T.C.
YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

K - ORTALAMALAR ALGORİTMASINA DAYALI KÜMELEME ANALİZİ SİSTEMİ VE PERAKENDECİLİK SEKTÖRÜNDE UYGULAMASI

MERVE ÜSTÜNEL

YÜKSEK LİSANS TEZİ
MATEMATİK MÜHENDİSLİĞİ ANABİLİM DALI
MATEMATİK MÜHENDİSLİĞİ PROGRAMI

DANIŞMAN
PROF. DR. AYLA ŞAYLI

İSTANBUL, 2018

T.C.
YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

K - ORTALAMALAR ALGORİTMASINA DAYALI KÜMELEME ANALİZİ SİSTEMİ VE PERAKENDECİLİK SEKTÖRÜNDE UYGULAMASI

Merve ÜSTÜNEL tarafından hazırlanan tez çalışması 27/06/2018 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü Matematik Mühendisliği Anabilim Dalı'nda **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Tez Danışmanı

Prof. Dr. Ayla ŞAYLI

Yıldız Teknik Üniversitesi

Jüri Üyeleri

Prof. Dr. Ayla ŞAYLI

Yıldız Teknik Üniversitesi _____

Prof. Dr. İbrahim EMİROĞLU

Yıldız Teknik Üniversitesi _____

Dr. Öğr. Üyesi Mustafa Zahid GÜRBÜZ

Doğuş Üniversitesi _____

ÖNSÖZ

Veri madenciliği, günümüzde karar verme sürecine ihtiyaç duyulan pek çok alanda uygulanması ile birlikte büyük önem kazanmıştır. Firmalarda biriken verinin her geçen gün artmasıyla, verilerin doğru analiz edilmesi ve kullanılabilir hale getirilmesi mevcut ve ileriye dönük planlamalara öngörü sağlamaktadır. Veri madenciliğindeki temel teknikler; sınıflandırma, kümeleme ve birliktelik kuralları olarak gruplandırılmıştır. Bu tez çalışmamızda kümeleme esas alınmıştır.

Kümeleme yöntemleri arasından seçilen K-Ortalamalar algoritması ile bu algoritmanın daha doğru uygulanması için gereken metotların belirlenmesiyle oluşan veri tabanı üzerinde Java dili kullanılarak bir veri analiz sistemi oluşturulmuştur. Bu bağlamda, bu çalışmayla birlikte çalışmaya başlarken belirlediğim amaçlarıma büyük ölçüde ulaştığımı söyleyebilirim. Tüm bu çalışmalar; Java dilini öğrenme, uygulama tasarımı, algoritma geliştirme ve çeşitli metotların algoritma kalitesine etkilerini belirleme gibi birbirini bütünleyen birçok konuda tecrübe kazandırmıştır. Diğer yandan, yazılım alanındaki şüphelerimden kurtulup, bu alanı çalışma sektörü olarak tercih etmemi de sağlamıştır.

Tez çalışmasında kullanılan veri dosyası Migros Ticaret A.Ş. ye ait olup, gerçek verilerden oluşmaktadır. Çalışma sonunda analiz sonuçları firmaya sunulmuştur. Söz konusu sonuçların firmanın karar verme süreçlerine katkı sağlaması ve firmanın marka değerinde artış olması en büyük temennimizdir.

Tez çalışmamın planlanmasında, araştırılmasında, yürütülmesinde ve oluşumunda ilgi ve desteğini esirgemeyen, engin bilgi ve tecrübelerinden yararlandığım saygıdeğer hocam Prof. Dr. Ayla Şaylı'ya sonsuz teşekkürlerimi sunarım.

Çalışmamda desteğini ve bana olan güvenini benden esirgemeyen diğer üniversite hocalarıma ve benim için çok değerli arkadaşlarıma bana kazandırdıkları için teker teker teşekkür ederim. Son olarak, beni bu günlere sevgi ve saygı kelimelerinin anlamlarını bilecek şekilde yetiştirerek getiren ve desteğini hiçbir zaman esirgemeyen, bu hayattaki en büyük şansım olan aileme sonsuz teşekkürü bir borç bilirim.

Haziran, 2018

Merve ÜSTÜNEL

İÇİNDEKİLER

Sayfa

SİMGE LİSTESİ

X	Veri dosyası
n	Veri dosyasındaki kayıt sayısı
x	Veri dosyasında bulunan herhangi bir kayıt
m	Veri dosyasındaki öz nitelik sayısı
$\overline{x}$	Veri dosyasındaki belirli bir öz nitelik için kayıtların aritmetik ortalaması
σ_x	Veri dosyasındaki belirli bir öz nitelik için kayıtların standart sapması
x^*	Veri dosyasındaki herhangi bir kaydın normalleştirilmiş hali
$x_{\max}$	Veri dosyasındaki belirli bir öz nitelik için en büyük kayıt
$x_{\min}$	Veri dosyasındaki belirli bir öz nitelik için en küçük kayıt
x_i	Veri dosyasındaki i'inci kayıt
d	Uzaklık fonksiyonu
μ	Ortalama vektör
S^{-1}	Varyans-kovaryans matrisinin tersi
$SLINK(X,Y)$	Tek bağlantı uzaklık ölçüsü
$CLINK(X,Y)$	Tam bağlantı uzaklık ölçüsü
C	Küme
k	Küme sayısı
c	Küme merkezi
CH	Calinski-Harabasz indeksi
WSS	Kümeler içi kareler toplamı
BSS	Kümeler arası kareler toplamı
$DIFF$	Kümeler içi kareler toplamlarının farkı
KL	Krzanowski-Lai indeksi
$SilC$	Silhouette indeksi
$Gap_n(k)$	Fark istatistiği değeri
$a(i)$	i'inci kaydın bulunduğu kümedeki tüm kayıtlara uzaklıklarının ortalaması
$b(i)$	i'inci kaydın diğer kümelerdeki tüm kayıtlara olan ortalama uzaklıklarının minimumu
v	Özvektör
p	C kümesinin v özvektörü üzerindeki izdüşümü

KISALTMA LİSTESİ

AIC	Akaike Bilgi Kriteri (**A**kaike **I**nformation **C**riterion)
BIC	Bayesian Bilgi Kriteri (**B**ayesian **I**nformation **C**riterion)
CRM	Müşteri İlişkileri Yönetimi (**C**ustomer **R**elationship **M**anagement)
GUI	Grafiksel Kullanıcı Arayüzü (**G**raphical **U**ser **I**nterface)
JDBC	Java Veri Tabanı Bağlantısı (**J**ava **D**ata**b**ase **C**onnectivity)
PCA	Temel Bileşenler Analizi (**P**rincipal **C**omponent **A**nalysis)
ROBIN	Sağlam Başlatma (**Rob**ust **In**itialization)
SQL	Yapısal Sorgulama Dili (**S**tructured **Q**uery **L**anguage)

ŞEKİL LİSTESİ

Sayfa

ÇİZELGE LİSTESİ

Sayfa

ÖZET

K - ORTALAMALAR ALGORİTMASINA DAYALI KÜMELEME ANALİZİ SİSTEMİ VE PERAKENDECİLİK SEKTÖRÜNDE UYGULAMASI

Merve ÜSTÜNEL

Matematik Mühendisliği Anabilim Dalı

Yüksek Lisans Tezi

Tez Danışmanı: Prof. Dr. Ayla ŞAYLI

Gelişen ve değişen çevre koşulları, internetin küreselleşme derecesinin yükselmesi, farklı Ar-Ge (Araştırma Geliştirme) ve pazarlama yöntemleri ile rekabetin belirgin bir şekilde artması ve müşterileri memnun etmenin zorlaşması, veriden çıkarılacak bilginin önemini her geçen gün artırmaktadır. Bilginin bazı yöntemler kullanılarak analiz edilmesi ve elde edilen sonuçların konunun uzmanı tarafından yorumlanmasıyla geçmiş verilerden gelecek tahminleri yapma işlemi veri madenciliği (data mining) olarak belirtilebilir. Firmalar ve işletmeler için veri madenciliği, karar vericilerin karar vermelerini kolaylaştıran ve hızlı karar almasını sağlayan önemli ve stratejik bir araçtır.

Verilerin benzer gruplara ayrılması, verilerin kümelenmesi, veri madenciliğindeki en temel tekniklerden biridir. Bu tez çalışmasında, hiyerarşik olmayan kümeleme yöntemlerinden biri olan K-Ortalamalar algoritmasından faydalanılarak müşteri satın alma davranışları analiz edilecektir. Kümelenen veriler ile hangi müşteri profilinin hangi markayı, hangi ürünü, ne zaman ve ne miktarda tercih ettiği belirlenecektir. Yapılan çalışmada amaç, müşteri tercihleri dikkate alınarak firma için hem talep yaratma, hem de doğru talebi doğru zamanda karşılama ve sunma gibi avantajların sağlanacağı bir sistem oluşturmak ve bu sistemden yararlanarak firma için veri analizi gerçekleştirmektir. Tez çalışması sırasında analiz için kullanılacak olan sistem Java dilinde geliştirilmiştir ve analiz sonuçları grafik ve tablo ile görselleştirilmiştir. Böylece, K-Ortalamalar algoritması için dinamik bir kümeleme analiz sistemi oluşturulmuştur. Analizde kullanılacak olan veri dosyası dünya perakende sektörü listesinde yer alan

Migros Ticaret A.Ş. ye ait olup, gerçek verilerden oluşmaktadır. Veriler MS-SQL veri tabanında bir tabloda tutularak, tüm veri hazırlama işlemleri bu tablo üzerinde gerçekleştirilmiştir.

Tez çalışmasından önce ilk olarak, aynı veri dosyası ile "Brand Loyalty Analysis System Using K-Means Algorithm" adlı müşterilerin marka bağımlılığını inceleyen bir makale üzerinde çalışılmıştır. Makale çalışmamızda, veri analiz için küme sayısı bir metoda bağlı olmaksızın tahmini olarak seçilmiştir. Ayrıca, başlangıç merkezlerin seçimi de rastgele gerçekleştirilmiştir. Sonuçlar; genel marka bağımlılığı, ürün bazında marka bağımlılığı ve kategori bazında marka bağımlılığı olarak üç farklı şekilde analiz edilmiştir ve uluslararası bir dergide yayınlanmıştır.

Bununla birlikte veri analiz sistemi, tez çalışmasında kullanılmak üzere, küme sayısının seçiminde ve başlangıç merkezlerin seçiminde gereken bazı metotlar kullanılarak iyileştirilmiştir. Tez çalışmasında, k değeri 2 ile 20 arasında seçilerek, her bir k değeri için hata hesaplanmıştır ve veri dosyasının kaç kümeye ayrılması gerektiğine (optimal k değerinin belirlenmesi) Elbow metodu kullanılarak karar verilmiştir. Belirlenen k değeri için, başlangıç merkezlerin belirlenmesi amacıyla Maximin, Katsavounidis, PCA-Part, Var-Part ve K-Ortalamalar++ metotları kullanılmıştır. Optimal k değerini belirleme yöntemi olarak seçilen Elbow metodu ile farklı başlangıç merkezleri seçiminin kümelemeye etkisi araştırılmıştır. Kümeleme sonuçları, kümeleme değerlendirme kriterleri olan Silhouette ve Calinski-Harabasz indeksleri kullanılarak değerlendirilmiştir ve sonuçlar firmaya sunulmuştur. Geliştirilen analiz sistemi, diğer firmalar ve işletmeler için de bir karar destek sistemi olarak kullanılabilir.

Anahtar Kelimeler: Veri Madenciliği, Kümeleme Analizi, K-Ortalamalar Algoritması, Elbow Metodu, Başlangıç Merkezlerin Seçimi, Kümeleme Değerlendirme Kriterleri

ABSTRACT

CLUSTERING ANALYSIS SYSTEM BASED ON K-MEANS ALGORITHM AND ITS APPLICATION IN THE RETAIL SECTOR

Merve ÜSTÜNEL

Department of Mathematical Engineering

MSc. Thesis

Adviser: Prof. Dr. Ayla ŞAYLI

Developing and changing environmental conditions, globalization of the internet, competition with different research and development activities and marketing methods, and difficulties in customers' satisfaction are increasing the importance of information obtained from data day by day. The analysis of the information using some methods, the interpretation of the obtained results by the subject matter experts and making future forecasts from historical data can be stated as data mining. Data mining for companies and businesses is an important and strategic tool that facilitates decision making and allows decision makers to make quick decisions.

Separating of data into similar groups, clustering of data, is one of the most basic methods in data mining. In this thesis, customer buying behaviors will be analyzed using the K-Means algorithm which is one of the non-hierarchical clustering methods. With the clustered data, it will be determined which brand, which product, when and how much is preferred by different customer profiles. The aim of this thesis is to create a system that will provide advantages such as both creating demand and meeting the right demand at the right time considering the customer preferences, and also realize data analysis using this system for the company. The system to be used for the analysis during the thesis was developed in the Java language and the obtained results were visualized by graphics and tables. Thus, a dynamic clustering analysis system was

established for the K-Means algorithm. The data file to be used in the analysis belongs to Migros Ticaret A.S. on the global powers of retailing and consists of actual data. The data were stored in a table in the MS-SQL database, and all data preparation operations were performed on this table.

Before the thesis, an article reviewing the brand loyalty of the customers named "Brand Loyalty Analysis System Using K-Means Algorithm" with the same data file was studied first. In our article, the number of clusters for data analysis was estimated, regardless of a method. In addition, the selection of the initial centers was conducted randomly. The results that are general brand loyalty, brand loyalty based on item and brand loyalty based on category were published in an international journal.

However, the analysis system has been improved by using some methods for selecting the number of clusters and selecting the initial centers. In this thesis, the error for each k value is calculated by choosing k values from 2 to 20, and how many clusters of the data should be separated (determining the optimal k value) has been determined using the Elbow method. For the determined k value; Maximin, Katsavounidis, PCA-Part, Var-Part and K-Means++ methods have been used to find the initial centers. With the Elbow method which is chosen as the method of determining the optimal k value, the effect of cluster selection of different initial centers has been investigated. Clustering results have been evaluated using the Silhouette and Calinski-Harabasz criterions and were presented to the company. The developed analysis system can also be used as a decision support system for other companies and businesses.

Keywords: Data Mining, Clustering Analysis, K-Means Algorithm, Elbow Method, Selection of Initial Centers, Clustering Validation Criterions

BÖLÜM 1

GİRİŞ

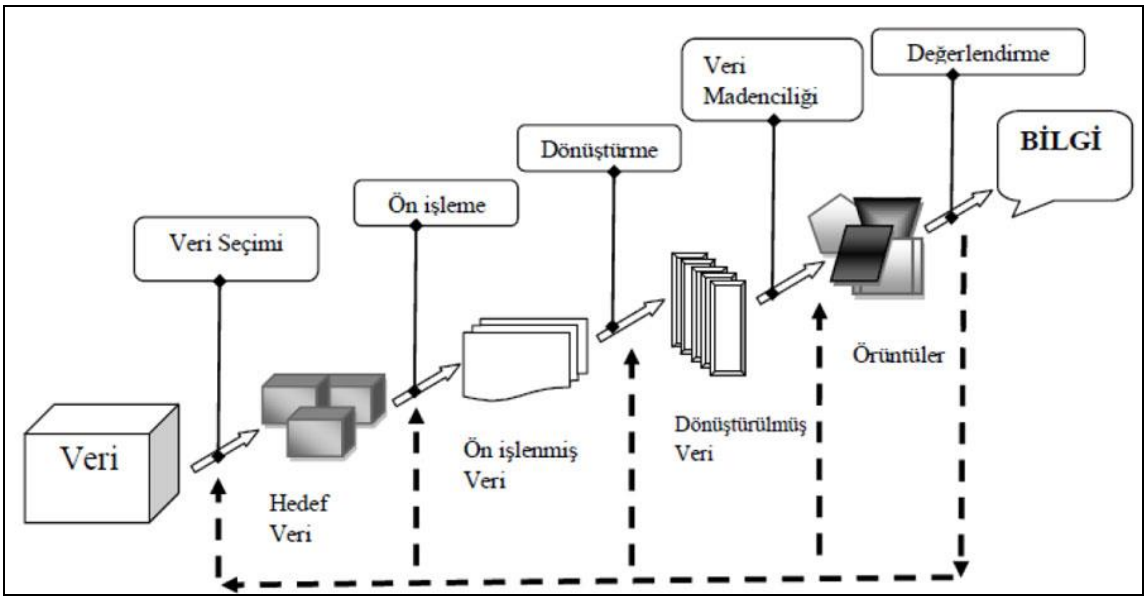
1.1 Literatür Özeti

Veri madenciliğinin son yıllarda öne çıkması, büyük miktarda verinin geniş çapta kullanılabilir yararlı bilgiye dönüştürülmesi ihtiyacından kaynaklanmaktadır. Veri madenciliği bilgi teknolojilerinin gelişiminin doğal bir sonucu olarak görülebilir. 1960 ların başında ilkel dosya işleme sistemleri ile başlayan veriyi toplama ve veri tabanı oluşturma sürecinin evrimsel gelişimi, 1970 lerde veri tabanı yönetimi sistemlerinin gelişmesi ve 1980 lerin sonundan günümüze kadar gelen veri madenciliği ve bilgi keşfi (verinin analizi ve veriyi anlama) olarak görülmüştür. Bu gelişim sırasında her bir adım önceki adımdan tetiklenmiştir ya da her bir adım sonraki adıma ortam hazırlamıştır. Örneğin, veriyi toplama ve veri tabanı oluşturma adımlarının gelişimi; veriyi depolama ve geri çağırma, veri sorgulama ve işleme gibi adımlar için bir ön ortam hazırlamıştır. Veriyi sorgulama ve işleme imkânı sunan çok sayıda veri tabanı sistemiyle de verinin analizi ve veriyi anlama adımları hedef haline gelmiştir [1].

Veri madenciliği, veriler arasındaki ilişkileri ortaya koymak ve gerektiğinde ileriye yönelik tahminlerde bulunmak amacıyla, büyük ölçekli veriler arasından "değeri olan" bilginin elde edilmesi işidir [1] ve veri madenciliği, Şekil 1.1 de gösterilen bilgi keşfi sürecindeki önemli adımlardan bir tanesidir. Süreçte izlenen adımlar şu şekilde sıralanabilir [2]:

- **İhtiyacın tanımlanması:** Çalışmanın hangi iş ihtiyacı için yapılacağı belirlenir.

- **Verilerin hazırlanması:** Sonucu doğrudan etkileyecek olan bu adımda; “toplama”, “değer biçme”, “birleştirme ve temizleme”, “örneklem seçimi” ve “dönüştürme” gibi veri hazırlık aşamaları gerçekleştirilir.
- **Modelin kurulması ve değerlendirilmesi:** Tanımlanan iş ihtiyacı için en iyi olduğu düşünülen model elde edilinceye kadar yinelenir ve en uygun model seçilir.
- **Modelin kullanılması:** Kurulan ve geçerliliği kabul edilen model, sonuçları elde etmek ve gerekli değerlendirmeleri yapmak üzere kullanılır.
- **Modelin izlenmesi:** Sistemlerin ve ürettikleri verilerin zaman içinde değişmesiyle birlikte, kurulan model devamlı olarak izlenir.



Şekil 1. 1 Bilgi keşfi sürecinde veri madenciliğinin yeri [2]

Veri madenciliği her geçen gün daha fazla yaygınlaşmaktadır; veri madenciliği, firmalara ve işletmelere mevcut veri tabanlarındaki modellerin ve eğilimlerin ortaya çıkarılmasını sağlamaktadır [3]. Bu bağlamda, veri madenciliği firma ve işletmeler için karar destek sistemleri açısından önemlidir ve verilerin her geçen gün artmasıyla daha da önemli bir konuma gelecektir.

Veri madenciliği başlıca aşağıda belirtilen uygulama alanlarında tercih edilmektedir:

Pazarlama

- Müşterilerin demografik özellikleri arasındaki bağlantıların ortaya konulması

- Müşterilerin satın alma alışkanlıklarının belirlenmesi
- Mevcut müşterilerin elde tutulması, yeni müşterilerin kazanılması
- Müşteri değerlendirmesi ve müşteri ilişkilerinin yönetilmesi
- Pazar sepet analizi ve satış tahmini yapılması

Bankacılık

- Farklı finansal göstergeler arasındaki ilişkilerin belirlenmesi
- Kredi kartı harcamalarına göre müşteri gruplarının belirlenmesi
- Kredi taleplerinin değerlendirilmesi

Sigortacılık

- Yeni poliçe talep ederek müşterilerin tahmin edilmesi
- Sigorta dolandırıcılıklarının tespiti
- Riskli müşteri gruplarının belirlenmesi

Elektronik Ticaret

- Saldırıların çözümlenmesi
- e-CRM uygulamalarının yönetilmesi
- Web sayfalarına yapılan ziyaretçilerin çözümlenmesi

Veri madenciliğinin günümüzde yaygınlaşmasına paralel olarak, veri madenciliği konusundaki araştırmaların ve çalışmaların sayısında da artış görülmektedir [4]. Özellikle Türkiye genelinde veri madenciliği üzerine yapılan çalışmalar ve gerçekleştirilen uygulamaların incelendiği birçok çalışma bulunmaktadır.

Savaş, Topaloğlu ve Yılmaz yaptıkları çalışmada, Türkiye'de yapılan veri madenciliği çalışmalarını araştırmıştır [2]. Çalışma başlıkları; mühendislik alanında gerçekleştirilen veri madenciliği uygulamaları, tıp alanında gerçekleştirilen veri madenciliği uygulamaları, bankacılık ve borsa alanında gerçekleştirilen veri madenciliği uygulamaları, eğitim alanında gerçekleştirilen veri madenciliği uygulamaları, ticari alanda gerçekleştirilen veri madenciliği uygulamaları ve telekomünikasyon alanında

gerçekleştirilen veri madenciliği uygulamaları şeklindedir. Çalışmanın sonuçlarına bakıldığında, firma ve işletmelerin çoğunun öncelikli olarak müşteri analizlerine yöneldiği görülmüştür.

Doğan yaptığı çalışmada "veri madenciliği" anahtar sözcüklerini kullanarak şimdiye kadar yazılmış tezleri tarayıp, bu tezlerin yüksek lisans/doktora tezleri olarak dağılımlarını göstermiştir [5]. Ayrıca, söz konusu tezlerin hangi veri madenciliği teknikleri kullanılarak hazırlandığına değinmiştir. Çalışma sonuçları incelendiğinde, veri madenciliği tekniklerinden biri olan kümelemeyi kullanan tez sayısının diğer teknikleri kullanan tez sayısına göre daha fazla olduğu görülmüştür.

Somut veya soyut kayıtları, benzer kayıtların sınıflarına ayırma işlemine kümeleme denir. Kümelemede kayıtlar, sınıf içi benzerliği en üst düzeye çıkarmak ve sınıflar arası benzerliği en aza indirgeme ilkesine dayalı olarak gruplandırılmıştır yani kümelenmiştir. Kümeler, bir küme içindeki kayıtların birbirine göre yüksek benzerliğe sahip olmaları ve diğer kümelerdeki kayıtlar ile düşük benzerlik taşımaları amacıyla oluşturulmuştur [6].

Kümeleme tekniği; biyoloji, arkeoloji, botanik, tıp, psikoloji, coğrafya, pazarlama ve görüntü işleme gibi birçok konu alanında kullanılmaktadır [7]. Etkili bir kümeleme tekniği, veri tabanı büyüklüğü ayırt etmeden her veri tabanı için elverişli olmalıdır. Bu durum aynı zamanda kümeleme algoritmasının ölçeklenebilirlik özelliğine sahip olup olmadığını göstermektedir. Temel olarak iyi bir kümeleme algoritması; uygulanması kolay, yorumlanabilir, fonksiyonel ve anlaşılır olmalıdır [8, 9].

Veri madenciliği için kullanılan tekniğe bağlı olmaksızın, verinin bilgi keşfine hazırlanması süreci genellikle aynıdır. Bölüm 2 de veri temizleme, veri bütünleştirme, veri indirgeme ve veri dönüştürme şeklindeki aşamalardan oluşan veri hazırlama işlemleri ve veri madenciliği teknikleri ele alınacaktır. Bölüm 3 te tez çalışmasında ele alınacak veri madenciliği tekniklerinden kümeleme analizi ve kümeleme analizinde önemli görülen bazı metotlar ele alınacaktır. Bölüm 4 te kümelemede yaygın olarak kullanılan K-Ortalamalar kümeleme algoritması detaylı olarak incelenecektir. Bölüm 5 de tez çalışmamızın esas başlığı olan "K-Ortalamalar Algoritmasına Dayalı Kümeleme Analizi Sistemi ve Perakendecilik Sektöründe Uygulaması" için çalışmada kullanılacak olan veri dosyası ve analiz sistemi tanıtılacaktır; analiz sisteminin çalışma şekli, İris veri

dosyası kullanılarak detaylı olarak gösterilecektir. Bölüm 6 da veri dosyasının analizinde kullanılan K-Ortalamalar kümeleme algoritmasının uygulama sonuçlarına yer verilecektir. Bölüm 7 de uygulama sonuçları değerlendirilerek yorumlanacaktır. Son olarak, ileriye dönük önerilere Bölüm 8 de yer verilecektir.

1.2 Tezin Amacı

Bu tez çalışmasının amacı, Java dilini kullanarak dinamik bir analiz sisteminin geliştirilmesi; bu sistem üzerinde, veri madenciliği tekniklerinden kümeleme için tercih edilen K-Ortalamalar algoritmasıyla birlikte küme sayısının seçimi ve başlangıç merkezlerin seçimi için farklı metotlar kullanılarak verilerin doğru, etkin ve hızlı bir şekilde kümelenmesi ve analiz edilmesidir.

1.3 Hipotez

Gerçek satış verisinin veri madenciliğinde kümeleme tekniği ile müşteri, marka, ürün ve kategori bazında yakınlık ve uzaklıklara göre farklı kümelemelere ayrılabilmesi ve bu kümelemelerin en az hata değeriyle nasıl oluşacağını ortaya koymaktır.

Elde edilen kümeleme sonuçlarının müşterinin marka bağımlılığı, müşterinin ürün ve kategori tercihleri gibi farklı ticari hedefler için kullanılabilir olduğunu göstermektir.

BÖLÜM 2

VERİ MADENCİLİĞİ

Gartner Group'a göre veri madenciliği, örüntü tanıma teknolojilerinin yanı sıra istatistiksel ve matematiksel teknikler kullanılarak, depolarda saklanan büyük miktarda veriyi eleyerek anlamlı yeni korelasyonları, örüntüleri ve eğilimleri keşfetme sürecidir [10]. Veri madenciliği ve bilginin keşfi alanlarında devam eden kayda değer büyüme, çeşitli faktörler tarafından tetiklenmiştir [11]:

- Veri toplamada patlayıcı artış,
- Verilerin veri ambarlarında depolanması,
- İnternetten veya kurum içi ağlardan verilere erişimin artması,
- Küreselleşen ekonomide pazar payını artırmaya yönelik rekabet,
- Hazır veri madenciliği yazılım paketlerinin geliştirilmesi,
- Hesaplama gücü ve depolama kapasitesindeki büyüme.

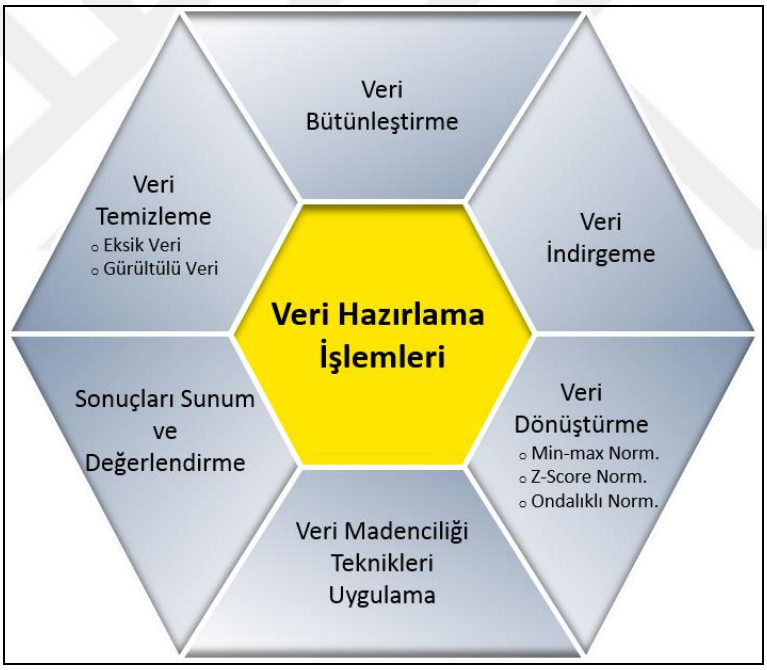
Tüm bu faktörler düşünüldüğünde, veri madenciliği için bir başka popüler alan olan doğrudan pazarlama akla gelmektedir. Promosyon teklifleri oldukça kârlı bir yanıt oranına sahiptir. Promosyon tekliflerinde, daha küçük bir örnekten yararlanılarak aynı veya neredeyse aynı yanıtı elde eden her yöntem değerlidir [12].

2.1 Veri Hazırlama İşlemleri

Veri tabanlarında bulunan ham verilerin çoğu işlenmemiş, eksik ve gürültülüdür. Veri tabanlarının, veri madenciliği amaçları için yararlı olabilmesi için veri temizliği ve veri

dönüşümü gibi ön işlemlere tabi tutulması gerekir. Veri madenciliği yıllar boyunca bakılmayan verilerle ilgilenir; bu nedenle veri tabanları, verilerin çoğunun süresi dolmuş, artık alakalı olmayan veya veri madenciliği teknikleri için uygun formda olmayan, aykırı, eksik veya gereksiz olan değerler içerir [11]. Bu durumun bir nedeni de, veri tabanını oluşturan büyük boyutlu ve heterojen kaynaklardır. Düşük kaliteli veriler, düşük kaliteli madencilik sonuçlarına yol açacaktır [12]. Veri madenciliğinde analiz edilecek olan veriler ne kadar kaliteli olursa, elde edilecek sonuçlar da bir o kadar kaliteli olacaktır.

Veri madenciliği bütün bir süreç olarak ele alınır. Dorian Pyle "Veri Madenciliğinde Veri Hazırlama" adlı kitabında veri hazırlama işlemlerinin, tüm veri madenciliği sürecinde tek başına tüm zaman ve çabanın %60 ını oluşturduğunu belirtmiştir [13]. Şekil 2.1 de gösterilen veri hazırlama işlemleri bu bölümde detaylı olarak ele alınacaktır.



Şekil 2. 1 Veri hazırlama işlemleri

2.1.1 Veri Temizleme

Analiz edilecek veya üzerinde çözümleme yapılacak verilerin bazı durumlarda istenen özelliklere sahip olmadığı görülebilir. Tutarsız ve/veya hatalı diyebileceğimiz bu verilere gürültülü veri denir. Gürültülü verilerin söz konusu olan istenmeyen sorunlardan

temizlenmesi gerekir. Diğer yandan, eksik verilerin yerine ise yenileri belirlenerek konulmalıdır. Bu bağlamda, aşağıda belirtilen yöntemlerden biri kullanılabilir [1]:

- Eksik değer içeren kayıtlar veri dosyasından çıkarılabilir.
- Kayıp değerlerin yerine genel bir sabit kullanılabilir. Ancak tüm öz niteliklerde kayıp değerlerin yerine aynı sabit değerin kullanılması doğru olmayabilir.
- İlgili öz nitelik için tüm veriler kullanılarak verilerin ortalaması hesaplanır ve eksik değer yerine hesaplanan yeni değer kullanılabilir.
- İlgili öz nitelikte tüm verilerin yerine sadece bir sınıfa ait örneklerin değişken ortalaması hesaplanarak, eksik değer yerine hesaplanan yeni değer kullanılabilir.
- Verilere uygun bir tahmin yapılarak (örn. karar ağacı modeli kurularak) eksik değer tahmin edilebilir ve bu tahmin eksik değer yerine kullanılabilir.

2.1.1.1 Eksik Veri

Analiz edilecek veri dosyasındaki verilerden bir ya da birkaçının eksik olmasıdır. Eksik verilerin tamamlanabilmesi için eksik verilerin girişinin elle yapılması, verilerin sabit bir değer ile doldurulması, verilerin ortalama ve/veya ortanca değerinin atanması ya da verilerin olası değerler ile doldurulması gibi yöntemler uygulanır. Bazı durumlarda ise, eksik değer bir hataya yol açmayabilir; hataya yol açması durumunda ise, eksik verilerin yok sayılması da farklı bir yöntem olarak kullanılabilir [6].

Eksik veriler giderilse de bir diğer önemli olan konu, iyi bir veri tabanı yönetimi ve veri girişi prosedürü tasarımıdır. Bu durum, eksik değerlerin ve/veya hataların sayısını en aza indirmeye yardımcı olacaktır.

2.1.1.2 Gürültülü Veri

Gürültülü veri, genellikle ölçüm yanlışlarından kaynaklanan tutarsız ve/veya hatalı veri girişleri yapıldığında karşılaşılan verilerdir. Verinin kullanıcı tarafından yanlış girilmesi, veri tabanı için uygun formatta olmaması, birim ya da duyarlılık gibi farklılıklardan kaynaklanan problemler nedeniyle gürültülü veri oluşabilmektedir.

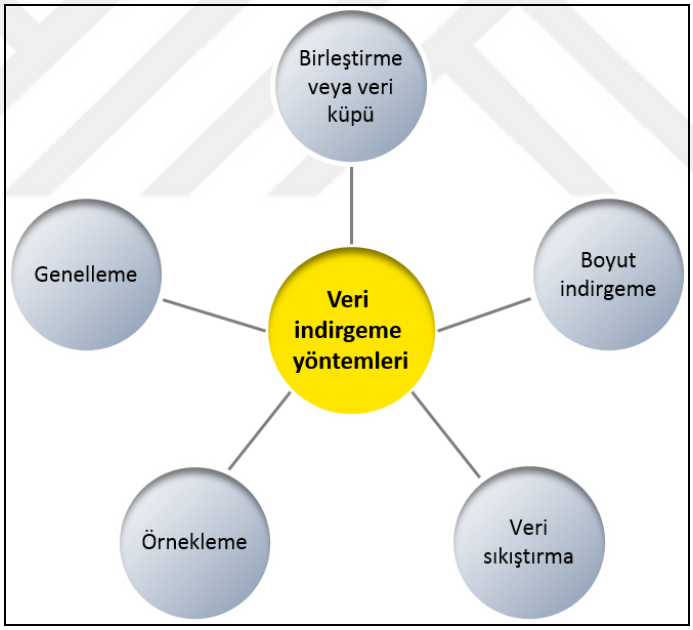
İyi bir veri madenciliği süreci için gürültülü verinin giderilmesi önem taşımaktadır. Bu anlamda, verilerin düzeltilmesi (smoothing) için en çok bilinen yöntemler; kutulama (binning), kümeleme ve doğrusal regresyondur [12].

2.1.2 Veri Bütünleştirme

Veri bütünleştirme, farklı veri kaynaklarında bulunan verilerin birlikte değerlendirilmesi amacıyla, farklı türdeki verilerin tek bir ortak türe dönüştürülmesi işlemidir.

2.1.3 Veri İndirgeme

Veri madenciliğinde verinin çözümlenmesi işlemi bazı durumlarda uzun zaman almaktadır. Eğer çözümlemeden elde edilecek olan sonuçların değişmeyeceği düşünülüyorsa, kayıt sayısı ve/veya öz nitelik sayısı azaltılabilir. Şekil 2.2 de gösterildiği gibi veri indirgeme farklı yöntemler kullanılarak uygulanabilir:



Şekil 2. 2 Veri indirgeme yöntemleri

2.1.4 Veri Dönüştürme

Birden çok öz nitelikli analizlerde çoğu zaman birimleri farklı olan öz niteliklerle ilgilenilir. Bazı durumlarda verileri olduğu gibi almak doğru olmayabilir; bu anlamda verilerin aynı birimde olması daha iyi sonuç verecektir. Bu amaçla, öz nitelik değerleri standartlaştırılarak aynı birime dönüştürülür. Diğer yandan, değişkenlerin ortalamaları

ve varyansları birbirinden büyük ölçüde farklı olduğu durumlarda ise yüksek ortalama ve varyansa sahip öz niteliklerin diğer öz nitelikler üzerinde etkisi büyüktür ve bu durum, diğer öz niteliklerin rollerinin büyük ölçüde azalması anlamına gelir. Örneğin; kümeleme analizinde uzaklıkların hesaplanması adımında, genellikle değişkenlerin standartlaştırılması yoluna gidilerek dönüştürme uygulanır. Standartlaştırma aynı zamanda, birimleri farklı değişkenlerin yapılarının aynı grafik üzerinde gösterilebilmesi ve yorumlanabilmesi için de uygulanmaktadır.

2.1.4.1 Min-Max Normalleştirmesi

Verilerin 0 ile 1 arasındaki sayısal değerlere dönüştürülmesi işlemidir. Min-max normalleştirilmesi yöntemi, veriler arasındaki en büyük ve en küçük sayısal değerlere sahip verilerin belirlenerek diğerlerinin buna uygun şekilde dönüştürülmesi esasına dayanır. Söz konusu dönüştürme formülü aşağıdaki gibidir:

$$x^* = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{2.1}$$

Burada; x kaydın değerlerini, x^* dönüştürülen değerleri, x_{max} en büyük kayıt değerini ve x_{min} en küçük kayıt değerini ifade etmektedir.

Örnek olarak $X = \{165,180,150,192,168,172\}$ boy değerleri verilsin. Bu değerler 0 ile 1 arasındaki sayısal değerlere dönüştürülmek istendiğinde, öncelikle $x_{max} = 192$ ve $x_{min} = 150$ olarak belirlenir. Dönüştürme formülü uygulandığında elde edilen yeni küme $X = \{0.3571,0.7143,0,1,0.4286,0.5238\}$ olarak elde edilir.

Bu tez çalışmasında bazı öz nitelikler için min-max normalleştirmesi kullanılmıştır.

2.1.4.2 Z-Score Normalleştirmesi

Verilerin ortalaması ve standart sapması göz önüne alınarak yeni sayısal değerlere dönüştürülmesi işlemidir. Söz konusu dönüştürme formülü aşağıdaki gibidir:

$$x^* = \frac{x - \bar{x}}{\sigma_x} \tag{2.2}$$

Burada; x kaydın değerlerini, x^* dönüştürülen değerleri, $\overline{x}$ verilerin aritmetik ortalamasını ve σ_x standart sapmayı ifade etmektedir.

Örnek olarak, bir önceki örnekte verilen $X = \{165, 180, 150, 192, 168, 172\}$ boy değerlerini ele alalım. Bu değerlere z-score standartlaştırma formülünü uygulamak için öncelikle $\overline{x}$ değerini ve σ_x standart sapmayı belirlemek gerekir. Buradan,

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{6}(165+180+150+192+168+172) = 171{,}17$$

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}} = 14{,}20$$

olarak bulunur. Dönüştürme formülü uygulandığında elde edilen yeni küme $X = \{-0.4341, 0.6219, -1.4901, 1.4667, -0.2229, 0.0587\}$ olarak elde edilir.

2.1.4.3 Ondalıklı Normalleştirme

Ondalık ölçeklendirme, ondalık noktayı hareket ettirir ancak yine de orijinal rakam değerinin büyük kısmını korur. Tipik ölçek, değerleri −1 ile 1 aralığında korur [14]. Söz konusu dönüştürme formülü aşağıdaki gibidir:

$$x^* = \frac{x}{10^a} \tag{2.3}$$

Burada; x kaydın değerlerini, x^* dönüştürülen değerleri, a ise $\max\left(\left|x^*\right|\right) < 1$ olacak şekilde bir tam sayı değerini ifade etmektedir.

Örnek olarak yine $X = \{165,180,150,192,168,172\}$ boy değerlerini ele alalım. Bu değerler ondalık değerlere dönüştürülmek istendiğinde, öncelikle $a = 3$ seçilmesi gerektiği görülür. Dönüştürme formülü uygulandığında elde edilen yeni küme $X = \{0.165, 0.180, 0.150, 0.192, 0.168, 0.172\}$ olarak elde edilir.

2.1.5 Veri Madenciliği Tekniklerini Uygulama

Veri madenciliği teknikleri uygulanmadan önce, veri analize hazırlanma adımları yapılmalıdır. Veri hazır hale getirildikten sonra ilgili veri madenciliği teknikleri uygulanabilir.

2.1.6 Sonuçları Sunum ve Değerlendirme

İlgili veri madenciliği teknikleri uygulandıktan sonra sonuçlar değerlendirilmek üzere hazırlanır. Sonuçlar genellikle grafikler veya tablolar aracılığıyla desteklenir.

Bu tez çalışmasında verinin analizinden elde edilen sonuçlar hem grafik hem tablo ile desteklenerek görselleştirilmiştir.

2.2 Veri Madenciliği Teknikleri

2.2.1 Sınıflandırma

Veri içindeki gizli örüntülerin ortaya çıkarılması amacıyla verilerin sınıflandırılması için mevcut veri dosyasının bir kısmı eğitim amacıyla kullanılarak sınıflandırma kurallarının oluşturulması sağlanır. Oluşturulan kurallar yardımıyla yeni bir durum ortaya çıktığında nasıl karar verileceği belirlenir [1]. Birçok sınıflandırma yöntemi vardır; en yaygın olan yöntem karar ağaçlarıdır. Karar ağaçları oluşturmak için ID3 ve C4.5 algoritmaları gibi yöntemler kullanılmaktadır. Örneğin, mevcut kredi müşterilerinin risk durumu karar ağaçları yardımıyla belirlenerek bir kural tablosu oluşturulabilir. Eğitim kümesinden elde edilen bu kural tablosu kullanılarak, yeni bir müşterinin risk durumu hakkında değerlendirme yapılabilir.

2.2.2 Kümeleme

Küme tanımı genellikle birbirine benzer ya da birbirine yakın kayıtların oluşturduğu topluluklar için yapılmaktadır. Kümeleme analizi ise gruplanmamış bir veri dosyasındaki kayıtları, sahip oldukları öz nitelikler çerçevesinde kümelemek için geliştirilmiş yöntemler topluluğudur [15].

Kümeleme analizi Bölüm 3 te detaylı olarak ele alınacaktır.

2.2.3 Birliktelik Kuralları

Birliktelik kuralları; kayıtların birbirleriyle olan ilişkisi ele alınarak, hangi olayların eş zamanlı olarak birlikte gerçekleşebileceklerini ortaya koymaya çalışan bir veri madenciliği tekniğidir. İlişkilerin belirlenmesiyle "birliktelik kuralları" elde edilir. "Pazar sepet analizi" adı verilen uygulamalar da bu tür veri madenciliği tekniklerine dayanmaktadır [1].

Pazar sepet analizleri yardımıyla bir müşteri bir ürünü aldığında, başka hangi ürünleri tercih edeceği belirli bir olasılık hesaplaması ile belirlenir. Örneğin, eş zamanlı alınan ürünler belirlenirse, müşterilerin ürünlere ulaşımlarının kolaylaştırılması amacıyla mağaza düzenlemeleri sağlanabilir.

BÖLÜM 3

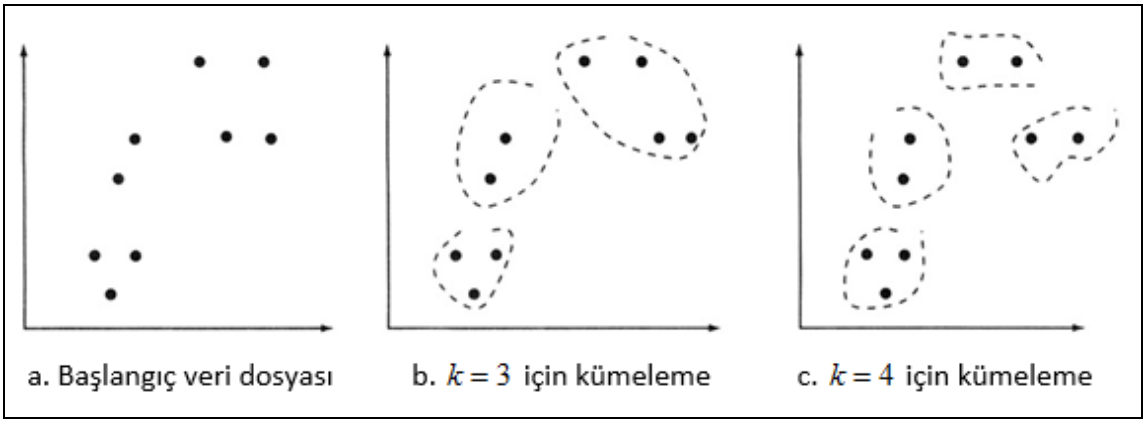
KÜMELEME ANALİZİ

Veri madenciliği tekniklerinden biri olan kümeleme analizinde, belirlenen kriterlere göre birbirine çok benzeyen kayıtlar aynı küme içinde sınıflandırılır. Kümeler, farklı öz nitelik değerlerine sahip kayıtlardan ayrılan benzer öz niteliklere sahip kayıtların kombinasyonlarıdır. Kümelerin kendi içlerinde oldukça homojen, kendi aralarında ise olabildiğince heterojen bir yapıda olması beklenmektedir. Başarılı bir şekilde tamamlanan kümeleme analizinde, aynı/farklı küme içindeki kayıtlar geometrik olarak işaretlendiğinde birbirlerine/birbirlerinden oldukça yakın/uzak olacaktır [16].

Kümelemede kullanılan değişken ifadesi, özellik ve öz nitelik ifadeleri ile aynı anlamdadır. Benzer şekilde; nesne ve gözlem ifadeleri, veri dosyasındaki kayıtları ifade etmek için kullanılabilir. Bu tez çalışmasında; veri dosyasındaki her bir satır kayıt, kaydın sahip olduğu değişkenler ise öz nitelik olarak adlandırılacaktır.

3.1 Kümeleme Nedir?

Kümeleme “Sınıf içi benzerliği en üst düzeye çıkarma ve sınıflar arası benzerliği en aza indirgeme” ilkesine bağlı kalarak, Şekil 3.1 de gösterildiği gibi belirli bir kümeleme sayısı için kayıtları gruplara ayırmada kullanılan bir veri madenciliği tekniğidir [11]. Oluşturulan her küme kuralların türetilebileceği bir sınıf olarak görülebilir. Kümelenme aynı zamanda taksonominin oluşumunu da kolaylaştırabilir ve kayıtların, benzer olayları birlikte gruplandıran ilgili sınıfların hiyerarşisine girmesini sağlayabilir [17].



Şekil 3. 1 İki öz nitelik içeren kayıtlar için örnek kümeleme analizi

Kümeleme analizi; veri analizi, desen tanıma, görüntü işleme ve pazar araştırması gibi çok sayıda uygulamada yaygın olarak kullanılmaktadır. Özellikle veri madenciliği, istatistik, makine öğrenmesi, biyoloji, pazarlama vb. alanlarda farklı vurgu ve stratejilerle uygulanmaktadır. Örneğin, müşterilerin satın alma alışkanlıkları baz alınarak, farklı müşteri gruplarının keşfedilebilmesi veya diğer bir alandan örnek olarak bitki ve hayvan sınıflarının türetilmesi, benzer fonksiyonlara sahip genlerin kategorize edilmesi ve popülasyonlara özgü özellikleri kavramak amacıyla kullanılabilir [1].

Kümeleme analizinin başarılı bir şekilde tamamlanması ve nihai sonuçların elde edilebilmesi için birkaç temel unsur bulunmaktadır [18]:

- Veri dosyasının belirlenmesi
- Kayıtların ve kayıtlara ait öz niteliklerin belirlenmesi
- Kümeleme yapılacak öz niteliklerin seçimi
- Başlangıç merkezlerin seçimi
- Uzaklık ölçümünün seçimi
- Kümeleme kriterlerinin belirlenmesi
- Kümeleme algoritmaları ve bilgisayar uygulaması
- Küme sayısının seçimi
- Sonuçların yorumlanması

Kümeleme analizi, potansiyel uygulamaların kendi özel gereksinimlerini oluşturduğu bir araştırma alanı olmakla birlikte, veri madenciliğinde kümelemenin tipik gereksinimlerinden bazıları şu şekildedir:

- **Ölçeklenebilirlik:** Birçok kümeleme algoritması az sayıda kayıt içeren veri dosyaları ile iyi çalışmaktadır ancak büyük veri tabanları milyonlarca kayıt içerebilmektedir.
- **Farklı tipteki verileri destekleme:** Birçok kümeleme algoritması nümerik verilerin kümelenmesi için tasarlanmıştır ancak bazı uygulamalar kategorik, ikili, sıralı ya da söz konusu veri tiplerinin karması olan veri tiplerinin kümelemesine de ihtiyaç duymaktadır.
- **Farklı büyüklüklerdeki kümeleri belirleyebilme:** Birçok kümeleme algoritması Öklid ve Manhattan uzaklık ölçümlerini temel alarak kümeleri belirlemektedir. Söz konusu uzaklıklar, kümeleme sonucunda benzer boyut ve yoğunluklara sahip küresel kümeler elde etme eğilimindedir.
- **Girdi parametrelerini belirlemeye yönelik alan bilgisi için minimal gereksinimler:** Kümeleme analizinde birçok kümeleme algoritması kullanıcıların belirli parametreleri girmelerine ihtiyaç duymaktadır. Kümeleme sonuçları giriş parametrelerine karşı oldukça hassastır; o nedenle, özellikle yüksek boyutlu veri dosyalarında bu parametrelerin belirlenmesi büyük önem taşımaktadır.
- **Gürültü ile başa çıkabilme:** Veri tabanlarında birçok eksik, bilinmeyen veya hatalı veri bulunabilmektedir. Bu bağlamda, bazı kümeleme algoritmaları söz konusu verilere karşı oldukça hassas olabilir ve bu durum, kümelemenin kalitesinin düşük olmasına neden olmaktadır.
- **Girdi kayıtlarını düzenlemek için duyarsızlık:** Bazı kümeleme algoritmaları girilen verinin sıralamasına karşı oldukça hassastır. Örneğin, aynı veri dosyası farklı sıralama ile gösterildiğinde kümeleme analizi sonucunda büyük ölçüde farklı kümelerin oluşması söz konusu olabilir. Bu anlamda algoritmaların giriş sıralarına karşı duyarsız olarak geliştirilmesi önemlidir.
- **Yüksek boyutluluk:** Bir veri tabanı ve/veya veri ambarı birçok boyut/öz nitelik içerebilir. Bu bağlamda birçok kümeleme algoritması iki ya da üç boyutlu veriler

için iyi çalışmaktadır. Özellikle yüksek boyutlu verilerin çok seyrek ve oldukça çarpık olabileceğini göz önünde bulundurarak, kayıtların yüksek boyutlu uzayda kümelenmesi zordur.

- **Kısıtlamaya dayalı kümeleme:** Uygulamalarda kümeleme analizinin çeşitli türlerde kısıtlamalar altında gerçekleştirilmesine ihtiyaç duyulabilmektedir.
- **Yorumlanabilirlik ve kullanışlılık:** Kullanıcılar kümeleme sonuçlarının yorumlanabilir, anlaşılır ve kullanılabilir olmasını beklemektedir. Bu bağlamda bir kümeleme yöntemi seçiminin uygulamaya nasıl katkı sağlayabileceğini araştırmak önemlidir.

3.2 Uzaklık Ölçüleri

Kümeleme sonucunu doğrudan etkileyen faktörlerden biri de kayıtlar arasındaki uzaklık ölçümüdür. Uzaklıklar genel olarak, iki veri kaydı arasındaki benzerlikleri ve/veya farklılıkları ölçmek amacıyla kullanılır. Kümelerde bulunan kayıtların arasındaki benzerlik arttıkça, belirli bir kaydın belirli bir kümeye ait olma olasılığı artar. Uzaklık ölçüleri için kullanılacak veri yapısı aşağıda tanımlanmıştır (Çizelge 3.1):

Çizelge 3. 1 Uzaklık ölçüleri için veri yapısı [15]

Kayıt	Öz Nitelik				
	x_1	x_2	.	.	x_m
1	x_{11}	x_{12}	.	.	x_{1m}
.	.	.	.	.	.
i	x_{i1}	x_{i2}	.	.	x_{im}
j	x_{j1}	x_{j2}	.	.	x_{jm}
.	.	.	.	.	.
.	.	.	.	.	.
n	x_{n1}	x_{n2}	.	.	x_{nm}

Kümeleme için kullanılan uzaklık ölçümlerinden bazıları aşağıdaki alt başlıklarda tanımlanmıştır [19,20]:

3.2.1 Öklid Uzaklık Ölçüsü

Öklid (Euclidean) uzaklık ölçüsü, uzaklık ölçüleri arasında en yaygın olarak kullanılan ölçülerden biridir. En önemli özelliklerinden biri; iki kayıt arasındaki mesafenin, analize yeni kayıtların eklenmesiyle etkilenmemesidir. Boyutları benzer ölçeklere sahip kayıtlarda kullanılmak için iyi bir uygulamadır [21].

İki kayıt arasındaki Öklid mesafesi, karşılık gelen değerler arasındaki farklarının karelerinin toplamının karekökünü hesaplamayı içerir. Karekökün derecesi iki olduğunda Öklid uzaklık ölçüsü formülü aşağıdaki gibidir:

$$d_{ij} = \sqrt{\sum_{k=1}^{m} \left(x_{ik} - x_{jk} \right)^2} \tag{3.1}$$

Tez çalışmasında uygulanacak kümeleme analizinde Öklid uzaklık ölçüsü kullanılmıştır.

3.2.2 Kare Öklid Uzaklık Ölçüsü

Kare Öklid (Squared Euclidean) uzaklık ölçümü, iki kayıt arasında kayıtlara karşılık gelen değerlerin farklarının karelerinin toplamını hesaplamayı içerir. Öklid uzaklık ölçüsü ile aynı denklemi kullanmakla birlikte, Öklid uzaklığının karesidir. Kare Öklid uzaklık ölçüsü formülü aşağıdaki gibidir:

$$d_{ij} = \sum_{k=1}^{m} \left(x_{ik} - x_{jk} \right)^2 \tag{3.2}$$

Kare Öklid uzaklık ölçüsü birbirinden daha fazla ayrı olan kayıtlar arasındaki mesafeyi büyütür.

3.2.3 Manhattan Uzaklık Ölçüsü

City-Block uzaklık ölçüsü, mutlak değer uzaklığı veya L_1 uzaklığı olarak da bilinir. Manhattan uzaklık ölçüsü, bir üçgenin hipotenüs olmayan kenarları boyunca bir rota

izleyen bir uzaklık hesaplamasını içerir. Kesikli sayısal veriler için daha fazla önerilir. Manhattan uzaklık ölçüsü formülü aşağıdaki gibidir:

$$d_{ij} = \sum_{k=1}^{m} \left| x_{ik} - x_{jk} \right| \tag{3.3}$$

3.2.4 Mahalanobis Uzaklık Ölçüsü

Mahalanobis uzaklık ölçüsü, çok değişkenli veri dosyasındaki bir kaydın verinin merkezine olan uzaklığıdır. Mahalanobis uzaklığı öz nitelikler arasındaki ilişkileri (varyans/kovaryans) dikkate alır ve kovaryans matrisi kullanıldığı için öz nitelikler arasındaki korelasyonları da dikkate alır.

$$d_{ij}{}^{2} = \left[\left(\mu_i - \mu_j \right)' S^{-1} \left(\mu_i - \mu_j \right) \right] \tag{3.4}$$

Burada; μ_i $i.$kümenin ortalama vektörü ve S^{-1} kovaryans matrisin tersidir.

3.2.5 Minkowski Uzaklık Ölçüsü

Minkowski uzaklık ölçüsü, Öklid uzaklık ölçüsünün ve Manhattan uzaklık ölçüsünün bir genelleştirilmesidir. Minkowski uzaklık ölçüsü formülü aşağıdaki gibidir:

$$d_{ij} = \left(\sum \left(\left| x_{ij} - x_{jk} \right| \right)^{1/q} \right)^{q} \tag{3.5}$$

Burada, q pozitif bir tamsayıdır. Formül, q = 2 olduğunda Öklid uzaklık ölçüsünü; $q = 1$ olduğunda ise Manhattan uzaklık ölçüsünü vermektedir. Chebyshev uzaklık ölçüsü ise, $q = \infty$ (limiti alınarak) olan Minkowski uzaklık ölçüsünün bir çeşididir.

3.2.6 Chebyshev Uzaklık Ölçüsü

Tchebyshev uzaklık ölçüsü, maksimum uzaklık değeri veya L_∞ metriği olarak da bilinir. İki vektör arasındaki uzaklığın, herhangi bir koordinat boyutu boyunca farklarının en büyük olduğu bir vektör uzayı üzerinde tanımlanan metriktir. Chebyshev uzaklık ölçüsü hem sıralı hem de sayısal değişkenler için kullanılabilir. İki vektör arasındaki Chebyshev uzaklık ölçüsü formülü aşağıdaki gibidir:

$$d_{ij} = \max \left| x_{ik} - x_{jk} \right| \tag{3.6}$$

3.2.7 Cosine Uzaklık Ölçüsü

Cosine uzaklık ölçüsü, göreceli farklılığı dikkate alan bir tür Pearson ölçüsüdür. Bazı durumlarda, Cosine uzaklık ölçüsü özellikle verilerin normal olarak dağıtılmadığı durumlarda daha iyi sonuçlar verir. A ve B iki vektör olarak göz önüne alındığında, nokta çarpımdan yararlanılarak hesaplanan Cosine uzaklık ölçüsünün formülü aşağıdaki gibidir:

$$Similarity = \cos(\theta) = \frac{AB}{|A||B|} \tag{3.7}$$

Aralarındaki açının kosinüsünü bularak, genellikle m boyutundaki iki vektör arasındaki benzerlik ölçüsüdür ve genellikle metin madenciliğinde belgeleri karşılaştırmak için kullanılır. Ayrıca, veri madenciliği tekniklerinden olan kümelemede, küme içindeki uyumu ölçmek için kullanılır.

3.2.8 Bray Curties Uzaklık Ölçüsü

Sorensen uzaklık ölçüsü olarak da bilinir. Botanik ve çevre bilimi alanlarında yaygın olarak kullanılan bir normalizasyon yöntemidir. Bray Curties uzaklık ölçüsünde, eğer tüm koordinatlar pozitifse, uzaklık değeri 0 ile 1 arasındadır. Sıfır Bray Curties uzaklığı tam olarak benzeyen koordinatı temsil eder. Her iki kayıt da sıfır koordinatlarındaysa, Bray Curties uzaklığı tanımsızdır. Normalizasyon, aşağıdaki formülde gösterildiği gibi toplamı ile farkın mutlak değeri kullanılarak yapılır:

$$d_{ij} = \frac{\sum_{k=1}^{n} \left| x_{ik} - x_{jk} \right|}{\sum_{k=1}^{n} \left(x_{ik} + x_{jk} \right)} \tag{3.8}$$

3.2.9 Canberra Uzaklık Ölçüsü

Canberra uzaklık ölçüsü, bir çift kaydın koordinatları arasındaki bir kesir farkı serisinin toplamını inceler. Her kesir farkı terimi 0 ile 1 arasında bir değere sahiptir. Eğer koordinatın biri 0 ise, terim diğer değere bakılmaksızın tek olur. Eğer koordinatları sıfırsa, mesafe (0/0 = 0) olarak tanımlanmalıdır. Aksi takdirde, mesafe sonsuz değer

alacaktır. Canberra uzaklık ölçüsü, her iki koordinat da sıfıra yakın olduğunda küçük değişimlere karşı oldukça duyarlıdır. Canberra uzaklık ölçüsü formülü aşağıdaki gibidir:

$$d_{ij} = \sum_{k=1}^{n} \frac{\left| x_{ik} - x_{jk} \right|}{\left(x_{ik} + x_{jk} \right)} \tag{3.9}$$

3.2.10 Korelasyon Uzaklık Ölçüsü

Korelasyon uzaklık ölçüsü kapsamında dört tane uzaklık ölçüsü bulunur:

$$d_{ij} = \left(1 - r_{ij}\right) / 2 \tag{3.10}$$

$$d_{ij} = 1 - \left| r_{ij} \right| \tag{3.11}$$

$$d_{ij} = 1 - r_{ij}^{\,2} \tag{3.12}$$

$$d_{ij} = 1 - r_{ij} \tag{3.13}$$

Söz konusu dört ölçü arasında en fazla dördüncü ölçü ile karşılaşılır. Korelasyon katsayısı -1 ile +1 arasında değişirken; dördüncü eşitlikten elde edilen uzaklık ölçüsü 0 ile 2 arasındadır [15].

3.2.11 Karl Pearson Uzaklık Ölçüsü

Standartlaştırılmış Öklid uzaklık ölçüsü olarak da bilinir. Öklid uzaklık ölçüsündeki farkların $\frac{1}{\sigma_k^2}$ ile düzeltilmesiyle/standartlaştırılmasıyla elde edilen uzaklık ölçüsüdür. Karl Pearson uzaklık ölçüsü formülü aşağıdaki gibidir:

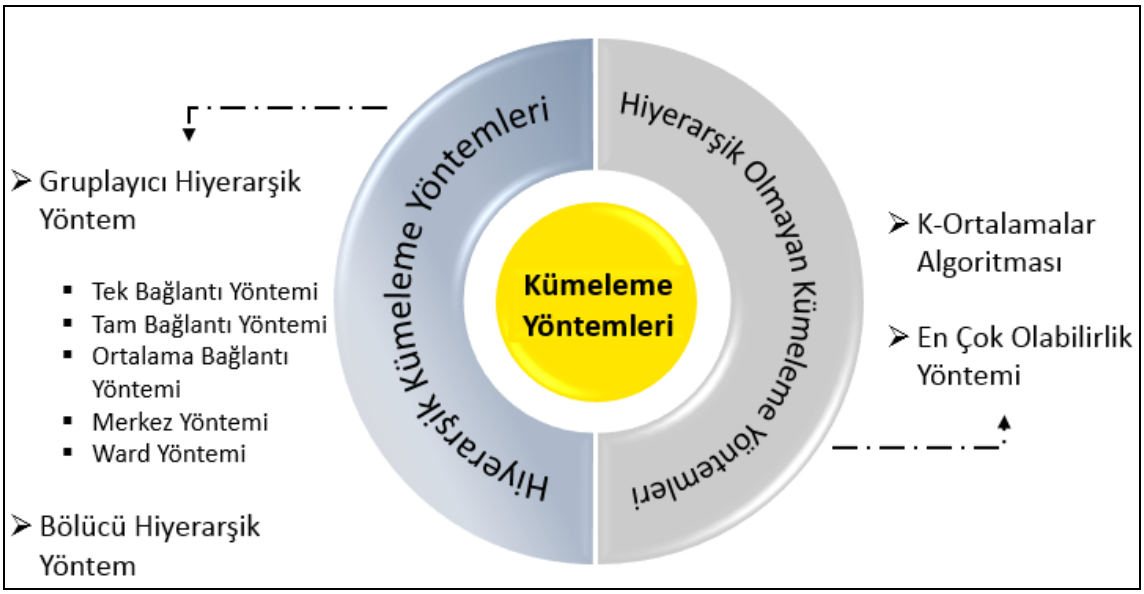
$$d_{ij} = \sqrt{\sum_{k=1}^{m} \frac{1}{\sigma_k^2} \left(x_{ik} - x_{jk} \right)^2} \tag{3.14}$$

Burada, σ_k^2 standart sapmanın karesidir.

3.3 Kümeleme Yöntemleri

Kayıtların kümelenmesinde pek çok yöntem kullanılmaktadır. Kümeleme yöntemleri Şekil 3.2 de gösterildiği gibi hiyerarşik ve hiyerarşik olmayan yöntemler olmak üzere iki gruba ayrılmaktadır [22]. Hiyerarşik kümeleme algoritmaları da uygulanma stratejisine

göre, yukarıdan aşağı veya aşağıdan yukarı olmak üzere kendi içinde ikiye ayrılır. Başlangıçta veri dosyası tek bir küme olarak ele alınırken, son adımda veri dosyasındaki tüm kayıtlar ayrı birer küme halini alır. Diğer yandan; hiyerarşik olmayan yöntemlerin en bilineni ise, K-Ortalamalar kümeleme algoritmasıdır ve bu tez çalışmasında K-Ortalamalar kümeleme algoritması esas alınmıştır. Söz konusu algoritma, hiyerarşik kümeleme algoritmalarından farklı olarak sonuç küme sayısını da girdi olarak almaktadır.



Şekil 3. 2 Kümeleme yöntemleri

3.3.1 Hiyerarşik Kümeleme Yöntemleri

Aşamalı kümeleme yöntemi olarak da bilinir. Gruplayıcı ve bölücü olmak üzere, iki farklı hiyerarşik yöntem bulunmaktadır. Hiyerarşik yöntemlerin ağaç diyagramları ile gösterilen sonuçlarına dendrogram denir.

3.3.1.1 Gruplayıcı Hiyerarşik Yöntem

Gruplayıcı (Agglomerative) hiyerarşik yöntemde her kayıt başlangıçta bir küme olarak kabul edilir. Daha sonra uzaklık ölçümüne göre en yakın iki küme birleştirilir. Böylece her adımda küme sayısı bir azaltılarak uzaklık ölçümü yeniden oluşturulur ve t birim aşamalı olarak sırasıyla $t,(t-1),(t-2),...(t-r),...3,2,1$ küme oluşturulur. Gruplayıcı yöntemler arasından tek bağlantı yöntemi, tam bağlantı yöntemi, ortalama bağlantı yöntemi ve varyans yöntemi alt başlıklar olarak anlatılacaktır.

3.3.1.1.1 Tek Bağlantı Yöntemi

Tek bağlantı yöntemi (single linkage method), uzaklık ölçümünü kullanarak birbirine en yakın kayıtları birleştirmeye dayanmaktadır. Bu yöntemde önce birbirine en yakın iki kayıt bir kümeye yerleştirilir. Daha sonra diğer en yakın uzaklık tespit edilerek kayıt ilk oluşturulan bu kümeye eklenir. Bu işlem tüm kayıtlar bir kümeye yerleştirilinceye kadar devam etmektedir.

X_1, X_2 iki küme olmak üzere, $D(X_1, X_2) = SLINK(X_1, X_2) = \min_{x_1 \in X_1, x_2 \in X_2} d(x_1, x_2)$ olarak tanımlanır.

3.3.1.1.2 Tam Bağlantı Yöntemi

Tam bağlantı yöntemi (complete linkage method), tek bağlantı yönteminin tam tersi bir yöntemdir. Bu yöntem en uzak kayıtların birleştirilmesine dayanmaktadır. İlk adımda her kayıt kendi kümesine aitken son adımda bütün kayıtları içeren tek bir kümeye ulaşılır.

X_1, X_2 iki küme olmak üzere, $D(X_1, X_2) = CLINK(X_1, X_2) = \max_{x_1 \in X_1, x_2 \in X_2} d(x_1, x_2)$ olarak tanımlanır.

3.3.1.1.3 Ortalama Bağlantı Yöntemi

Ortalama bağlantı yöntemi, aşırı uç kayıtlardan başlamaz. Bu yöntemde iki küme arası uzaklık olarak, tek bağlantı ve tam bağlantı yöntemiyle hesaplanan uzaklıkların ortalaması alınmaktadır.

X_1, X_2 iki küme olmak üzere, $D(X_1, X_2) = \frac{SLINK(X_1, X_2) + CLINK(X_1, X_2)}{2}$ olarak tanımlanır.

3.3.1.1.4 Merkez Yöntemi

Merkez yöntemi, kümeyi oluşturan kayıtların ortalamasını esas almaktadır. Eğer bir kümede sadece bir kayıt varsa onun değeri merkez kabul edilmektedir.

3.3.1.1.5 Ward Yöntemi

Ward yöntemi, iki küme arasındaki uzaklığın hesaplanmasında merkezden sapmaları yani varyansı esas alır. Bu yönteme en küçük varyans yöntemi de denilmektedir. Ward yönteminde amaç, kümeler içi kareler toplamını minimize etmektir. Yönteme her birinin içinde tek bir kayıt bulunan k tane küme ile başlanır. Yöntemin ilk basamağında her kayıt bir küme olduğundan, kümeler içi kareler toplamı sıfır olmaktadır. Her aşamada iki alt küme bir sonraki seviyeyi oluşturmak için birleştirilir.

3.3.1.2 Bölücü Hiyerarşik Yöntem

Bölücü (Divistive) hiyerarşik yöntemi, gruplayıcı hiyerarşik yöntemin tam tersidir. Bu yöntemde, tüm kayıtlardan oluşan büyük bir küme ile başlanır. Benzer olmayan kayıtlar ayıklanarak daha küçük kümeler oluşturulur. Her kayıt tek başına bir küme oluşturana kadar işleme devam edilir.

3.3.2 Hiyerarşik Olmayan Kümeleme Yöntemleri

Bazı durumlarda küme sayısı önceden bellidir ve kullanıcı bu küme sayısına göre çözümler üretmektedir. Hiyerarşik olmayan kümeleme yönteminde, küme sayısı belirli bir değer olarak verilebilir. Hiyerarşik olmayan kümeleme yöntemlerinden en sık kullanılan algoritma K-Ortalamalar kümeleme algoritmasıdır. Hiyerarşik olmayan kümeleme yöntemleri, hiyerarşik kümeleme yöntemlerine göre daha büyük veri dosyalarına uygulanabilir.

3.3.2.1 K-Ortalamalar Algoritması

MacQueen tarafından 1967 yılında önerilen K-Ortalamalar kümeleme algoritması (K-Means clustering algorithm), kayıtları önceden belirlenen küme sayısına göre gruplandırmakla işleme başlar. Böylece her biri tek kayıttan oluşan k tane küme ile analize başlanır ve her bir yeni kayıt en yakın olan kümeye atanır. Kümeye yeni bir kayıt eklendikten sonra küme ortalaması yeniden hesaplanır. Bütün bu süreç, tüm kayıtlar ilgili kümelere atanıncaya kadar devam eder. Tüm kayıtlar gruplara atandıktan sonra, atandıkları küme ortalamasından daha yakın küme ortalaması varsa kayıtların yerleri

yeniden değiştirilmektedir. Tez çalışmasında esas alınacak bu algoritma, Bölüm 4 te detaylı olarak anlatılacaktır.

3.3.2.2 En Çok Olabilirlik Yöntemi

En çok olabilirlik yönteminde (maximum likelihood method), her bir kayıt en büyük olabilirlik değerini verecek şekilde daha önceden belirlenen kümelere atanır. En çok olabilirlik yöntemi yaygın olarak kullanılmamaktadır.

BÖLÜM 4

K-ORTALAMALAR KÜMELEME ALGORİTMASI

K-Ortalamalar kümeleme algoritması büyük verileri işlemedeki başarısı ile veri madenciliğinde en yaygın kullanılan algoritmalardan biridir. K-Ortalamalar kümeleme algoritmasının uygulama aşamaları aşağıdaki şekilde özetlenebilir:

Adım 1. Başlangıç küme merkezlerini belirlemek için k tane merkez seçilir. Küme merkezleri rastgele ya da çeşitli metotlar yardımıyla seçilebilir.

Adım 2. Her kaydın seçilen merkezlere uzaklığı hesaplanır. Elde edilen sonuçlara göre tüm kayıtlar k kümeden kendilerine en yakın olan kümeye atanır.

Adım 3. Oluşan kümelerin yeni merkezleri, kümedeki tüm kayıtların ortalaması alınarak yeniden hesaplanır.

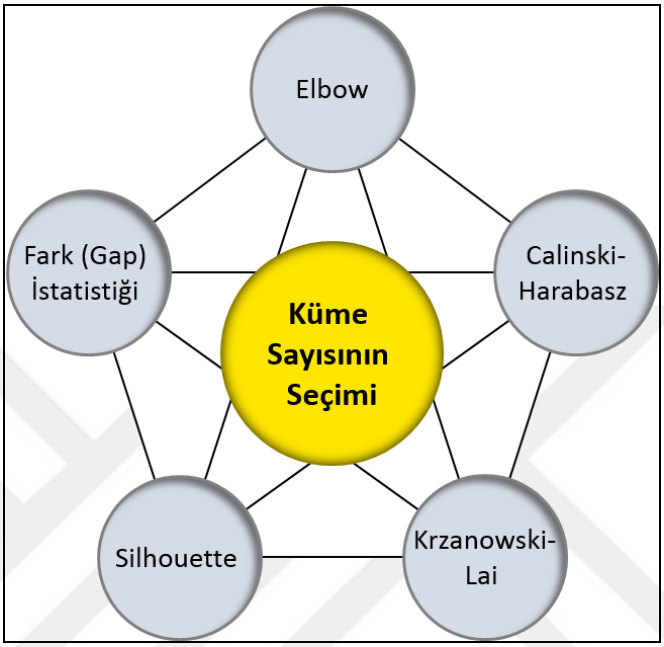
Adım 4. Adım 2 ve Adım 3 küme merkezleri değişmeyene kadar tekrarlanır.

K-Ortalamalar kümeleme algoritmasında, aksi belirtilmedikçe, Öklid uzaklık ölçüsü kullanılır.

4.1 Küme Sayısının Seçimi

Küme sayısının seçimi, kümeleme analizindeki en önemli konulardan biridir. Doğru kümelemenin yapılabilmesi oldukça önemlidir; bu bağlamda, doğru bir kümeleme yapabilmek amacıyla farklı k değerleri için deneme yanılma yönteminin uygulanması gerekmektedir [23, 24, 25]. Küme sayısının seçimi için Şekil 4.1 de gösterildiği gibi birçok metot bulunmaktadır; söz konusu metotlar, bu bölümde detaylı olarak

anlatılacaktır. Bu metotlardan Silhouette ve Calinski-Harabasz indeksleri, aynı zamanda kümelemenin geçerliliğinin (kalitesinin) değerlendirilmesi amacıyla kümeleme değerlendirme kriterleri olarak da kullanılmaktadır [26]. Bu tez çalışmasında Elbow metodu ile belirlenen küme sayısı, söz konusu değerlendirme kriterleri kullanılarak karşılaştırılmıştır.



Şekil 4. 1 Küme sayısının seçimi

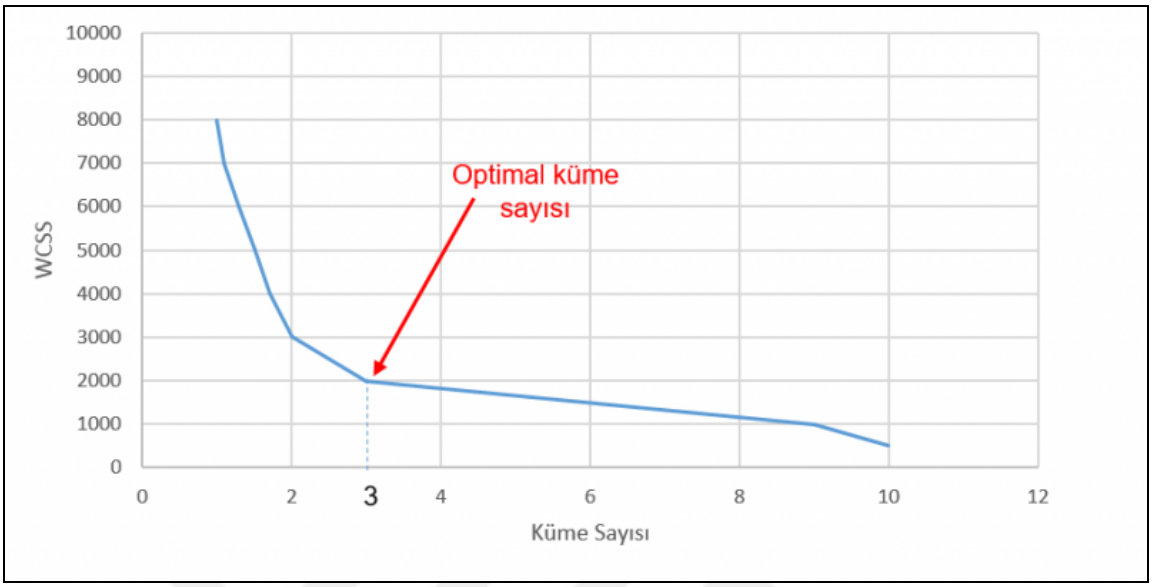
4.1.1 Elbow Metodu

Küme sayısının belirlenmesinde kullanılan en eski ve en yaygın metot olarak bilinir, Dirsek metodu olarak da adlandırılmaktadır. Metodun bilinirliği 1953 yılına kadar uzanmaktadır [27].

Kümelemede, birbirine benzeyen (birbirine yakın olan) kayıtların aynı kümede; birbirine benzemeyen (birbirine uzak olan) kayıtların ise farklı kümelerde olması gibi temel bir mantık bulunmaktadır. Elbow metodundaki bu mantık, kümeler içi kareler toplamı olarak işlemektedir. Kümeleme için belirlenen aralıkta k sayıları tek tek ele alınarak, her bir k değeri için küme içi kareler toplamı hesaplanır.

Şekil 4.1 de küme sayısı ile kümeler içi kareler toplamı arasındaki ilişkiyi gösteren grafik verilmiştir. İyi kurgulanan bir modelde kümeler içi kareler toplamının daha düşük olması beklenmektedir. Şekilde $k=3$ e kadar kümeler içi kareler toplamı önemli

ölçüde bir düşüş göstermektedir ve $k=3$ ten sonra ise grafik yatay olarak seyretmektedir. Bu durum modelin yorumlanabilirliğini azaltmaktadır. Şekil 4.2 den $k=3$ değerinin optimal küme sayısı olduğu anlaşılmaktadır.



Şekil 4. 2 Optimal küme sayısının seçimi

Elbow metodu kullanımında veri dosyasının iyi tanınması gerekmektedir. Algoritmanın hesaplanmasında kümeleme aralıkları kullanıcı tarafından verileceğinden, Elbow metodu küme sayısına karar vermede yardımcı olabilecek etkili bir yöntem olarak kabul edilebilir. Bu tez çalışmasında küme sayısının seçimi için bu metot kullanılmıştır.

4.1.2 Calinski-Harabasz İndeksi

Calinski-Harabasz tarafından 1974 yılında önerilen indekste k kümeye sahip bir kümelemenin geçerliliği (kalitesi) aşağıda verilen formül ile değerlendirilmektedir [28]:

$$CH(k)=\frac{BSS(k)/(k-1)}{WSS(k)/(n-k)} \tag{4.1}$$

Burada,

$$WSS(k)=\frac{1}{2}\sum_{l=1}^{k}\sum_{i,j\in C_l} d_{ij} \tag{4.2}$$

$$BSS(k)=\frac{1}{2}\sum_{l=1}^{k}\sum_{\substack{i\in C_l\\ j\notin C_l}} d_{ij} \tag{4.3}$$

olmak üzere; $WSS(k)$ ve $BSS(k)$ sırasıyla, kümeler içi ve kümeler arası kareler toplamlarıdır. $WSS(k)$ ve $BSS(k)$ değerleri hesaplanırken Kare Öklid uzaklık ölçüsü kullanılmaktadır. Calinski-Harabasz indeksine göre maksimum CH değerine ulaşılan küme sayısı, optimal küme sayısı olarak seçilmektedir.

4.1.3 Krzanowski-Lai İndeksi

Krzanowski ve Lai tarafından 1985 yılında önerilen yöntemde, kümeler içi kareler toplamı değerinin azalışı dikkate alınmıştır ve aşağıdaki eşitlikle tanımlanmıştır [29]:

$$DIFF(k) = (k-1)^{2/p}\ WSS(k-1) - (k)^{2/p}\ WSS(k) \tag{4.4}$$

Eşitliğe bağlı olarak, tanımlanan indeksin formülü aşağıdaki gibidir:

$$KL(k) = \left| \frac{DIFF(k)}{DIFF(k+1)} \right| \tag{4.5}$$

Burada,

$$DIFF(k) = k^{2/p}\ WSS(k) \tag{4.6}$$

olarak hesaplanır. KL değerinin maksimum olduğu durumda küme sayısı, optimal küme sayısı olarak seçilir. Birden fazla eşdeğer maksimum değer çıkması durumunda ise söz konusu değerler kendi içlerinde değerlendirilmelidir. Örneğin, $k=5$ küme sayısına karşılık gelen maksimum KL değeri ile birlikte, küme sayısının artırılması ile birlikte $k=10$ küme sayısı için de eşdeğer bir maksimum değer elde edilebilir. Burada elde edilen maksimum değer, genellikle çok küçük bir $DIFF(k)$ ile daha küçük bir $DIFF(k+1)$ bölümünün sonucu olacaktır ve bu gibi durumlar genelde göz ardı edilir.

4.1.4 Silhouette İndeksi

Kaufman ve Rousseeuw tarafından 1990 yılında önerilen indeks ile her bir kaydın bulunduğu kümeye uygunluğu tanımlanmıştır. $sil(i)$ değeri, $a(i)$ ve $b(i)$ değerleri dikkate alınarak elde edilir [30]:

$$sil(i) = \begin{cases} 1 - a(i)/b(i) & eğer\, a(i) < b(i)\, ise, \\ 0 & eğer\, a(i) = b(i)\, ise, \\ b(i)/a(i) - 1 & eğer\, a(i) > b(i)\, ise. \end{cases} \quad (4.7)$$

Burada $a(i)$; i. kaydın bulunduğu kümedeki tüm kayıtlara olan ortalama uzaklıkları ve $b(i)$; i. kaydın diğer kümelerdeki tüm kayıtlara olan ortalama uzaklıkların minimumunu göstermektedir. İndeksin formül olarak gösterimi aşağıdaki gibidir:

$$sil(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (4.8)$$

$sil(i)$ değeri -1 ile +1 arasında değerler almaktadır. $sil(i)$ değerinin 1'e yaklaşması, i. kaydın atanmış olduğu kümeye uygun olduğunu; $sil(i)$ değerinin 0'a yaklaşması veya negatif değer alması ise i. kaydın atanmış olduğu kümeye uygun olmadığını göstermektedir. $sil(i)$ değerinin 0 olması durumunda ise $a(i)$ ve $b(i)$ değerleri eşittir ve bu durumda i. kaydın hangi kümeye uygun olduğu belirlenememektedir.

Tüm kümelemenin geçerliliği (kalitesi), tüm kayıtlar için $sil(i)$ değeri toplamının ortalama değeri alınarak hesaplanır ve maksimum $sil(C)$ değerinin elde edildiği küme sayısı, uygun küme sayısı olarak seçilir:

$$sil(C) = \frac{1}{n} \sum_{s_i \in S} sil(i) \quad (4.9)$$

4.1.5 Fark İstatistiği

Stanford Üniversitesi öğrencileri Robert Tibshirani, Guenther Walther ve Trevor Hastie tarafından 2001 yılında önerilen bu metot fark istatiği (gap statistic) olarak adlandırılmaktadır ve her kümeleme yöntemi için uygulanabilmektedir [31]. Fark istatistiğinde, uzaklık ölçütü olarak Öklid uzaklık ölçütü kullanılmaktadır.

i ve i' kayıtları arasındaki uzaklık $d_{ii'}$ olarak gösterilsin. k kümeye ayrılan veri dosyası için kümeler $C_1, C_2, \ldots, C_k$ olarak ve herhangi bir C_r kümesindeki kayıtların sayısı $n_r = |C_r|$ olarak verilsin. C_r kümesindeki tüm kayıtların ikili olarak uzaklıklarının toplamı formülü aşağıdaki gibidir:

$$D_r = \sum_{i,i' \in C_r} d_{ii'} \tag{4.10}$$

Küme içindeki tüm kayıtlar için toplam formülü ise aşağıdaki gibidir:

$$W_k = \sum_{r=1}^{k} \frac{1}{2n_r} D_r \tag{4.11}$$

En uygun küme sayısı, $\log(W_k)$'nın referans eğrisinin en altında kaldığı k değeridir ve aşağıdaki formülden yararlanılarak hesaplanır. $Gap_n(k)$ değerini maksimum yapan k değeri, uygun küme sayısı olarak seçilir:

$$Gap_n(k) = E_n^* \{\log(W_k)\} - \log(W_k) \tag{4.12}$$

Burada, E_n^* referans dağılımdan n boyutlu bir örneklemin altındaki beklenen değerdir ve $\log(W_k)$'nın yaklaşık olarak aşağıdaki ifadeye denk gelmesi beklenmektedir:

$$\log(pn/12) - (2/p)\log(k) + sabit \tag{4.13}$$

Küme sayısının seçiminde kullanılan diğer başlıca metotlar ise aşağıdaki şekildedir:

- Everitt tarafından 1974 yılında önerilen ilk yaklaşımlardan en çok bilineni olan bu metodun daha çok küçük örneklemli çalışmalarda kullanılması önerilmektedir. Basit ve her tipteki veri dosyasına uygulanabilen metodun formülü aşağıdaki şekildedir:

  $$k = \sqrt{\frac{n}{2}} \tag{4.14}$$

- Mariott tarafından 1971 yılında önerilen metot, M harfi ile gösterilmektedir ve formülü aşağıdaki gibidir:

  $$M = k^2 |W| \tag{4.15}$$

  Burada, W kümeler içi kareler toplamıdır ve formülü aşağıdaki gibidir:

  $$W = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)(x_{ij} - \bar{x}_j)' \tag{4.16}$$

Burada, n_j : j. kümedeki kayıt sayısını; k küme sayısını; x_{ij} : j. kümedeki i. kaydın değerlerini ve $\bar{x}_j$: j. kümenin ortalama vektörünü göstermektedir. M değerini minimum yapan k değeri, uygun küme sayısı olarak seçilir.

- Bilgi kriteri yaklaşımı, farklı sayıda parametreye sahip modeller arasından seçim yapmak için kullanılır. Her parametre için bir ceza terimi getirerek, ek parametrelerden kaynaklanan olasılığın artışının dengelenmesi amaçlanmaktadır. Akaike tarafından 1974 yılında önerilen Akaike Bilgi Kriteri (Akaike Information Criterion), modelin örneklem büyüklüğünü ve değişken sayısını dikkate alıp bazı düzenlemeler yaparak elde edilen değer sayesinde farklı modeller arasında en uygununu seçmeye yarayan bir kriterdir. Formülü aşağıdaki gibidir:

$$AIC = \log\left(\frac{KKT}{n}\right) + \frac{2(K+1)}{n} \tag{4.17}$$

Akaike ve Schwarz tarafından 1978 yılında önerilen Bayes Bilgi Kriteri (Bayes formülü ise aşağıdaki gibidir:

$$BIC = \log\left(\frac{KKT}{n}\right) + \log(N)\frac{(K+1)}{n} \tag{4.18}$$

4.2 Başlangıç Merkezlerin Seçimi

Başlangıç merkezlerin seçimi, kümeleme analizindeki kritik konulardan bir diğeridir ve Şekil 4.3 te gösterildiği gibi bu konuda birçok metot bulunmaktadır [32]. Bu tez çalışmasında, başlangıç merkezlerin seçimi için farklı metotlar ele alınmıştır ve bu yöntemlerin kümeleme analizi sonucuna iterasyon sayısı ve süre gibi konularda yaptığı etkisi belirlenmiştir. Söz konusu metotlar bu bölümde detaylı olarak anlatılacaktır.

Şekil 4. 3 Başlangıç merkezlerin seçimi

4.2.1 Maximin Metodu

Maximin (Min-Max) metodu yaygın olarak kullanılan metotlardan biridir.

Uygulama aşamaları aşağıdaki şekildedir [33]:

Adım 1. Kayıtlar arasından keyfi olarak bir ilk merkez c_1 seçilir.

Adım 2. Aşağıdaki eşitliği sağlayan x_j kaydı bir sonraki merkez $c_i \left(i \in \{2,3,\ldots,\ k\}\right)$ olarak seçilir.

$$j = \arg \max_{j \in \{1,2,\ldots N\}} \left(\min_{k \in \{1,2,\ldots,i-1\}} \left\| x_j - c_k \right\|_2^2 \right) \tag{4.19}$$

Adım 3. Adım 2 $\left(\mathrm{k}-1\right)$ kez tekrarlanır.

4.2.2 Katsavounidis Metodu

Katsavounidis metodu, maximin metodu ile çok benzerdir [34]. Farklı olarak, ilk merkezin seçimi kayıtlar arasından en büyük L_2 norma sahip olanının seçilmesidir.

Uygulama aşamaları aşağıdaki şekildedir:

Adım 1. Kayıtlar arasından en büyük L_2 norma sahip kayıt ilk merkez c_1 olarak seçilir.

Adım 2. Aşağıdaki eşitliği sağlayan x_j kaydı bir sonraki merkez $c_i \left(i \in \{2, 3, ..., k\}\right)$ olarak seçilir.

$$j = \arg \max_{j \in \{1,2,...N\}} \left\| x_j \right\|_2^2 \tag{4.20}$$

Adım 3. Adım 2 $(k-1)$ kez tekrarlanır.

4.2.3 Temel Bileşenler Analizi Metodu

Temel Bileşenler Analizi (PCA-Part) 1991 yılında Karl Pearson tarafından önerilmiştir ve hiyerarşik bir yaklaşım kullanır. Tüm X veri dosyasını içeren bir başlangıç küme ile başlayarak, en büyük kare hata toplamına sahip olan kümeyi dikkate alarak iki alt kümeye böler [33, 35, 36, 37].

Uygulama aşamaları aşağıdaki şekildedir:

Adım 1. C_i en büyük kare hataya sahip kümedir ve c_i bu kümenin merkezidir. İlk iterasyon için, $C_1 = X$ ve $c_1 = x$ olarak kabul edilir.

Adım 2. p, C_i nin v_i temel özvektörü üzerindeki c_i nin izdüşümüdür; örnek olarak $p = c_i \bullet v_i$ dir; burada, "$\bullet$" nokta çarpımı ifade etmektedir.

Adım 3. C_i aşağıdaki kurala göre C_{i1} ve C_{i2} olarak iki alt kümeye bölünür: Her bir x_j eleman C_i için eğer $x_j \bullet v_i \le p$ ise; o halde x_j, C_{i1} e atanır; aksi takdirde, C_{i2} ye atanır.

Adım 4. Adım 1-3 $(k-1)$ kez tekrarlanır.

4.2.4 Varyans Bileşenleri Analizi Metodu

Varyans bileşenleri (Var-Part) metodunda zaman karmaşıklığı her aşamada $O(nd)$ dir ve veri dosyasını k kümeye ayırmak için söz konusu aşamaların $(k-1)$ kez tekrarlanması gerekir. Bu durumda toplam zaman karmaşıklığı $O(ndk)$ dır. K-Ortalamalar kümeleme algoritması için bir iterasyonun zaman karmaşıklığı da $O(ndk)$ dır; buradan, Var-Part metodunda başlangıç merkezlerin seçiminin K-

Ortalamalar kümeleme algoritmasındaki bir iterasyonunun çalıştırılmasına eş değer olduğu görülür [38, 39].

Uygulama aşamaları aşağıdaki şekildedir:

Seçilen C_j kümesine göre;

Adım 1. Her bir öz nitelik için varyans hesaplanır ve en yüksek varyans değerine sahip öz nitelik $\left(d_p\right)$ seçilir.

Adım 2. d_p öz niteliğineki x_i kaydının değeri x_{ip} olarak gösterilsin. μ_{jp} , d_p özelliğindeki C_j nin ortalaması olarak verilsin. C_j aşağıdaki kurala göre C_{j1} ve C_{j2} olarak iki alt kümeye bölünür: Eğer $x_{ip} \le \mu_{jp}$ ise x_{ip} , C_{j1} e atanır; aksi takdirde, C_{j2} ye atanır.

4.2.5 K-Ortalamalar++ Metodu

K-Ortalamalar++ (K-Means++) metodu 2007 yılında David Arthur ve Sergei Vassilvitskii tarafından önerilmiştir. K-Ortalamalar kümeleme algoritmasından farkı, kümelerin ilk merkezlerinin belirlenmesindeki işlemdir. K-Ortalamalar kümeleme algoritmasında ilk merkezler genel olarak rastgele belirlenirken, K-Ortalamalar++ metodunda sadece birinci merkez rastgele belirlenir; geri kalan tüm merkezler birinci merkez referans alınarak ve olasılık hesaplamalarından faydalanılarak bulunur. K-Ortalamalar++ metodunun uygulanmasında zaman karmaşıklığı $O\left(\log k\right)$ dır [40, 41].

K-Ortalamalar++ metodunda $D\left(x\right)$ bir kaydın seçilmiş mevcut en yakın merkeze/ merkezlere olan uzaklıklarının en küçük değerini göstermektedir [42].

Uygulama aşamaları aşağıdaki şekildedir:

Adım 1. Kayıtlar arasından keyfi olarak bir ilk merkez c_1 seçilir.

Adım 2. Bir sonraki merkez $x \in X$ için aşağıdaki eşitliği olasılık hesaplaması yardımıyla seçilir:

$$\frac{D(X)^2}{\sum_{x \in X} D(X)^2} \tag{4.21}$$

Adım 3. Adım 2 k tane merkez oluşturulana kadar tekrarlanır.

4.2.6 Forgy Metodu

1965 yılında Forgy tarafından önerilen metot, her kaydı k tane kümeden birine rastgele atar. Merkezler daha sonra bu ilk kümelerin merkezleri hesaplanarak elde edilir. Bu metodun kuramsal bir temeli yoktur çünkü bu tür rastgele kümeler içsel homojenliğe sahip değildir.

4.2.7 Jancey Metodu

1966 yılında Jancey tarafından önerilen metot, her merkeze veri uzayı içinde rastgele oluşturulmuş bir kayıt atar. Veri dosyası uzayı doldurmadıkça, merkezlerin bazıları herhangi bir kayıttan oldukça uzak olabilir ve bu da boş kümelerin oluşmasına yol açabilir.

4.2.8 MacQueen Metodu

1967 yılında MacQueen tarafından iki farklı metot önerilmiştir. Birinci metot, veri dosyasındaki ilk k kaydın merkezler olarak alınmasıdır. İkinci metot ise, merkezleri veri dosyasından rastgele seçmektir ancak bu metotta aykırı değerleri seçmekten kaçınmak için herhangi bir kontrol bulunmamaktadır.

4.2.9 Ball-Hall Metodu

Ball-Hall metodu, veri dosyasının ilk merkezini aşağıdaki formül ile belirler:

$$\bar{X} = \frac{1}{N} \sum_{j=1}^{N} x_j \tag{4.22}$$

İlk merkez belirlendikten sonra kayıtları rastgele sırayla geçer ve k merkezleri elde edilene kadar önceden seçilmiş merkezlerin dışındaki en az T birimini ise, bir merkez olarak bir kayıt alır. Uzaklık eşik değeri olan T'nin amacı, kayıtların iyi ayrılmasını sağlamaktır. Bununla birlikte, T için uygun bir değere karar vermek zordur.

4.2.10 Basit Küme Arama Metodu

Basit küme arama metodu, Ball-Hall metodu ile aynıdır; tek fark olarak veri dosyasındaki ilk kayıt ilk merkez olarak alınır.

4.2.11 Spath Metodu

Spath metodu, Forgy metodu ile benzerdir; istisna olarak, kayıtlar kümelere döngüsel olarak, yani j-inci $\left(j \in \{1, 2, ..., N\} \right)$ kayıt $\left(j - 1 \left(mod\ k \right) + 1 \right)$ inci kümeye atanır. Forgy metodunun aksine bu metot daha duyarlıdır.

4.2.12 Al-Daoud Metodu

Al-Daoud yoğunluğa dayalı metot ilk olarak veri uzayını M ayrık hiperküplere ayırır. Daha sonra, küme merkezlerini elde etmek için hiperküpten CN_m / N kayıtlarını rastgele seçer. Burada N_m, m hiperküpündeki kayıtların sayısıdır. Bu metodun iki temel dezavantajı vardır. Birincisi, M için uygun bir değere karar vermek zordur. İkincisi ise, metot $O(2BD)$ depolama karmaşıklığına sahiptir; burada B, her bir öz nitelik için tahsis edilen bitlerin sayısıdır.

4.2.13 Lu Metodu

Lu metodu, iki fazlı bir piramidal yaklaşım kullanır. Her bir kaydın öz nitelikleri ilk olarak 2^Q-seviyesi niceleme kullanılarak tamsayı olarak kodlanır. Burada Q bir çözünürlük parametresidir. Bu kayıtların piramidin 0'ıncı seviyesinde olduğu kabul edilir. Aşağıdan yukarıya fazda; 0 seviyesinden başlayarak, en az $20K$ kayıt elde edilene kadar $k+1$ seviyesinde ağırlıklandırılmış kayıtlar elde etmek için komşu kayıtların k $\left(k \in \{0, 1, ...\} \right)$ seviyesinde ortalaması alınır. En yüksek seviyedeki kayıtlar, en büyük ağırlıklara sahip k kayıtları ile başlatılan K-Ortalamalar kümeleme algoritması kullanılarak düzenlenir. Yukarıdan aşağıya fazda; en yüksek seviyeden başlayarak, $k+1$ seviyesindeki merkezler k seviyesine yansıtılır ve daha sonra k-ıncı seviye kümelemeyi başlatmak için kullanılır. Yukarıdan aşağıya fazı, seviye 0'a ulaşıldığında sona erer. Bu seviyedeki

merkezler daha sonra nihai merkezleri elde etmek için ters nicelenir. Bu metodun performansı artan boyutsallık ile azalır.

4.2.14 Onoda Metodu

Onoda metodu, ilk olarak X in k bağımsız bileşenlerini hesaplar ve daha sonra i-inci $\left(i \in \{1, 2, ..., k\}\right)$ merkezini i-inci bağımsız bileşeninden en az kosinüs uzaklığına sahip olan kayıt olarak seçer.

4.2.15 Hartigan Metodu

Hartigan metodunda ilk olarak kayıtlar $\bar{X}$ e olan uzaklıklarına göre sıralanır ve daha sonra i-inci $\left(i \in \{1, 2, ..., k\}\right)$ merkez $\left(1+(i-1)N/k\right)$ -ıncı kayıt olarak seçilir. Bu metot, MacQueen'in ilk metoduna göre daha gelişmiştir. Hesaplama maliyeti $O(\mathrm{Nlog}N)$ dir.

4.2.16 Al-Daoud Varyansa Dayanan Metot

Al-Daoud varyansa dayanan metodu, ilk olarak en büyük varyansa sahip öz nitelik üzerindeki kayıtları sıralar (metot diğer öz nitelikleri göz ardı eder) ve ardından aynı boyuttaki k tane gruba ayırır. Merkezler daha sonra bu grupların medyanlarına karşılık gelen kayıtlar olarak seçilir.

4.2.17 Redmond ve Heneghan Metodu

Redmond ve Heneghan metodu ilk olarak yoğunluk tahmini gerçekleştirmek için kayıtların bir karar ağacını oluşturur ve daha sonra yoğun nüfuslu yaprak kovalarından k merkezlerini seçmek için değiştirilmiş bir maximin metodu kullanır. Bu metodun hesaplama maliyeti $O(\mathrm{Nlog}N)$ dir ve karar ağacı yapısının karmaşıklığından yararlanılarak belirlenebilir.

4.2.18 ROBIN Metodu

ROBIN (ROBust INitialization) metodu, merkezler olarak aykırı kayıtları seçmekten kaçınmak için yerel bir dış faktör kullanır. Metot ilk olarak kayıtların daha önce seçilen merkezlere minimum uzaklıklarını azalan olarak sıralar. Daha sonra, sıralanan kayıtları çaprazlar ve i-inci merkez olarak 1'e yakın bir dış faktör değerine sahip olan ilk kaydı seçer. Hesaplama maliyeti $O(\mathrm{N}\log N)$ dir.

4.2.19 Astrahan Metodu

Astrahan metodu d_1 ve d_2 olarak iki uzaklık eşik değeri kullanır. Metot ilk olarak, her bir kaydın yoğunluğunu d_1 in bir uzaklığı içindeki kayıtların sayısı olarak hesaplar. Kayıtlar, yoğunluklarına göre azalan sıraya göre sıralanır ve en yüksek yoğunluk noktası ilk merkez olarak seçilir. Sonraki merkezler, her yeni merkezin daha önce seçilen merkezlerden en azından d_2 lik bir uzaklıkta olması koşuluyla, azalan yoğunluk sırasına göre seçilir. Bu prosedür daha fazla merkez seçilmeden devam eder. Son olarak, eğer k merkezlerinden daha fazlası seçilirse, merkezlerin sadece k tanesi kalmayıncaya kadar gruplamak için hiyerarşik kümeleme kullanılır. Metodun temel problemi, d_1 ve d_2 değerlerine çok duyarlı olmasıdır. Örneğin; d_1 çok küçükse, sıfır yoğunluğa sahip izole edilmiş birçok kayıt olabilir; eğer çok büyükse, birkaç merkez tüm veri kümesini kapsayacaktır.

4.2.20 Kaufman-Rousseeuw Metodu

Kaufman-Rousseeuw metodunda ilk merkez $\bar{X}$ olarak alınır ve i-inci $\left(i \in \{2, 3, \ldots, k\}\right)$ merkez hatayı en çok azaltan kayıt olarak seçilir. Kayıtlar arasındaki ikili uzaklıkların her iterasyonda hesaplanması gerektiğinden, zaman karmaşıklığı $O\left(N^2\right)$ dir.

Bu tez çalışmasında, küme sayısının seçimi için Elbow, Calinski-Harabasz, Krzanowski-Lai ve Silhouette metotları; başlangıç merkezlerin seçimi için Maximin, Katsavounidis, PCA-Part, Var-Part ve K-Means++ metotları kullanılacaktır.

BÖLÜM 5

K - ORTALAMALAR ALGORİTMASINA DAYALI KÜMELEME ANALİZİ SİSTEMİ VE PERAKENDECİLİK SEKTÖRÜNDE UYGULAMASI

Perakendecilik sektöründeki müşteri davranışlarının dünya genelinde değişmiş olmasıyla gelinen noktada müşteriler, birçok kanal aracılığıyla her türlü bilgiye ve her türlü alışveriş imkânına sahiptir. Bu bağlamda, tek kanaldan faaliyet gösteren perakende firmaları geleneksel alışveriş tehdidi altındadır. Bu tez çalışmasında; Türkiye'nin 81 ilinde var olan mağazaları ve buna paralel olarak sunduğu sanal alışveriş imkânıyla, dünya perakende sektörü listesine girmiş bir firma olan Migros Ticaret A.Ş. verileri kullanılmıştır.

Müşteriler, rakip firmadan ürün veya hizmet satın almakla, bir firma üzerinde gücünü kullanabilmektedir. Hizmeti veya ürünü sağlayabilecek az sayıda müşteri ve rakip varsa bu güç oldukça büyük olacaktır. Firmaların, müşterilerin davranışlarını anlamak ve izlemek için gerekli bilgi ve süreçlere ulaşma konusundaki son gelişmeler, müşteri ilişkileri yönetimi (CRM) olarak adlandırılmıştır [43]. Veri tabanlarında yer alan müşteri detayları ve aktiviteleri gibi potansiyel bilgileri kullanmak için analiz teknikleri ve gelişmiş yazılım araçları bulunmaktadır. Firmalar, gündelik çözümler yerine daha sistematik ve kalıcı çözümler uygulamak amacıyla söz konusu teknolojileri kullanmaya başladığında, veri madenciliği süreci başlamış olacaktır [44]. Veri madenciliğini en başarılı şekilde uygulayan alanlar; perakendecilik, bankacılık ve sigortacılıktır [45, 46].

Büyük sistemlerde veri madenciliği için pazarda rekabet eden SPSS, Stata, R, SAS, Weka ve Statistica Data Miner gibi birçok paket program vardır. Özellikle SAS ve SPSS tüm

sistemlerde en yaygın kullanılan paketlerdir [16]. Programların genelinde kümeleme için çeşitli algoritmalar uygulanmıştır ve hemen hemen tüm paketler, tez çalışmasında ele alınan K-Ortalamalar algoritmasını içermektedir [7].

Bu tez çalışmasında hazır paket programlarını kullanmak yerine, K-Ortalamalar algoritmasına dayalı yeni bir kümeleme analiz sistemi geliştirilmiştir ve çalışmada elde edilen tüm sonuçlar, geliştirilen bu yeni analiz sistemi üzerinden sağlanmıştır. Sistemin tasarımı ve çalışma şekli bu bölümde detaylı olarak anlatılacaktır. Sistemin çalışma şekli ayrıca, İris veri dosyası üzerinde yapılan uygulama ile gösterilmiştir.

Kümeleme analizinde kullanılan bilgisayar; Intel Core i7 işlemci, 8 GB bellek, Windows 10 (64 bit) işletim sistemi ve 8 çekirdeğe sahip dizüstü bir bilgisayardır.

5.1 Veri Hazırlama İşlemleri

Tez çalışmasında kullanılan veri dosyası, 814 sadık müşterinin 2013 yılına ait satın almalarını içermektedir. Veri dosyasında bulunan 1.015.590 kaydın her biri; müşteri hesap numarası, ürün, marka, kategori, zaman (sezon), farklı gün (alışverişin bir ayda kaç farklı günde yapıldığı), toplam satış (adet) ve ağırlık (KG) öz niteliklerine sahiptir. Veri dosyasında 1547 farklı marka, 18.208 farklı ürün ve 9 farklı ana kategori bulunmaktadır.

Kümeleme analizinde kullanılacak veri dosyasında; veri temizleme ve veri dönüştürme hazırlık işlemleri uygulanmıştır. İşlemlerin tümü MS-SQL veri tabanında bulunan "MRKTBL" tablosunda gerçekleştirilmiştir. Başlangıçta bulunan 1.048.575 kayıttan;

- Toplam satış özelliği için eksi (-) ve sıfır (0) değer içeren 762 kayıt,
- Ağırlık özelliği için sıfır değer içeren 31.153 kayıt,
- Kategori özelliği için tanımsız olarak bulunan ve sıfır değeri alan 1070 kayıt

silinerek, veri dosyasındaki kayıtların yaklaşık %3 ü temizlenmiştir ve işlem sonucunda kümeleme analizi için kullanılacak olan 1.015.590 kayıt elde edilmiştir.

Bir diğer hazırlık işlemi olarak, veri dönüştürme uygulanmıştır. Veri dosyasında bulunan toplam satış, ağırlık ve farklı gün öz nitelikleri için min-max normalleştirmesi

uygulanarak tüm değerler 0 ile 1 aralığına getirilmiştir. Zaman ve kategori öz nitelikleri için ise kukla (dummy) değişkenler atanmıştır.

5.1.1 Java Programlama Dili

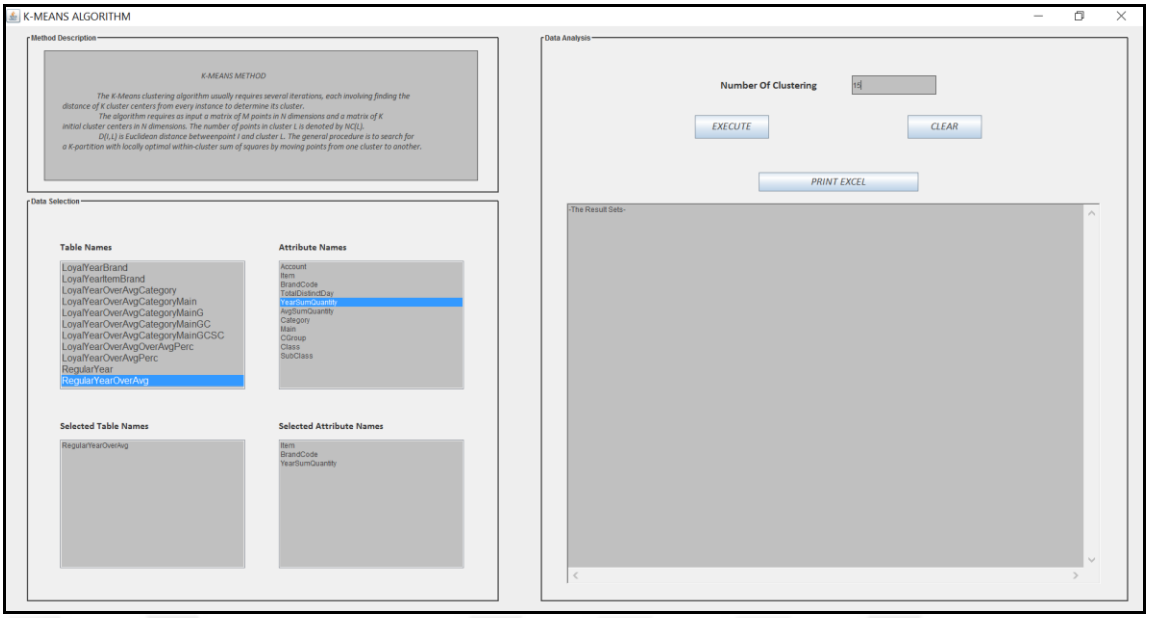
Kümeleme analizi sisteminin geliştirilmesinde Java dili kullanılmıştır. Java seçiminin nedeni olarak, Java da yazılan programların yeniden derlenmeye gerek kalmadan hemen hemen her bilgisayarda çalıştırılabilmesi gösterilebilir. İkinci bir neden ise, kullanıcı ile program arasında köprü görevi yapan, kullanıcının programa veri girmesini ve programın ürettiği verileri kullanıcıya iletmesini sağlayan (programla kullanıcı arasındaki iletişimi sağlayan) arayüz kısmıdır. Analiz sisteminde arayüz tasarımı için Swing kullanılmıştır. Swing, günümüz uygulamalarında kullanıcı arabirimini (GUI – Graphical User Interface) tanımlayan sınıflar koleksiyonudur.

5.1.2 MS-SQL Veri Tabanı

Java ile ilişkisel bir veri tabanına erişmek ve veri tabanında bulunan veriler ile işlem yapmak için programlama ve veri tabanı arasına JDBC sürücüsü eklenir. JDBC ile herhangi bir veri tabanına bağlanarak, sorgular aracılığıyla verilere erişilebilir [47]. Tez çalışmasında kullanılacak veri dosyasını içeren tablo MS-SQL veri tabanında tutulmaktadır. Çalışma sırasında Java programı üzerinden MS-SQL veri tabanına bağlanılarak, kümeleme analizi için gereken veri dosyasını içeren tabloya erişilmiştir.

5.2 Rastgele Seçilen k Sayısı ile K-Ortalamalar Algoritmasına Dayalı Analiz Sistemi

Tez çalışmasının başlangıç aşamasında aynı veri dosyası kullanılarak, “Brand Loyalty Analysis System Using K-Means Algorithm” isimli, müşterilerin marka bağımlılığını araştıran bir makale hazırlanmıştır. Makale çalışmamızda veri analizi, geliştirilen kümeleme analiz sistemi üzerinde gerçekleştirilmiştir. Ancak makalede küme sayısının seçimi, bir metoda bağlı olmaksızın tahmini olarak gerçekleştirilmiştir. Diğer yandan, başlangıç merkezler ise rastgele seçilmiştir. Makalede veri analizi için kullanılan sistem Şekil 5.1 deki gibidir:



Şekil 5. 1 Rastgele seçilen küme sayısı ile kümeleme analiz sistemi

Analiz sonuçları; genel marka bağımlılığı, ürüne bağlı marka bağımlılığı ve kategorik bazda marka bağımlılığı olarak üç farklı şekilde analiz edilmiştir. Makale 2016 yılında "Journal of Engineering Technology and Applied Sciences" adlı dergide yayınlanmıştır. Makale çalışma sonrasında analiz sistemi, küme sayısının seçimi ve başlangıç merkezlerin seçimi için çeşitli metotlar kullanılarak geliştirilmiştir.

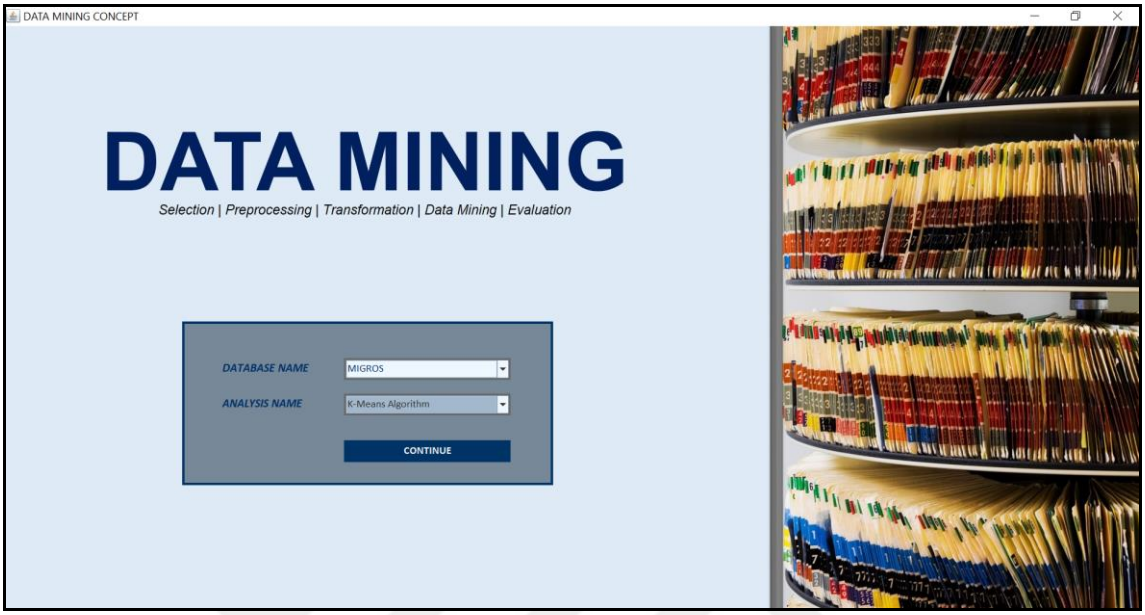
5.3 Sistematik Seçilen k Sayısı ile K-Ortalamalar Algoritmasına Dayalı Analiz Sistemi

Kümeleme analizi için bir sistemin geliştirilmesi, safhalar içinde başarılan bir süreçtir. Tüm bu süreçte, çalışma için olabildiğince iyi bir sistemin oluşturulması amaçlanmıştır. Sistemin geliştirilmesindeki adımlar şu şekilde özetlenebilir:

- Amacın tanımlanması,
- Amaca yönelik gereksinimlerin belirlenmesi,
- Sistem ihtiyaçlarının analiz edilmesi,
- Tüm bu bilgilerden faydalanılarak sistemin tasarlanması,
- Tasarımı yapılan sistemin yazılımının geliştirilmesi,
- Sistemin ihtiyaçları karşılayıp karşılamadığının test edilmesi,
- Sistemin uygulamaya alınması ve izlenmesi.

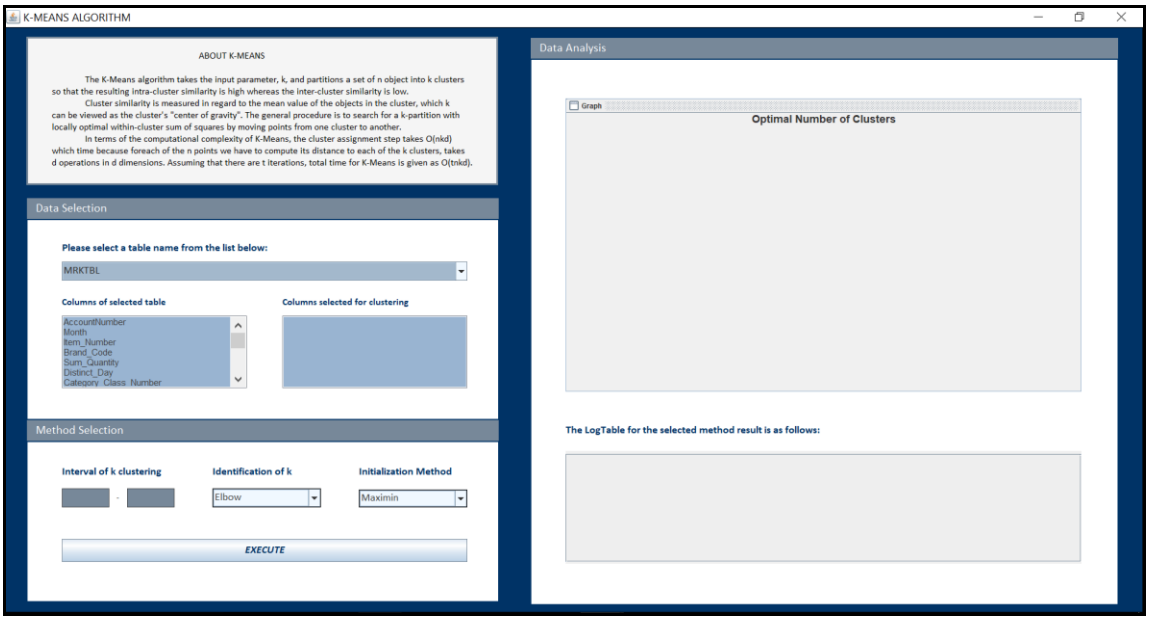
Bu adımlar arasında en kritik önemi taşıyan adım, amacın belirlenmesidir. Hedef yapının doğru şekilde belirlenmesi, diğer tüm adımlar için kolaylık sağlayacaktır.

Kümeleme analizi için iyileştirilen analiz sisteminde arayüzler aşağıdaki gibidir:



Şekil 5. 2 Kullanıcı giriş arayüzü

Şekil 5. 2 deki arayüz aracılığıyla; kümeleme analizi yapılarak veri tabanı ve kümeleme analizi yöntemi seçilir. Bu tez çalışmasında kullanılacak olan veri tabanı "MIGROS" veri tabanıdır ve analiz için kullanılacak kümeleme algoritması "K-Means Algorithm" dir. Giriş arayüzünde, ilgili seçimlerin yapılmasıyla birlikte açılan yeni arayüz şu şekildedir:



Şekil 5. 3 Kümeleme analizi arayüzü

Şekil 5. 3 teki arayüz aracılığıyla kümelemeye dair veri seçimi ve metot seçimi yapılarak, veri analizi sonuçları grafik ve tablo yardımıyla yorumlanır. Arayüzün kullanım adımları aşağıdaki gibidir:

- Kümeleme yapılacak veri dosyasını içeren tablo açılan listeden seçilir. Bu liste, kullanıcı giriş arayüzünde seçilen veri tabanında yer alan tablolardan oluşmaktadır.
- Seçilen tablo için tabloya ait tüm sütunlar otomatik olarak hemen aşağıda bulunan listeyi dolduracaktır. Bu sütunlar, veri dosyasına ait öz nitelikleri ifade etmektedir. Öz nitelikler arasından kümeleme için kullanılacak olanlar seçilir; seçilen öz nitelikler, otomatik olarak sağda bulunan listeyi dolduracaktır. Bu adımla birlikte, kümeleme analizi için veri seçimi tamamlanmış olacaktır.
- Metot seçiminde ilk adım olarak kümeleme için aralık değerleri girilir. Aralık değerleri değiştirilerek, deneme yanılma yoluyla optimal küme sayısının seçilmesi amaçlanır.
- Optimal küme sayısının seçimi için; Bölüm 4 te anlatılan Elbow, Calinski-Harabasz, Krzanowski-Lai ve Silhouette metotlarının algoritmaları analiz sistemi üzerinde geliştirilmiştir. Açılan liste aracılığıyla tercih edilen metot seçilir.
- Başlangıç merkezlerin seçimi için ise; yine Bölüm 4 te anlatılan Maximin, Katsavounidis, PCA-Part, Var-Part, K-Means++ metotlarının algoritmaları geliştirilmiştir. Açılan liste aracılığıyla tercih edilen metot seçilir.
- Veri seçimi ve metot seçimi adımlarının tamamlanmasıyla birlikte, “EXECUTE” butonu kullanılarak kümeleme analizi işlemi başlatılır.
- İşlem tamamlandığında kümeleme sonuçları, seçilen metotlara bağlı olarak grafik ve tablo olarak elde edilecektir. Sonuçlar aynı zamanda veri tabanında bir tabloya da kaydedilmektedir. Optimal küme sayısı ve başlangıç merkezlerin seçimi oluşan grafik ve tablo yardımıyla yapılır; verinin analizi ise veri tabanında oluşturulan tablo üzerinde çeşitli sorgularla gerçekleştirilir.

5.3.1 Analiz Sistemini Oluşturan Bileşenler

5.3.1.1 K-Ortalamalar Hakkında Özet Bilgi

Analiz sisteminde ilk olarak K-Ortalamalar kümeleme algoritması hakkında özet verilerek, kullanıcının bilgilendirilmesi sağlanmıştır.

5.3.1.2 Veri Seçimi

Analiz sisteminde veri seçiminin yapıldığı arayüz, Şekil 5. 3 te gösterilmiştir. İlgili seçimlerin yapılmasıyla birlikte açılacak olan kümeleme analizi arayüzünde ilk olarak veri dosyasını içeren "MRKTBL" tablosu seçilir. Tablonun seçilmesiyle birlikte tabloda bulunan sütunlar, yani veri dosyasına ait öz nitelikler listeyi doldurur. Listeden kümeleme yapmak amacıyla gereken öz nitelikler seçildiğinde, seçilen bu öz nitelikler diğer listeyi oluşturur. Bu işlemler ile birlikte veri seçimi tamamlanır.

5.3.1.3 Metot Seçimi

Analiz sisteminde ikinci temel adım metot seçimidir. Metot seçiminde ilk olarak kümeleme aralığı için giriş yapılır. Bu tez çalışmasında, firmaya ait veri dosyası için kullanılacak kümeleme aralığı $k = 2$ ile $k = 20$ arasındadır. Küme sayısının seçimi için kullanılan metotlar Bölüm 4 te anlatılmıştır. Veri dosyasının analizinde küme sayısının seçimi için kullanılacak olan metot ise Elbow dur. Diğer metotların kullanılmama nedeni, artan veri boyutuna bağlı olarak veri analizinin uzun sürmesidir. Diğer yandan, söz konusu metotlar arasından Silhouette ve Calinski-Harabasz metotları aynı zamanda kümeleme değerlendirme kriterleri olduklarından, bu metotlar kümeleme aralığını daraltma işleminden sonra belirlenen k değerleri için uygulanarak, kümeleme değerlendirme kriterleri olarak ele alınmıştır. Kümeleme aralığının daraltılması işleminde, Elbow metodu her bir başlangıç merkez seçimi metodu ile birlikte ayrı ayrı uygulanmıştır. Başlangıç merkezlerin seçimi için ise Maximin, Katsavounidis, PCA-Part, Var-Part ve K-Means++ metotları kullanılmıştır.

5.3.1.4 Grafik Aracılığıyla Veri Analizi

Sonuçların görsel olarak sunulması, kullanıcılar tarafından her zaman daha fazla tercih edilmektedir. Bu tez çalışmasında kümeleme analiz sonuçları grafik olarak verilmiştir; tez çalışmasında kullanılan Elbow metodu ile optimal küme sayısının seçiminde grafik etkin rol oynamıştır.

5.3.1.5 Tablo Aracılığıyla Veri Analizi

Sonuçların tablo olarak sunulması en fazla kullanılan diğer bir yöntemdir. Çalışmada analiz sonuçlarından elde edilen küme, hata ve iterasyon sayısı gibi bilgiler tablo olarak verilmiştir; tablo kullanıcının görmek istediği bilgilere göre tasarlanabilmektedir.

5.4 İris Veri Dosyası Uygulaması

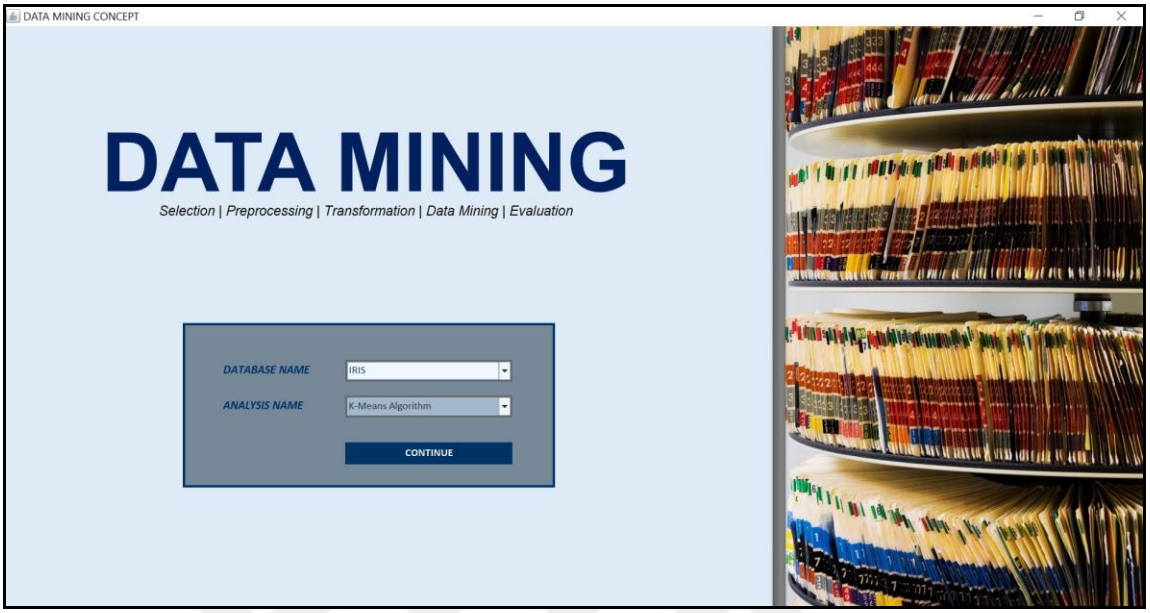
İris veri dosyası, İngiliz istatistikçi ve biyolog Ronald Fisher'ın 1936 yılındaki bir makalesinde sunulmuştur [48]. Makine öğrenmesi alıştırmalarında sıklıkla kullanılan en popüler veri dosyalarından olan İris veri dosyası toplam 150 kayda sahiptir. Her bir kayıt için; taç yaprak uzunluğu, taç yaprak genişliği, çanak yaprak uzunluğu ve çanak yaprak genişliği olmak üzere toplam dört öz nitelikle tanımlanmıştır. Veri dosyasında her bir bitki ayrı bir kaydı ifade ederken; bitki türleri bağımlı değişkenleri, bitkilerde ölçülen dört öz nitelik ise bağımsız değişkenleri ifade etmektedir [49].

İris veri dosyasına ait örnek veriler aşağıdaki gibidir (Çizelge 5.1):

Çizelge 5. 1 İris – Örnek veriler

TAÇ YAPRAK UZUNLUĞU	TAÇ YAPRAK GENİŞLİĞİ	ÇANAK YAPRAK UZUNLUĞU	ÇANAK YAPRAK GENİŞLİĞİ
5.1	3.5	1.4	0.2
4.9	3.0	1.4	0.2
4.7	3.2	1.3	0.2
4.6	3.1	1.5	0.2
5.0	3.6	1.4	0.2

Analiz sistemi üzerinde ilk olarak, Şekil 5. 4 te verilen kullanıcı giriş arayüzünden “IRIS” veri tabanı ve kümeleme analizi için kullanılacak olan “K-Means Algorithm” algoritması seçilir ve işleme devam edilir.

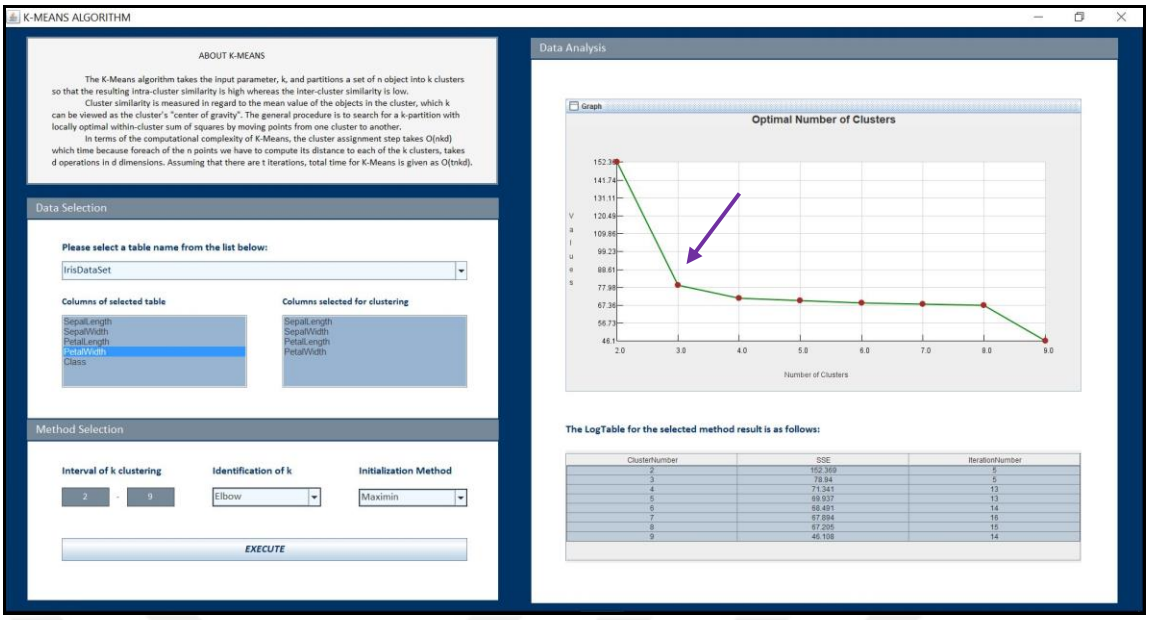


Şekil 5. 4 İris veri dosyası için veri tabanı ve kümeleme algoritması seçimi

İris veri dosyası uygulamasında, küme sayısının seçimi olarak; Elbow, Calinski-Harabasz, Krzanowski-Lai ve Silhouette metotları ile bu metotların her biri için farklı başlangıç merkez seçimi metotları ele alınarak sonuçlar bulunur.

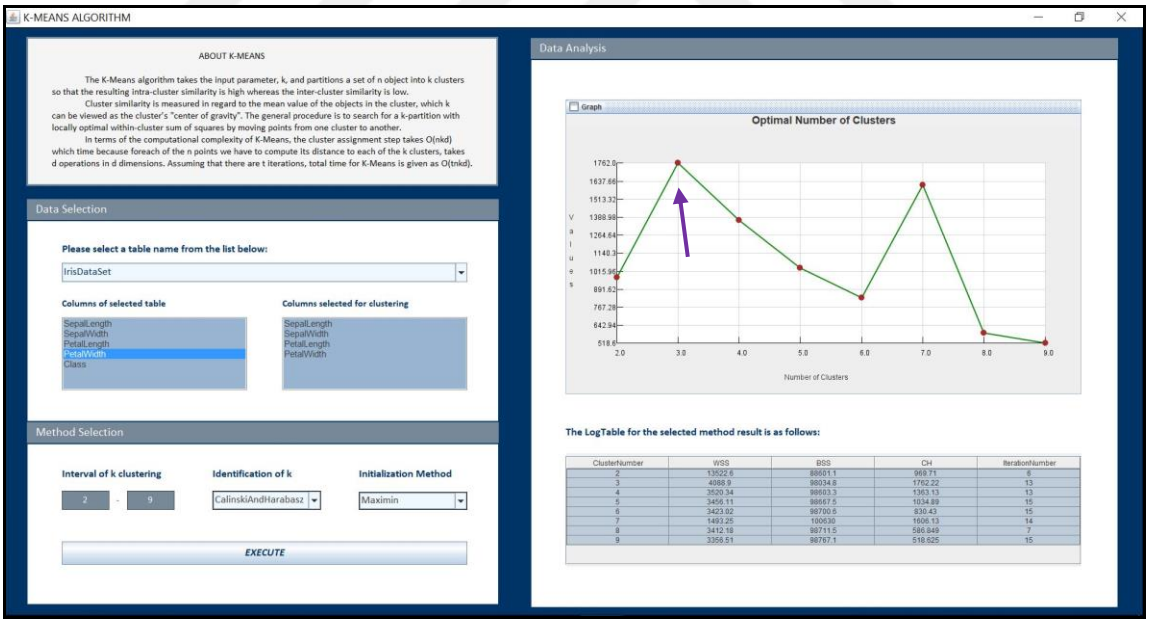
5.4.1 İris Veri Dosyası – Küme Sayısının Seçimi

Küme sayısının seçimi için kümeleme aralığı $k=2$ ile $k=9$ arasında seçilmiştir. Küme sayısının seçimi için kullanılacak tüm metotlar belirlenen kümeleme aralığında çalıştırılmıştır. Kümeleme aralığını daraltmayı sağlamak için, başlangıç merkezlerin seçimi metotlarından biri olan Maximin kullanılmıştır. Kümeleme analizi sonuçları aşağıdaki gibidir:



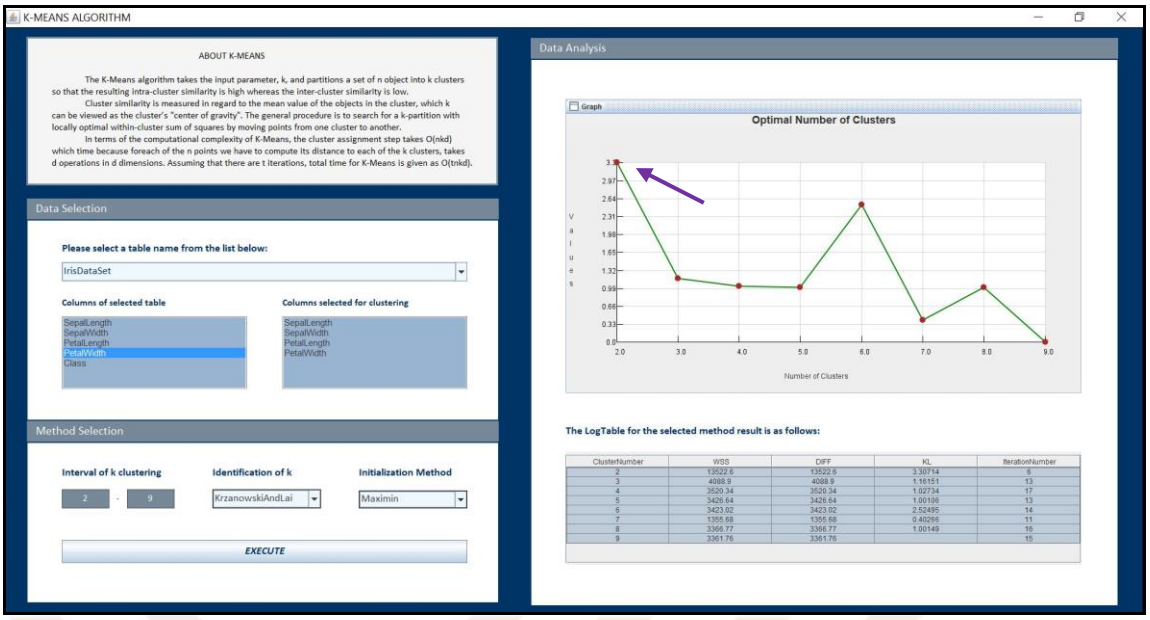
Şekil 5. 5 İris – Elbow ve Maximin metotları ile kümeleme analizi sonucu

Küme sayısının seçimi için Elbow metodu ve başlangıç merkezlerin seçimi için Maximin metodu kullanıldığında; Şekil 5. 5 te elde edilen grafik ve tablodan, küme sayısı $k = 3$ olarak belirlenmiştir.



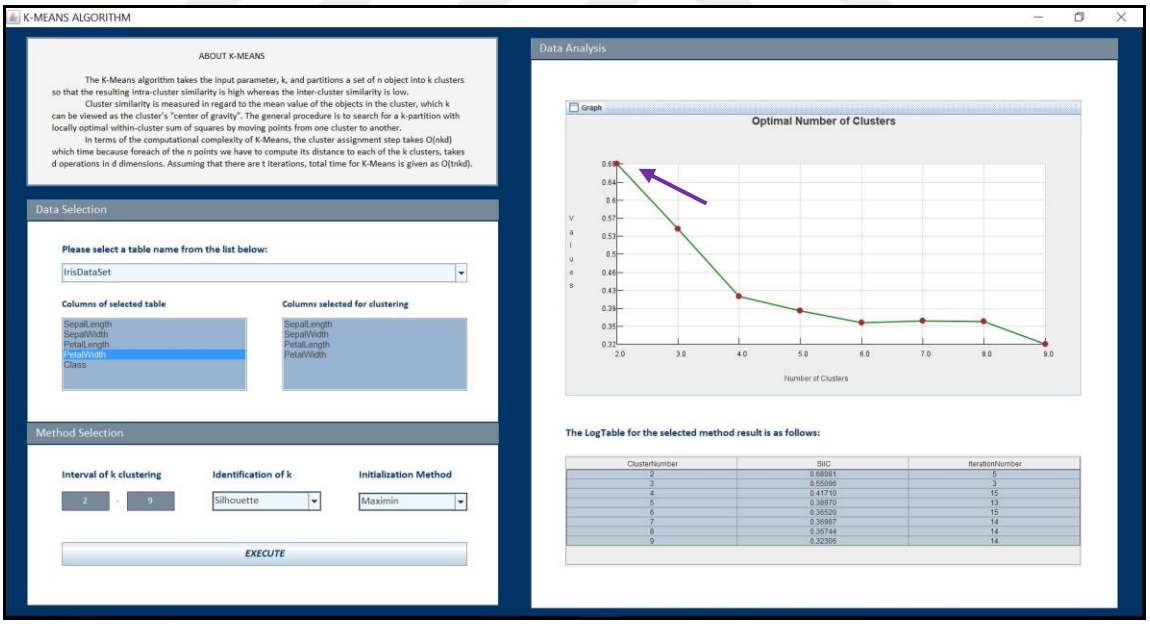
Şekil 5. 6 İris – CalinskiAndHarabasz ve Maximin metotları kümeleme analizi sonucu

Küme sayısının seçimi için Calinski-Harabasz metodu ve başlangıç merkezlerin seçimi için Maximin metodu kullanıldığında; Şekil 5. 6 da elde edilen grafik ve tablodan, küme sayısı yine $k = 3$ olarak belirlenmiştir.



Şekil 5. 7 İris – KrzanowskiAndLai ve Maximin kümeleme analizi sonucu

Küme sayısının seçimi için Krzanowski-Lai metodu ve başlangıç merkezlerin seçimi için Maximin metodu kullanıldığında; Şekil 5. 7 de elde edilen grafik ve tablodan, küme sayısı $k = 2$ olarak belirlenmiştir.



Şekil 5. 8 İris – Silhouette ve Maximin metotları ile kümeleme analizi sonucu

Küme sayısının seçimi için Silhouette metodu ve başlangıç merkezlerin seçimi için Maximin metodu kullanıldığında; Şekil 5. 8 de elde edilen grafik ve tablodan, küme sayısı yine $k = 2$ olarak belirlenmiştir.

Tüm bu analizler sonucunda, küme sayısı seçimi $k=2$ ya da $k=3$ olacak şekilde kümeleme aralığı daraltılmıştır (Çizelge 5.2):

Çizelge 5. 2 İris – Küme sayısı seçiminde kullanılan metotlar için sonuçlar

KÜME SAYISININ SEÇİMİ / KÜMELEME SAYISI	ELBOW	CALINSKI–HARABASZ	KRZANOWSKI–LAI	SILHOUETTE
	WCSS	CH	KL	SilC
2	152.369	969.710	**3.307.138**	**0.681**
3	**78.94**	**1.762.223**	1.161.507	0.551
4	71.341	1.363.133	1.027.344	0.417
5	69.937	1.034.891	1.001.057	0.390
6	68.491	830.429	2.524.946	0.365
7	67.894	1.606.133	0.402664	0.370
8	67.205	586.848	1.001.490	0.367
9	46.108	518.625	-	0.323

5.4.2 İris Veri Dosyası – Kümelemenin Değerlendirilmesi

$k=2$ ve $k=3$ küme sayıları için analiz sonuçları; başlangıç hata, final hata, iterasyon sayısı, başlangıç merkezlerin seçimi için geçen süre ve toplam süre şeklinde detaylı olarak incelenmiştir.

5.4.2.1 Kümelemenin Elbow Metodu ile Değerlendirilmesi

Küme sayısının seçimi için Elbow metodu ve başlangıç merkezlerin seçimi için Maximin metodu seçildiğinde sonuç aşağıdaki gibidir (Çizelge 5.3):

Çizelge 5. 3 İris – Elbow ve Maximin metotları için analiz sonuçları

BAŞLANGIÇ MERKEZLERİN SEÇİMİ / KÜMELEME SAYISI	MAXIMIN				
	BAŞLANGIÇ HATA	FİNAL HATA	İTERASYON SAYISI	BAŞLANGIÇ SÜRE (SN)	TOPLAM SÜRE (SN)
2	0	152.369	5	0.001648605	0.100944687

Çizelge 5. 3 İris – Elbow ve Maximin metotları için analiz sonuçları (devamı)

3	0	78.94	5	0.000587596	0.072664369

Küme sayısının seçimi için Elbow metodu ve başlangıç merkezlerin seçimi için Katsavounidis metodu seçildiğinde sonuç aşağıdaki gibidir (Çizelge 5.4):

Çizelge 5. 4 İris – Elbow ve Katsavounidis metotları için analiz sonuçları

BAŞLANGIÇ MERKEZLERİN SEÇİMİ / KÜMELEME SAYISI	KATSAVOUNIDIS				
	BAŞLANGIÇ HATA	FİNAL HATA	İTERASYON SAYISI	BAŞLANGIÇ SÜRE (SN)	TOPLAM SÜRE (SN)
2	0	152.369	6	0.00150919	0.100941694
3	0	78.945	15	0.000439628	0.066829032

Küme sayısının seçimi için Elbow metodu ve başlangıç merkezlerin seçimi için PCA-Part metodu seçildiğinde sonuç aşağıdaki gibidir (Çizelge 5.5):

Çizelge 5. 5 İris – Elbow ve PCA-Part metotları için analiz sonuçları

BAŞLANGIÇ MERKEZLERİN SEÇİMİ / KÜMELEME SAYISI	PCA-PART				
	BAŞLANGIÇ HATA	FİNAL HATA	İTERASYON SAYISI	BAŞLANGIÇ SÜRE (SN)	TOPLAM SÜRE (SN)
2	156.043	152.369	3	0.022456957	0.099858019
3	100.101	78.945	4	0.004599844	0.060971458

Küme sayısının seçimi için Elbow metodu ve başlangıç merkezlerin seçimi için Var-Part metodu seçildiğinde sonuç aşağıdaki gibidir (Çizelge 5.6):

Çizelge 5. 6 İris – Elbow ve Var-Part metotları için analiz sonuçları

BAŞLANGIÇ MERKEZLERİN SEÇİMİ / KÜMELEME SAYISI	VAR-PART				
	BAŞLANGIÇ HATA	FİNAL HATA	İTERASYON SAYISI	BAŞLANGIÇ SÜRE (SN)	TOPLAM SÜRE (SN)
2	160.06	152.369	4	0.002355945	0.109748369
3	104.324	78.94	4	0.000965215	0.062907447

Küme sayısının seçimi için Elbow metodu ve başlangıç merkezlerin seçimi için K-Means++ metodu seçildiğinde sonuç aşağıdaki gibidir (Çizelge 5.7):

Çizelge 5. 7 İris – Elbow ve K-Means++ metotları için analiz sonuçları

BAŞLANGIÇ MERKEZLERİN SEÇİMİ / KÜMELEME SAYISI	K-MEANS++				
	BAŞLANGIÇ HATA	FİNAL HATA	İTERASYON SAYISI	BAŞLANGIÇ SÜRE (SN)	TOPLAM SÜRE (SN)
2	0	152.369	6	0.001686667	0.100214255
3	**0**	**78.94**	**6**	**0.000725728**	**0.059493487**

Küme sayısının seçimi için seçilen Elbow metodu için elde edilen detaylı sonuçlara göre, final hata her metot için aynı hesaplanmıştır ve hatanın en az olduğu küme sayısı $k = 3$ olarak seçilmiştir. İterasyon sayısı ve süreler incelendiğinde, bu metotlar arasından en uygun metodun K-Means++ olduğu görülmüştür.

5.4.2.2 Kümelemenin Calinski-Harabasz Metodu ile Değerlendirilmesi

Küme sayısının seçimi için Calinski-Harabasz metodu ve başlangıç merkezlerin seçimi için Maximin metodu seçildiğinde sonuç aşağıdaki gibidir (Çizelge 5.8):

Çizelge 5. 8 İris – Calinski-Harabasz ve Maximin metotları için analiz sonuçları

BAŞLANGIÇ MERKEZLERİN SEÇİMİ / KÜMELEME SAYISI	MAXIMIN				
	BAŞLANGIÇ HATA	FİNAL HATA	İTERASYON SAYISI	BAŞLANGIÇ SÜRE (SN)	TOPLAM SÜRE (SN)
2	0	152.369	6	0.002211826	0.12371041
3	0	78.945	13	0.000602564	0.08884636

Küme sayısının seçimi için Calinski-Harabasz metodu ve başlangıç merkezlerin seçimi için Katsavounidis metodu seçildiğinde sonuç aşağıdaki gibidir (Çizelge 5.9):

Çizelge 5. 9 İris – Calinski-Harabasz metodu ve Katsavounidis metotları için analiz sonuçları

BAŞLANGIÇ MERKEZLERİN SEÇİMİ / KÜMELEME SAYISI	KATSAVOUNIDIS				
	BAŞLANGIÇ HATA	FİNAL HATA	İTERASYON SAYISI	BAŞLANGIÇ SÜRE (SN)	TOPLAM SÜRE (SN)
2	0	152.369	6	0.002220379	0.111086925
3	0	78.945	15	0.000396007	0.079381527

Küme sayısının seçimi için Calinski-Harabasz metodu ve başlangıç merkezlerin seçimi için PCA-Part metodu seçildiğinde sonuç aşağıdaki gibidir (Çizelge 5.10):

Çizelge 5. 10 İris – Calinski-Harabasz ve PCA-Part metotları için analiz sonuçları

BAŞLANGIÇ MERKEZLERİN SEÇİMİ / KÜMELEME SAYISI	PCA-PART				
	BAŞLANGIÇ HATA	FİNAL HATA	İTERASYON SAYISI	BAŞLANGIÇ SÜRE (SN)	TOPLAM SÜRE (SN)
2	156.043	152.369	3	0.023757452	0.111148080
3	100.101	78.945	4	0.005125431	0.066731527

Küme sayısının seçimi için Calinski-Harabasz metodu ve başlangıç merkezlerin seçimi için Var-Part metodu seçildiğinde sonuç aşağıdaki gibidir (Çizelge 5.11):

Çizelge 5. 11 İris – Calinski-Harabasz ve Var-Part metotları için analiz sonuçları

BAŞLANGIÇ MERKEZLERİN SEÇİMİ / KÜMELEME SAYISI	VAR-PART				
	BAŞLANGIÇ HATA	FİNAL HATA	İTERASYON SAYISI	BAŞLANGIÇ SÜRE (SN)	TOPLAM SÜRE (SN)
2	160.060	152.369	4	0.002479964	0.109660272
3	104.324	78.94	4	0.000750533	0.071174423

Küme sayısının seçimi için Calinski-Harabasz metodu ve başlangıç merkezlerin seçimi için K-Means++ metodu seçildiğinde sonuç aşağıdaki gibidir (Çizelge 5.12):

Çizelge 5. 12 Iris – Calinski-Harabasz metodu ve K-Means++ metodu için analiz sonuçları

BAŞLANGIÇ MERKEZLERİN SEÇİMİ / KÜMELEME SAYISI	K-MEANS++				
	BAŞLANGIÇ HATA	FİNAL HATA	İTERASYON SAYISI	BAŞLANGIÇ SÜRE (SN)	TOPLAM SÜRE (SN)
2	0	152.369	4	0.002116031	0.097845052
3	**0**	**78.945**	**4**	**0.000593156**	**0.054892359**

Küme sayısının seçimi için seçilen Calinski-Harabasz metodu için elde edilen detaylı sonuçlara göre final hata her metot için aynı hesaplanmıştır ve hatanın en az olduğu küme sayısı $k = 3$ olarak seçilmiştir. İterasyon sayısı ve süreler incelendiğinde, bu metotlar arasından en uygun metodun yine K-Means++ olduğu görülmüştür.

5.4.2.3 Kümelemenin Krzanowski-Lai Metodu ile Değerlendirilmesi

Küme sayısının seçimi için Krzanowski-Lai metodu ve başlangıç merkezlerin seçimi için Maximin metodu seçildiğinde sonuç aşağıdaki gibidir (Çizelge 5.13):

Çizelge 5. 13 İris – Krzanowski-Lai ve Maximin metotları için analiz sonuçları

BAŞLANGIÇ MERKEZLERİN SEÇİMİ / KÜMELEME SAYISI	MAXIMIN				
	BAŞLANGIÇ HATA	FİNAL HATA	İTERASYON SAYISI	BAŞLANGIÇ SÜRE (SN)	TOPLAM SÜRE (SN)
2	0	152.369	6	0.001418100	0.097550399
3	0	78.945	13	0.000528152	0.063870524

Küme sayısının seçimi için Krzanowski-Lai metodu ve başlangıç merkezlerin seçimi için Katsavounidis metodu seçildiğinde sonuç aşağıdaki gibidir (Çizelge 5.14):

Çizelge 5. 14 İris – Krzanowski-Lai ve Katsavounidis metotları için analiz sonuçları

BAŞLANGIÇ MERKEZLERİN SEÇİMİ / KÜMELEME SAYISI	KATSAVOUNIDIS				
	BAŞLANGIÇ HATA	FİNAL HATA	İTERASYON SAYISI	BAŞLANGIÇ SÜRE (SN)	TOPLAM SÜRE (SN)
2	0	152.369	6	0.001569062	0.098387317
3	0	78.945	15	0.000419100	0.073545763

Küme sayısının seçimi için Krzanowski-Lai metodu ve başlangıç merkezlerin seçimi için PCA-Part metodu seçildiğinde sonuç aşağıdaki gibidir (Çizelge 5.15):

Çizelge 5. 15 İris – Krzanowski-Lai ve PCA-Part metotları için analiz sonuçları

BAŞLANGIÇ MERKEZLERİN SEÇİMİ / KÜMELEME SAYISI	PCA-PART				
	BAŞLANGIÇ HATA	FİNAL HATA	İTERASYON SAYISI	BAŞLANGIÇ SÜRE (SN)	TOPLAM SÜRE (SN)
2	156.043	152.369	3	0.022699009	0.100504632
3	100.101	78.945	4	0.004791433	0.056605540

Küme sayısının seçimi için Krzanowski-Lai metodu ve başlangıç merkezlerin seçimi için Var-Part metodu seçildiğinde sonuç aşağıdaki gibidir (Çizelge 5.16):

Çizelge 5. 16 İris – Krzanowski-Lai ve Var-Part metotları için analiz sonuçları

BAŞLANGIÇ MERKEZLERİN SEÇİMİ / KÜMELEME SAYISI	VAR-PART				
	BAŞLANGIÇ HATA	FİNAL HATA	İTERASYON SAYISI	BAŞLANGIÇ SÜRE (SN)	TOPLAM SÜRE (SN)
2	160.06	152.369	4	0.001526724	0.101437772
3	**104.324**	**78.94**	**4**	**0.000777902**	**0.055155794**

Küme sayısının seçimi için Krzanowski-Lai metodu ve başlangıç merkezlerin seçimi için K-Means++ metodu seçildiğinde sonuç aşağıdaki gibidir (Çizelge 5.17):

Çizelge 5. 17 İris – Krzanowski-Lai ve K-Means++ metotları için analiz sonuçları

BAŞLANGIÇ MERKEZLERİN SEÇİMİ / KÜMELEME SAYISI	K-MEANS++				
	BAŞLANGIÇ HATA	FİNAL HATA	İTERASYON SAYISI	BAŞLANGIÇ SÜRE (SN)	TOPLAM SÜRE (SN)
2	0	152.369	5	0.001839340	0.099720742
3	0	78.94	7	0.000945970	0.088006449

Küme sayısının seçimi için seçilen Krzanowski-Lai metodu için elde edilen detaylı sonuçlara göre final hata her metot için aynı hesaplanmıştır ve hatanın en az olduğu küme sayısı $k = 3$ olarak seçilmiştir. İterasyon sayısı ve süreler incelendiğinde, bu metotlar arasından en uygun metodun Var-Part olduğu görülmüştür.

5.4.2.4 Kümelemenin Silhouette Metodu ile Değerlendirilmesi

Küme sayısının seçimi için Silhouette metodu ve başlangıç merkezlerin seçimi için Maximin metodu seçildiğinde sonuç aşağıdaki gibidir (Çizelge 5.18):

Çizelge 5. 18 İris – Silhouette ve Maximin metotları için analiz sonuçları

BAŞLANGIÇ MERKEZLERİN SEÇİMİ / KÜMELEME SAYISI	MAXIMIN				
	BAŞLANGIÇ HATA	FİNAL HATA	İTERASYON SAYISI	BAŞLANGIÇ SÜRE (SN)	TOPLAM SÜRE (SN)
2	0	152.369	5	0.002304198	0.123496155
3	**0**	**78.945**	**3**	**0.000608123**	**0.059035470**

Küme sayısının seçimi için Silhouette metodu ve başlangıç merkezlerin seçimi için Katsavounidis metodu seçildiğinde sonuç aşağıdaki gibidir (Çizelge 5.19):

Çizelge 5. 19 İris – Silhouette ve Katsavounidis metotları için analiz sonuçları

BAŞLANGIÇ MERKEZLERİN SEÇİMİ / KÜMELEME SAYISI	KATSAVOUNIDIS				
	BAŞLANGIÇ HATA	FİNAL HATA	İTERASYON SAYISI	BAŞLANGIÇ SÜRE (SN)	TOPLAM SÜRE (SN)
2	0	152.369	6	0.002284099	0.109313872
3	0	78.945	15	0.000283962	0.086215863

Küme sayısının seçimi için Silhouette metodu ve başlangıç merkezlerin seçimi için PCA-Part metodu seçildiğinde sonuç aşağıdaki gibidir (Çizelge 5.20):

Çizelge 5. 20 İris – Silhouette ve PCA-Part metotları için analiz sonuçları

BAŞLANGIÇ MERKEZLERİN SEÇİMİ / KÜMELEME SAYISI	PCA-PART				
	BAŞLANGIÇ HATA	FİNAL HATA	İTERASYON SAYISI	BAŞLANGIÇ SÜRE (SN)	TOPLAM SÜRE (SN)
2	156.043	152.369	3	0.022269644	0.098549398
3	100.101	78.945	4	0.003352378	0.068618764

Küme sayısının seçimi için Silhouette metodu ve başlangıç merkezlerin seçimi için Var-Part metodu seçildiğinde sonuç aşağıdaki gibidir (Çizelge 5.21):

Çizelge 5. 21 İris – Silhouette ve Var-Part metotları için analiz sonuçları

BAŞLANGIÇ MERKEZLERİN SEÇİMİ / KÜMELEME SAYISI	VAR-PART				
	BAŞLANGIÇ HATA	FİNAL HATA	İTERASYON SAYISI	BAŞLANGIÇ SÜRE (SN)	TOPLAM SÜRE (SN)
2	160.060	152.369	4	0.002717312	0.101971484
3	104.324	78.94	4	0.001229504	0.061346939

Küme sayısının seçimi için Silhouette metodu ve başlangıç merkezlerin seçimi için K-Means++ metodu seçildiğinde sonuç aşağıdaki gibidir (Çizelge 5.22):

Çizelge 5. 22 İris – Silhouette ve K-Means++ metotları için analiz sonuçları

BAŞLANGIÇ MERKEZLERİN SEÇİMİ / KÜMELEME SAYISI	K-MEANS++				
	BAŞLANGIÇ HATA	FİNAL HATA	İTERASYON SAYISI	BAŞLANGIÇ SÜRE (SN)	TOPLAM SÜRE (SN)
2	0	152.369	5	0.002022374	0.103347674
3	0	142.859	9	0.000947681	0.070018047

Küme sayısının seçimi için seçilen Silhouette metodu için elde edilen detaylı sonuçlara göre, final hataların en az olduğu metotlar $k = 3$ için; Maximin, Katsavounidis ve PCA-Part ve Var-Part metotlarıdır. İterasyon sayısı ve süreler incelendiğinde, bu metotlar arasından en uygun metodun Maximin olduğu görülmüştür.

5.4.3 İris Veri Dosyası – Kümeleme Sonuçları

Elde edilen sonuçlar birlikte değerlendirildiğinde en uygun başlangıç merkezlerin seçimi metodu, Silhouette metodu için Maximin; Elbow ve Calinski-Harabasz metotları için K-Means++; Krzanowski-Lai metodu için ise Var-Part olarak belirlenmiştir. Buradan; İris

veri dosyası için optimal küme sayısı $k = 3$ olarak seçilmiştir. Başlangıç merkezlerin seçimi için ise, K-Means++ metodu kullanılmıştır. Söz konusu metotlar kullanılarak veri analiz sisteminden elde edilen kümeleme sonucu aşağıdaki gibidir (Çizelge 5.23):

Çizelge 5. 23 İris – Kümeleme sonuçları

KÜME NO	TAÇ YAPRAK UZUNLUĞU	TAÇ YAPRAK GENİŞLİĞİ	ÇANAK YAPRAK UZUNLUĞU	ÇANAK YAPRAK GENİŞLİĞİ
1	7.0	3.2	4.7	1.4
1	6.9	3.1	4.9	1.5
1	6.7	3.0	5.0	1.7
1	6.3	3.3	6.0	2.5
1	7.1	3.0	5.9	2.1
1	6.3	2.9	5.6	1.8
1	6.5	3.0	5.8	2.2
1	7.6	3.0	6.6	2.1
1	7.3	2.9	6.3	1.8
1	6.7	2.5	5.8	1.8
1	7.2	3.6	6.1	2.5
1	6.5	3.2	5.1	2.0
1	6.4	2.7	5.3	1.9
1	6.8	3.0	5.5	2.1
1	6.4	3.2	5.3	2.3
1	6.5	3.0	5.5	1.8
1	7.7	3.8	6.7	2.2
1	7.7	2.6	6.9	2.3
1	6.9	3.2	5.7	2.3
1	7.7	2.8	6.7	2.0
1	6.7	3.3	5.7	2.1
1	7.2	3.2	6.0	1.8

Çizelge 5. 23 İris – Kümeleme sonuçları (devamı)

1	6.4	2.8	5.6	2.1
1	7.2	3.0	5.8	1.6
1	7.4	2.8	6.1	1.9
1	7.9	3.8	6.4	2.0
1	6.4	2.8	5.6	2.2
1	6.1	2.6	5.6	1.4
1	7.7	3.0	6.1	2.3
1	6.3	3.4	5.6	2.4
1	6.4	3.1	5.5	1.8
1	6.9	3.1	5.4	2.1
1	6.7	3.1	5.6	2.4
1	6.9	3.1	5.1	2.3
1	6.8	3.2	5.9	2.3
1	6.7	3.3	5.7	2.5
1	6.7	3.0	5.2	2.3
1	6.5	3.0	5.2	2.0
1	6.2	3.4	5.4	2.3
2	5.1	3.5	1.4	0.2
2	4.9	3.0	1.4	0.2
2	4.7	3.2	1.3	0.2
2	4.6	3.1	1.5	0.2
2	5.0	3.6	1.4	0.2
2	5.4	3.9	1.7	0.4
2	4.6	3.4	1.4	0.3
2	5.0	3.4	1.5	0.2
2	4.4	2.9	1.4	0.2
2	4.9	3.1	1.5	0.1

Çizelge 5. 23 İris – Kümeleme sonuçları (devamı)

2	5.4	3.7	1.5	0.2
2	4.8	3.4	1.6	0.2
2	4.8	3.0	1.4	0.1
2	4.3	3.0	1.1	0.1
2	5.8	4.0	1.2	0.2
2	5.7	4.4	1.5	0.4
2	5.4	3.9	1.3	0.4
2	5.1	3.5	1.4	0.3
2	5.7	3.8	1.7	0.3
2	5.1	3.8	1.5	0.3
2	5.4	3.4	1.7	0.2
2	5.1	3.7	1.5	0.4
2	4.6	3.6	1.0	0.2
2	5.1	3.3	1.7	0.5
2	4.8	3.4	1.9	0.2
2	5.0	3.0	1.6	0.2
2	5.0	3.4	1.6	0.4
2	5.2	3.5	1.5	0.2
2	5.2	3.4	1.4	0.2
2	4.7	3.2	1.6	0.2
2	4.8	3.1	1.6	0.2
2	5.4	3.4	1.5	0.4
2	5.2	4.1	1.5	0.1
2	5.5	4.2	1.4	0.2
2	4.9	3.1	1.5	0.1
2	5.0	3.2	1.2	0.2
2	5.5	3.5	1.3	0.2

Çizelge 5. 23 İris – Kümeleme sonuçları (devamı)

2	4.9	3.1	1.5	0.1
2	4.4	3.0	1.3	0.2
2	5.1	3.4	1.5	0.2
2	5.0	3.5	1.3	0.3
2	4.5	2.3	1.3	0.3
2	4.4	3.2	1.3	0.2
2	5.0	3.5	1.6	0.6
2	5.1	3.8	1.9	0.4
2	4.8	3.0	1.4	0.3
2	5.1	3.8	1.6	0.2
2	4.6	3.2	1.4	0.2
2	5.3	3.7	1.5	0.2
2	5.0	3.3	1.4	0.2
3	6.4	3.2	4.5	1.5
3	5.5	2.3	4.0	1.3
3	6.5	2.8	4.6	1.5
3	5.7	2.8	4.5	1.3
3	6.3	3.3	4.7	1.6
3	4.9	2.4	3.3	1.0
3	6.6	2.9	4.6	1.3
3	5.2	2.7	3.9	1.4
3	5.0	2.0	3.5	1.0
3	5.9	3.0	4.2	1.5
3	6.0	2.2	4.0	1.0
3	6.1	2.9	4.7	1.4
3	5.6	2.9	3.6	1.3
3	6.7	3.1	4.4	1.4

Çizelge 5. 23 İris – Kümeleme sonuçları (devamı)

3	5.6	3.0	4.5	1.5
3	5.8	2.7	4.1	1.0
3	6.2	2.2	4.5	1.5
3	5.6	2.5	3.9	1.1
3	5.9	3.2	4.8	1.8
3	6.1	2.8	4.0	1.3
3	6.3	2.5	4.9	1.5
3	6.1	2.8	4.7	1.2
3	6.4	2.9	4.3	1.3
3	6.6	3.0	4.4	1.4
3	6.8	2.8	4.8	1.4
3	6.0	2.9	4.5	1.5
3	5.7	2.6	3.5	1.0
3	5.5	2.4	3.8	1.1
3	5.5	2.4	3.7	1.0
3	5.8	2.7	3.9	1.2
3	6.0	2.7	5.1	1.6
3	5.4	3.0	4.5	1.5
3	6.0	3.4	4.5	1.6
3	6.7	3.1	4.7	1.5
3	6.3	2.3	4.4	1.3
3	5.6	3.0	4.1	1.3
3	5.5	2.5	4.0	1.3
3	5.5	2.6	4.4	1.2
3	6.1	3.0	4.6	1.4
3	5.8	2.6	4.0	1.2
3	5.0	2.3	3.3	1.0

Çizelge 5. 23 İris – Kümeleme sonuçları (devamı)

3	5.6	2.7	4.2	1.3
3	5.7	3.0	4.2	1.2
3	5.7	2.9	4.2	1.3
3	6.2	2.9	4.3	1.3
3	5.1	2.5	3.0	1.1
3	5.7	2.8	4.1	1.3
3	5.8	2.7	5.1	1.9
3	4.9	2.5	4.5	1.7
3	5.7	2.5	5.0	2.0
3	5.8	2.8	5.1	2.4
3	6.0	2.2	5.0	1.5
3	5.6	2.8	4.9	2.0
3	6.3	2.7	4.9	1.8
3	6.2	2.8	4.8	1.8
3	6.1	3.0	4.9	1.8
3	6.3	2.8	5.1	1.5
3	6.0	3.0	4.8	1.8
3	5.8	2.7	5.1	1.9
3	6.3	2.5	5.0	1.9
3	5.9	3.0	5.1	1.8

BÖLÜM 6

SONUÇLAR

Analiz sisteminde kullanıcı giriş arayüzünden seçilen “MIGROS” veri tabanı ve “K-Means Algorithm” kümeleme algoritması ile açılan analiz arayüzünde yapılan seçimler, kümeleme analizi için belirlenen dosyalara göre değişiklik göstermektedir.

6.1 Veri Dosyalarının Tanımlanması

Bu tez çalışmasında kümeleme analizi için üç farklı dosya ele alınmıştır. İlk dosyada; toplam satış ve ağırlık öz nitelikleri; ikinci dosyada toplam satış, ağırlık, farklı gün ve kategori öz nitelikleri; son dosyada ise toplam satış, ağırlık, zaman ve kategori öz nitelikleri kullanılarak kümeleme analizleri gerçekleştirilmiştir.

6.2 Dosya 1 Uygulaması

Dosya 1 de kümeleme için toplam satış olarak “Sum_Quantity_Norm” ve ağırlık olarak “Unit_KG_Norm” şeklinde 2 öz nitelik seçilmiştir. Bu kümeleme analizinde, müşterilerin ürünü satın alma miktarı ve ürünün ağırlığının etkisi analiz edilmektedir.

Dosya 1 e ait örnek veriler aşağıdaki gibidir (Çizelge 6.1):

Çizelge 6. 1 Dosya 1 – Örnek veriler

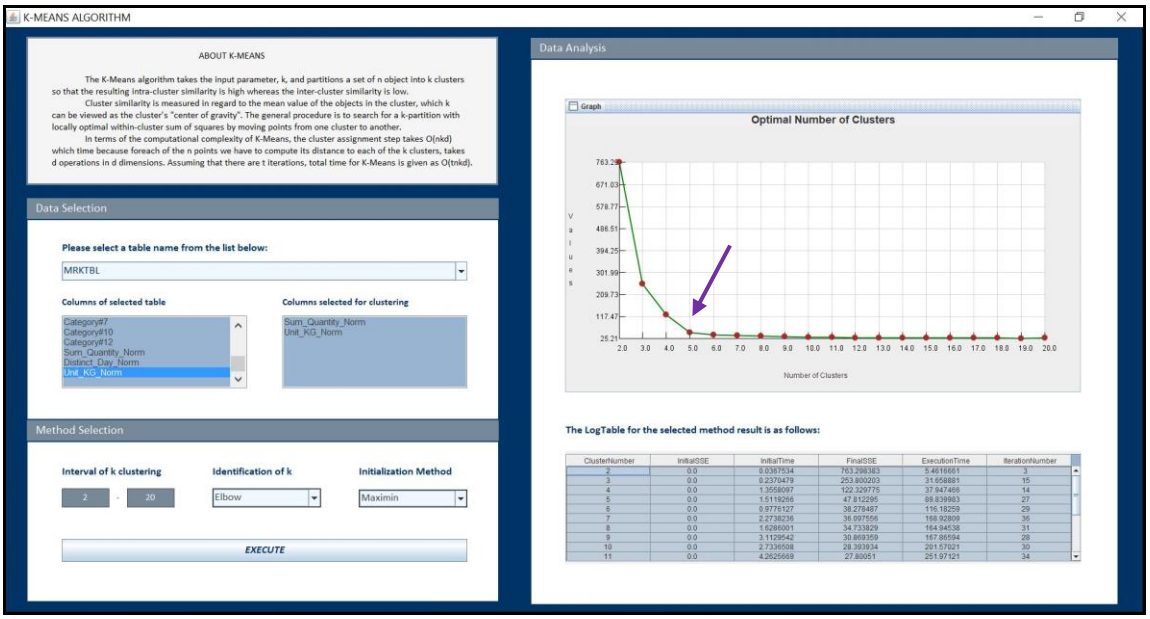
Sum Quantity Norm	UnitKG Norm
0.0005	0.0005

Çizelge 6. 1 Dosya 1 – Örnek veriler (devamı)

0.0005	0.0005
0.0005	0.0002
0.0005	0.0002
0.0005	0.0004

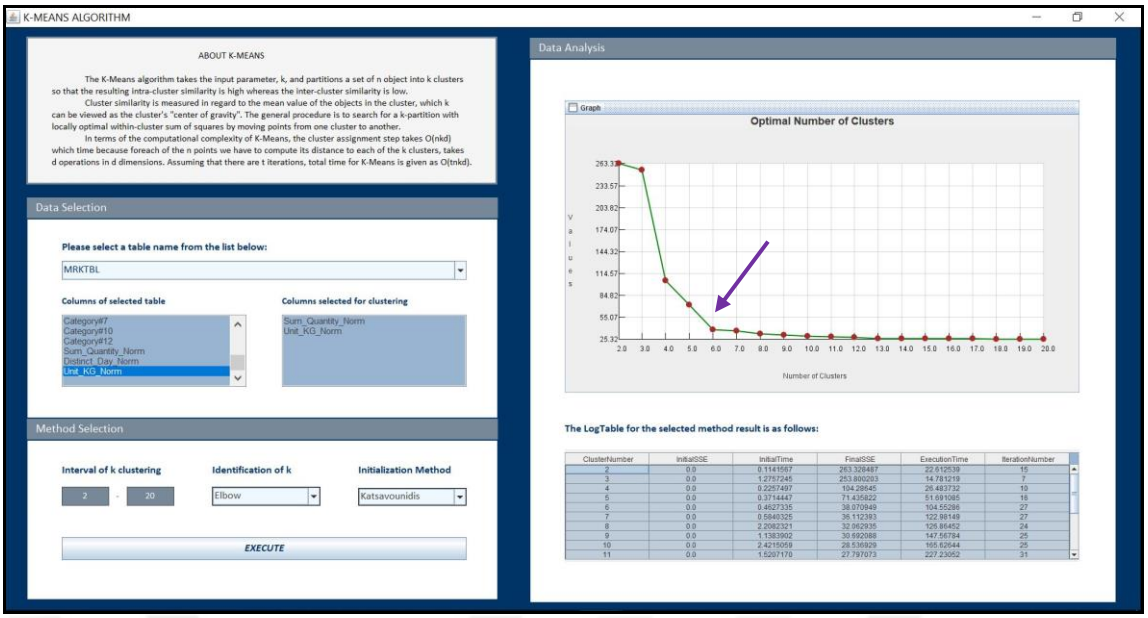
6.2.1 Dosya 1 – Küme Sayısının Seçimi

$k = 2$ ile $k = 20$ kümeleme aralığında, seçilen Elbow ve Maximin metotları ile elde edilen sonuç aşağıdaki gibidir:



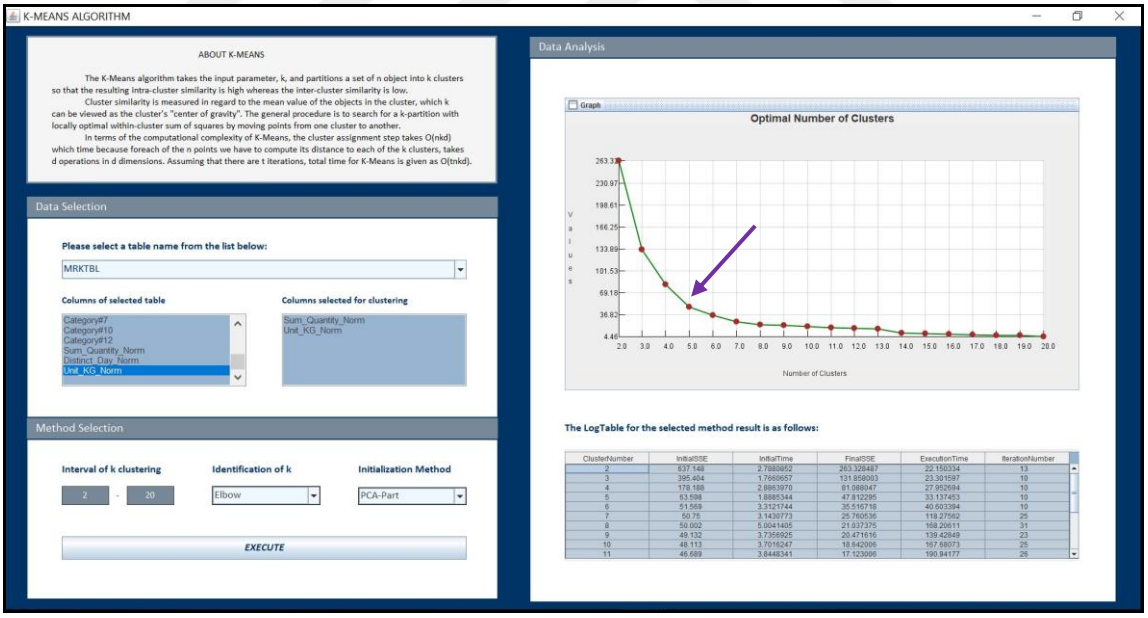
Şekil 6. 1 Dosya 1 – Elbow ve Maximin metotları ile küme sayısının seçimi

Şekil 6. 1 de grafik ve tablo incelendiğinde, Elbow metoduna göre optimal k değerinin 5 olduğu görülmüştür. $k = 2$ ile $k = 20$ kümeleme aralığında, seçilen Elbow ve Katsavounidis metotları ile elde edilen sonuç aşağıdaki gibidir:



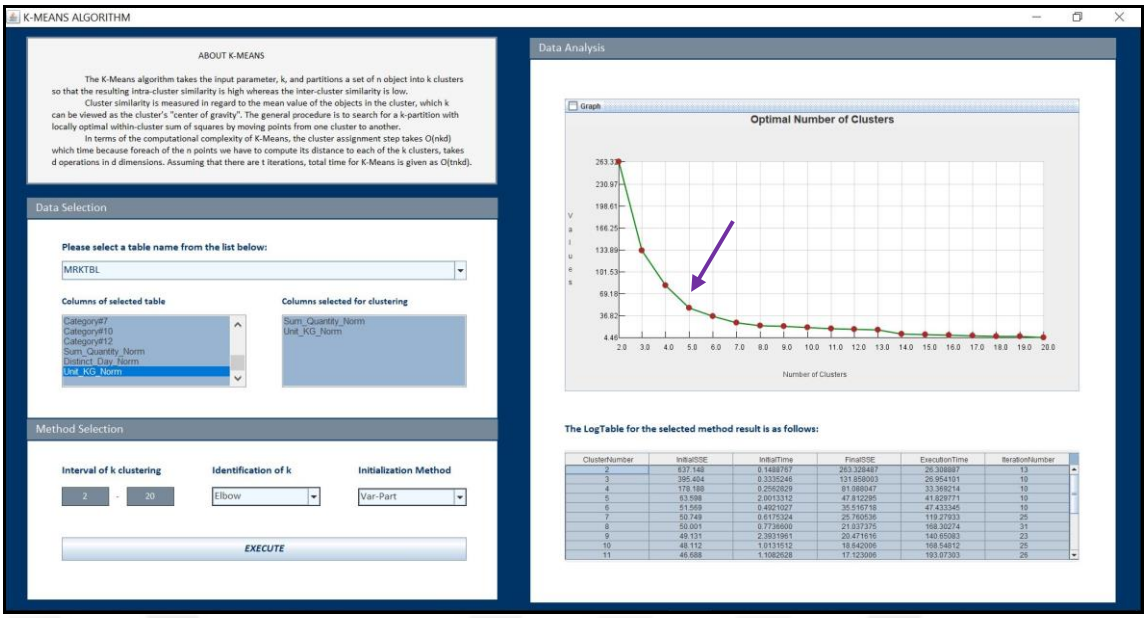
Şekil 6. 2 Dosya 1 – Elbow ve Katsavounidis metotları ile küme sayısının seçimi

Şekil 6. 2 de grafik ve tablo incelendiğinde, Elbow metoduna göre optimal k değerinin 6 olduğu görülmüştür. $k=2$ ile $k=20$ kümeleme aralığında, seçilen Elbow ve PCA-Part metotları ile elde edilen sonuç aşağıdaki gibidir:



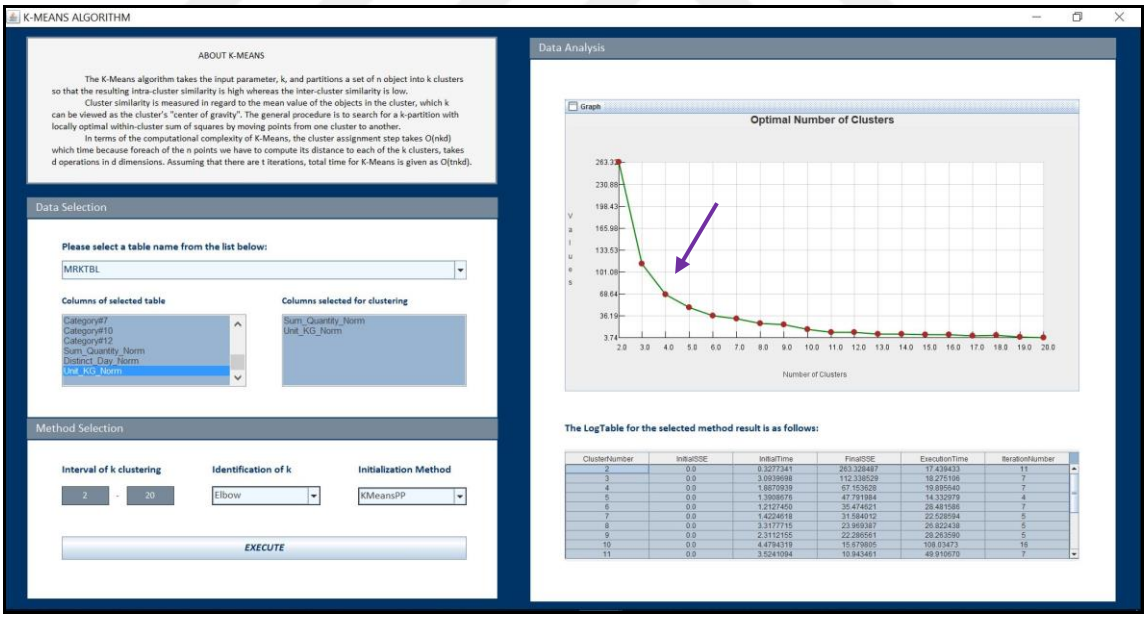
Şekil 6. 3 Dosya 1 – Elbow ve PCA-Part metotları ile küme sayısının seçimi

Şekil 6. 3 te grafik ve tablo incelendiğinde, Elbow metoduna göre optimal k değerinin yine 5 olduğu görülmüştür. $k=2$ ile $k=20$ kümeleme aralığında, seçilen Elbow ve Var-Part metotları ile elde edilen sonuç aşağıdaki gibidir:



Şekil 6. 4 Dosya 1 – Elbow ve Var-Part metotları ile küme sayısının seçimi

Şekil 6. 4 te grafik ve tablo incelendiğinde, Elbow metoduna göre optimal k değerinin yine 5 olduğu görülmüştür. $k=2$ ile $k=20$ kümeleme aralığında, seçilen Elbow ve K-Means++ metotları ile elde edilen sonuç aşağıdaki gibidir:



Şekil 6. 5 Dosya 1 – Elbow ve K-Means++ metotları ile küme sayısının seçimi

Şekil 6. 5 te grafik ve tablo incelendiğinde, Elbow metoduna göre optimal k değerinin 4 olduğu görülmüştür. Tüm bu analizler sonucunda küme sayısı seçiminin $k=4$ ile $k=6$ arasında olacağı görülerek seçim için aralık daraltılmıştır (Çizelge 6.2):

Çizelge 6. 2 Dosya 1 – Küme sayısının seçimi

KÜME SAYISI SEÇİMİ / KÜMELEME SAYISI	ELBOW				
	MAXIMIN	KATSAVOUNIDIS	PCA-PART	VAR-PART	K-MEANS++
2	763.298.383	263.328.487	263.328.487	263.328.487	263.328.487
3	253.800.203	253.800.203	131.858.003	131.858.003	112.338.529
4	122.329.775	104.286.450	81.088.047	81.088.047	**67.153.628**
5	**47.812.295**	71.435.822	**47.812.295**	**47.812.295**	47.791.984
6	38.278.487	**38.070.949**	35.516.718	35.516.718	35.474.621
7	36.097.556	36.112.393	25.760.536	25.760.536	31.584.012
8	34.733.829	32.062.935	21.037.375	21.037.375	23.969.387
9	30.869.359	30.692.088	20.471.616	20.471.616	22.286.561
10	28.393.934	28.536.929	18.642.006	18.642.006	15.679.805
11	27.800.510	27.797.073	17.123.006	17.123.006	10.943.461
12	26.648.586	27.673.178	15.982.758	15.982.758	1.161.128
13	25.648.697	25.959.198	15.435.728	15.435.728	8.652.302
14	25.403.389	25.749.516	8.917.452	8.917.452	8.336.638
15	25.409.057	25.373.314	8.681.187	8.681.187	7.614.684
16	25.319.468	25.623.867	7.572.556	7.572.556	7.414.578
17	25.378.164	25.623.858	6.331.319	6.331.319	5.718.154
18	25.346.123	25.322.731	6.039.181	6.039.181	6.512.144
19	25.214.101	25.322.731	5.660.670	5.660.670	4.174.093
20	25.336.388	25.322.731	4.464.704	4.464.704	3.746.434

6.2.2 Dosya 1 – Kümelemenin Değerlendirilmesi

$k=4$, $k=5$ ve $k=6$ küme sayıları için analiz sonuçları; başlangıç hata, final hata, iterasyon sayısı, başlangıç merkezlerin seçimi için geçen süre ve toplam süre şeklinde detaylı olarak incelenmiştir.

Küme sayısının seçiminde $k = 4$ için sonuç aşağıdaki gibidir (Çizelge 6.3):

Çizelge 6. 3 Dosya 1 – $k = 4$ için analiz sonuçları

BAŞLANGIÇ MERKEZLERİN SEÇİMİ İÇİN METOTLAR	BAŞLANGIÇ HATA	FİNAL HATA	İTERASYON SAYISI	BAŞLANGIÇ SÜRE (SN)	TOPLAM SÜRE (SN)
MAXIMIN	0.0	122.329.775	14	1.355.810	37.947.466
KATSAVOUNIDIS	0.0	104.286.450	10	0.225750	26.483.732
PCA-PART	178.188	81.088.047	10	2.886.397	27.952.695
VAR-PART	178.188	81.088.047	10	0.256283	33.369.214
K-MEANS++	0.0	67.153.628	7	1.887.094	19.895.640

Küme sayısının seçiminde $k = 5$ için sonuç aşağıdaki gibidir (Çizelge 6.4):

Çizelge 6. 4 Dosya 1 – $k = 5$ için analiz sonuçları

BAŞLANGIÇ MERKEZLERİN SEÇİMİ İÇİN METOTLAR	BAŞLANGIÇ HATA	FİNAL HATA	İTERASYON SAYISI	BAŞLANGIÇ SÜRE (SN)	TOPLAM SÜRE (SN)
MAXIMIN	0.0	47.812.295	27	1.511.927	89.839.984
KATSAVOUNIDIS	0.0	71.435.822	16	0.371445	51.691.085
PCA-PART	63.598	47.812.295	10	1.888.534	33.137.454
VAR-PART	63.598	47.812.295	10	2.001.331	41.829.772
K-MEANS++	0.0	47.791.984	4	1.390.868	14.332.980

Küme sayısının seçiminde $k = 6$ için sonuç aşağıdaki gibidir (Çizelge 6.5):

Çizelge 6. 5 Dosya 1 – $k = 6$ için analiz sonuçları

BAŞLANGIÇ MERKEZLERİN SEÇİMİ İÇİN METOTLAR	BAŞLANGIÇ HATA	FİNAL HATA	İTERASYON SAYISI	BAŞLANGIÇ SÜRE (SN)	TOPLAM SÜRE (SN)
MAXIMIN	0.0	38.278.487	29	0.977613	116.182.599
KATSAVOUNIDIS	0.0	38.070.949	27	0.462734	104.552.867
PCA-PART	51.569	35.516.718	10	3.312.174	40.603.394
VAR-PART	51.569	35.516.718	10	0.492103	47.433.346

Çizelge 6. 5 Dosya 1 – $k = 6$ için analiz sonuçları (devamı)

K-MEANS++	**0.0**	**35.474.621**	**7**	**1.212.745**	**28.481.586**

Kümeleme değerlendirme kriterlerinde $k = 4$ için sonuç aşağıdaki gibidir (Çizelge 6.6):

Çizelge 6. 6 Dosya 1 – $k = 4$ için kümeleme değerlendirme kriterleri

DEĞERLENDİRME KRİTERLERİ / BAŞLANGIÇ MERKEZLERİN SEÇİMİ İÇİN METOTLAR	SILHOUETTE	CALINSKI-HARABASZ
MAXIMIN	0.916814	2.411.719
KATSAVOUNIDIS	0.959468	7.058.658
PCA-PART	0.916814	2.411.719
VAR-PART	0.916814	2.411.719
K-MEANS++	0.916812	2.411.933

Kümeleme değerlendirme kriterlerinde $k = 5$ için sonuç aşağıdaki gibidir (Çizelge 6.7):

Çizelge 6. 7 Dosya 1 – $k = 5$ için kümeleme değerlendirme kriterleri

DEĞERLENDİRME KRİTERLERİ / BAŞLANGIÇ MERKEZLERİN SEÇİMİ İÇİN METOTLAR	SILHOUETTE	CALINSKI-HARABASZ
MAXIMIN	0.919407	1.855.853
KATSAVOUNIDIS	0.919407	1.661.299
PCA-PART	0.903349	2.372.814
VAR-PART	0.903349	2.372.814
K-MEANS++	0.910606	1.014.600

Kümeleme değerlendirme kriterlerinde $k = 6$ için sonuç aşağıdaki gibidir (Çizelge 6.8):

Çizelge 6. 8 Dosya 1 – $k = 6$ için kümeleme değerlendirme kriterleri

DEĞERLENDİRME KRİTERLERİ / BAŞLANGIÇ MERKEZLERİN SEÇİMİ İÇİN METOTLAR	SILHOUETTE	CALINSKI-HARABASZ
MAXIMIN	0.899385	2.271.741
KATSAVOUNIDIS	0.910925	1.610.076
PCA-PART	0.903542	1.902.733
VAR-PART	0.903542	1.902.733
K-MEANS++	**0.911181**	1.536.641

Başlangıç merkezlerin seçimi ve kümeleme değerlendirme kriterleri dikkate alındığında; Çizelge 6. 5 te gösterilen final hatanın, toplam sürenin ve iterasyon sayısının daha az ve diğer yandan, Çizelge 6. 8 de değerlendirme kriteri olarak Silhouette değerinin daha yüksek olduğu görülmüştür. Buradan, veri analizi için en uygun başlangıç merkezlerin seçimi metodu $k = 6$ küme sayısı için K-Means++ olarak belirlenmiştir.

6.2.3 Dosya 1 – Kümeleme Sonuçları

Analiz sistemi üzerinde $k = 6$ küme sayısı için Elbow ve K-Means++ metotları seçilerek, kümeleme sonuçları veri tabanında bir tabloya kaydedilmiştir.

Kümeleme dağılımlarına bakıldığında dağılımlar aşağıdaki gibidir (Çizelge 6.9):

Çizelge 6. 9 Dosya 1 – Kümeleme dağılımları

KÜME NO	KAYIT SAYISI
1	975.379
2	10.023
3	197
4	1313
5	26.276
6	2402

Kümeleme sonuçları analiz edildiğinde; toplam satış ve ağırlık öz nitelikleri dikkate alınarak gerçekleştirilen kümeleme işleminde en fazla tercih edilen ürünler, şarküteri pazarlama kategorisinde bulunan sade sütler ve kuru gıda kategorisinde bulunan kolalı içeceklerdir. Bu bağlamda, toplam satışlar incelendiğinde kümelemede en üst sırayı iki farklı markanın paylaştığı görülmüştür. En fazla tercih edilen diğer ürünler ise sırasıyla; kuru gıda kategorisinde bulunan meyve suları, tablet çikolata, ramazan kolisi ve sular olmuştur. Tüm bu tercihler incelendiğinde ise altı farklı markanın diğer tüm markalara göre daha fazla tercih edildiği belirlenmiştir.

6.3 Dosya 2 Uygulaması

Dosya 2 de kümeleme için toplam satış olarak "Sum_Quantity_Norm", ağırlık olarak "Unit_KG_Norm", farklı gün olarak "Distinct_Day_Norm" ve kategori olarak "Category#1-9" şeklinde on iki öz nitelik seçilmiştir. Bu kümeleme analizinde, müşterilerin ürünü satın alma miktarına ve ürünün ağırlığına ek olarak, bu ürünün kaç farklı günde satın alındığı ve en çok tercih edilen ürün kategorisi belirlenmiştir.

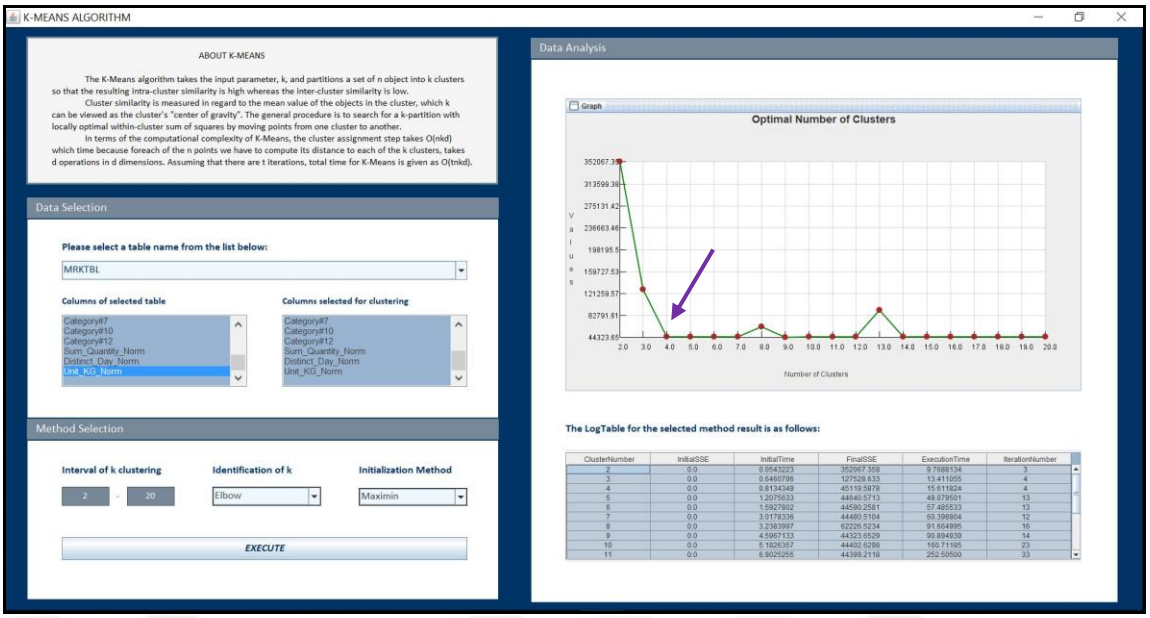
Dosya 2 ye ait örnek veriler aşağıdaki gibidir (Çizelge 6.10):

Çizelge 6. 10 Dosya 2 – Örnek veriler

Sum Quantity Norm	UnitKG Norm	Distinct Day Norm	Category #1	Category #2	Category #3	Category #4	Category #5	Category #6	Category #7	Category #10	Category #12
0.0005	0.0005	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
0.0005	0.0005	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
0.0005	0.0002	0.0333	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
0.0005	0.0002	0.0333	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
0.0005	0.0004	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0

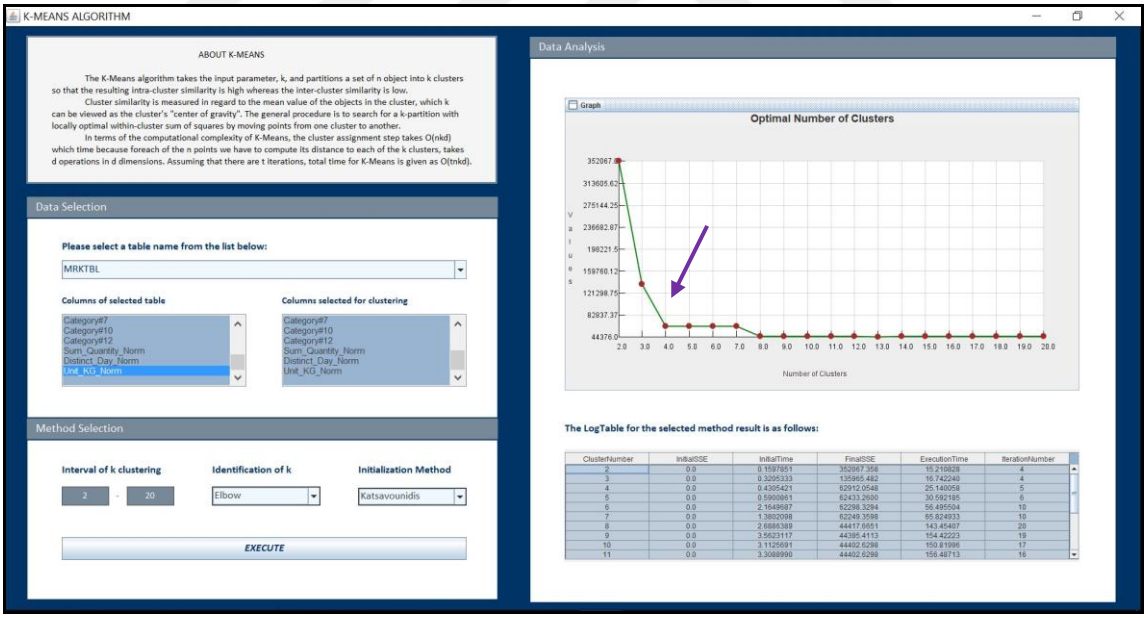
6.3.1 Dosya 2 – Küme Sayısının Seçimi

$k = 2$ ile $k = 20$ kümeleme aralığında, seçilen Elbow ve Maximin metotları ile elde edilen sonuç aşağıdaki gibidir:



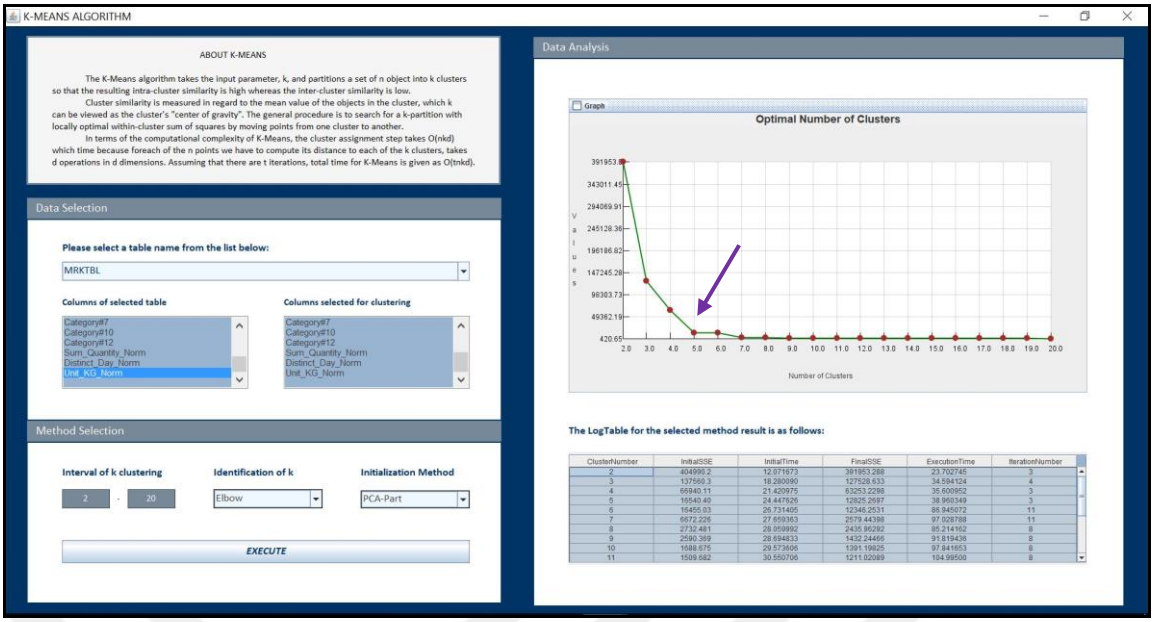
Şekil 6. 6 Dosya 2 – Elbow ve Maximin metotları ile küme sayısının seçimi

Şekil 6. 6 da grafik ve tablo incelendiğinde, Elbow metoduna göre optimal k değerinin 4 olduğu görülmüştür. $k=2$ ile $k=20$ kümeleme aralığında, seçilen Elbow ve Katsavounidis metotları ile elde edilen sonuç aşağıdaki gibidir:



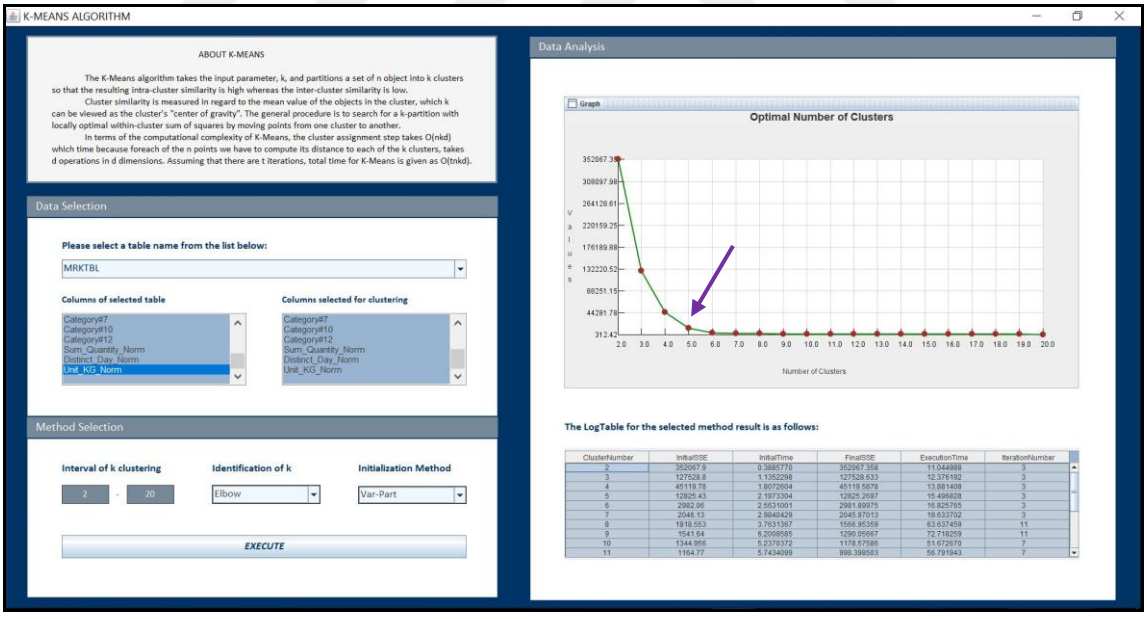
Şekil 6. 7 Dosya 2 – Elbow ve Katsavounidis metotları ile küme sayısının seçimi

Şekil 6. 7 de grafik ve tablo incelendiğinde, Elbow metoduna göre optimal k değerinin yine 4 olduğu görülmüştür. $k=2$ ile $k=20$ kümeleme aralığında, seçilen Elbow ve PCA-Part metotları ile elde edilen sonuç aşağıdaki gibidir:



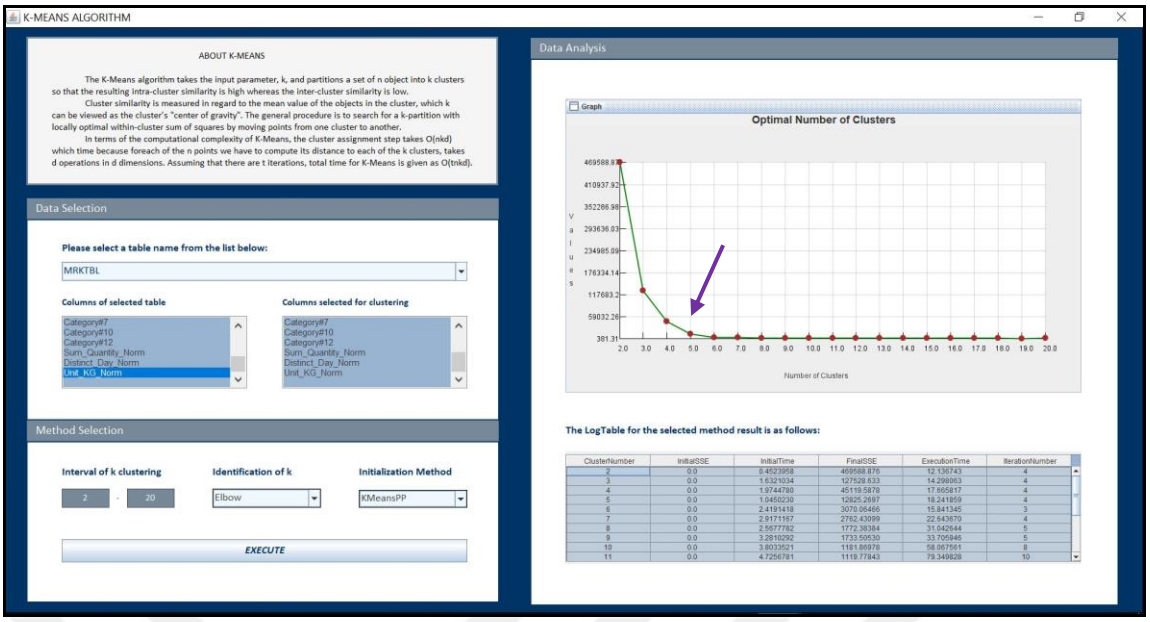
Şekil 6. 8 Dosya 2 – Elbow ve PCA-Part metotları ile küme sayısının seçimi

Şekil 6. 8 de grafik ve tablo incelendiğinde, Elbow metoduna göre optimal k değerinin 5 olduğu görülmüştür. $k=2$ ile $k=20$ kümeleme aralığında, seçilen Elbow ve Var-Part metotları ile elde edilen sonuç aşağıdaki gibidir:



Şekil 6. 9 Dosya 2 – Elbow ve Var-Part metotları ile küme sayısının seçimi

Şekil 6. 9 da grafik ve tablo incelendiğinde, Elbow metoduna göre optimal k değerinin yine 5 olduğu görülmüştür. $k=2$ ile $k=20$ kümeleme aralığında, seçilen Elbow ve K-Means++ metotları ile elde edilen sonuç aşağıdaki gibidir:



Şekil 6. 10 Dosya 2 – Elbow ve K-Means++ metotları ile küme sayısının seçimi

Şekil 6. 10 da grafik ve tablo incelendiğinde, Elbow metoduna göre optimal k değerinin yine 5 olduğu görülmüştür. Analizler sonucunda, küme sayısı seçiminin $k=4$ ya da $k=5$ olacağı görülerek seçim için aralık daraltılmıştır (Çizelge 6.11):

Çizelge 6. 11 Dosya 2 – Küme sayısının seçimi

KÜME SAYISI SEÇİMİ / KÜMELEME SAYISI	ELBOW				
	MAXIMIN	KATSAVOUNIDIS	PCA-PART	VAR-PART	K-MEANS++
2	352.067.359	352.067.359	391.953.288	352.067.359	469.588.877
3	127.528.634	135.965.482	127.528.634	127.528.634	127.528.634
4	**45.119.588**	**62.912.055**	63.253.230	45.119.588	45.119.588
5	44.640.571	62.433.260	**12.825.270**	**12.825.270**	**12.825.270**
6	44.590.258	62.298.329	12.346.253	2.981.900	3.070.065
7	44.480.510	62.249.360	2.579.444	2.045.970	2.762.431
8	62.226.523	44.417.665	2.435.963	1.566.954	1.772.384
9	44.323.653	44.395.411	1.432.245	1.290.057	1.733.505
10	44.402.630	44.402.630	1.391.198	1.178.576	1.181.870
11	44.399.212	44.402.630	1.211.021	998.399	1.119.778
12	44.398.365	44.377.341	934.124	951.061	913.946

Çizelge 6. 11 Dosya 2 – Küme sayısının seçimi (devamı)

13	91.104.068	44.376.254	808.189	825.125	714.887
14	44.397.680	44.397.663	698.263	699.440	768.048
15	44.397.541	44.397.501	692.639	567.990	652.191
16	44.375.548	44.397.398	689.155	489.106	606.331
17	44.397.472	44.397.385	538.258	474.339	556.730
18	44.397.460	44.397.465	531.699	437.057	551.567
19	44.397.457	44.397.457	452.838	376.564	381.316
20	44.375.288	44.397.457	420.654	312.423	389.567

6.3.2 Dosya 2 – Kümelemenin Değerlendirilmesi

$k = 4$ ve $k = 5$ küme sayıları için analiz sonuçları; başlangıç hata, final hata, iterasyon sayısı, başlangıç merkezlerin seçimi için geçen süre ve toplam süre şeklinde detaylı olarak incelenmiştir.

Küme sayısının seçiminde $k = 4$ için sonuç aşağıdaki gibidir (Çizelge 6.12):

Çizelge 6. 12 Dosya 2 – $k = 4$ için analiz sonuçları

BAŞLANGIÇ MERKEZLERİN SEÇİMİ İÇİN METOTLAR	BAŞLANGIÇ HATA	FİNAL HATA	İTERASYON SAYISI	BAŞLANGIÇ SÜRE (SN)	TOPLAM SÜRE (SN)
MAXIMIN	0.0	45.119.588	4	0.813435	15.611.825
KATSAVOUNIDIS	0.0	62.912.055	5	0.430542	25.140.058
PCA-PART	66.940.112	63.253.230	3	21.420.976	35.600.952
VAR-PART	45.119.782	45.119.588	3	1.807.260	13.881.409
K-MEANS++	0.0	45.119.588	4	1.974.478	17.665.818

Küme sayısının seçiminde $k = 5$ için sonuç aşağıdaki gibidir (Çizelge 6.13):

Çizelge 6. 13 Dosya 2 – $k = 5$ için analiz sonuçları

BAŞLANGIÇ MERKEZLERİN SEÇİMİ İÇİN METOTLAR	BAŞLANGIÇ HATA	FİNAL HATA	İTERASYON SAYISI	BAŞLANGIÇ SÜRE (SN)	TOPLAM SÜRE (SN)
MAXIMIN	0.0	44.640.571	13	1.207.563	49.079.502
KATSAVOUNIDIS	0.0	62.433.260	6	0.590086	30.592.185
PCA-PART	16.540.408	12.825.270	3	24.447.627	38.960.350
VAR-PART	12.825.438	12.825.270	3	2.197.330	15.496.829
K-MEANS++	**0.0**	**12.825.270**	**4**	**1.045.023**	**18.241.859**

Kümeleme değerlendirme kriterlerinde $k = 4$ için sonuç aşağıdaki gibidir (Çizelge 6.14):

Çizelge 6. 14 Dosya 2 – $k = 4$ için kümeleme değerlendirme kriterleri

DEĞERLENDİRME KRİTERLERİ / BAŞLANGIÇ MERKEZLERİN SEÇİMİ İÇİN METOTLAR	SILHOUETTE	CALINSKI-HARABASZ
MAXIMIN	0.939029	1.363.138
KATSAVOUNIDIS	0.921835	1.363.138
PCA-PART	0.921471	1.112.560
VAR-PART	0.939029	6.109.771
K-MEANS++	0.924009	4.807.824

Kümeleme değerlendirme kriterlerinde $k = 5$ için sonuç aşağıdaki gibidir (Çizelge 6.15):

Çizelge 6. 15 Dosya 2 – $k = 5$ için kümeleme değerlendirme kriterleri

DEĞERLENDİRME KRİTERLERİ / BAŞLANGIÇ MERKEZLERİN SEÇİMİ İÇİN METOTLAR	SILHOUETTE	CALINSKI-HARABASZ
MAXIMIN	0.928312	4.719.615
KATSAVOUNIDIS	0.911551	1.029.184
PCA-PART	0.970014	2.105.986

Çizelge 6. 15 Dosya 2 – $k = 5$ için kümeleme değerlendirme kriterleri (devamı)

VAR-PART	0.970014	2.105.986
K-MEANS++	**0.967735**	**5.942.184**

Başlangıç merkezlerin seçimi ve kümeleme değerlendirme kriterleri dikkate alındığında; Çizelge 6. 13 te final hatanın daha az olduğu ve diğer yandan Çizelge 6. 15 teki değerlendirme kriterlerine bakıldığında, Silhouette ve Calinski-Harabasz değerlerinin ikisinin de yüksek olduğu görülmüştür. Buradan, veri analizi için en uygun başlangıç merkezlerin seçimi metodu $k = 5$ küme sayısı için K-Means++ olarak belirlenmiştir.

6.3.3 Dosya 2 – Kümeleme Sonuçları

Analiz sistemi üzerinde $k = 5$ küme sayısı için Elbow ve K-Means++ metotları seçilerek, kümeleme sonuçları veri tabanında bir tabloya kaydedilmiştir.

Kümeleme dağılımlarına bakıldığında dağılımlar aşağıdaki gibidir (Çizelge 6.16):

Çizelge 6. 16 Dosya 2 – Kümeleme dağılımları

KÜME NO	KAYIT SAYISI
1	52.197
2	435.957
3	200.864
4	26.801
5	299.771

Kümeleme sonuçları analiz edildiğinde; toplam satış, ağırlık, farklı gün ve kategori öz nitelikleri dikkate alınarak gerçekleştirilen kümeleme işleminde en fazla tercih edilen ürün kuru gıda kategorisinde bulunan kolalı içeceklerdir. Bu ürünü en fazla tercih eden bir müşterinin her ay düzenli olarak markete geldiği ve bu ürünün yanında yine kuru gıda kategorisinde bulunan meyve suları, ayçiçek yağı, toz şeker, baldo pirinçler, bisküvi ve dökme çayları tercih ettiği görülmüştür. Bu müşteri ayrıca, şarküteri pazarlama kategorisinden sade sütü de tercih etmektedir.

Diğer yandan, 814 sadık müşteri arasından markete en fazla gelen iki farklı müşteri belirlenmiştir. Bu iki müşteri alışverişlerini Mart ve Mayıs aylarında yapmıştır. Birinci müşterinin devamlı olarak aldığı ürünler; şarküteri pazarlama kategorisinde bulunan sade sütlerdir ve gıda dışı kategorisinde bulunan türkçe gazetelerdir. Diğer müşterinin ise Mart ayında her gün gelerek gıda dışı kategorisinde bulunan türkçe gazetelerden aldığı görülmüştür. Müşterilerin günlük olarak en sık aldıkları diğer ürünler ise sırasıyla; kuru gıda kategorisinde bulunan form bisküvidir ve şarküteri pazarlama kategorisinde bulunan kültürlü peynirdir.

6.4 Dosya 3 Uygulaması

Dosya 3 te kümeleme için toplam satış olarak "Sum_Quantity_Norm", ağırlık olarak "Unit_KG_Norm", zaman olarak "Season#1-2" ve kategori olarak "Category#1-9" şeklinde on üç öz nitelik seçilmiştir. Bu kümeleme analizinde, müşterilerin ürünü satın alma miktarının ve ürünün ağırlığının yanında, hangi kategorideki hangi ürünlerin hangi sezonda daha fazla ya da daha az tercih edildiği görülmüştür.

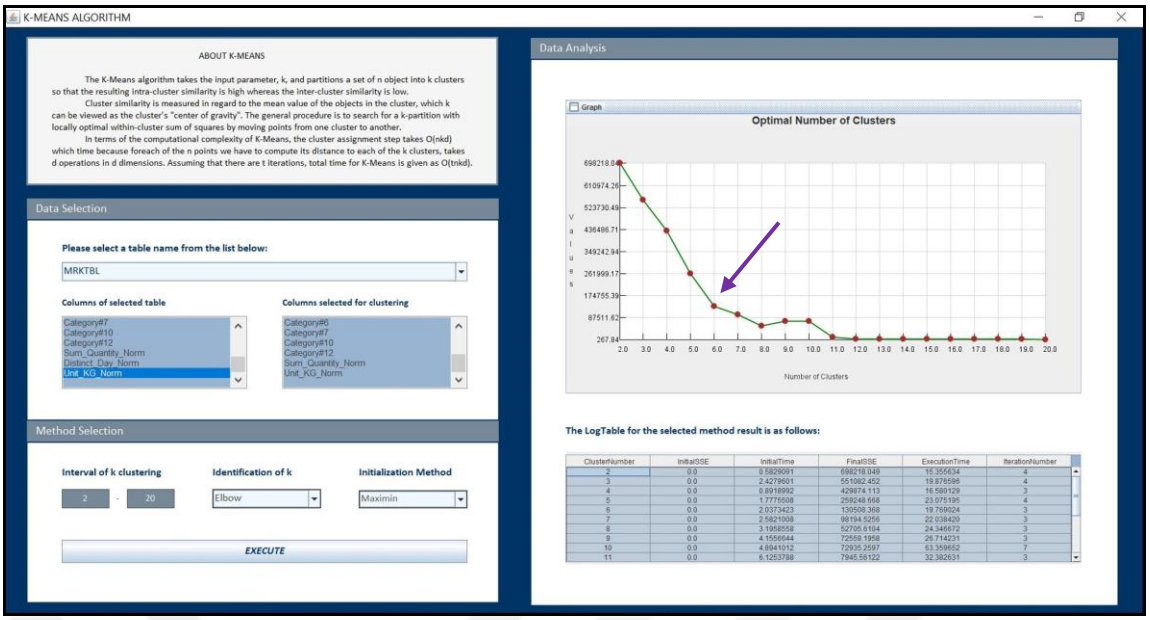
Dosya 3 e ait örnek veriler aşağıdaki gibidir (Çizelge 6.17):

Çizelge 6. 17 Dosya 3 – Örnek veriler

Sum Quantity Norm	UnitKG Norm	Category #1	Category #2	Category #3	Category #4	Category #5	Category #6	Category #7	Category #10	Category #12	Season#1	Season#2
0.0005	0.0005	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
0.0005	0.0005	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
0.0005	0.0252	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
0.0011	0.0005	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
0.0011	0.0005	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0

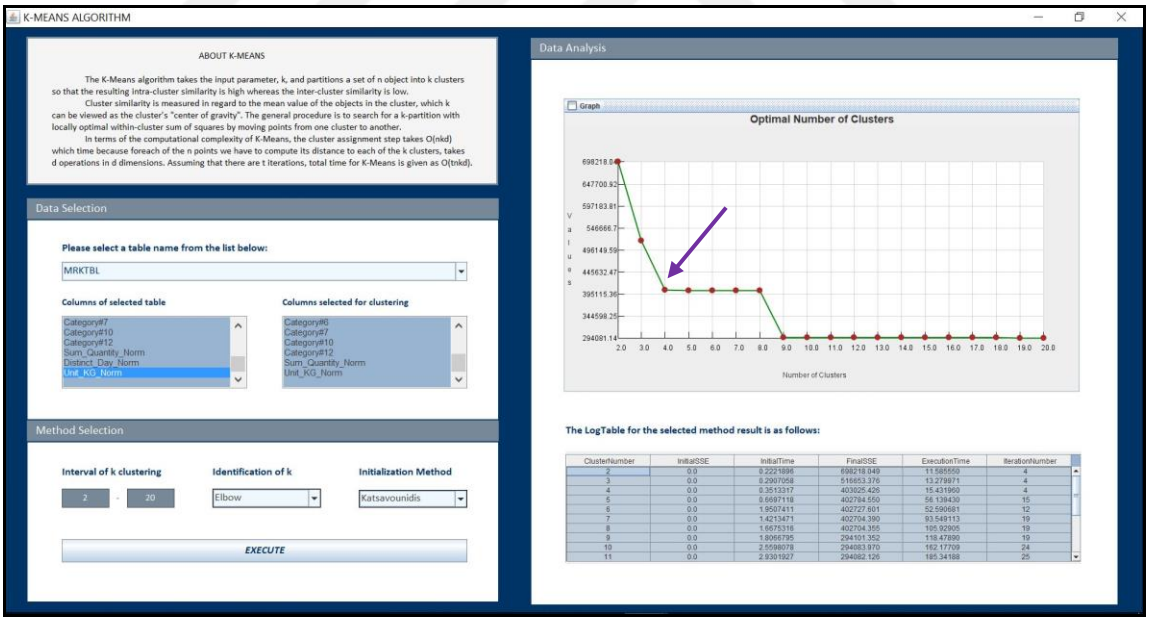
6.4.1 Dosya 3 – Küme Sayısının Seçimi

$k = 2$ ile $k = 20$ kümeleme aralığında, seçilen Elbow ve Maximin metotları ile elde edilen sonuç aşağıdaki gibidir:



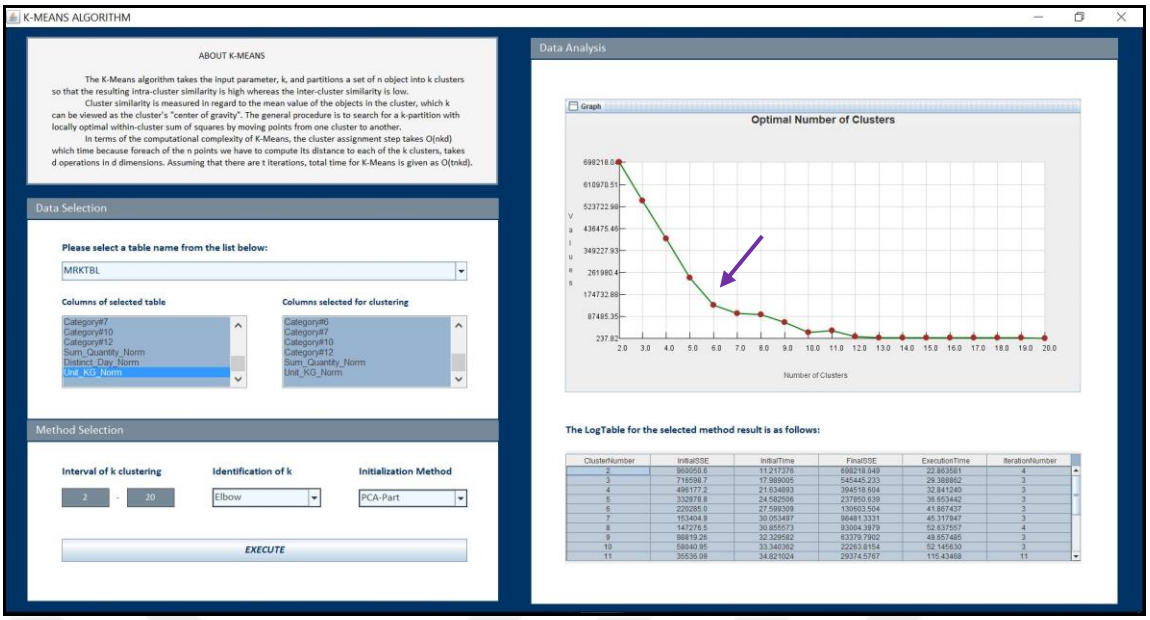
Şekil 6. 11 Dosya 3 – Elbow ve Maximin metotları ile küme sayısının seçimi

Şekil 6. 11 de grafik ve tablo incelendiğinde, Elbow metoduna göre optimal k değerinin 6 olduğu görülmüştür. $k=2$ ile $k=20$ kümeleme aralığında, seçilen Elbow ve Katsavounidis metotları ile elde edilen sonuç aşağıdaki gibidir:



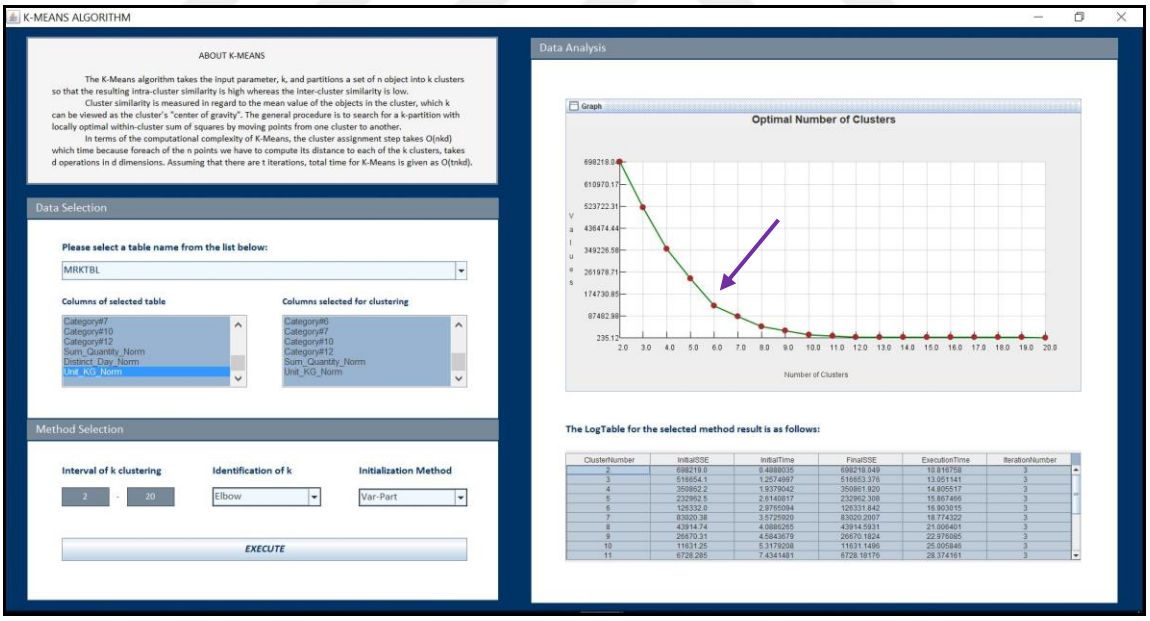
Şekil 6. 12 Dosya 3 – Elbow ve Katsavounidis metotları ile küme sayısının seçimi

Şekil 6. 12 de grafik ve tablo incelendiğinde, Elbow metoduna göre optimal k değerinin 4 olduğu görülmüştür. $k=2$ ile $k=20$ kümeleme aralığında, seçilen Elbow ve PCA-Part metotları ile elde edilen sonuç aşağıdaki gibidir:



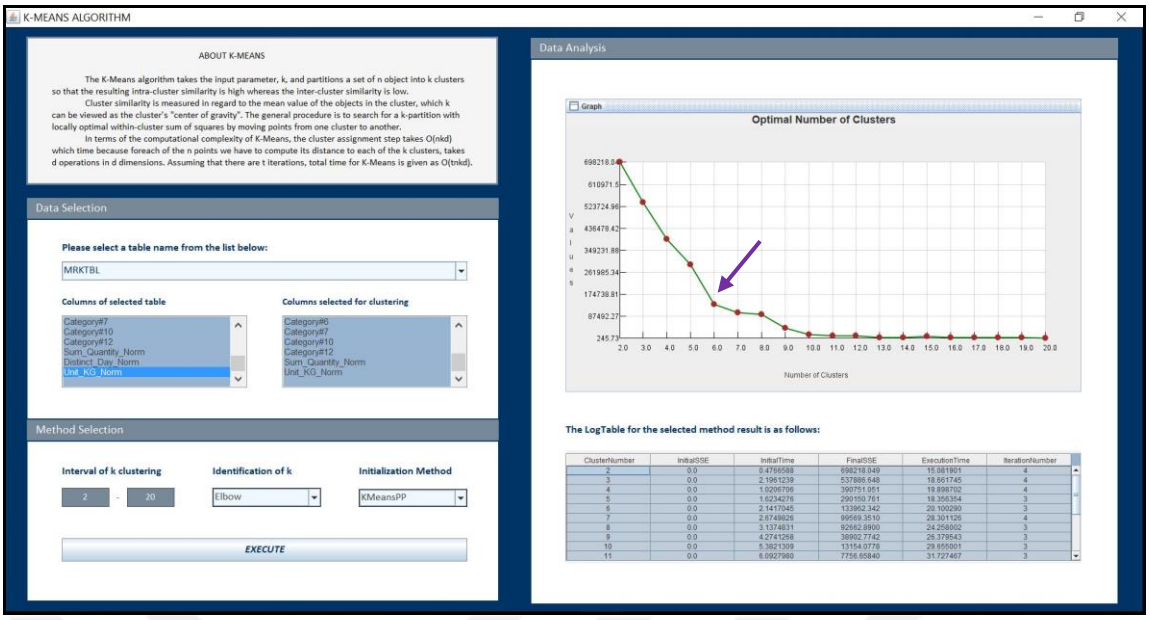
Şekil 6. 13 Dosya 3 – Elbow ve PCA-Part metotları ile küme sayısının seçimi

Şekil 6. 13 te grafik ve tablo incelendiğinde, Elbow metoduna göre optimal k değerinin yine 6 olduğu görülmüştür. $k=2$ ile $k=20$ kümeleme aralığında, seçilen Elbow ve Var-Part metotları ile elde edilen sonuç aşağıdaki gibidir:



Şekil 6. 14 Dosya 3 – Elbow ve Var-Part metotları ile küme sayısının seçimi

Şekil 6. 14 te grafik ve tablo incelendiğinde, Elbow metoduna göre optimal k değerinin yine 6 olduğu görülmüştür. $k=2$ ile $k=20$ kümeleme aralığında, seçilen Elbow ve K-Means++ metotları ile elde edilen sonuç aşağıdaki gibidir:



Şekil 6. 15 Dosya 3 – Elbow ve K-Means++ metotları ile küme sayısının seçimi

Şekil 6. 15 te grafik ve tablo incelendiğinde, Elbow metoduna göre optimal k değerinin yine 6 olduğu görülmüştür. Tüm bu analizler sonucunda, küme sayısı seçiminin $k=4$ ile $k=6$ arasında olacağı görülerek seçim için aralık daraltılmıştır (Çizelge 6.18):

Çizelge 6. 18 Dosya 3 – Küme sayısının seçimi

KÜME SAYISI SEÇİMİ / KÜMELEME SAYISI	ELBOW				
	MAXIMIN	KATSAVOUNIDIS	PCA-PART	VAR-PART	K-MEANS++
2	698.218.050	698.218.050	698.218.050	698.218.050	698.218.050
3	551.082.453	516.653.377	545.445.233	516.653.377	537.886.648
4	429.874.113	**403.025.427**	394.518.605	350.861.920	390.751.051
5	259.248.668	402.784.551	237.850.639	232.962.309	290.150.761
6	**130.508.368**	402.727.601	**130.603.505**	**126.331.843**	**133.962.343**
7	98.194.526	402.704.391	98.481.333	83.020.201	99.569.351
8	52.705.610	402.704.356	93.004.398	43.914.593	92.662.890
9	72.559.196	294.101.352	63.379.790	26.670.182	38.902.774
10	72.935.260	294.083.971	22.263.815	11.631.150	13.154.078
11	7.945.561	294.082.126	29.374.577	6.728.182	7.756.658
12	1.897.033	294.082.088	6.482.096	1.799.892	7.425.121

Çizelge 6. 18 Dosya 3 – Küme sayısının seçimi (devamı)

13	1.658.821	294.081.523	1.579.128	1.194.133	1.658.583
14	1.599.832	294.081.509	1.301.132	872.780	1.590.875
15	1.281.346	294.081.504	646.255	631.903	6.525.423
16	560.429	294.081.273	405.378	393.691	557.771
17	1.016.264	294.081.269	338.652	336.742	818.467
18	566.564	294.081.261	270.844	270.015	336.695
19	274.987	294.081.144	249.517	246.812	270.783
20	267.848	294.081.143	237.828	235.123	245.736

6.4.2 Dosya 3 – Kümelemenin Değerlendirilmesi

$k=4$, $k=5$ ve $k=6$ küme sayıları için analiz sonuçları; başlangıç hata, final hata, iterasyon sayısı, başlangıç merkezlerin seçimi için geçen süre ve toplam süre şeklinde detaylı olarak incelenmiştir.

Küme sayısının seçiminde $k=4$ için sonuç aşağıdaki gibidir (Çizelge 6.19):

Çizelge 6. 19 Dosya 3 – $k=4$ için analiz sonuçları

BAŞLANGIÇ MERKEZLERİN SEÇİMİ İÇİN METOTLAR	BAŞLANGIÇ HATA	FİNAL HATA	İTERASYON SAYISI	BAŞLANGIÇ SÜRE (SN)	TOPLAM SÜRE (SN)
MAXIMIN	0.0	429.874.113	3	0.891899	16.580.130
KATSAVOUNIDIS	0.0	403.025.427	4	0.351332	15.431.961
PCA-PART	496.177.257	394.518.605	3	21.634.894	32.841.241
VAR-PART	350.862.281	350.861.920	3	1.937.904	14.805.518
K-MEANS++	0.0	390.751.051	4	1.020.671	19.898.702

Küme sayısının seçiminde $k=5$ için sonuç aşağıdaki gibidir (Çizelge 6.20):

Çizelge 6. 20 Dosya 3 – $k = 5$ için analiz sonuçları

BAŞLANGIÇ MERKEZLERİN SEÇİMİ İÇİN METOTLAR	BAŞLANGIÇ HATA	FİNAL HATA	İTERASYON SAYISI	BAŞLANGIÇ SÜRE (SN)	TOPLAM SÜRE (SN)
MAXIMIN	0.0	259.248.668	4	1.777.551	23.075.195
KATSAVOUNIDIS	0.0	402.784.551	15	0.669712	56.139.430
PCA-PART	332.878.833	237.850.639	3	24.582.506	36.653.442
VAR-PART	232.962.597	232.962.309	3	2.614.082	15.867.467
K-MEANS++	0.0	290.150.761	3	1.623.428	18.356.355

Küme sayısının seçiminde $k = 6$ için sonuç aşağıdaki gibidir (Çizelge 6.21):

Çizelge 6. 21 Dosya 3 – $k = 6$ için analiz sonuçları

BAŞLANGIÇ MERKEZLERİN SEÇİMİ İÇİN METOTLAR	BAŞLANGIÇ HATA	FİNAL HATA	İTERASYON SAYISI	BAŞLANGIÇ SÜRE (SN)	TOPLAM SÜRE (SN)
MAXIMIN	0.0	130.508.368	3	2.037.342	19.769.024
KATSAVOUNIDIS	0.0	402.727.601	12	1.950.741	52.590.681
PCA-PART	220.285.072	130.603.505	3	27.599.310	41.867.437
VAR-PART	**126.332.068**	**126.331.843**	**3**	**2.976.509**	**16.903.016**
K-MEANS++	0.0	133.962.343	3	2.141.705	20.100.291

Kümeleme değerlendirme kriterlerinde $k = 4$ için sonuç aşağıdaki gibidir (Çizelge 6.22):

Çizelge 6. 22 Dosya 3 – $k = 4$ için kümeleme değerlendirme kriterleri

DEĞERLENDİRME KRİTERLERİ / BAŞLANGIÇ MERKEZLERİN SEÇİMİ İÇİN METOTLAR	SILHOUETTE	CALINSKI-HARABASZ
MAXIMIN	0.450660	2.032.214
KATSAVOUNIDIS	0.618197	2.032.214
PCA-PART	0.622726	3.082.342
VAR-PART	0.652761	3.725.171

Çizelge 6. 22 Dosya 3 – $k = 4$ için kümeleme değerlendirme kriterleri (devamı)

K-MEANS++	0.631937	3.140.917

Kümeleme değerlendirme kriterlerinde $k = 5$ için sonuç aşağıdaki gibidir (Çizelge 6.23):

Çizelge 6. 23 Dosya 3 – $k = 5$ için kümeleme değerlendirme kriterleri

DEĞERLENDİRME KRİTERLERİ / BAŞLANGIÇ MERKEZLERİN SEÇİMİ İÇİN METOTLAR	SILHOUETTE	CALINSKI-HARABASZ
MAXIMIN	0.614024	6.076.827
KATSAVOUNIDIS	0.614024	1.524.464
PCA-PART	0.763169	5.892.979
VAR-PART	0.767903	5.323.106
K-MEANS++	0.645616	3.782.980

Kümeleme değerlendirme kriterlerinde $k = 6$ için sonuç aşağıdaki gibidir (Çizelge 6.24):

Çizelge 6. 24 Dosya 3 – $k = 6$ için kümeleme değerlendirme kriterleri

DEĞERLENDİRME KRİTERLERİ / BAŞLANGIÇ MERKEZLERİN SEÇİMİ İÇİN METOTLAR	SILHOUETTE	CALINSKI-HARABASZ
MAXIMIN	0.606958	7.356.659
KATSAVOUNIDIS	0.606958	1.219.624
PCA-PART	0.867280	1.116.278
VAR-PART	**0.871991**	1.381.422
K-MEANS++	0.761169	1.983.774

Başlangıç merkezlerin seçimi ve kümeleme değerlendirme kriterleri dikkate alındığında; Çizelge 6. 21 de gösterilen final hatanın ve toplam sürenin daha az, diğer yandan Çizelge 6.24 de değerlendirme kriterleri olarak Silhouette değerinin daha yüksek

olduğu görülmüştür. Buradan, veri analizi için en uygun başlangıç merkezlerin seçimi metodu $k = 6$ küme sayısı için Var-Part olarak belirlenmiştir.

6.4.3 Dosya 3 – Kümeleme Sonuçları

Sistem üzerinde $k = 6$ küme sayısı için Elbow ve Var-Part metotları seçilerek, veri dosyası analiz sonuçları veri tabanında bir tabloya kaydedilmiştir.

Kümeleme dağılımlarına bakıldığında dağılımlar aşağıdaki gibidir (Çizelge 6.25):

Çizelge 6. 25 Dosya 3 – Kümeleme dağılımları

KÜME NO	KAYIT SAYISI
1	245.722
2	143.552
3	197.755
4	227.697
5	106.240
6	94.624

Kümeleme sonuçları analiz edildiğinde; toplam satış, ağırlık, zaman ve kategori öz nitelikleri dikkate alınarak gerçekleştirilen kümeleme işleminde, kış sezonunda şarküteri pazarlama kategorisinde yer alan farklı iki markaya ait sade sütlerin en fazla tercih edilen ürünler olduğu görülmüştür. Diğer ürünler ise sırasıyla; kuru gıda kategorisinde bulunan tablet çikolata, meyve suları, sular ve ayçiçek yağı olarak belirlenmiştir. Diğer yandan, yaz sezonunda en çok tercih edilen ürünler, kuru gıda kategorisinde bulunan kolalı içeceklerdir ve meyve sularıdır.

Kış sezonunda en sık tercih edilen ürünler gıda dışı kategorisinde yer alan türkçe gazeteler ve şarküteri pazarlama kategorisinde yer alan sade sütler iken; diğer yandan, yaz sezonunda en sık tercih edilen ürünlerin gıda dışı kategorisinde yer alan türkçe gazeteler ve tütün alkollü içecek kategorisinde yer alan sigaralar olduğu görülmüştür.

Yaz sezonundaki satışların kış sezonuna göre yaklaşık %8 oranında daha fazla olduğu da tespit edilmiştir.

BÖLÜM 7

SONUÇLARIN YORUMLANMASI

Bu tez çalışmasında, K-Ortalamalar algoritmasıyla birlikte farklı metotlar kullanılarak veriler doğru, etkin ve hızlı bir şekilde kümelenmiş ve analiz edilmiştir.

Kümeleme analizi için K-Ortalamalar algoritmasına dayalı yeni bir sistem geliştirilmiştir. Veri analizi için kullanılacak veri dosyasının hazırlık işlemleri MS-SQL veri tabanında yapılarak, tüm analiz sonuçları geliştirilen sistem üzerinden sağlanmıştır. Kümeleme analizinde, küme sayısının seçimi için Elbow metodu; başlangıç merkezlerin seçimi için Maximin, Katsavounidis, PCA-Part, Var-Part ve K-Means++ metotları; kümelemelerin değerlendirilmesi için ise Silhouette ve Calinski-Harabasz metotları kullanılmıştır.

Sonuçlar incelendiğinde Elbow metodunun hata, toplam süre ve iterasyon anlamında en iyi sonuçları K-Means++ ve Var-Part metotlarıyla kullanıldığında verdiği görülmüştür. Sonuçların doğruluğu Silhouette ve Calinski-Harabasz metotlarıyla değerlendirilmiştir ve tüm analizler bu doğrultuda gerçekleştirilmiştir. Çalışmada, 814 sadık müşterinin genel satışlarda en fazla tercih ettikleri markaların, tüm markaların yaklaşık %99 unu; ürünlerin, tüm ürünlerin yaklaşık %98 ini oluşturduğu görülmüştür. Toplam satışlarda en çok tercih edilen kategori şarküteri pazarlamadır; farklı gün öz niteliği seçildiğinde ise kuru gıda olduğu belirlenmiştir. Sezon bazında incelendiğinde kış sezonunda şarküteri pazarlama, yaz sezonunda kuru gıda en fazla tercih edilen kategorilerdir.

Bu tez çalışması ile birlikte, Migros Ticaret A.Ş. nin müşteri-ürün-marka planlamasında etkili olabilecek sonuçlar ve güçlü oldukları ürün-marka-kategori hakkında bilgi sahibi olmaları sağlanmıştır.

BÖLÜM 8

ÖNERİLER

Tez çalışmasında kullanılan veri dosyası üzerinde ihtiyaca göre farklı analizler de yapılabilir. Kümeleme için toplam satış, ağırlık ve kategori öz nitelikleri seçilerek, müşterilerin kategori bazında ne kadar alışveriş yaptıkları ve bununla birlikte ağırlık öz niteliğinin bu analizde ne derece etkili olduğu incelenebilir.

Yine farklı bir analiz olarak; toplam satış, ağırlık, kategori ve zaman (ay) öz nitelikleri seçilerek, müşterilerin tercih ettikleri ürün-marka-kategori grupları ay düzeyinde incelenebilir. Bu analiz ile birlikte müşterinin alışveriş yapma sıklığı da belirlenebilir.

Bu tez çalışmasında kullanılan kümeleme tekniği, veri madenciliği tekniklerinden yalnızca bir tanesidir. Çalışmada kullanılan veri dosyasına uygulanabilecek bir diğer veri madenciliği tekniği ise birliktelik kurallarıdır. Birliktelik kuralları uygulanarak da sadık müşterilerin alışveriş kayıtlarının birbirleriyle olan ilişkileri incelenebilir ve alışveriş sırasında hangi olayların eş zamanlı olarak gerçekleştikleri ve/veya gerçekleşebilecekleri belirlenebilir. Bu bağlamda, birliktelik kurallarından yola çıkılarak müşterilerin alışveriş alışkanlıkları belirlenmeye çalışılabilir.

KAYNAKLAR

[1] Özkan, Y., (2013). Veri Madenciliği Yöntemleri, İkinci Baskı, Papatya Yayıncılık Eğitim, İstanbul.

[2] Savaş, S., Topaloğlu, N. ve Yılmaz, M., (2012). "Veri Madenciliği ve Türkiye'deki Uygulama Örnekleri", İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi, 21: 1–23.

[3] Liao, S., Chu, P. ve Hsiao, P., (2012). "Data Mining Techniques and Applications – A decade review from 2000 to 2011", Expert Systems with Applications, 39: 11303–11311.

[4] Berry, M.J.A. ve Linoff, G.S., (2004). Data Mining Techniques for Marketing, Sales, and Customer Relationship Management, Second Edition, John Wiley and Sons, Canada.

[5] Doğan, O., (2017). "Türkiye'de Veri Madenciliği Konusunda Yapılan Lisanüstü Tezler Üzerine Bir Araştırma", Gazi Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 19(3): 929–951.

[6] Han, J., Kamber, M. ve Pei, J., (2012). Data Mining: Concepts and Techniques, Third Edition, Morgan Kaufmann Publishers, San Francisco.

[7] Everitt, B., (2011). Cluster Analysis, Fifth Edition, John Wiley and Sons, Londra.

[8] Pasin, Ö., (2015). Sağlık Alanında Yapılan Araştırmalarda Kümeleme Algoritmalarının Kullanımı: Bir Uygulama, Yüksek Lisans Tezi, Düzce Üniversitesi Sağlık Bilimleri Enstitüsü, Düzce.

[9] Zaki, M. J. ve Meira, W., (2014). Data Mining and Analysis Fundamental Concepts and Algorithms, Cambridge University Press, New York.

[10] Gartner Group, IT Glossary, https://www.gartner.com/it-glossary/data-mining, 10 Mayıs 2018.

[11] Larose, D.T., (2005). Discovering Knowledge in Data: An Introduction to Data Mining, John Wiley and Sons, New Jersey.

[12] Chakrabarti, S., Cox. E., Frank, E., Güting, R. H., Han J., Jiang, X., Kamber, M., Lighstone, S. S., Nadeau, T. P., Neapolitan R. E., Pyle, D., Refaat, M., Schneider,

M., Teorey, T. J. ve Witten I. H., (2009). Data Mining Know It All, First Edition, Morgan Kaufmann.

[13] Pyle, D., (1999). Data Preparation for Data Mining, Morgan Kaufmann Publishers, San Francisco.

[14] Kantardzic, M., (2003). Data Mining: Concepts, Models and Algorithms, IEEE Press and John Wiley, New York.

[15] Alpar, R., (2017). Uygulamalı Çok Değişkenli İstatistiksel Yöntemler, Beşinci Baskı, Detay Yayıncılık, Ankara.

[16] Tuffery, S., (2011). Data Mining and Statistics for Decision Making, John Wiley and Sons, London.

[17] Jain, A.K., (2010). "Data clustering: 50 years beyond K-means", Pattern Recognition Letters, 31: 651–666.

[18] Ayramo, S. ve Karkkainen, T., (2006). "Introduction to partitioning-based clustering methods with a robust example", Reports of the Department of Mathematical Information Technology Series C. Software and Computational Engineering, No.1.

[19] Sakthivel, E. ve Kannan, K.S., (2013). "Clustering Algorithms using Different Distance Measures", CiiT International Journal Data Mining and Knowledge Engineering, 5(4): 140–143.

[20] Bora, D.J. ve Dr. Gupta, A.K., (2014). "Effect of Different Distance Measures on the Performance of K-Means Algorithm: An Experimental Study in Matlab", International Journal of Computer Science and Information Technologies, 5(2): 2501–2506.

[21] Singh, A., Yadav, A. ve Rana, A., (2013). "K-means with Three Different Distance Metrics", International Journal of Computer Applications (0975 – 8887), Vol. 67, No. 10.

[22] Gürcü, Ö., (2014). Kümeleme Analizi, https://prezi.com/_qvp7mj981sx/kumeleme-analizi/, 24 Nisan 2018.

[23] Charrad, M., Ghazzali, N., Boiteau, V. ve Niknafs, A., (2014). "NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set", Journal of Statistical Software, Vol. 61, Issue 6.

[24] Kodinariya, T.M. ve Makwana, P.R., (2013). "Review on determining number of Cluster in K-Means Clustering", International Journal of Advance Research in Computer Science and Management Studies, www.ijarcsms.com, ISSN: 2321–7782.

[25] Atbaş, A.C.G., (2008), "Kümeleme Analizinde Küme Sayısının Belirlenmesi Üzerine Bir Çalışma", Yüksek Lisans Tezi, Ankara Üniversitesi Fen Bilimleri Enstitüsü, Ankara.

[26] Rezaei, M., (2016). Clustering Validation, Doktora Tezi, Publications of the University of Eastern Finland Dissertations in Forestry and Natural Sciences, No. 225, East Finland.

[27] Elbow Method (Clustering), https://en.wikipedia.org/wiki/Elbow_method_(clustering), 22 Ocak 2018.

[28] Calinski, T. ve Harabasz, J., (1974). "A dendrite method for cluster analysis", Communications in Statistics, 3(1): 1–27.

[29] Krzanowski, W.J. ve Lai, Y.T., (1988). "A Criterion for Determining the Number of Groups in a Data Set Using Sum-of-Squares Clustering", International Biometric Society, 44(1): 23–34.

[30] Rousseeuw, P.J., (1987). "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis", Journal of Computational and Applied Mathematics, North-Holland, 53–65.

[31] Tibshirani, R., Walther G. ve Hastie, T., (2001). "Estimating the number of clusters in a data set via gap statistic", Royal Statistical Society, 411–423.

[32] Çelebi, M.E., Kingravi, H.A. ve Vela, P.A., (2013). "A Comparative Study of Efficient Initialization Methods for the K-Means Clustering Algorithm", Expert Systems with Applications, 40(1): 200–210.

[33] Çelebi, M.E. ve Kingravi H.A., (2014). "Linear, Deterministic, and Order-Invariant Initialization Methods for the K-Means Clustering Algorithm", Partitional Clustering Algorithms, Springer, 79–98.

[34] Katsavounidis, I., Kuo, C.-C.J ve Zhang, Z., (1994). "A New Initialization Technique for Generalized Lloyd Iteration", IEEE Signal Processing Letters, Vol. 1, No. 10, October.

[35] Shlens, J., (2014). "A tutorial on Principle Component Analysis", Cornell University Library.

[36] Anbazhagan, N., (2011). "Improving the Performance of K-Means Clustering for High Dimensional Data Set", International Journal on Computer Science and Engineering (IJCSE), Vol. 3, No. 6.

[37] Smith, L.I., (2002). "A tutorial on Principal Component Analysis", February 26.

[38] Su, T. ve Dy J.G., (2007). "In Search of Deterministic Methods for Initializing K-Means and Gaussian Mixture Clustering", Intelligent Data Analysis, 11(4): 319–338.

[39] Peres-Neto, P.R., Legendre, P., Dray, S. ve Borcard, D., (2006). "Variation Partitioning of Species Data Matrices: Estimation and Comparison of Fractions", Ecological Society of America, 87(10): 2614–2625.

[40] Arthur, D. ve Vassilvitskii, S., (2007). "K-Means++: The Advantages of Careful Seeding", Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, 1027–1035.

[41] Arthur, D. ve Vassilvitskii, S., (2006). "How Slow is the k-Means Method?", Proceedings of the twenty-second annual symposium on Computational geometry, 144–153.

[42] Bahmani, B., Moseley, B., Vattani, A., Kumar, R. ve Vassilvitskii, S., (2012). "Scalable k-means++", Proceedings of the VLDB Endowment, 5(7): 622–633.

[43] Curtis, G. ve Cobham, D., (2009). Business Information Systems: analysis, design and practice, Sixth Edition, FT Press, İngiltere.

[44] Çınar, A. ve Silahtaroğlu, G., (2012). "Veri Madenciliği Teknikleri ile Müşteri Memnuniyetine Etki Eden Gizli Nedenlerin Keşfi", Marmara Üniversitesi, 309-330.

[45] Petre, R., (2013). "Data Mining Solutions for the Business Environment", Database Systems Journal, 4(4): 21–29.

[46] Kusrini, K., (2015). "Grouping of Retail Items by Using K-Means Clustering", The Third Information Systems International Conference, 72(2015): 495–502.

[47] Elmasri, R. ve Navathe, S.B., (2003). Fundamentals of Database Systems, Fourth Edition.

[48] Fisher, R.A., (1936). "The Use of Multiple Measurements in Taxonomic Problems", Annals of Eugenics, 7: 179–188.

[49] İris Veri Dosyası, https://archive.ics.uci.edu/ml/datasets/iris, 12 Şubat 2018.

ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Adı Soyadı : Merve ÜSTÜNEL

Doğum Tarihi ve Yeri : 09.05.1992 / Üsküdar

Yabancı Dili : İngilizce

E-posta : ustunelmerve@gmail.com

ÖĞRENİM DURUMU

Derece	**Alan**	**Okul/Üniversite**	**Mezuniyet Yılı**
Lisans	Matematik	Yıldız Teknik Üniversitesi	2014
Lise	Fen Bilimleri	Hayrullah Kefoğlu Anadolu Lisesi	2010

İŞ TECRÜBESİ

Yıl	**Firma/Kurum**	**Görevi**
2016 – devam ediyor	EY Kurumsal Finansman Danışmanlık A.Ş.	Teknoloji Danışmanı

YAYINLARI

Makale

1. Saylı, A., Ozturk, I., Ustunel, M. (2016). "Brand Loyalty Analysis System using K-Ortalamalar Algorithm", Journal of Engineering Technology and Applied Sciences, 1 (3): 107-126.

ÖDÜLLERİ

1. ITIL® Foundation Certificate in IT Service Management (2017). PeopleCert, GR750303822MU.