasıl uygulanabiliceğine dair bir yöntem önerilmiştir. Karışık verileri kümelemek bölüm 3.2'deki gibi bir yol izlenmiştir. Bölüm 3.2'deki yollar takip edilirken dikkat edilmesi gereken üç önemli husus vardır. Bunlardan birincisi ilk küme merkezleri bulunurken klasik K-Means algoritması gibi rastgele değil bölüm 3.1'de açıklandığı gibi K-Means algoritmasının kategorik verilerde çalıştırma mantığına göre olmasıdır. Bunun nedeni klasik K-Means algoritmasına göre ilk küme merkezlerinin öneminin bulunmayıp K-Means algoritması içinde sayısal olmayan verileri çalıştırma mantığında öneminin olmasıdır. İkinci olarak kategorik verileri K-Means algoritmasına göre kümelerken, eldeki verilerin optimum olarak kaç kümeye ayrılabileceği iyi ayarlanmalı ve Dunn Geçerlilik İndeksi, Davies-Bouldin İndeksi, Silhoutte Geçerlilik Yöntemi, C İndexi gibi algoritmalardan faydalanılmalıdır.[14]. Bunun nedeni ise ilk küme merkezlerinin bölüm 3.1'de açıklanan algoritmada öneminin olduğu için eldeki veriler fazla kümeye bölünmek istenirse ilk küme merkezlerinde sapmalar yaşanabilir. Son önemli husus ise verilerde normalizasyon yapılırken, normalizasyonun çok iyi yapılması gerektiğidir. Bu çalışmada normalizasyon için çok uygun bir yöntem geliştirilememiş olup decimal normalizasyona benzer bir normalizasyon yapılmıştır. Eğer bu algoritmaya uygun bir normalizasyon işlemi bulunursa algoritmadan çok başarılı sonuçlar alınabilir. Bahsedilen yolla karışık veri kümelerinde kümeleme yapılmak istendiği takdirde çok yüksek bir oranda doğru bir şekilde kümeleme yapılabilmektedir.

## KAYNAKLAR


1. Ada, M., Altunay, F., Civelek, M., Kaplan, S., et al: Kümeleme Analizi

2. Fosca Giannotti, Dino Pedreschi (Eds.), Mobi-lity, Data Mining and Privacy, p-3, 47 http://www.mlplatform.nl/what-is-machine-learning/

3. Hair, Jr. F.J., Anderson, E. R., Tatham, L. R., et al.: Multivariate Data Analysis With Readings, 5. Ed., Prentice-Hall, USA, 1998.

4. Han, J., & Kamber M. Second Edition. Data Mining Concepts and technques.

5. Hopfield, J.J. 1982. Neural networks and physical systems with emergent collective computational abilities. Proc.of the National Academy of Sciences, vol. 79, pp.2554-2558.

6. Kalıkov, A., (2006), Veri Madenciliği ve Bir E-Ticaret Uygulaması, Yüksek Lisans Tezi, Gazi Üniversitesi, Fen Bilimleri Enstitüsü.

7. Kaufman, L. and Rousseeuw, P.J. 1990. Finding Groups in Data: An Introduction to Cluster Analysis, John Wiley and Sons.

8. Kohonen, T. 1982. Self-organized formation of topologically correct feature maps. Biological Cybernetics, vol. 43, pp. 59-69.

9. OĞUZLAR Ayşe , KÜMELEME ANALİZİNDE YENİ BİR YAKLAŞIM: KENDİNİ DÜZENLEYEN HARİTALAR (KOHONEN AĞLARI)

10. Widrow, B. and Hoff, M.E. 1960. Adaptive switching circuits. IRE Wescon Convention Record: Part 4, Computers: Man-machine systems, pp. 96-104, Los Angeles.
11. www.ist.yildiz.edu.tr/dersler/dersnotu/Kum-Analiz.doc

12. P. Jaccard. The distribution of flora in the alpine zone. 1912. New Phytologist. 11, 37-50.

13. Saharkhiz Aresh,( 3 Jan 2009), K-Means Clustering Used in Intention Based Scoring Projects
14. Silahtaroğlu, G., Veri Madenciliği Kavram ve Algoritmaları. 2013. S-163, 208, 215.

# ÖZGEÇMİŞ

10 Mart 1986 tarihi, İstanbul ili Fatih ilçesi doğumluyum. ilköğretim ve liseyi Avcılar ilçesinde tamamladıktan sonra, Doğu Akdeniz Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümüne kaydoldum. Bu bölümden 2009 yılında mezun olduktan sonra, 5 ay Login Yazılım A.Ş.'de çalıştım. Askerlik görevimi, Hatay Yayladağı Dağardı Hudut Karakolunda Karakol Komutanı olarak tamamladım.
2011 yılından beri, Vakko Hazır Giyim Tekstil A.Ş. 'nin Bilgi İşlem Merkezinde görevime devam etmekteyim.

2012 yılında da, Beykent Üniversitesi, Bilgisayar Mühendisliği Anabilim Dalında yüksek lisans eğitimine basladım.

Özel ilgi alanlarım, kitap okumak, rafting, yelken, tenis, futbol, basketbol, yüzmek, sinema-tiyatro, zeka ve strateji oyunları, puzzle, tamirat, akvaryum ile uğraşmak, balık tutmak, seyahat etmek, bilek güreşi, bahçe işleri ile uğraşmak.

Yabancı dilim İngilizce olup, evliyim.

**Aday:**