T.C. İSTANBUL KÜLTÜR ÜNİVERSİTESİ

LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

K-MEANS ALGORİTMASININ SİGORTA SEKTÖRÜNDE UYGULANMASI

YÜKSEK LİSANS TEZİ

Emre Akgöz

1510010405

Anabilim Dalı: İşletme

Programı: İşletme Uzaktan Eğitim

Tez Danışmanı: Prof. Dr. Müge Çetiner

MAYIS 2019

T.C. İSTANBUL KÜLTÜR ÜNİVERSİTESİ

LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

K-MEANS ALGORİTMASININ SİGORTA SEKTÖRÜNDE UYGULANMASI

YÜKSEK LİSANS TEZİ

Emre Akgöz

1510010405

Anabilim Dalı: İşletme

Programı: İşletme Uzaktan Eğitim

Tez Danışmanı : Prof. Dr. Müge Çetiner
Jüri Üyeleri : Doç. Dr. Ceyda Aysuna Türkyılmaz
Dr. Öğr. Üyesi Çağla Arıker

MAYIS 2019

ÖZET

Veri madenciliği bir veri ambarında, veriler üzerindeki gizli bağlantıları ortaya çıkarmak için kullanılmaktadır. Verinin bulunduğu her ortamda ham veriler anlamlı hale getirilerek bilgi keşfi işlemine veri madenciliği denir.

Günümüzde ilerleyen teknoloji ve bilgiye erişiminin artması ile birlikte kurum ve kuruluşlar bu bilgilerden faydalanarak üretim, satış ve kaynakların yönetilmesi gibi konularda daha doğru, düzgün ve yeni kararlar verebilmek için önemli bir yol olarak veri madenciliğinden yararlanmanın faydalı olabileceğini düşünmüşlerdir. Bununla birlikte veri madenciliğinde yeni yöntemler, algoritmalar, düşünceler gelişmiş ve sektörlerin için önem arz eden bir durum haline gelinmiştir.

Bilgi miktarının büyük oranlarda arttığı bu bilgi çağında büyük hacimlerdeki verilerden anlamlı bilgilerin elde edilmesi bir süreç gerektirmektedir. Bu sürecin en önemli adımı ise veri madenciliğidir.

Bu çalışmada bilginin ortaya çıkarılması süreci araştırılmış ve incelenmiştir. Yine aynı şekilde sürecin en önemli adımı olan veri madenciliği adımı da ayrıntılı olarak incelenmiştir.

Tez içerisinde anlamlı bilgiye ulaşma süreci aşamalarıyla açıklanmıştır. Sürecin en önemli adımı olan veri madenciliği adımı ayrıntılı olarak ele alınmış ve veri madenciliği tekniklerinden bahsedilmiştir. Veri madenciliği tekniklerinden olan kümeleme analizi ayrıntılı olarak incelenmiştir

Tezin son kısmında veri madenciliği algoritmalarından biri olan k-means algoritması sigorta veri tabanına uygulanmış ve sonuçlar ortaya konulmuştur.

ABSTRACT

Data mining is used to find hidden connections on the data via data warehouse.Knowledge discovery means gathering meaningful data, on any environment where the data is found, the process is called data mining.

With the increasing technology and access to information nowadays, institutions and organizations have benefited from this information and thought to benefit from data mining as an important way to make more accurate, smooth and new decisions on issues such as production, sales and management of resources. In addition, new methods, algorithms, ideas developed in data mining have become important for the sectors.

In this information age, where the amount of information has increased greatly, it is necessary to obtain meaningful information from large volumes of data. The most important step of this process is data mining.

In this study, the process of revealing information has been researched and examined. Likewise, the most important step of the process, the data mining step, has also been examined in detail.

The process of reaching meaningful information in the thesis is explained with the stages. Data mining step, which is the most important step of the process, has been discussed in detail and data mining techniques are mentioned. Clustering analysis, which is one of the data mining techniques, has been examined in detail.

In the last part of the thesis, k-means algorithm, which is one of the data mining algorithms, is applied to the insurance database and the results are presented.

İÇİNDEKİLER

İKİNCİ BÖLÜM
KÜMELEME ANALİZİ

ÜÇÜNCÜ BÖLÜM

UYGULAMA:BÖLÜMLEMELİ YÖNTEMLERDEN K-MEANS ALGORİTMASININ SİGORTA SEKTÖRÜNDE UYGULANMASI

KISALTMALAR

TDK :Türk Dil Kurumu
VTYS :Veri Tabanı Yönetim Sistemleri
MS : Microsoft
SQL :Yapılandırılmış Sorgulama Dili (Structured Query Language)
VM :Veri Madenciliği
OLAP :Çevrim İçi Analitik İşleme (Online Analytical Processing)
ARFF :İlişkilendirilmiş Dosya Formatı(Relationship File Format)
IBM :Uluslararası İş Makineleri (International Business Machines)
BKZ :Bakınız
NO :Numara
S. :Sayfa
DVYTS :Dağıtık Veri Tabanı Yöntemi Sistemleri
İVYTS :İlişkisel Veritabanı yönetim sistemi

TABLO LİSTESİ

ŞEKİL VE GRAFİK LİSTESİ

ÖNSÖZ

Günümüzde binlerce veri büyük boyutlarda veri tabanlarını oluşturmaktadır. Bu veri yığınlarındaki gizli bilgilerin açığa çıkarılmasını sağlayan ve geniş bir kullanım alanı olan veri madenciliği teknikleri oldukça önem kazanmıştır. Bu nedenle tez çalışmasında yöntem olarak veri madenciliği yöntemleri tercih edilmiştir.

Veri madenciliği teknikleri birçok farklı alanda faydalı bilgiler sunmaktadır. Bu tez çalışmasında veri madenciliği yöntemleri diğer alanlardan farklı olarak sigortacılık sektöründe kullanılmıştır.Bu kapsamda müşteri memnuniyetsizliğin altında yatan ana ve gizli nedenler ortaya çıkartılmaya çalışılmıştır.İncelenen veriler üzerinde K-Means algoritması kullanılmış olup grafiksel analiz yapılmıştır.Çalışma sonucunda tutarlı sonuçlar elde edilmiştir.

Bu çalışmanın hazırlanmasında destek veren tez danışmanım Prof. Dr. Müge Çetiner'e, uygulama sırasında yardımlarından dolayı değerli yöneticim Eureko Sigorta Analitik Sigortacılık Direktörü Özlem Odar 'a , her daim desteklerinden ötürü aileme ve sevgili eşime teşekkürlerimi sunarım.

GİRİŞ

Son yıllarda bilişim teknolojilerinde meydana gelen gelişmeler oldukça hızlı bir şekilde ilerlemektedir. Bu ilerlemelerden birçok alanda yararlanılmaktadır. Yapılan her işlemin bilişim teknolojilerinin ilerlemesiyle kayıt altına alınabilmesi ve biriktirilmesi sonucu belli yöntem ve tekniklerle yaralı bilgilere ulaşılmaktadır. Bilgi ve iletişim teknolojilerinin ilerlemesiyle elde edilen çok sayıda veri içinde işletmeler, ülkeler, kurum ve kuruluşlar vb. için gizli ve faydalı bilgiler yer alabilmektedir. Özellikle bilişim alanında yer alan birçok matematiksel ve istatistiksel yöntem ve teknikler kullanılarak bu gibi birimler için birçok önemli bilgiler sunabilmektedir. Büyük miktardaki verilerin işlenmesiyle gizli bilgileri ve ilişkileri ortaya çıkaran veri madenciliği, bilişim alanındaki yöntemlerden biridir.

Günümüzde oldukça yaygın kullanım alanına sahip olan veri madenciliği teknikleri veri tabanlarının içinde bulunan önemli bilgileri kullanıcılara sunmaktadır. Ayrıca veriler arasındaki görülemeyen ilişkileri de ortaya çıkaran bir tekniktir. Veri madenciliği veri tabanlarındaki yararlı bilgiyi ortaya çıkarırken birçok istatistiksel ve matematiksel yöntemleri kullanmaktadır. Veri madenciliği birçok veriyle işlevini yerine getirirken karar ağaçları ve çok değişkenli istatistiksel yöntemlerden de faydalanmaktadır. Bu yöntemlerin başında kümeleme analizi, diskriminant analizi, regresyon analizi gelmektedir.

Veri madenciliği araçlarıyla ortaya çıkarılan bilgi birçok farklı alanda yaygın olarak kullanılmaktadır. Pazarlama, finans, mühendislik, lojistik, uzay bilimleri, tıp ve psikoloji gibi alanlar veri madenciliğinin kullanım alanlarından bazılarıdır. Bu alanların yanı sıra veri madenciliği tekniklerinin kullanımı yeni alanlarda da denenmektedir. Bu tez çalışmasında bahsedilen alanlardan finans alanlarından sigorta sektöründe veri madenciliği teknikleri uygulanmıştır.

Tez çalışmasının amacı veri madenciliği tekniklerinin ele alınarak; temel sigorta değişkenlerin ve müşteri demografik değişkenleri kullanılarak müşteri memnuniyetsizliğin altında yatan ana ve gizli nedenlerin ortaya çıkartılmaya çalışılması ve sonucunda anlamlı bilginin analiz edilmesidir.

Tez çalışmasının ilk bölümünde veri madenciliği kavramı üzerinde durularak süreci göz önüne alınmaktadır. Ayrıca veri madenciliği alanında yapılmış çalışmalar incelenerek, yöntemleri ayrıntılı bir şekilde ele alınmaktadır.

Tezin ikinci bölümünde ise veri madenciliği tekniklerinden ve uygulama kısmında da kullanılacak olan kümeleme analizine yer verilmektedir. Bununla birlikte kümeleme analizinin amacı, kullanımı ve yöntemleri de açıklanmaktadır..

Tezin üçüncü bölümü ise uygulama kısmı olup öncelikle tüm değişkenler ve uygulama ortamı incelenmektedir.Veri madenciliği tekniklerinden K-Means algoritması kullanılarak birbirine benzer nitelikte müşterilerin özellikleri kümelenmiştir.

BİRİNCİ BÖLÜM
VERİ MADENCİLİĞİ

1.1.Veri

Veri, veri madenciliğinin temel yapıtaşıdır. Veri olmadan veri madenciliği yapılamaz bu sebeple veriyi tanımlamak önemlidir.

TDK (2018) veriyi "bir araştırmanın, bir tartışmanın, bir muhakemenin temeli olan ana öğe, sonuç çıkarmak, çıkarsama yapmak, ya da bir incelemeyi sürdürmek için gerekli olaylara, ilişkilere ve sayısal ham bilgilere verilen ad olarak tanımlamıştır.

Veri bilginin ham maddesi, bilgiyi oluşturan en küçük parça, enformasyon oluşumundaki yapıtaşıdır. [1]

Genel anlamda veri bağımsız ve soyut gerçeklerin, ölçümlerin, sembol ve sayısal karakterlerin bir dizisidir. Veri işlenmeden tek başına bir anlam ifade etmez. Bir verinin anlamlandırılması için analiz edilerek yorumlanması gerekir.[2] Veri madenciliğinin ana öğesi olan verinin analiz edilebilmesi için ise bilişim teknolojileri vasıtasıyla kayıt altına alınması önemlidir.

1.1.1. Veri Tabanı Yönetim Sistemleri

Bir veri tabanına veri eklenebilir veya silinebilir. Bu tip işlemlerin yapılması için VTYS'ye ihtiyaç vardır. VTYS'leri veri tabanını yaratmak, tanımlamak, kullanmak, değiştirmek vb. işlemlerin yapılması amacıyla hazırlanmış yazılımlardır. Yeni bir veritabanı oluşturmak, geliştirmek, düzenlemek, bakımını yapmak gibi karmaşık işlemlerin yapıldığı yazılımlar VTYS olarak adlandırılmaktadır .[3]

En bilinir VTYS yazılımları[4]:

- MS (Microsoft) SQL Server,
- My SQL,
- Oracle,

[1] Atlı, 2014, s.632
[2] Yılmaz, 2013, s.243
[3] Vural ve Sağıroğlu, 2010, s.73
[4] Kaplan vd., 2017; Abrar ve Wahyudi, 2016; Yağcı vd., 2015; Deperlioğlu ve Sarpkaya, 2009

•Access ‘dir.

Veri madenciliğinin gerçekleştirilebilmesi için verinin kayıt altına alınması gerekir. Bu noktada VTYS yazılımları önemlidir. VTYS ile belli bir düzene göre kayıt altına alınan veri madenciliği uygulamasının en temel öğesi olan veriyi kayıt altına alır.

1.2.Veri Madenciliği Tanımı ve Gelişimi

Gelişen dünyada sürekli bilgi oluşumu ve akışı meydana gelmektedir. Son yıllarda kullanılabilir bilginin artması, kullanıcılar tarafından bu bilgilerin stoklanarak daha faydalı bilgiler oluşturmasını sağlamaktadır. Günümüzde biriken bilgi stokları birçok veri tabanını oluşturmakla birlikte bu verilerin dijital ortamda saklanması veri tabanlarının oluşumunu daha da hızlandırmaktadır. Dijital ortamlarda oluşturulan bu veri tabanları içerisinde, işlendiğinde ortaya çıkacak birçok faydalı bilgi yer almaktadır. Veri tabanları arasındaki gizli ve faydalı bilgileri ortaya çıkarmak giderek daha da önem kazanmaktadır. Bu noktada veri madenciliği kavramı devreye girmekledir ve birçok farklı yöntemle gizli ve faydalı bilgileri ortaya çıkarmaktadır.

1990'ların ortasından beri veriden daha fazla bilgi elde etmeyi vurgulayan birbiriyle ilgili yeni alan bilgi sistemlerinde ve bilgi teknolojisinde meydana gelmiştir. Bu alanlar veri ambarı, bilgi yönetimi ve veri madenciliğidir. Bu alanlardan veri madenciliği, verilerdeki kabul edilebilir, özgün, potansiyel fayda sağlayan ve açık olan korelasyonları ve modelleri tanımlamayı amaçlamaktadır. Bilgisayar donanımı ve yazılımındaki ilerlemelerle birlikte birçok veri madenciliği uygulamaları önceki zamanlardansa şimdi firmalara daha erişilebilir ve daha satın alınabilir özelliktedirler.[5]

Veri madenciliği büyük miktarlardaki veriyi farklı metotlarla hem anlaşılabilir hem de faydalı hale getiren ve beklenmedik ilişkileri bulmak için gözlemsel veri setlerinin analiz edilmesini sağlayan bir yöntemdir.[6] Daha önceleri tespit edilememiş, geçerliliği olan ve uygulanabilen bilgilerin çoğunun büyük veri tabanlarından temin edilmesi ve bu bilgilerin

[5] Hian Chye Koh, Wei Chin Tan, Goh Chwee Peng, "Credit Scoring Using Data MiningTechniques", Singapore Management Review 26.2., 2004, s. 101.

6 David J. Hand, Heikki Mannila, Padhraic Smyth, Principles of Data Mining, MIT Press, 2001, s. 6.

organizasyon kararları alınırken kullanılmasıdır. Burada üzerinde durulması gereken olan önemli noktalardan biri temin edilecek bilgiler hakkında daha önceleri bilinmezlik durumunun olmasıdır.

Literatür incelendiğinde çeşitli yaklaşımlarla veri madenciliği tanım ve sürecinin açıklanmaya çalışıldığı gözlemlenebilmektedir.[7] Bu tanımlardan birkaçı aşağıda verilmiştir.

Gartner Group şirketine göre, "Veri madenciliği depolarda (veri ambarında) saklanan çok büyük boyuttaki verileri inceleyerek anlamlı yeni korelasyonların, örüntülerin ve eğilimlerin keşfedilmesi sürecidir." [8]

"Verilerin içerisindeki örüntülerin, ilişkilerin, değişimlerin, düzensizliklerin, kuralların ve istatistiksel olarak önemli olan yapıların keşfedilmesidir" .[9]

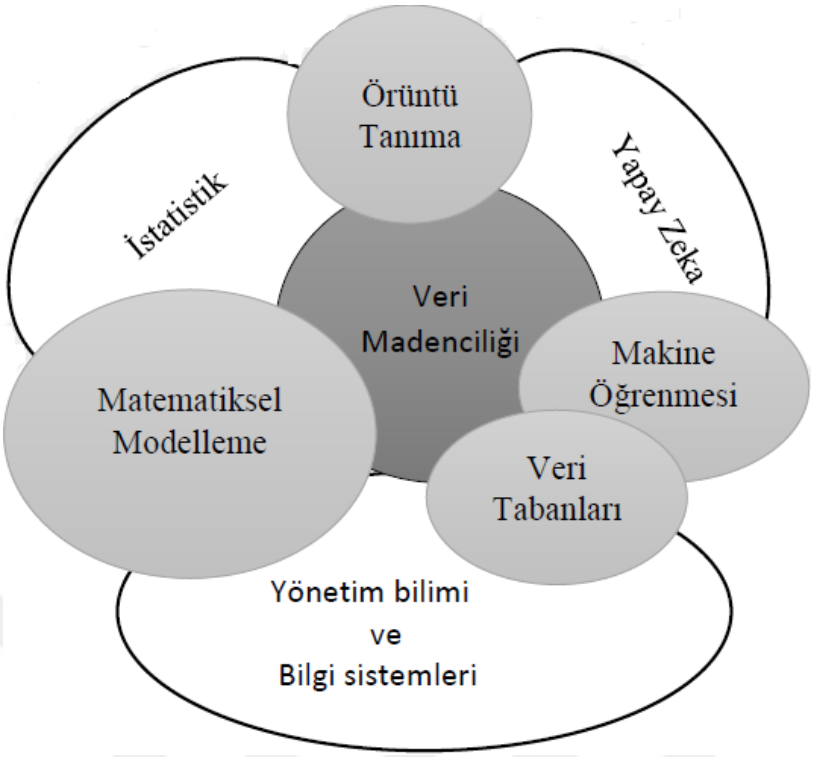
Pirinççiler ve Şen'e göre ise veri madenciliği; "Büyük miktarlardaki verinin otomatik ya da yarı otomatik araçlarla, veri içerisindeki kullanışlı desenleri (model) ortaya çıkarmak için yapılan keşif ve analizdir" .[10]

"Veri madenciliği büyük veri yapılarından değerli bilgilerin çıkartılması süreci/ yöntemleri olarak tanımlanabilir".[11]

Savaş ve arkadaşları ise veri madenciliği için "Veri madenciliği, çok büyük miktarda bilginin depolandığı veri tabanlarından, amacımız doğrultusunda, gelecek ile ilgili tahminler yapmamızı sağlayacak, anlamlı olan veriye ulaşma ve veriyi kullanma işidir." tanımını yapmışlardır.[12]

Veri madenciliğinin gelişim süreci incelendiğinde veri madenciliği yöntemlerinin 1980'li yılların sonundan itibaren uygulanmaya başlandığı görülmektedir. 1990'lı yıllara gelindiğinde veri tabanı sayısında beliren artış ile bu veri tabanlarından yararlı bilginin nasıl elde edileceği düşünülmeye başlanmıştır. 2000'li yıllarda veri madenciliği istikrarlı bir gelişim göstererek birçok alanda kullanılmaya başlanmıştır. [13] Veri madenciliğinin gelişimine farklı birçok disiplinin de katkısı olduğundan veri madenciliği diğer disiplinlerle ilişki içerisindedir. Söz konusu ilişki Şekil 1'de şematize edilmiştir.

7 Balaban & Kartal, 2016.
8 Larose,2005 ,Çataloluk, 2012.
9 Ulucan & Pektekin, 2009.
10 Pirinççiler & Şen, 2012.
11 Altunkaynak, Veri Madenciliği Yöntemleri ve R Uygulamaları, 2017
12 Savaş, Topaloğlu, & Yılmaz, 2012.
13 Savaş, Topaloğlu, & Yılmaz, 2012.



Şekil 1.1. Veri Madenciliği Ve Diğer Disiplinler Arasındaki İlişkiler.[14]

Üzerinde veri madenciliği yöntemleriyle analizler yapılacak bu "alt veri kümesine" veri ambarı denilmektedir.[15] Fakat bazen büyük veri ambarları oluşturmak çok uzun zaman alan ve oldukça maliyetli bir işlem olabilir. Böyle bir durumda veri madenciliği yöntemlerini uygulamak için bir veya daha fazla operasyonel veya işlemsel veri tabanından salt okunabilir bir veri tabanı oluşturmak yeterlidir[16].

[14] (Turban, Sharda, & Delen, 2011)
[15] Altunkaynak, 2017
[16] Two Crows, 1999

1.3. Veri Madenciliğinin Amacı

Veri madenciliği, büyük yoğunluktaki tarihsel veri üzerinde analiz ve kümeleme sonucunda, anlamlı örüntülerin ve ilişkilerin keşfedilme sürecidir.[17]Veriler işletimsel sistemlerde iş süreçlerini düzgün biçimde yürütmek amaçlı tutulur.İşletimsel sistemlerde analiz raporlama vb. süreçler yürütülmemelidir. Verilerdeki gizli bağlantıları, ilişkisel anlamlı örüntüleri bulma süreçlerinin geneline veri madenciliği denir.Veri madenciliği bir veri ambarı sistemi üzerinde yapılır.

Veri madenciliği kavramı, onlarca yıl öncesinde ortaya çıkmış olmasına karşın kullanımı her işlemin dijital cihazlar ile kayıt edilebildiği - son yıllarda yaygınlaşmıştır.[18] Günümüzde dünya üzerinde oluşan veri miktarının her geçen gün arttığı görülmektedir. Bilişim teknolojilerindeki gelişim ve teknolojik cihazların bellek kapasitelerindeki artış daha çok veriyi saklamaya olanak tanımaktadır. Bu sebeple artan verinin analiz edilerek işlenmesi karar merkezleri için önemli hale gelmiştir.[19] Veriler yalnızca belli bir amaç doğrultusunda işlenirse anlamlıdır. Bu amaç doğrultusunda günümüzde en çok kabul gören yöntem veri madenciliğidir.[20]Veri madenciliği, veri tabanlarında depolanan veriden, yararlı ve aralarında ilişki olabilecek veriyi keşfederek, bu verinin anlaşılır ve kullanılabilir bilgilere dönüştürülmesine olanak tanıyan yöntemler topluluğudur.

1.4. Veri Madenciliğinin Uygulama Alanları

Verinin bulunduğu her yerde veri madenciliği yapılabilir. Veri madenciliğinin amacı verilerin arasındaki gizli bağlantıları bulabilmek, veriyi daha anlamlı hale getirebilmektir. İşletimsel sistemlerde veriler ilişkisel olarak tutulsa dahi verinin tüm yapı içerisindeki hareketini ve anlamını ortaya çıkarmak bir veri madenciliği sürecidir. Burada önemli olan veri madenciliği sürecinde veriye en kapsamlı halde hakim olmaktır.

Veri madenciliğinin faydalarının anlaşılmasıyla kullanım alanı da genişlemiştir. Çok miktarda veri içeren her alanda veri madenciliği kullanılabilir. Tıp, pazarlama, telekomikasyon, bankacılık, eğlence sektörü vb. birçok alan veri madenciliği kullanım

17 Seidman,2001.
[18] Silahtaroğlu ve Ergül, 2016, s.103-104
[19] Albayrak, 2017, s.751-752z
[20] Çeşmeli vd., 2015, s.37

alanlarına örnek gösterilebilir. Aşağıdaki veri madenciliği kullanan bazı alanların hangi konular için veri madenciliğine ihtiyaç duydukları gösterilmiştir.[21]

Veri Madenciliği Uygulama Alanı	Uygulama Konusu
TIP	Organ nakli gereken durumlarda donör-organ uyumu tahmininde
	Genom biliminde
Bilgisayar Donanımı ve Yazılımı	Disket sürücüleri hatalarını öngörmek
	İstenmeyen web içeriklerini ve e-mailleri saptama ve filtrelemek
	Güvenli olmayan yazılım ürünlerini belirlemek vb.
Pazarlama	Müşterilerin elde tutulması ve yeni müşterilerin kazanılması
	En çok kazanç sağlanan müşterilerin saptanması vb.
Bankacılık & Sigortacılık	Kredibilite ve risk değerlendirilmesi
	Online banka işlemleri ve kredi kartında yapılan sahtekarlık ve dolandırıcılık tespiti vb.
	Sigorta hasar rezervlerini hesaplama,hasar anında müşteri memnuniyetsizliğin altında yatan nedenlerin tespiti,kaza-hasar adetine göre müşteri skorlama
Üretim	Makine hatalarının önceden tahmini
	Üretim kapasitesinin optimizasyonu amacıyla üretimdeki normal dışlıkların ve ortaklıkların belirlenmesi
	Üretim kalitesinin belirlenmesi ve geliştirilmesi adına farklı örüntüler keşfetmek vb.
Turizm	Gelir yönetimi yapmak
	Talep tahmini ile sınırlı kaynakların doğru yerlere tahsisi
	En çok kazanç sağlayan müşterileri saptayarak kalıcılıklarını sağlamak

Tablo 1.1. Veri Madenciliğinde Kullanılan Bazı Alanlar

[21] Turban, Sharda, & Delen, 2011

1.5. Veri Madenciliğinin Geleceği

Veri madenciliği yeni ve gelecek vaat eden bir teknolojidir. Dünya üzerindeki veri miktarları her yıl daha da büyümektedir. Yakın geçmişte terabayte ile ölçülen bir verinin büyük miktarda bir veri olduğu düşünülürken artık günümüzde petabyte boyutlu bir verinin daha büyük olduğu kabul edilmektedir. Ancak bunun da ötesinde exebayte ve zetabyte veri ölçü birimleri göz önünde bulundurulmaya başlanmıştır.[22]

Sosyal medyanın yaygınlaşması ve insanların mobil cihazlar sayesinde dijital dünyaya daha çok dâhil olması, büyük miktarda verinin dijital olarak depolanmasını sağlamıştır. Kullandığımız birçok mobil uygulama vasıtasıyla depolanan bu veri, dinlediğimiz müzikten sevdiğimiz restorana, aldığımız ayakkabıdan gittiğimiz şehirlere kadar birçok konuyu içerebilmektedir. Bu verilerin kayıt altına alınması, gelişen mobil uygulamalar ve hızlanan internet teknolojisi sayesinde çok kolaylaşmıştır ve bu verilerin daha da artacağı ön görülmektedir.

Amerika birleşik devletlerinde IDC Analyze The Future tarafından yapılan dijital evren çalışmalarına göre dünya üzerindeki verilerin, 2012-2020 yılları arasında 898 exebayttan 6,6 zetabayta çıkacağı veya her yıl verinin %25'ten fazla büyüyeceği öngörülmektedir. Yani bu durum verinin her 3 yılda bir 2'ye katlanacağı anlamına gelmektedir . [23]

Dünya çapında toplanan bu veri, veri yığınlarına dönüşerek ’büyük veri’ (big data) kavramını oluşturmakta ve veri madenciliği sürecini etkilemektedir. Büyük veri, tipik veritabanı yazılım araçlarının yakalama, depolama, yönetme ve analiz etme yeteneğinin ötesinde olan [24]analiz ve görselleştirilmesinin güç olduğu, çok karmaşık yapıya sahip[25] veri kümelerini ifade eder. Son yıllarda, öğrenme ortamları tarafından (facebook, Youtube, Twitter, cep telefonu konum verileri vb.) üretilen veriler akabinde büyük veri teknolojilerine ve bunları ele alacak araçlara duyulan gereksinimi arttırmaya başlamıştır.[26] Bu verilerin anlamlı ve kullanılabilir veri haline getirilmesi için veri madenciliğinin çeşitli yöntemleri kullanılmakta ve böylelikle’’büyük veri madenciliği‟ kavramı oluşmaktadır.

[22] Doğan, 2014, s.6
[23] Gantz vd, 2012, s.1
[24] (Brown vd., 2011, s.1),
[25] (Sağıroğlu ve Sinanç, 2013, s.42)
[26] Sin ve Muthu, 2015, s.1035

Büyük veri madenciliği, büyük veri kümelerinden veya veri akışlarından yararlı bilgiler çıkarma yeteneğidir. Bu tür verilerin hacim, değişkenlik ve hızından dolayı yeni madencilik teknikleri ise gerekli olmaktadır .[27] Bu durum ortaya koymaktadır ki veri toplama ve veri depolama teknolojilerinin gelişmesiyle yeni veri madenciliği teknik ve kavramlarının oluşması kaçınılmazdır. Ayrıca gelecekte oluşması ön görülen veri büyüklükleri ve verinin toplanma hızları düşünüldüğünde; birçok kaynaktan depolanan ve dinamik olarak çoğalan heterojen verinin analizi için günümüzde kullanılan veri madenciliği tekniklerinin yetersiz kalacağı, daha hızlı ve akıllı algoritmaların geliştirilmesi gerekliliği öngörülebilir.

1.6. Veri Madenciliğinin Önemi

İçinde bulunulan bilişim çağı sayesinde veri toplama araçlarında ve veri depolama teknolojilerinde olağan üstü gelişmeler yaşanmaktadır. Bu gelişim daha fazla ve detaylı veri toplanmasına olanak tanımaktadır. Günümüzde sağlık, finans, eğitim, lojistik, bankacılık, endüstri, haberleşme, ticaret, tıp, kütüphanecilik vb. birçok alanda bilişim teknolojileri yaygın biçimde kullanılmakta ve bu alanlarla ilgili kayıtlar veri tabanlarında depolanmaktadır. Ancak depolanan bu veriler birer veri yığınına dönüşmekle birlikte, bu verilerde dezenformasyon (bilgiyi çarpıtma) ve bilgi kirliliği yüksek olabilmektedir. Teknolojinin sürekli olarak ilerlediği ve kendini yenilediği bilişim çağında doğru bilgiye ulaşmak önemlidir; ancak dezenformasyon ve bilgi kirliği doğru bilgiye ulaşmayı zorlaştırmaktadır.

Veri bilginin ham halidir ve bir verinin bilgiye dönüşmesi için işlenerek (toplama, birleştirme, temizleme, dönüştürme, analiz, yorumlama vb.) anlamlandırılması gerekir. Kurumlar veriyi etkin biçimde işleyerek yönettikleri zaman, dezenformasyondan uzak, kaliteli bilgiye ulaşıp doğru kararlar alabilirler.[28]

Veri tabanlarındaki veri artışı, bilgi ve bilgiyi akıllıca kullanan teknolojilerin geliştirilmesi ihtiyacını doğurmuştur. Bu nedenle, veri madenciliği giderek daha önemli bir araştırma alanı haline gelmiştir.[29]

Ayrıca çeşitli alanlarda bilgi teknolojilerinin kullanımı yeni veri biçimleri (video, ses kaydı, resim, belge gibi) ortaya çıkarmıştır ve çeşitli türde büyük miktarda veri depolanmasına olanak tanımıştır. Farklı alanlardan toplanan veriyi kullanarak daha iyi ve

[27] Fan ve Bifet, 2013, s.1
[28] Dinçerden, 2017, s.137
[29] Liao vd., 2012, s.11303

doğru karar vermek amacıyla bilgi çıkarımı yapılması gereklidir. Veri madenciliğinin temel işlevi, veri kalıplarını keşfetmek ve örüntüler çıkartmak için çeşitli yöntem ve algoritmaları uygulamaktır. Veri madenciliği karar vermede yardımcı gizli kalıpları ve ilişkileri keşfetmek adına veri yığınlarını inceler.[30] Çünkü büyük miktardaki veriyi manüel olarak işlemek veri fazlalığı, çeşitliliği ve kirliliği nedeniyle neredeyse imkânsızdır. Klasik istatistiksel yöntemler büyük veri kümelerini işlemede yetersiz kalabilmektedir. Bu nedenle büyük ölçekli veri kümelerini işleyebilmek için daha özel yöntemler gerekir. Klasik yöntemlerin aksine veri madenciliği araştırmacıya veri yığınlarını analiz etmede kolaylık sağlar. [31]

Veri madenciliği araştırmacılara veri kalıplarını keşfetmek için daha akıllı ve çeşitli veri işleme algoritmaları sunmaktadır.[32] Verileri bilgiye dönüştürürken hangi algoritmaların kullanılacağı önemlidir. Araştırma amacına uygun olarak toplanan verilerin yine araştırma amacına uygun tekniklerle analiz edilmesi gerekir. Klasik istatistiksel yöntemlerde analiz teknikleri sınırlı sayıda olmasına karşın veri madenciliğinin kullandığı teknikler çeşitlilik göstermektedir.

1.7. Veri Madenciliğinde Kullanılan Yazılımlar

Veri madenciliği algoritmalarının uygulanması için güçlü yazılımlara ihtiyaç duyulmaktadır.[33] Veri madenciliği algoritmalarında başarıya ulaşmak için çok sayıda veri madenciliği yazılımı geliştirilmiştir. Bu yazılımların bazıları ücretli bazıları ise ücretsiz olarak kullanıcılara sunulmaktadır.[34]

Veri madenciliği uygulamaları için geliştirilen ve yaygın olarak kullanılan bazı yazılımlar şunlardır: Weka, RapidMiner, R, Orange, Knime açık kaynak kodlu ve ücretsiz, SPSS Clementine, SAS Enterprise Miner, MS SQL Server Analysis Services[35] ise kapalı kaynak kodlu (ticari) ve ücretli yazılımlardır. Bu yazılımlar ilerleyen paragraflarda ayrıntılı olarak açıklanmıştır.

[30] Ahmed ve Elaraby, 2014, s.43-44
[31] Alan, 2012, s.165
[32] Dangare ve Apte, 2012, s.44
[33] Mikut ve Reischl, 2011, s.431
[34] Doğan, 2017, s.88
[35] Dener vd., 2009, s.788

1.7.1. Weka

WEKA (Waikato Environment for Knowledge Analysis), Yeni Zelanda Waikato Üniversitesi'nde[36] Java programlama kullanılarak geliştirilmiş, 1996 yılında ilk resmi sürümü yayınlanmış, tüm modern bilgisayarlarda çalışabilen açık kaynak kodlu bir yazılımdır. Kullanıcı dostu bir ara yüze sahiptir. Veri analizinin tahmin edici modelleri için görselleştirme araçlarını ve algoritmalar topluluğunu içinde barındırır. WEKA, basit makine öğrenme, veri ön izleme, kümeleme, sınıflandırma, birliktelik analizi, ilişkilendirme, görselleştirme gibi çeşitli standart veri madenciliği görevlerinin hazır olarak geldiği bir veri madenciliği yazılımıdır.[37]

Bu yazılım SQL (Structured Query Language) veri tabanlarına Java Database Connectivity kullanarak erişir. Dosya uzantısı olarak .arff (Attribute Relationship File Format) kullanmaktır. Bu uzantı WEKA'ya özeldir.[38] WEKA eğitim, tarım, sağılık, endüstri ve birçok çeşitli alanda uygulanan, sınıflandırma[39], kümeleme[40], birliktelik analizi (Apriori) algoritmasını içinde barındıran kullanımı yaygın veri madenciliği araçlarındandır.

1.7.2. RapidMiner

RapidMiner, Dortmund Teknoloji Üniversitesinin Yapay Zeka biriminde Ingo Mierswa, Ralf Klinkenberg ve Simon Fischer tarafından geliştirilmiş açık kaynak kodlu bir yazılımdır.[41] İş analizi, tahmin edici analiz, metin madenciliği vb. konularla ilgilenen makine öğrenimi üzerine kurulu bir araçtır. Oluşturulan modellerin değerlemesi için çapraz doğrulama ve bağımsız doğrulama kümelerini kullanır.[42]

RapidMiner kullanımı kolay kullanıcı dostu bir ara yüze sahiptir. Ayrıca açık kaynak kodlu yazılım olması dolayısıyla geliştiriciler temel yazılıma ek işlevsellik katabilir.[43] Java programlama dili ile oluşturulan bu yazılım, kullanıcının kod yazdırmasını gerektirmeden hazır veri analiz süreçleri olan veri yükleme, veri ön izleme, görselleştirme, öngörülü analitik,

[36] Tamayo vd., 2018, s.76
[37] Wu vd., 2018, s.102; Aher ve Lobo, 2011, s.23
[38] Kaya ve Özel, 2014, s.49-50
[39] C4.5, J48, SVM, ID3, Naive Bayes) (Abidin vd., 2017, s.1223
[40] Canopy, Cobweb, EM, Farthest First, Filtered Clusterer, Hierarchical Clusterer, Make Density Based Clustere, Simple K-Means
[41] Kaya ve Özel, 2014, s.50
[42] Baruah ve Choudhary, 2017, s.270
[43] Prekopcsak vd., 2013, s.4

istatistiksel modelleme, değerleme, dağıtım gibi makine öğrenme prosedürleri sağlar. Excel, Access, Oracle, IBM (International Business Machines), Db2, MS SQL, Sybase vb. veri tabanlarından veri çekebilen, çok katmanlı veri görünümü konseptine sahip yüksek boyutlu grafik çizimleri yapabilen bir veri madenciliği aracıdır.[44]

1.7.3. R

R, Auckland Üniversitesi istatistik bölümü bilim adamlarından Robert Gentleman ve Ross Ihaka tarafından, çok çeşitli istatistiksel hesaplama ve grafik oluşturmak için geliştirilen, açık kaynak kodlu veri madenciliği yazılımıdır[45]. R, hem akademik çalışmalarda hem de endüstride yaygın olarak kullanılan, açık kaynak kodlu yazılımın tüm avantajlarına sahip, çok kullanışlı bir yazılımdır.[46] Bu yazılım ile uygulanabilen bazı veri madenciliği görevleri şunlardır [47]:

- Makine Öğrenimi ve İstatistiksel Öğrenme,
- Küme Analizi ve Sonlu Karışım Modelleri,
- Zaman Serisi Analizi,
- Çok Değişkenli istatistik,
- Mekânsal Verilerin Analizi.

1.7.4. Orange

Orange, Makine Öğrenmesi tabanlı, açık kaynak kodlu bir Veri Madenciliği yazılımıdır. Slovenya Ljubljana Üniversitesi'nde Bilgisayar Fakültesi, Bilgi ve Bilişim Bilimleri Bioinformatics Laboratuarında Python ve C++ dillerinde geliştirilmektedir. Keşifsel veri analizi için kullanılan bir programdır. Aynı zamanda bir Python Kütüphanesi olarak da kullanılabilir. Basit veri ön izleme, görselleştirme, öğrenme algoritmalarının ampirik değerlendirmesi ve tahmini modelleme, model değerlendirme bileşenlerine sahiptir.[48] Bu

[44] Gupta ve Malhotra, 2015, s.24; Ristoski vd, 2015, s.142
[45] Kaya ve Özel, 2014, s.49; Shmueli vd., 2017, s.11
[46] Olson ve Wu, 2017, s.119
[47] Zhao, 2013, s.2
[48] Naik ve Samant, 2016, s.664; Kranjc vd., 2017, s.40

bileşenlere ‘widge’ adı verilmektedir[49]. Ayrıca metin madenciliği ve biyoinformatik eklentileri de mevcuttur.[50]

1.7.5. Knime

KNIME (Konstanz Information Miner)[51], Eclipse platformunu temel alan genel amaçlı bir veri madenciliği yazılımıdır.İlk kez 2004 yılında Konstanz Üniversitesi tarafından geliştirilmeye başlanan bu yazılımın ilk sürümü 2006 yılında piyasaya sürülmüştür. KNIME açık kaynak kodlu bir yazılım olmasına karşın ticari lisans ile daha üst seviyeye yükseltilebilen bir yazılımdır.[52]

KNIME, yeni algoritmaların, veri bağlantılarının ve görselleştirme yöntemlerinin kolay ilişkilendirilmesini sağlayan modüler yapıda bir yazılımdır. Yazılım, iş akışında verileri değiştirmek ve işlemek için kilit birim olarak adlandırılan düğümleri kullanır. Bu düğümler grafiksel bir ara yüz kullanılarak birbirine bağlanır. Bu ara yüz üzerinden faklı yöntemler veri yükleme, veri ön işleme, model oluşturma, veri görselleştirme seçilebilir, kullanıcı dostu sürükle bırak yöntemi ile öğeler ilişki düğümleri, giriş bağlantıları, çıkış bağlantıları oluşturulabilir ve iş akışı kolayca uygulanabilir. KNIME ile veri analizi, tahmini modelleme

ve değerlendirme, yapısal özellik analizi, kural tabanlı tarama vb. veri işleme teknikleri uygulanabilir.[53]

1.7.6. Ibm Spss Modeler (Spss Clementine)

SPSS Clementine olarak da bilinen IBM SPSS Modeler yazılımı veri madenciliği uygulamalarında en çok tercih edilen yazılımlardan bir tanesidir. IBM firması tarafından geliştirilmiş bu yazılım CRISP-DM (Cross Industry Standard Process Model for Data Mining) modeli çerçevesinde tahmin edici modeller geliştirir. Kullanıcıların programlama kodu

[49] Patel, 2015, s.209
[50] Yıldız ve ġeker, 2016, s.17
[51] Melagraki ve Afantitis, 2013, s.9
[52] Jovic vd., 2014, s.1117
[53] O'Hogan ve Kell, 2015, s.387-391; Steinmetz vd., 2015, s.173

yazmalarına gerek olmadığı, içinde veri madenciliği algoritmaları bulunan, görsel ara yüze sahip modüler bir veri madenciliği yazılımıdır.[54]

Bu yazılım veri toplama, veri ön işleme, dönüştürme, model oluşturma, modeli değerlendirme ve örüntü oluşturma gibi veri madenciliği süreçlerini desteklemektedir.[55] Farklı veri kaynaklarına ulaşabilen yazılım hızlı model oluşturan ve bu modelleri karşılaştıran kullanışlı bir yapıdadır.[56] Veriyi ön işleme, görselleştirme teknikleri, uygulama şablonları ve akıllı algoritmaları ile geniş bir veri madenciliği yelpazesi sunmaktadır.[57]

1.7.7. Sas Enterprise Miner

SAS Enterprise Miner, SAS Corporation tarafından üretilen SAS Analytics'in tahmini ve betimleyici modelleme, metin inceleme, tahmin etme, optimizasyon, simülasyon ve deneysel tasarım için entegre edilmiş[58], çok büyük miktardaki verinin analizinde kullanılan, son derece kesin tahminler ve açıklayıcı modeller oluşturan bir veri madenciliği yazılımıdır. Veri madenciliğinin veri ön izleme, ilişki oluşturma, tahmin etme, model seçimi, veri modelleme ve geliştirme gibi yöntemlerini içerir.
Bu yazılım zaman serileri, kümeleme analizi[59], karar ağaçları, sınıflandırma[60], doğrusal ve lojistik regresyon, yapay sinir ağları gibi teknikleri kullanır.

1.7.8. Ms Sql Server Analysis Services

MS SQL Server ilişkisel veri tabanı olmasının yanında veri madenciliği modelleri ile de ön plana çıkmaktadır. MS SQL Server Analysis Services, Microsoft araştırma ile birlikte SQL Server veri madenciliği ürün ekibi tarafından geliştirilen zengin algoritma kümesini içerir. Bu Algoritmalar Microsoft karar ağaçları, Microsoft kümeleme, Microsoft zaman serileri, Microsoft birliktelik kuralları, Microsoft sekans kümeleme, Microsoft bayes teoremi, Microsoft sinir ağları, Microsoft doğrusal regresyon ve Microsoft lojistik regresyon şeklindedir.[61] MS SQL Server Analysis Service veri madenciliği motoru ve OLAP (Online

[54] Özseven ve Ersöz, 2016, s.86-87; Aksoy ve Narlı, 2015, s.150
[55] Lu vd., 2013, s.4
[56] arlı vd., 2014, s.45
[57] Al Ghoson, 2010, s.64
[58] Al Ghoson, 2010, s.62
[59] Nakkeeran vd., 2012, s.1
[60] Piwczyński vd., 2013, s.19
[61] Aggarwal vd., 2012, s.167

Analytical Processing - çevrim içi analitik işleme) motoru olan gelişmiş özelliklere sahip iki önemli bileşeni mevcuttur. Bu iki bileşen veri madenciliği için önemli tekniklerdir ve birbirlerini tamamlarlar.[62]

1.7.9. Microsoft Power Bi

Microsoft Power BI bir çok bileşenden oluşan, rapor oluşturmanıza, raporları yayınlamamıza, verilerimiz üzerinde analiz yapmamıza ve daha bir çok özelliği olan iş analizi araçlarının bir araya gelmesidir.

Her ne kadar tek bir ürün gibi gözükse de içerinde bir çok veri analizi, veri kaynaklarına bağlanmayı sağlayan arabirimler ve servisler bulunduran, verinin sunulması, paylaşılması ve birlikte çalışabilirliği sağlayan bir çok bileşeni içerisinde bulundurur.

1.8.Veri Madenciliği Yöntemleri

Veri madenciliğinde elde bulunan veri setine ve uygulanacak alana göre farklı yöntemler kullanılmaktadır.[63]Veri analizinde hangi yöntem veya yöntemlerin kullanılacağı veritabanında bilgi keşfi sürecinde tanımlanan probleme göre belirlenir. Bu bağlamda veri madenciliğinde tahmin edici ve tanımlayıcı modeller olmak üzere 2 adet genel yöntem bulunmaktadır. Karar alma sürecinde önemli olan tahmin edici modeller bir veri setindeki olayın sonuç değerlerinden hareket ederek model geliştirilmesi ve kurulan bu model ile sonuçları bilinmeyen bir olayın sonuç değerinin tahmin edilmesi için geliştirilmiş modellerdir.[64] Veri madenciliğinde zaman serileri analizi, sınıflandırma ve regresyon tahmin edici modellerdir.

Tanımlayıcı modeller karar vermede rehberlik etmesi için mevcut veri örüntülerinin tanımlanmasını ve veri kümesi içerisinde ne tür ilişkilerin olduğunun anlaşılmasını sağlar.[65] Veri madenciliğinin birliktelik kuralları ve ardışık zamanlı örüntüler ile kümeleme analizleri tanımlayıcı modellerdir.

1.8.1. Sınıflama ve Regresyon

Sınıflandırma ve regresyon, önemli veri sınıflarını ortaya koyan, gelecekteki bir verinin bu sınıflardan hangisine ait olduğunu tahmin eden veri madenciliği yöntemleridir. Sınıflama,

[62] Tang vd., 2005, s.80
[63] Albayrak, 2017, s.755
[64] Alagöz vd., 2014, s.7
[65] Sevindik vd., 2012, s.188

kategorik değerleri tahmin eder, regresyon ise süreklilik gösteren değerlerin tahmin edilmesinde kullanılır. Regresyon analizi, bir veya daha fazla bağımsız değişken ile bağımlı değişken arasındaki ilişkiyi modellemek için kullanılır. Veri madenciliğinde bağımsız değişkenler zaten bilinen niteliklerdir, istenen ise bağımlı değişkenleri tahmin etmektir.[66] Sınıflama teknikleri ile veri seti içinde sınıfı belli olamayan gözlemlerin çeşitli özelliklerine göre, nitelikleri önceden bilinen, eğitim seti ile tanımlanan bir veri sınıfına ataması yapılır.[67] Bu yöntem ile veri kümesi içerisindeki kategorisi bilinmeyen bir verinin kategorik değerleri tahmin edilebilir.[68]

Sınıflandırma ve regresyon kapsamında ele alınabilecek bazı veri madenciliği problemleri şunlardır:

- Kredi- borç onayı tespiti,
- Dolandırıcılık tespiti,
- Hastalık teşhisi,
- Ayrılabilecek müşterilerin tahmini,
- Ses ve Karakter tanıma.

Ne yazık ki, birçok gerçek dünya problemi basitçe tahmin edilemez yapıdadır. Bu nedenle, gelecekle ilgili değerleri tahmin etmek için daha karmaşık teknikler gerekebilir. Bu sebeple aynı model tipleri hem regresyon hem de sınıflandırma için kullanılabilir. Örneğin, Yapay sinir ağları ve karar ağaçları ile hem sınıflandırma hem de regresyon modelleri oluşturabilir.[69]

Sınıflandırma ve regresyon problemleri için geliştirilmiş çeşitli veri madenciliği teknikleri bulunmaktadır. Bu tekniklerden en çok kullanılan algoritmalar karar ağaçları, yapay sinir ağları, bayes sınıflandırıcılar, k-en yakın komşu, genetik algoritmalar, destek vektör makineleri, doğrusal regresyon, lojistik regresyon şeklindedir.

[66] Baradwaj ve Pal, 2011, s.64; ÇalıŞ vd., 2014, s.5
[67] Vadim, 2018, s.235; Bilen vd., 2014, s.82
[68] Alagöz vd., 2014, s.8
[69] Baradwaj ve Pal, 2011, s.64

1.8.2. Zaman Serileri Analizi

Bir zaman serisi, zaman içindeki sıralı ölçümlerden elde edilen değerler topluluğunu temsil eder. Zaman serisi verileri, borsa günlük dalgalanmaları, bilimsel deneyler, tıbbi ve biyolojik deneysel gözlemler, sosyal ağlardan elde edilen çeşitli veriler, konum güncellemeleri gibi hemen hemen her uygulama alanından benzeri görülmemiş bir ölçekte ve oranda üretilmektedir. Bu sebeple son dönemlerde bu tür verileri sorgulama ve inceleme konusundaki ilgi de fazlasıyla artmaktadır. Günümüzde zaman serileri analizi, çeşitli araştırma alanlarındaki ekonomik tahminler, saldırı tespitleri, gen analizleri, tıbbi gözetim analizleri gibi gerçek yaşam problemlerinin çözümünde kullanılmaktadır.[70] Zaman serisi analizinde verilerin zamana bağlı olarak değişimleri incelenmektedir. Bu analiz hareketli ortalama, göreceli güç endeksi, momentum ve değişim oranı gibi yöntemlerle, verinin zaman üzerindeki farklılıklarını tespit edip aykırı değerleri yakalamak, zamana bağlı olarak değişen olaylarda gerçekleşmesi muhtemel olayları tahmin etmek, eksik verileri tamamlamak ve aykırı değerlerdeki hataları düzeltmek amacıyla kullanılabilir.[71]

1.8.3.Kümeleme

Küme analizi nesnelerin belirlenen özniteliklerine ve dolayısı ile öznitelik değerlerine göre alt gruplara, kümelere ayrılabilmesi için istatistik ve makine öğrenimi alanlarında geliştirilen çeşitli yöntem ve süreçlerdir. Algılama ve öğrenmenin temel öğesi, nesnelerin benzerliklerine göre sınıflandırmasıdır. Makine öğreniminde denetimsiz öğrenmenin temel araçlarından biri olan küme analizi, nesnelerin benzerlik ilişkilerine göre gruplandırması ile insan beyninin tipik bir akıl yürütme işlevini taklit etmektedir. Bu bakış açısına göre küme analizi gizli örüntülerin denetimsiz öğrenme yaklaşımı ile aranmasıdır. Örneğin bitkilerin hayvanlardan, köpeklerin kedilerden ayırt edilmesi çocukluğun erken dönemlerinde geliştirilen bilinçaltı öğrenme mekanizmaları ile gerçekleştirilen kümeleme süreçleridir.

[70] Esling ve Agon, 2012, s.13-14;Wang vd., 2013, s.276
[71] Seker, 2015b, s. 24-30

Kümeleme problemleri, genel olarak çıkarılmış olan özellikler kullanılarak bir veri setindeki değerlerin farklı kümeler arasında paylaşılmasıdır. Sınıflandırma ile arasındaki fark, bir veri kümesindeki verilerin arasında ilişki olup olmadığının sorgulanmasıdır.[72]
Kümeleme analizi çoğu disiplinle alakalı bir uygulamadır. Bu disiplinlerin en başında istatistik, bilgisayar, matematik gelmektedir. Bilgisayar alanında kümeleme analiziyle alakalı çalışmalara ses, resim tanıma ve makine öğrenmesi örnek olarak verilebilir. Kümeleme analizi istatistik alanında çok değişkenli tahminlerde ve örüntü tanıma konularında yardım sağlamaktadır.[73]Veri kümeleme, seyrek ve kalabalık alanları tanımlamakta ve veri kümesinin dağılım modellerini keşfetmektedir. Bununla birlikte türeyen kümeler orijinal veri kümesinden daha etkin ve etkili görselleştirebilmektedir.[74]

Kümeleme benzer amaçlara göre verilerin gruplara ayrılmasıdır. Küme diye adlandırılan her bir grup, diğer grupların birimlerine benzemeyen ancak birbirine benzeyen birimlerden oluşmaktadır. Makine öğrenmesi açısından kümeler gizli örüntülerle uyumlu olmaktadır. Kümelerin araştırılması denetimsiz öğrenmedir. Bununla birlikte neticelenen sistem bir veri anlayışını temsil etmektedir.Kümeleme bir sıklık tahmin problemi olarak görülebilmektedir.

Bu da geleneksel çok değişkenli istatistiksel tahminlerin konusu olmaktadır. Kümeleme hem de görüntü işlemedeki veri sıkıştırma içinde oldukça geniş bir şekilde kullanılmaktadır.[75]

1.8.4.Birliktelik Kuralları ve Ardışık Zamanlı Örüntüler

Birliktelik kuralları belli başlı türlerdeki veri bağlantılarını açıklayan bir model olarak tanımlanmaktadır. Herhangi bir ürünün satın alınması durumunda bu ürün yanında başka bir üründe satın alınmasıyla bir birliktelik kuralı oluşturulabilmektedir. Bu açıdan tanımlayıcı özelliklere sahip bir model olduğu görülmektedir. Belli bir ürün ve bu ürünle birlikte başka ürünlerinde alınması durumunda birliktelik kuralları gündeme gelmektedir. Dolayısıyla perakendecilik sektöründe faaliyet gösteren firmalarda daha fazla uygulanmaktadır. Örneğin bir süpermarkette yapılan alışverişin incelenip hangi ürünün hangi ürün ile satın alındığının belirlenmesi birliktelik kurallarını ilgilendirmektedir. Bunun dışında örneğin iletişim

[72] Şeker, a.g.e., s. 82.
[73] P. Arthur Dempster, Nan M. Laird, Donald B. Rubin, "Maximum Likelihood from Incomplete Data Viathe EM Algorithm", Journal of The Royal Statistical Society Series B (Methodological), 1977, s. 35
[74] 269Tian Zhang, Raghu Ramakrishnan, Miron Livny, "BIRCH: An Efficient Data Clustering Method for Very Large Databases", In: ACM Sigmod Record, 1996, s. 5.
[75] Pavel Berkhin, A Survey of Clustering Data Mining Techniques, In: Grouping Multidimensional Data, Springer Berlin, 2006, s. 30.

ağlarında meydana gelen hataların belirlenmesinde de kullanılabilmektedir.[76] Birliktelik kuralları bilgisayar bilimi alanında geliştirilmiştir ve pazar sepet analizi gibi önemli uygulamalarda kullanılmaktadır. Belli bir müşteri tarafından satın alınan ürünler arasındaki birliktelikleri ölçmek için kullanılmaktadır. Bir web kullanıcısı tarafından sıralı olarak bakılan sayfalar arasındaki birliktelikleri ölçmek için kullanılmaktadır. Amaç bütün işlemler kümesindeki birlikte meydana gelen eleman gruplarını vurgulamaktır. Veri tabanı her bir işlem için gerçekleşen öğe listesini içermektedir. Her bir bireyin veri kümesinde bir defadan daha fazla görülüp görülmediğine dikkat edilmelidir. Birliktelik kuralları özel model tipleri arasındaki kuralları dikkate almaktadır. Bu model tipleri öğe kümeleri olarak adlandırılmaktadır. Öğe kümesindeki her bir değişken çift değerlidir. Yani özel durum o değişken için doğruysa 1 değeri verilir yanlışsa 0 değeri verilmektedir. Birliktelik kuralları değişkenler arasındaki ilişkileri belirlemek için kullanılan en yaygın metottur. Oldukça kompleks ve dinamik olan küresel bir analiz için çok geniş veri kümelerini araştırıp çıkartma amacıyla kullanılmaktadır.[77]

Birliktelik kuralları öğe kümeleri arasında meydana gelen bağımlılıklar üzerinde faydalı bilgi sunan örüntülerdir. Apriori gibi mevcut birliktelik kuralları oldukça büyük kurallar kümesini ortaya çıkarmaktadır. Bu kuralları anlamlandırmak için faydalı örüntüleri vurgulayan belli biçimdeki grup kurallarına ihtiyaç duymaktadır.[78]

Birliktelik kuralları ile oluşturulan modeller aşağıdaki örneklerden de anlaşıldığı gibi aynı zamanda meydana gelen bağlantıların açıklanmasında kullanılmaktadır. Aşağıda sunulan örneklerde görüldüğü gibi eş zamanlı olarak gerçekleşen ilişkilerin tanımlanmasında kullanılır.

- Tüketiciler kola satın alırken, % 90 olasılıkla çekirdekte satın almaktadır.
- Az yağlı peynir ve diyet yoğurt alan tüketiciler, % 80 olasılıkla diyet bisküvi de satın almaktadır.[79]

Birliktelik kuralları aslında olaylar arasında gerçekleşen probabilastik ilişkiyi açıklamaktadır. Gerçekleşen olaylar arasındaki bağlantı ise sıklıkla bir arada gözlemlenen olaylardır. Veri tabanlarının ilk adımı veri tabanındaki birliktelik durumlarının açığa çıkarılmasıdır. Veri tabanındaki herhangi bir X'in aynı zamanda Y'yi içermesi bir birlikteliktir. Bu durum ile alakalı yaygın bir örnek olarak : "Mama içeren %70 alışverişin,

[76] Margaret H. Dunham, Data Mining-Introductory and Advanced Topics, Person Education, 2003, s.
[77] Grabmeıer ve Rudolph, a.g.e., s. 332.
[78] Koh, Tan ve Peng, a.g.e., s. 182.
[79] Akpınar, a.g.e., s.46

%10'u aynı zamanda çocuk bezi de içermektedir." Burada %70 güven düzeyini %10 ise bu güven seviyesinde bulunan desteği işaret etmektedir.[80]

Birliktelik kuralları için veri madenciliği 2 adımda gerçekleşmektedir. İlk adım bütün sık öğe kümelerini bulmakla meydana gelmektedir. Bu öğe kümeleri asgari destek olarak adlandırılan ve veri tabanlarında yer alan kullanıcıya ait belli ve özellikli sıklıktan oluşmaktadır. İkinci adım sık öğe kümeleri arasındaki kuralları düzenlemekle meydana gelmektedir.[81]

Benzeyenleri bir araya topluyor, daha sonra bunları en çok benzerden, en az benzere doğru sıralıyoruz. Örüntü tanıma amacıyla geliştirilen birden fazla algoritma bulunmaktadır. "K-en yakın komşu algoritması", "doğrusal sınıflayıcı", "üstel sınıflayıcı" bunlardan sadece birkaç tanesidir.[82]

Ardışık zaman örüntüleri, veri madenciliği ve perakendecilik sektöründen gelen yoğun ilgiyle gelişimini sürdürmektedir. Bir dizi olayın gelmesiyle birlikte o olaydaki alt dizilerden oluşan örüntüler ardışık zamanlı olarak isimlendirilmektedir.[83]Barkod teknolojisindeki gelişme, perakendecilik sektöründe faaliyette bulunan firmaların yüksek hacimlerde ve

Sıklıkla denilebilecek aralıklarla meydana gelen bilgilerin, hızla stoklanmasına imkan sağlamıştır. Satış miktarlarını gösteren bu verilere sepet verisi denilebilmektedir.[84] Bu veriler genel anlamda işlem tarihi, satılmış olan malları, işlem sıra numarası gibi bilgileri kapsamaktadır. İşlemin gerçekleştiği yerde ki burası tipik olarak bir süpermarkettir; üyelik kartları ve kulüp kartları diyebileceğimiz kartlar kullanılmaktaysa, bu bilgilerin içinde müşterilerin numaraları da bulunabilmektedir. Bu kayıtların incelenmesiyle, ardışık zaman örüntüleri tespit edilebilmektedir. Eğer müşterilerin bir kısmı, önce bir CD oynatıcısı, daha sonraki bir zamanda ABC ismindeki bir filmi, ardından XYZ ismindeki bir filmi satın alıyorlarsa bu bir ardışık örüntü oluşturacaktır. Başka müşteriler ABC filmi ile XYZ arasındaki zaman diliminde, farklı filmler ya da farklı ürünler alsalar bile onlarda bu örüntüye dâhildirler.[85]

[80] Silahtaroğlu, a.g.e., s. 50.
[81] Zaki vd.,"New Algorithms for Fast Discovery of Association Rules", KDD Proceedings Vol. 97, 1997, s. 284.
[82] Silahtaroğlu, a.g.e., s. 54.
[83] Marek Wojciechowski, "Interactive Constraint-Based Sequential Pattern Mining", East European Conference on Advances in Databases and Information Systems, Springer Berlin, 2001, s. 2.
[84] Rakesh Agrawal, Ramakrishnan Srikant, "Fast Algorithms for Mining Association Rules", Proc. 20th Int. Conf. VeryLarge Data Bases, VLDB. Vol. 1215, 1994, s. 490.
[85] [280]Silahtaroğlu, a.g.e., s. 55.

İKİNCİ BÖLÜM
KÜMELEME ANALİZİ

2.1. Kümeleme Analizi Tanımı

Günümüzde bireyler kritik kararlar vermeden önce geriye dönüp bakarak sorunun genelini anlamaya çalışmaktadırlar. Ancak çoğu zaman sorunun genelini çözebilmek çok karmaşık olmaktadır. Devasa büyüklükte veri tabanları fazla hacimde alanı kapsayabilmekte ve aşırı karmaşık bir yapısı bulunduğundan en iyi sonuç veren yöntemler bile bu veri yığınları arasından anlamlı neticeler çıkaramamaktadırlar. Fazla karmaşık olan problemlerin çözümünü bulmakta takip edilen teknik genel olarak problemi daha kolay çözülebilecek hale getirmekle gerçekleşmektedir yani problem alt sorunlara bölünerek ve her biri ayrı ayrı çözüldükten sonra elde edilen çözümler birleştirilerek sonuç elde edilmeye çalışılır. Ancak bazı koşullarda verilerin dağılımları, nereden bölünmeleri gerektiği konusunda kestirimde bulunmayı engellemektedir. Bu nedenle otomatik olarak küme tespit etme teknikleri bulunmuştur.[86]

İstatistiksel yöntemlerden bazıları değişkenler ve bu değişkenlere ait veriler arasında bulunan benzerlikler ya da farklılıkları bulmaya imkân sağlamaktadır. Özellikle çok değişkenli istatistik teknikler çeşitli niteliklere ait değişimlerin aynı zamanda değerlendirilmesine olanak sağlamaktadır. Verilerin içerisinde meydana gelen gruplamaları meydana çıkaran yöntem çok değişkenli istatistik tekniklerden olan kümeleme analizidir. Karışık veri setleri içinde var olan yapıların ortaya çıkarılmasını sağlayan kümeleme analizinin, kendi içerisinde farklı yöntemleri bulunmaktadır. Kümeleme analizinde hedef birbirlerine göre farklılıklar veya yüksek seviyede benzerlikler taşıyan verilerin veya değişkenlerin kümelerde gruplanmasını sağlamaktır. Her kümede yer alan birimlerin aynı hassasiyeti taşıması ve tüm veride daha çok homojen kümelerin oluşması beklenmektedir.[87]

---

[86] Özgür M. Dolgun, Tülin Özdemir, Doruk Oğuz, "Veri Madenciliğinde Yapısal Olmayan Verinin Analizi: Metin Ve Web Madenciliği", İstatistikçiler Dergisi: İstatistik Ve Aktüerya, 2.2.,2009, s.50
333Carme Hervada Sala, Eusebi Jarauta Bragulat,"A Program To Perform Ward's Clustering Method On Several Regionalized Variables", Computers & Geosciences, Vol.30 No.8 ,2004,s.883.

[87] Carme Hervada Sala, Eusebi Jarauta Bragulat,"A Program To Perform Ward's Clustering Method On Several Regionalized Variables", Computers & Geosciences, Vol.30 No.8 ,2004,s.883.

Tekrardan tanımlanan sınıflara bağlı olmamaktadır. Kümeleme yöntemi, denetimsiz öğrenmeyle gerçekleşmektedir.[88]

Kümeleme analizi çok değişkenli istatistiksel yöntemlerden olup, grup sayıları önceden bilinmeyen verilerin birbirlerine olan benzerlik durumlarına göre gruplandırılması için uygulanmaktadır. Birimler ve değişkenler açısından benzer olan veriler kümeleme analiziyle ayrık kümelere toplanmaktadır.[89]Kümeleme analizi belli bir gözlem setinin ya da kümelerin kısıtlı bir sayıya ayrılmasını hedefleyen çok değişkenli istatistik tekniklerdendir. Söz konusu ayrım aynı grupta yer alan gözlemlerin birbirlerine benzer durumları varken farklı grupta yer alan gözlemlerin birbirleri arasında farklı durumları olması kaydıyla yapılmaktadır.[90]

Kümeleme teknikleri; uzaklık matrisi veya benzerlik matrisi kullanarak birimlerin veya değişkenlerin birbirleri içerisinde homojen ve birbirleri aralarında heterojen gruplamalar oluşturmasına imkan sağlayan tekniklerdir.[91]

Kümeleme süreci heterojen bir yapıda olan kitlenin daha homojen alt gruplara veya kümelere bölünmesiyle gerçekleşmektedir. Kümeleme işleminde önceden belirlenen sınıflara göre bölünme yapılmamaktadır. Bu da sınıflama ile kümelemeyi ayıran en önemli farktır. Sınıflama işleminde ise veriler gerçekleştirilen bir eğitim sonucunda meydana gelen bir modele uygun olarak önceden belirlenen bir sınıfa atamayla gerçekleşmektedir. Kümelemede ise başlangıçta tanımlanan sınıf ya da örnekler yer almaktadır. Elde edilen sınıfların hangi anlamlara geldiği ise uygulamayı yapan kişiye bağlı kalmaktadır.[92]

2.2. Kümeleme Analizinin Özellikleri

Kümeleme analizi öncelikli hedef nesnelerin özelliklerini göz önüne alarak gruplama yapmak olan çok değişkenli istatistik yöntemlerden birisidir. Kümeleme analizinde söz konusu bu gruplama başlangıçta belirlenen kritere uygun olarak gerçekleştirilmektedir. Kümelerin kendi içerisinde fazla homojenlik, kendi aralarında ise fazla heterojenlik taşımasına dikkat edilmektedir. Kümeleme analizinin genel yapısı içerisinde rastlantı değişkeni önemli bir yer kapsamaktadır. Söz konusu değişken kümeleme analizinde

[88] Şebnem Koltan Yılmaz, Said Patır, "Kümeleme Analizi Ve Pazarlamada Kullanımı", Akademik Yaklaşımlar Dergisi, 2.1.,2012, s.94.
[89] Zeki Çakmak, Nevin Uzgören, Gülnur Keçek," Kümeleme Analizi Teknikleri İle İllerin Kültürel Yapılarına Göre Sınıflandırılması Ve Değişimlerinin İncelenmesi", Dumlupınar Üniversitesi Sosyal Bilimler Dergisi, 12.3.,2005, s.20
[90] Tim H. NEIL, Applied Multivariate Analysis, New York, 2002, s.515
[91] Kazım Özdamar, Paket Programlar ile İstatiksel Veri Analizi, İstanbul, Kaan Kitapevi, 5. Baskı, 2004, s.528.
[92] Metin Vatansever, "Görsel Veri Madenciliği Tekniklerinin Kümeleme Analizlerinde Kullanımı ve Uygulanması", Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü Yayınlanmamış Yüksek Lisans Tezi, 2008, s.53

nesnelerin karşılaştırılması amacıyla kullanılmakta olan nitelikleri temsil eden değişkenler kümesinden oluşmaktadır. Küme analizinde yer alan küme rastlantı değişkeni sadece nesnelerin karşılaştırılması amacıyla kullanılmakta olan değişkenleri kapsadığı için söz konusu nesnelere ait özellikleri belirlemektedir. Yalnızca nesneleri karşılaştırmak için kullanılan değişkenleri içerdiği için nesnelerin özelliğini belirlemektedir. Kümeleme analizinde, rastlantı değişkeni deneysel anlamda tahmin edilmektedir. Bunun yanı sıra bu değişkenin araştırmayı yapan tarafından belirlenmesi bu tekniği diğer çok değişkenli istatistik yöntemlerden ayırmaktadır. Kümeleme analizinde hedef rastlantı değişkenini tahmin etmek değildir ancak bu değişkene bağlı kalarak nesneleri kıyaslamaktır.[93] Kümeleme analizinin amaçları arasında radikal tipleri belirlemek, grupların içerisinde ön tahmin yapmak, hipotez testini uygulamak, verilerin yerine kümeleri değerlendirmek ve aykırı değerleri bulmakta yer almaktadır.[94]

İyi bir kümeleme analizinde aşağıda bulunan özellikler yer almalıdır:

- Ölçeklenebilme özelliği olmalıdır. Az sayıda kayıt bulunan veri tabanına da yüzbinlerce kaydı olan kümelere de uygulanma özelliği bulunmalıdır.
- Verilerin farklı çeşitleriyle kullanılabilir olmalıdır. Veri tabanının hem kategorik hem de sayısal verilerine uygulanabilmelidir.
- Düzgün şekle sahip olmayan kümelerde de uygulanabilir olmalıdır.
- Giriş değişkenlerine minimum sayıda gereksinim duyulmalıdır. Bir teknikte giriş değişkenine ne kadar az gereksinim duyuluyorsa, o kadar bağımsız sonuçlar elde edilir.
- Verilerde yer alan gürültülerle de kullanılabilme özelliği bulunmalıdır.
- Kümedeki veri kayıtlarının sıralaması analizden bağımsız olmalıdır. Analize kümedeki elemanlardan hangisiyle başlanırsa başlansın sonuç değişime uğramamalıdır.
- Devasa boyutlardaki veri tabanlarında da uygulanma özelliği olmalıdır.

[93] Joseph F. Hair, Rolph E. Anderson, Ronald L. Tatham Ve C. William Black,Multivariate Data Analysis With Readings, USA, 1995, 4th Edition, s. 423.

[94] Ömer Yılmaz, Sinan M. Temurlenk, Türkiye'deki İstatistik Bölgelerin Kışı Başına Düşen Gelir Açısından Hiyerarşik Ve Hiyerarşik Olmayan Kümeleme Analizi İle Değerlendirilmesi: 1965-2001, Atatürk Üniversitesi İktisadi ve İdari Bilimler Dergisi, 19.2.,2005, s.80.

- Sonuçlar kolay yorumlanabilmeli ve işlevsel özelliği bulunmalıdır.[95]

Normallik, sabit varyans, doğrusallık varsayımlarının kümeleme analizi içinde önemi çok az olmakla birlikte, asıl dikkat edilecek durumlar; örneklemin popülasyonu temsil etme başarısı, değişkenler arasındaki çoklu bağlantı ve gözlemlerdeki aşırı değerlerdir. Kümeleme analizinin temel varsayımı, veri matrislerinin analizden önce tahmin ve koşul değişkenlerinin alt matrislerine bölüştürülmemesidir. Diğer varsayımı ise verilerin bir kısmının homojen bir kısmının heterojen olmasıdır. çok değişkenli istatistiksel analizlerin en önemli varsayımlarından normallik kümeleme analizinde önemli görülmemektedir. Sadece uzaklık değerlerinin normal dağılması yeterlidir.[96]

Birimler p değişkene göre hesaplanmaktadır. Benzerliği ortaya çıkarma için kullanılan bazı ölçülerle kümeleme analizinde homojen gruplar oluşturulmaktadır. Kümeleme analizinin hedefler dört gruba ayrılmaktadır:

1) n sayıda birimin, nesnenin, oluşumun, p tane değişken değerine göre belirlenen niteliklere uygun mümkün olduğunca kendi içlerinde homojen ve kendi aralarında heterojen alt gruplara ayrılması

2) p tane değişken, n tane birimle belirlenen değerlere uygun ortak özelliklerin açıklandığı farz edilen alt kümelere ayrılması ve ortak faktörlerin meydana çıkarılması

3) Hem birimlerin hem değişkenlerin bir arada ele alınmasıyla, ortak n birimin p değişkene uygun ortak özelliklerle alt kümelere ayrılması,

4) Birimlerin, p adet değişkene uygun belirlenen değerleri için, takip ettikleri biyolojik ve tipolojik sınıflandırmayı ortaya çıkarmaktır.[97]

[95] Esra Dinçer, Veri Madenciliğinde K-Means Algoritması ve Tıp Alanında Uygulanması, Kocaeli Üniversitesi Fen Bilimleri Enstitüsü Yüksek Lisans Tezi, 2006, s.35.

[96] Hüseyin Tatlıdil, Uygulamalı Çok Değişkenli İstatistiksel Analiz, Ankara, Engin Yayınları, 1992, s.252.

[97] Yetiş Şazi Murat, Alper Şekerler, " Trafik Kaza Verilerinin Kümeleme Analizi Yöntemi İle Modellenmesi", Teknik Dergi, 20.98.,2009, s.4761.

2.3. Kümeleme İşlemi Süreci

Kümeleme analizinin uygulama adımları veri matrisini belirleme, benzerlik ya da farklılık matrisini belirleme, kümelere ayırma ve yorumlama gibi adımlardan oluşmaktadır.

- Veri matrisini belirleme: Birimlerin veya değişkenlerin doğal gruplanmalarıyla ilgili kesin bilgiler olmadığı popülasyonlardan elde edilen n tane birimin ve p adet değişkenin gözlemlerinin belirlenmesidir.

- Benzerlik ya da farklılık matrislerini belirleme: Birimler/değişkenler birbirleri arasındaki benzerlikleri veya farklılıkları ortaya çıkaran makul bir benzerlik ölçümüyle birimlerin/değişkenlerin birbirleriyle aralarında olan mesafelerin tespit edilmesidir.

- Kümelerin ayrılması: Uygun kümeleme tekniğinin yardımı sonucu benzerliklerle ya da farklılıklarla oluşturulan matrislere göre birimlerin veya değişkenlerin makul sayıda kümelere bölünmesidir.

- Sonuçların yorumlanması: Elde edilen kümeleri yorumlamak ve kümeleme yapısına bağlı kalarak ortaya çıkarılan hipotezleri onaylamak amacıyla gereken analitik tekniklerin uygulanmasıdır.[98]

Tekniğin kullanılması sonucunda nesnelerin kümelere ayrımı yapılmış olmaktadır. Analizin son aşaması ise kümeleme analiziyle elde edilen sonuçların anlamlılık değerlerinin yorumlanmasıdır. Kümeleme analizinin neticesinde kümeleri meydana getiren elemanların birbirlerine benzerlikleri varken, diğer küme elemanlarıyla farklılık göstermektedirler. Kümeleme süreci başarılı olarak sonuçlanırsa geometrik bir çizimle birimlerin küme içindeki konumları birbirine yakınken, kümeler birbirinden uzaklık gösterecektir.[99]

[98] Murat ve Şekerler, a.g.e., s.4765.
[99] Yılmaz ve Patır, a.g.e., s.93.

2.4. Uzaklık Veya Benzerlik Ölçüleri

Genellikle kullanılmakta olan kümeleme teknikleri birimlerin arasında yer alan mesafelere dayalı benzerlik veya farklılık matrisine uygun işlemler gerçekleştirmektedir. Çeşitli kümeleme teknikleri farklı mesafe ölçümlerine göre farklı neticeler verebilmektedir. Ayırmaya bağlı kümeleme teknikleri her veri setiyle her bir birimi sadece bir kümeye ayırmaktadır. Bu sayede aşamalı veya aşamalı olmayan kümeleme teknikleri her bir birim için karar vermekteler ve o birimi belli bir kümenin elemanı yapmaktadırlar. Sonuçlar açısından benzerlikler taşıyan yöntemler de belli başlı birimlerin farklı kümelerde bulunduğu da izlenmiştir.[100]

Küme, birbirleri arasında yakın mesafe bulunan birey ya da nesnelerin meydana getirdiği beraberlik olarak tanımlanmaktadır. Bu birliktelik çok boyutlu uzayda yer almaktadır. Dolayısıyla kümenin tanımlanması benzerlik ve farklılık kavramlarını da karşımıza çıkarmaktadır. Küme noktalarının geometrik gösterimi ikiden fazla boyutta gerçekleştirildiğinde, noktaların arasında yer alan mesafelerin birden fazla boyutta hesaplanmasına ihtiyaç duyulmaktadır.

Daha öncede belirtildiği gibi bir küme içindeki verilerin benzerliği en fazla, grup dışındaki verilerin uzaklığı ise en fazladır. Veri tipine bağlı olarak veri setler içinde uzaklık-benzerlik ölçüleri hesaplanır. Uzaklığı ve yakınlığı belirleyebilmek için en az 2 nokta olması gerekir. Benzerlik ve uzaklığı belirlemek için birçok algoritma var fakat bunların içinde en yaygın olanlardan bir tanesi mesafe için öklit benzerlik ölçütüdür. Veri matrisi X'de bulunan n birimle p değişkenin yer aldığı uzaklık matrisi D ile gösterilmektedir. Bu matrise ait elemanlar ise d_{ij} şeklinde gösterilmektedir. Diğer taraftan birimler arasındaki benzeme seviyeleri, benzerlik matrisi S, matrise ait elemanlar da s_{ij} biçiminde ifade edilir. Birimler arasındaki benzerlikler $s_{ij}=100(1-d_{ij}/maxd_{ij})$

Sayısal veriler için kullanılmakta olan uzaklık ölçütlerinin formülleri aşağıda verilmiştir.

[100] Murat ve Şekerler, a.g.e., s.90.

Öklid Uzaklık Ölçüsü

$$d_2(x_i, x_j) = \sum_{k=1}^{p} \left[\left| x_{ik} - x_{jk} \right|^2 \right]^{1/2}$$

Mahalanobis Uzaklık Ölçüsü

$$d_2(x_i, x_j) = \left[\sum_{k=1}^{p} w_k^2 (x_{ik} - x_{jk})^2 \right]^{1/2}$$

Mahalanobis Uzaklık Ölçüsü

$$d_2(x_i, x_j) = \left[\sum_{k=1}^{p} w_k^2 (x_{ik} - x_{jk})^2 \right]^{1/2}$$

2.5. Kümeleme Analizi Yöntemleri

Bireyleri ya da nesneleri kümelemede birçok farklı yöntem kullanılabilmektedir. Bu farklı teknikleri hiyerarşik ve bölümlemeli olarak ayrılmaktadır.

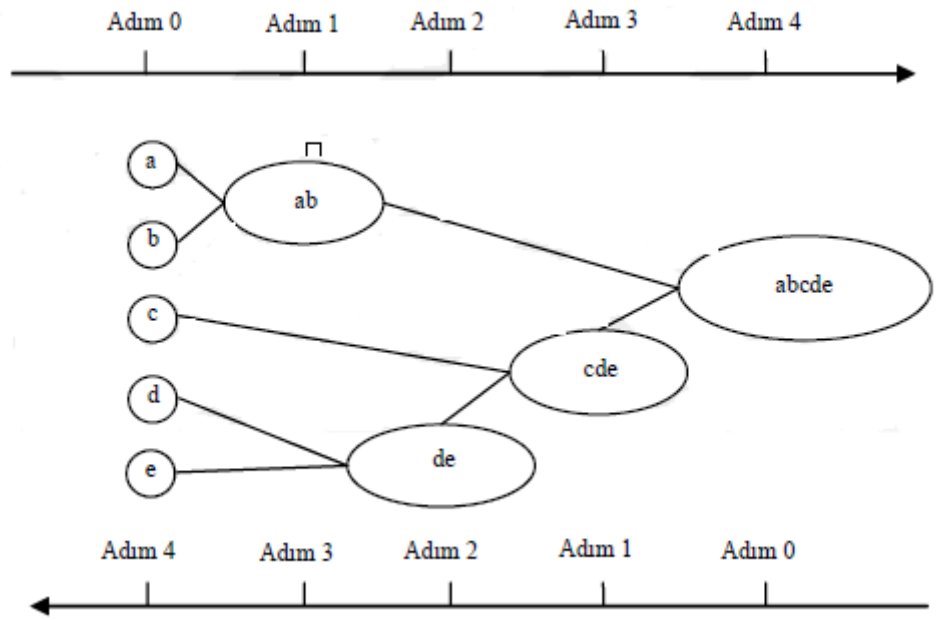
2.5.1.Hiyerarşik Yöntemler

Daha çok küçük veri tabanlarında tercih edilir. Hiyerarşik yöntemler karar ağacına sahiptir bu ağaç dallara ve oradan yapraklara ayrılır. Yapraklara ulaşınca ağaç biter. Ağacın dallara ayrıldığı yerlere düğüm adı verilir. Hiyerarşik yöntemler birleştirici ve ayrıştırıcı kümeleme algoritmaları olarak 2 ye ayrılır birleştirici kümeleme algoritmasında en başta her bir veriyi küme olarak görür ve bu kümeleri algoritmalara göre birleştirerek ayrı ayrı kümelere oluşturur. Ayrıştırıcı kümeleme algoritmasındaysa en başta tüm veriler tek küme olarak alınır daha verilerin farklarına başlangıçtaki tek olan küme ayrı ayrı kümelere bölünür. Hiyerarşik yöntemlerde en ünlü algoritmalar BIRCH, SLINK, CLUCDUH, CHAMELON ve CURE algoritmalarıdır.[101]

Kümeleme analizleri içerisinde yaygın olarak kullanılan hiyerarşik (aşamalı) kümeleme, parametrelerin belirsiz olduğu, kaç küme oluşturulacağının bilinmediği durumlarda aşağıdan yukarıya veya yukarıdan aşağıya doğru[102] küme ağacı şeklinde ayırma esasına dayanmaktadır. Hiyerarşik kümeleme algoritmalarında bir nesne beli bir kümeye atandıktan sonra başka bir kümeye dâhil olamaz[103]. Bu teknik, küme birimlerini birbirleri ile farklı aşamalarda bir araya getirip ardışık kümeler oluşturmayı ve oluşan bu kümelere dâhil olacak verilerin hangi uzaklık veya benzerlik seviyesinde olduğunu tespit etmeye yarayan bir algoritmadır. Hiyerarşik kümeleme algoritmalarında şekil 2.2. ve şekil 2.3.'te görünen dendrogram adı da verilen öbek ağaçlarından yararlanılır.

[101] Silahtaroğlu, G., Veri Madenciliği Kavram ve Algoritmaları. 2013. s-163, 208, 215.
[102] Seker, 2015, s.32
[103] Bilen vd., 2014, s.82

Birleştirici(AGNES)



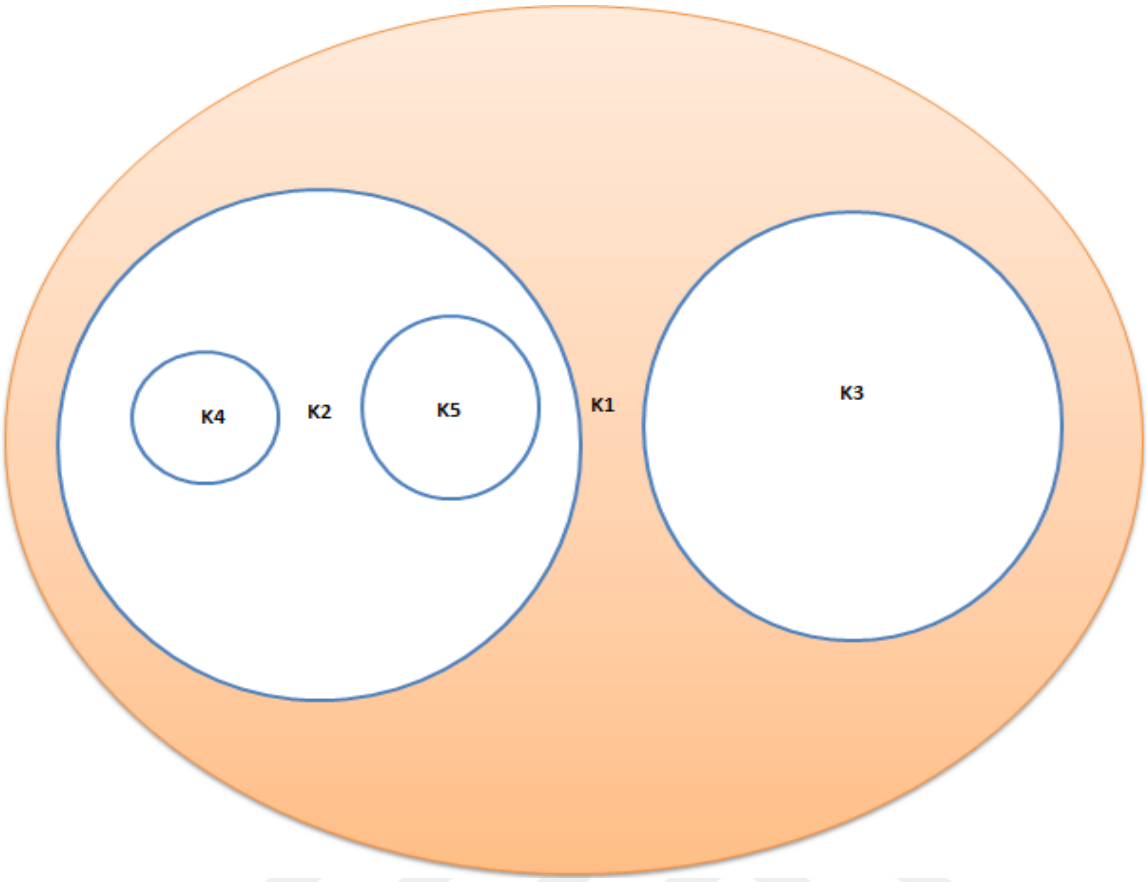
Ayırıcı(DIANA)

Şekil 2.1. Birleştirici ve ayırıcı hiyerarşik kümeleme[104]

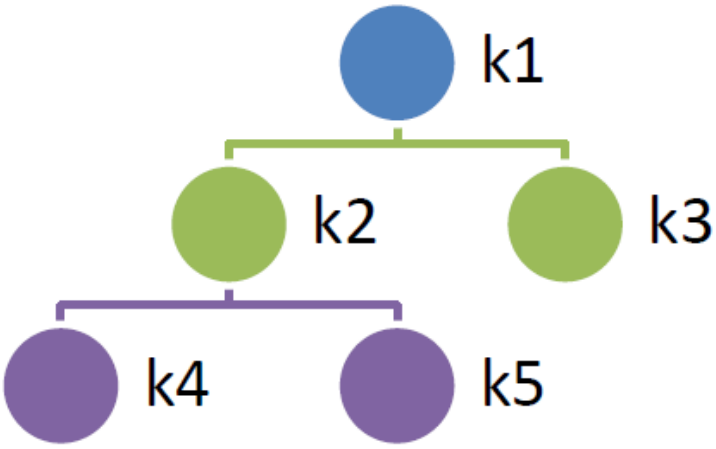
Kaufmann ve Rousseeuw (1990) tarafından ortaya atılan birleştirici kümeleme tekniği verileri kümelemek için aşağıdan yukarıya doğru çalışan bir yöntemdir.Şekil 2.1.'den de anlaşılacağı üzere birleştirici kümeleme yönteminde 0. adımda veri kümesindeki her nesne bir küme olarak kabul edilmektedir. Daha sonra birbirine en yakın iki nesne yeni bir küme olarak birleştirilir. Her aşamada gözlem sayısı azalmaktadır.[105] Nesneler arasındaki benzerlikler belirlenirken Öklid veya Manhattan uzaklık formülleri kullanılabilir.

Ayırıcı hiyerarşik kümeleme yukarıdan aşağıya doğru[106] çalışan bir strateji izler. Bu teknik şekil 2.1.'den de anlaşılacağı üzere 0. adımda veri setinde bulunan nesnelerin tümü bir küme olarak kabul edilir. Sonraki her adımda kendi aralarında benzer olan nesneler bir araya getirilerek kümeler oluşturulur. Algoritmanın her adımında kümeler daha küçük kümelere bölünür.[107] Bu işlem homojen kümeler elde edilene kadar devam eder.

[104] Selvi ve Çağlar 2017
[105] Özbek ve Atik, 2013, s.197
[106] Kovaleva ve Mirkin, 2005, s.416
[107] Selvi ve Çağlar, 2017, s.418



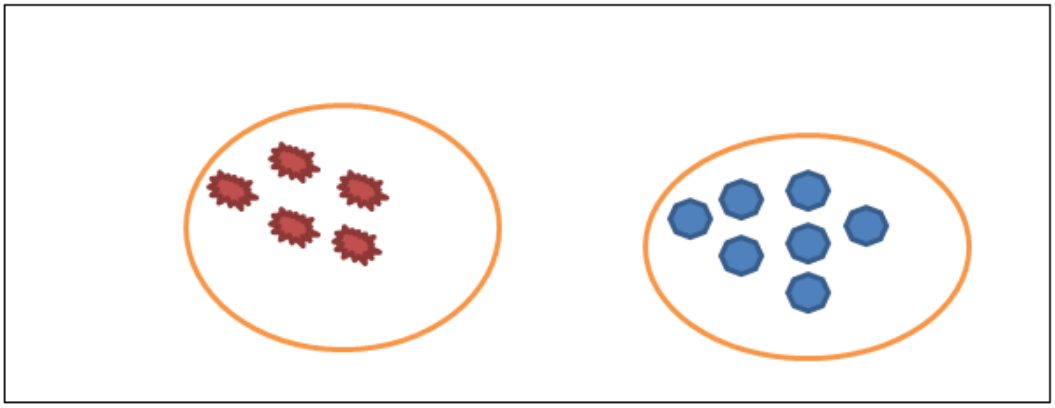
Şekil 2.2. Hiyerarşik kümeleme 1



Şekil 2.3. Hiyerarşik kümeleme 2

2.5.2.Bölümlemeli Yöntemler

Bölümlemeli yöntemlerde veri setindeki veriler daha önceden belirlenmiş kritere göre ayrılır. D veri tabanında m tane veri varsa bu veri tabanı en fazla m-1 kadar bölüme ayrılabilir. Bölümlemeli yöntemlere de hiyerarşik yöntemlerin aksine birbirine bağlı yöntemler yerine her veri bölünen bağımsız bir kümede yer alır. Uygulama çalışmasında bölümlemeli yöntemlerden K-Means algoritması ölçütü üzerinde durulacaktır.



Şekil 2.4. Bölümlemeli yöntemler

2.5.2.1. K Medoids Yöntemi

Medoid kümeleme yönteminde, küme içi gözlemlere ait n birimin kendi içinde benzer olmasıyla ve kümelerin arasındaki gözlem değerlerinin ise farklı olmasıyla k kümenin ayrımını yapmak amaçlanmaktadır. Bu amaç medoid ismi verilen k küme ve bu kümeleri tanıtacak olan çekirdek noktalar yardımıyla gerçekleşmektedir. Medoid bir kümedeki başka olgular arasında yer alan farklılıkların minimum hesaplandığı birimlerden oluşmaktadır.[108] Bu algoritmanın esasını, verinin farklı yapıdaki özellikleriyle meydana gelen k adet temsilci nesnenin bulunması oluşturmaktadır. 1987'de Kaufman ve Rousseeuw tarafından geliştirilen algoritma yaygın olarak kullanılmaktadır. Temsil yeteneği olan nesnenin kümedeki diğer nesnelerle olan ortalama mesafeyi minimum hale getirmesi onu kümedeki en merkezi nesne

[108] Özdamar, a.g.e., s.336.

yapmaktadır. Dolayısıyla bu ayırma yönteminde her bir nesnenin ve referans noktasının arasında var olan benzersizliklerin toplam değerini minimum yapma mantığı kullanılmaktadır. Kümeleme analizinde temsil yeteneği olan nesneler genellikle merkezi tipler olarak adlandırılmaktadır.[109]

2.5.2.2. Bulanık Kümeleme

Kümeleme analizinin temelini veri setine ait değişkenler dikkate alınarak birimler arasında yer alan mesafenin göz önüne alınması oluşturmaktadır. Ayrıca yeni bir birimin gruplardan hangisine dahil olacağının tahmin edilmesi de önemli bir yer tutmaktadır. Kümeleme analizi olasılıklı, katı ve bulanık diye üye kayıtları açısından incelenmektedir. Bütün gözlemler ya bir kümeye aittirler ya da ait değildirler. Bulanık kümeleme de üyelik kayıtları 0 ve 1 arasında değişmektedir. Verilerin eş zamanlı olarak birden çok kümeye üyeliği söz konusu olmaktadır. Bulanık kümeleme tekniğinde veri kayıtlarının kümeye üye olma olasılıkları toplamı 1'dir. Bu teknikte herhangi bir birimin bir kümede bulunma olasılık değeri bütün olası kümeler içinde 0 ve 1 arasında değişmektedir. Böylece verilere ait noktalar eş zamanlı olarak 1'den çok kümeye ait olabilmektedir. Kümelere bağlı üyelerin durumu bulanık olduğundan veriye ait noktanın hangi kümeye dâhil bulunduğunu gösteren bir tek değer bulunmaktadır. Bu tek bir değer yerine değerler kümesi bulunmaktadır. Herhangi bir birime ait belli bir bölüm bir kümeye üye iken, başka bir bölümüyle söz konusu kümenin dışında olabilmektedir. Hangi kümede olma olasılığı yüksekse o kümeye atanmaktadır. Bu nedenle klasik kümeleme tekniklerini bulanık kümelemenin alt konusu gibi incelemek uygun görülmektedir.[110]

2.5.2.3. K-Means Algoritması

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

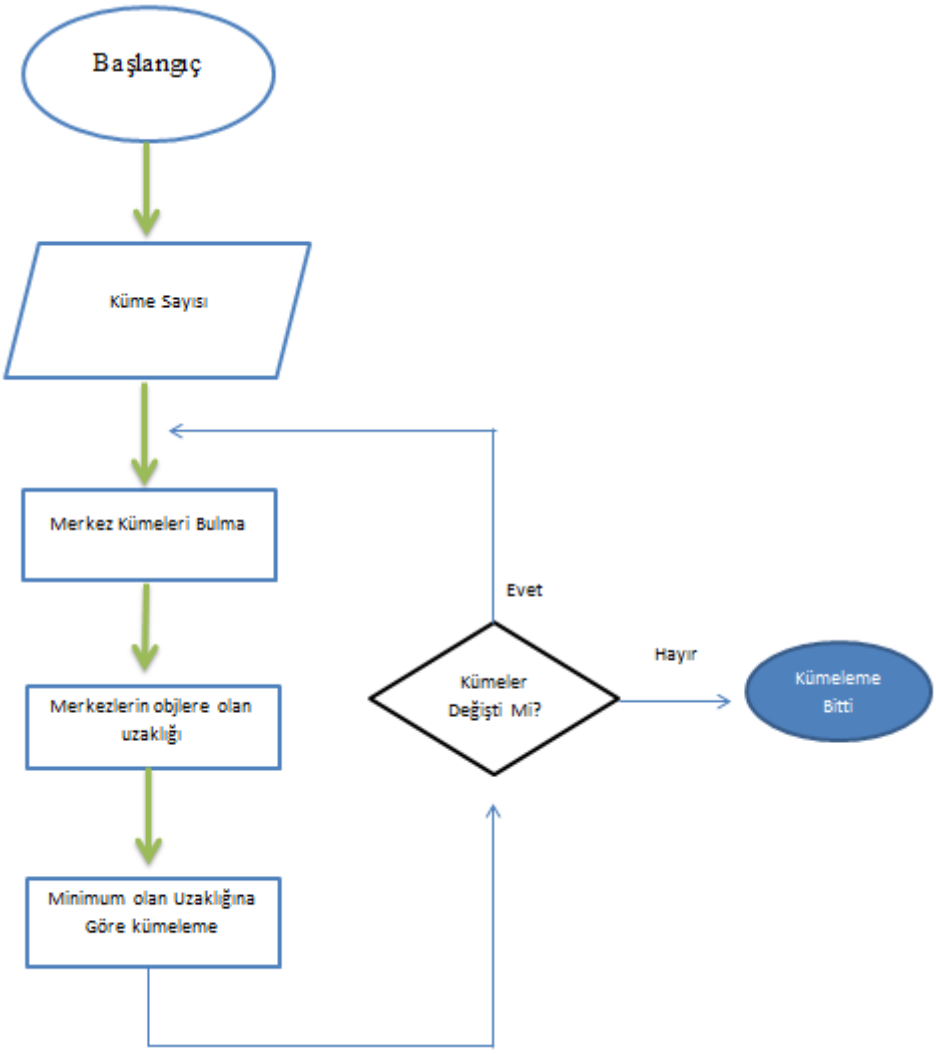
[109] Işık ve Çamurcu, a.ge.,s.76
[110] Yılancı, a.g.e., s.456

arasında ise uzaklık minimum olacak şekilde kümeleme işlemi yapılır. Öklit algoritmasının formülü şu şekildedir.

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

K-Means algoritmasının çalışma mantığı aşağıdaki gibidir.

- İlk veri setinin kaç kümeye ayrılacağı belirlenir.
- İlk sefere mahsus eldeki verilerin içinden küme sayısı kadar rastgele merkez kümeler seçilir. Seçilen bu merkezlerin her biri veri setinin ayrılacağı her bir kümeyi temsil etmektedir.
- Veri setinin içindeki her bir elemanın merkez kümelere göre öklit bağlantısı alınır ve öklitten çıkan sonuç en az olan o merkez kümenin kümesine eklenir.
- Veri setindeki her eleman merkez kümelerin temsil ettiği kümelere yerleştirildikten sonra yeni merkez kümeler bulunur. Yeni merkez kümelerin bulunması için kümeye eklenen her veri toplanıp kümedeki eleman sayısına bölünür.
- En sonunda bu işlemler bir döngüde devam ettirilir. Bir önceki döngüdeki kümeler bir sonradaki kümeler ile aynı ise işlem sonlandırılır. Veri seti K-Means algoritmasına göre kümelenmiş olur.



Şekil 2.5. K-Means Algoritması Akış Şeması

ÜÇÜNCÜ BÖLÜM

UYGULAMA:BÖLÜMLEMELİ YÖNTEMLERDEN K-MEANS ALGORİTMASININ SİGORTA SEKTÖRÜNDE UYGULANMASI

3.1. Uygulamanın Amacı

Bu uygulamanın amacı, müşterilerin memnuniyetsizliklerinin temel sebeplerini ortaya çıkarmak ve veri madenciliği yoluyla kurallar oluşturmak, ayrıca bu kuralların tüketicilerin kişisel özelliklerine (yaş, cinsiyet, ..vs) bağlı olarak değişimini incelemektir.

Teknolojik ilerlemeler insanlar arası ilişkileri etkilediği kadar şirketlerin müşterileri ile olan ilişkilerini, satış ve pazarlama sistemlerini, hatta tüm kurumsal organizasyonlarını etkilemiştir. Müşteri ilişkileri yönetimi kavramı, müşteri odaklı bir yaklaşımla uzun dönemli ilişki kurarak şirketin karlılığını arttırmayı hedefleyen bir yaklaşımdır. Bu kapsamda birbirine benzer nitelikte müşterilerin özelliklerini tanımlamak ve gruplandırmak önemli bir faaliyettir. Bu amaçla farklı kaynaklardan alınan müşterilerin verileri bir araya getirilir ve bu müşterilerin karakteristik özelliklerini belirlemek için analizler yapılır. Veri madenciliği (VM) bu karakteristik özelliklerin tespitinde kullanılabilecek veri analizi yöntemlerini içeren bir disiplindir.

Bu çalışmada, Türkiye'de faaliyet gösteren bir sigorta şirketinin müşterilerine ait veriler VM'nin en çok kullanılan kümeleme algoritmalarından K-Means algoritması ile analiz edilmiştir. Bu analiz ile elde edilen sonuçlar yardımıyla, şirketin benzer müşterilerinin özelliklerini tespit etmesi ve onlara uygun yeni pazarlama stratejileri geliştirebilmesi hedeflenmektedir.

3.2. Uygulamanın Kapsamı

Sigortacılık sektörü üzerine gerçekleştirilen uygulamada Türkiye'nin önemli sigorta şirketlerinden birinin farklı tarih aralıklarında satışı gerçekleştirilmiş olan poliçe verileri K-Means algoritması ile analiz edilmiştir.

2019.01.01-2019.01.31 tarihleri arasında işletme tarafından yapılan ankette 521 müşteri nin memnuniyet verileri alınıp veri tabanına işlenmiştir. Bu müşteri grubunun memnuniyet verisi ile birlikte müşterinin poliçe ürünü, hasar adedi, demografik özellikleri ile

ilişkilendirilerek memnuniyetsizliğin altında yatan nedenler araştırılmıştır.

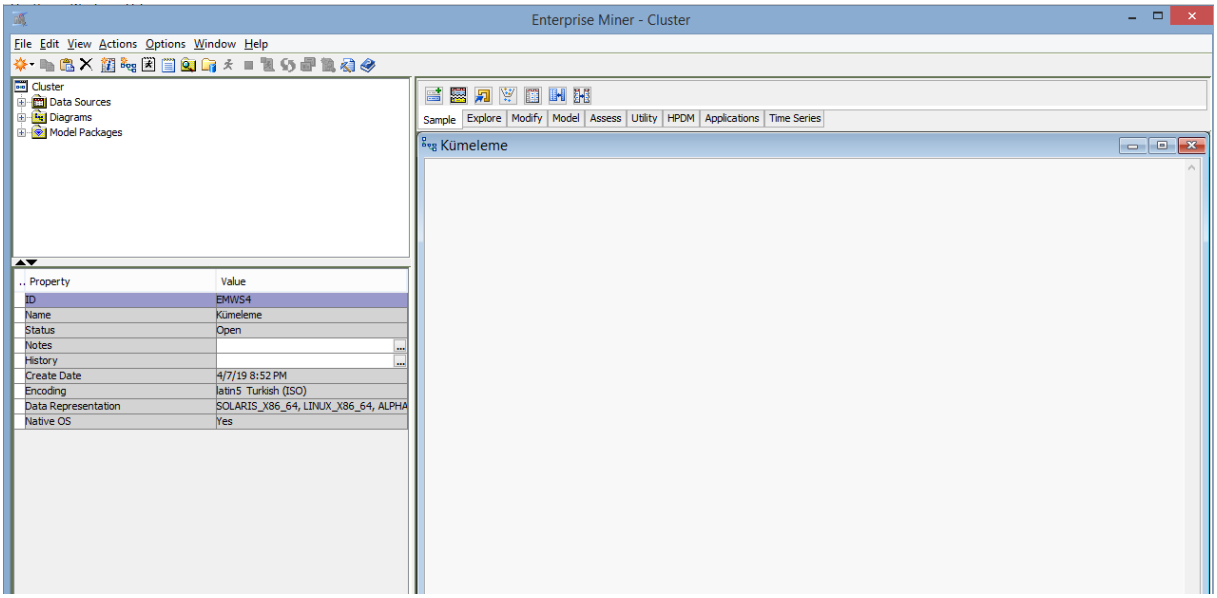
3.3. Uygulamanın Geliştirme Ortamı

Uygulamanın gerçekleştirileceği ortamları belirlerken bazı kriterler dikkate alınmıştır.Bu kriterleri aşağıdaki gibi ifade edebiliriz:

- Uygulamanın gerçekleşme süresinin kısalığı
- VTYS programlama dilinin VM uygulaması için uyumluluğu
- Kullanılacak olan VTYS sorgulama dilinin uyumluluğu
- Makine veya bilgisayarların donanım özellikleri

Uygulamanın gerçekleştirileceği ortam tezde detaylı olarak bahsedilen Sas Enterprise Miner yazılım programıdır.Bunun nedenlerini şu şekilde sıralayabiliriz:

- Bünyesinde çalışmış olduğum sigorta şirketinin bu yazılım programını kullanması
- Uygulamada kullanılacak olan veri madenciliği kümeleme yöntemlerinden K-Means algoritmasının bu yazılım ortamı içerisinde hazır fonksiyon olarak bulunması
- Verilerin tutulduğu VTYS ile Sas Enterprise Miner yazılımının uyum içerisinde kullanılabilmesi
- Uygulama sırasında kullanılan diğer fonksiyonlara veya özelliklere sahip olması



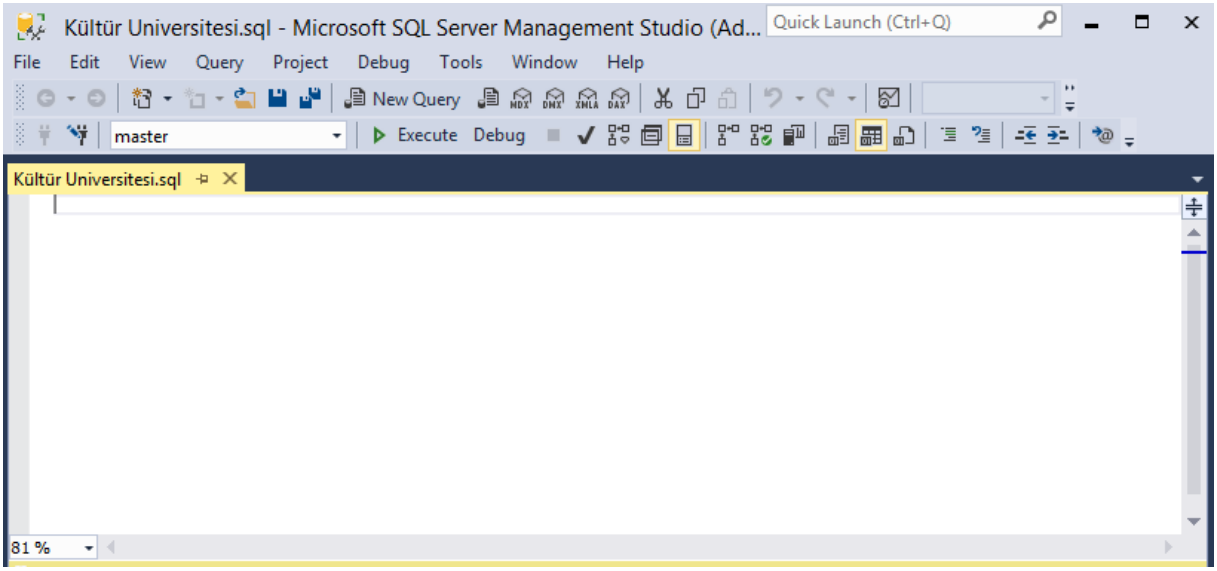
Şekil 3.1. Sas Enterprise Miner Yazılım Ortamı

Uygulamanın gerçekleştirilmesinde kullanılan veri tabanı yönetim sistemi Microsoft SQL Server 2017'dir.Bu veri tabanı yönetim sisteminin uygulamada kullanma sebeblerini şu şekilde sıralayabiliriz :

•Bünyesinde çalışmış olduğum sigorta şirketinin müşterilerin tüm verilerini Microsoft SQL Server 2017 veri tabanı yönetim sistemi üzerinde tutmaktadır.

•Sas Enterprise Miner yazılım programı ile beraber senkronize bir şekilde çalışabilmesi

• Büyük hacimdeki verilerin saklanması ,yönetilmesi ve kullanılmasındaki performansının iyi olması



Şekil 3.2. Microsoft SQL Server 2017

3.4. Verilerin Yapısı

Uygulama kapsamında müşteri verileri sigorta şirketinin veri tabanında 3 farklı tabloda tutulmaktadır. Müşteri tablosunda genel olarak müşterinin ürünü, aile bilgileri,nufüs bilgileri,yaş,cinsiyet,medeni durum,eğitim durumu gibi demografik bilgileri yer almaktadır.Mennuniyet tablosunda yapılan anketlerde müşterinin vermiş olduğu puanı,memnuniyet neden açıklaması gibi verileri yer almaktadır.Hasar tablosunda ise müşterilerin hasar dosya adetleri gibi hasara konu olan veriler tutulmaktadır. Bu 3 tablolunun alanları ve açıklamaları EK-1,EK-2 ve EK-3 'te verilmiştir.

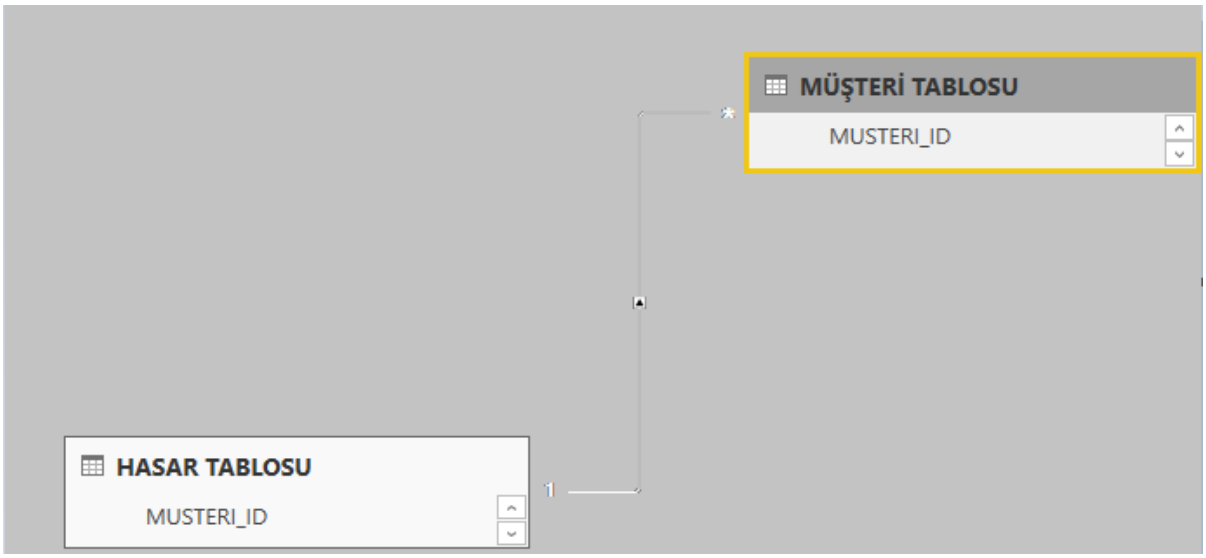
3.5. Uygulamanın Veri Madenciliği Süreçleri

Veri madenciliği uygulama süreci önceki bölümlerde de açıklandığı gibi küçük ver birbirleriyle bağlantılı adımlardan oluşan bir süreçtir.Bu süreçte önemli olan verilerin kendisi ve doğruluğudur. Doğru ve güvenilir sonuçlara ulaşmak için veri madenciliği sürecinde yer alan adımların doğru ve eksiksiz bir biçimde yerine getirilmesi gerekmektedir.

Veri madenciliğinin her bir aşaması sonraki diğer aşamasıyla bağlantılıdır. Veri madenciliği sürecinin birbirine bağımlı süreçlerden oluşması , veri yığınları arasında, soyut kazılar yapılarak veriyi ortaya çıkarmanın yanı sıra, bilgi keşfi sürecinde örüntülerin ayrıştırılarak süzülmesi ve bir sonraki adıma hazır hale getirilmesini ifade etmektedir.

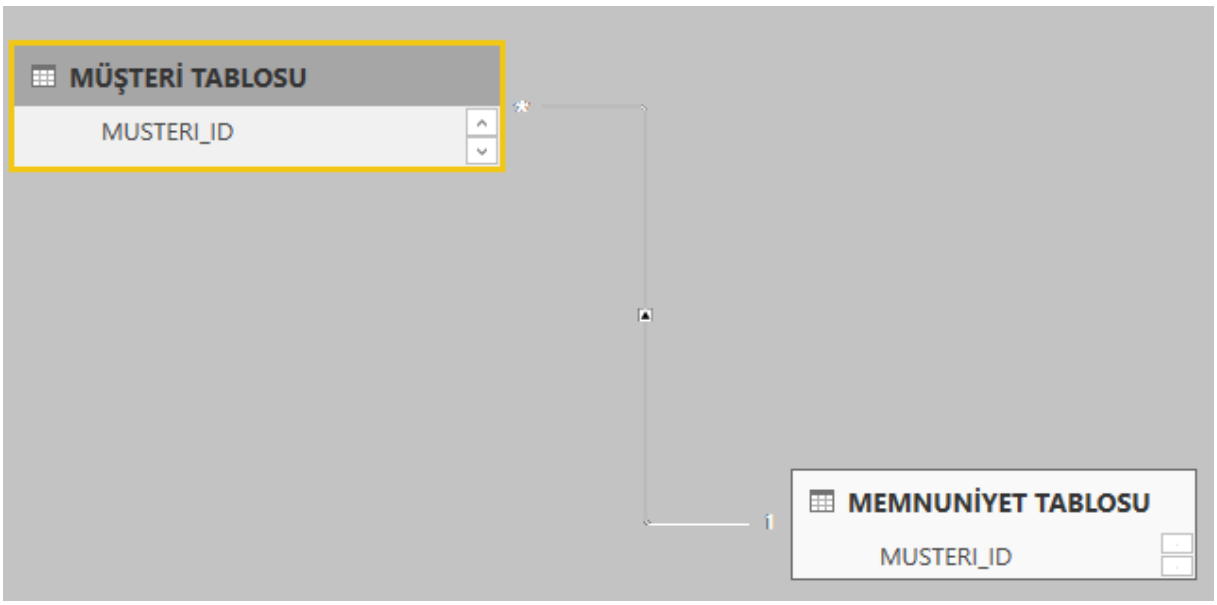
3.5.1. Veri Toplama ve Birleştirme

Bölüm 3.5 'te belirtildiği gibi uygulama çalışmasında yer alan veriler 3 tabloda tutulmaktadır. Uygulamanın bu adımında müşteri bilgilerini tutan müşteri tablosu ile müşteri hasar bilgilerini tutan hasar tablosu birleştirilmektedir.Aşağıda verilen şekilde olduğu gibi müşteri_id bilgisi üzerinden Microsoft SQL Server 2017 programlama dili kullanarak 2 tablo birleştirilmiştir.



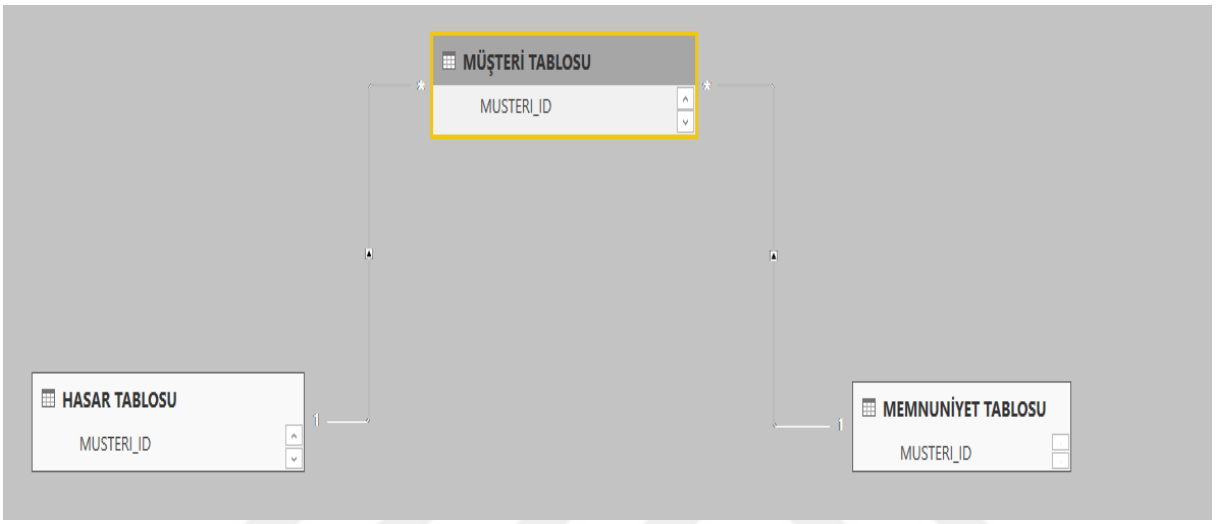
Şekil 3.3. Müşteri & Hasar Tablo Birleştirme

Benzer şekilde aşağıda verilen şekilde olduğu gibi müşteri_id bilgisi üzerinden müşteri tablosu ile müşteri memnuniyet bilgilerini tutan memnuniyet tablosu birleştirilmektedir.



Şekil 3.4. Müşteri & Memnuniyet Tablo Birleştirme

Sonuç olarak Müşteri tablosu müşteri id bilgisi üzerinden hasar ve memnuniyet tablolarına bileştirilerek aşağıdaki şekilde gösterildiği gibi veri modeli oluşmaktadır. Bu model sayesinde müşterinin hem demografik bilgileri hem memnuniyet bilgileri hem de hasar bilgileri elde edilmiş olmaktadır.



Şekil 3.5. Müşteri Veri Modeli

3.5.2. Veri Kalitesi Analizi

Birleştirme aşamasından sonra oluşan modeldeki veriler analiz edildiğinde müşterinin demografik bilgileri ,hasar bilgileri, memnuniyet bilgileri eksiksiz ve doğru olarak geldiği gözükmektedir. Bu adımda bahsedilen veri kalitesi analizinden sonra veri madenciliğinde kullanılacak olan tablo hazırlanmış oldu. Bu işlemlerden sonra tabloda 521 satır kayıt yer almaktadır

3.5.3. Veri Madenciliği

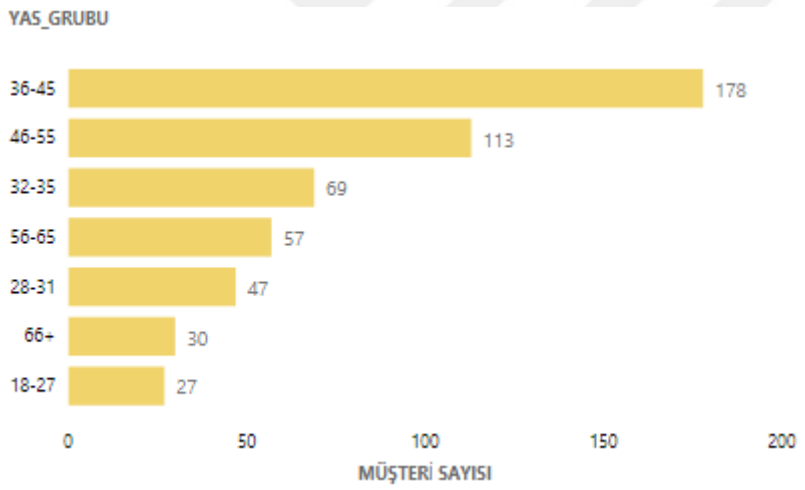
Uygulamanın bu bölümünde hazırlanmış olan veri modeli veri madenciliği işlemine sokulmaktadır. Veri madenciliği sürecine Şekil 3.5.'te müşteri veri modelinde yer alan 3 tablo girmektedir.

Veri madenciliği işlemi Sas Enterprise Miner yazılımı üzerinde gerçekleştirilmiştir.Uygulamının sonuçlarının gösterimi içim Microsoft Power BI yazılımı kullanılmıştır.

Öncelikle analizde kullanılacak olan tüm değişkenler veri madenciliği yazılımlarının grafiksel araçları ile incelenmiş ve aşağıdaki sonuçlar elde edilmiştir. Çalışmanın bu aşaması konu hakkında genel bir bilgi vermeyi de amaçlamaktadır.

Yaş Değişkeni

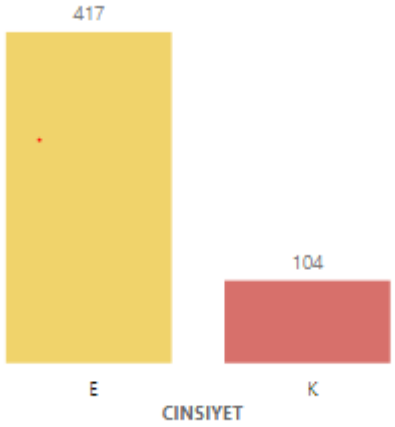
Aşağıdaki grafikte görüldüğü gibi, analize konu alan veri tabanında yer alan kişilerin büyük bir çoğunluğu 178 kişi olarak 36-45 yaş grubu içerisindedir. Bunu 113 kişi ile 46-55 yaş grubu takip etmektedir.



Grafik 3.1. Yaş Değişkeni Dağılımı

Cinsiyet Değişkeni

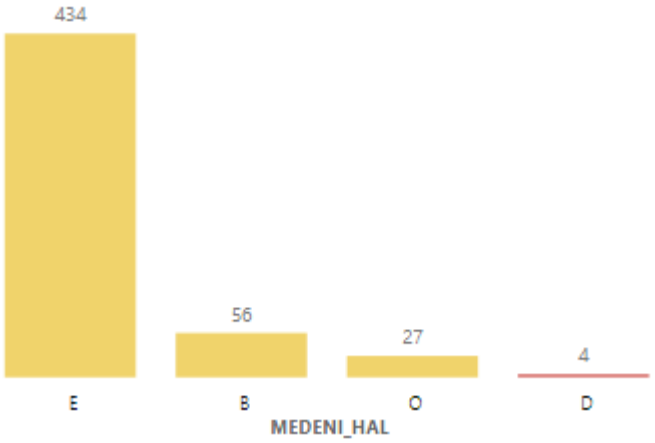
Veri tabanında yer alan müşterilerin cinsiyet dağılımı aşağıdaki grafikte gibi olup çoğunluğu 417 kişi ile erkekler oluşturmaktadır.



Grafik 3.2. Cinsiyet Değişkeni Dağılımı

Medeni Hal Değişkeni

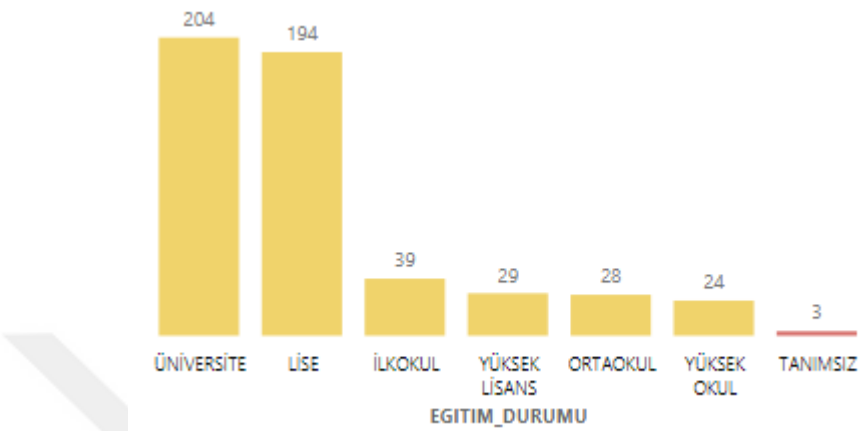
Veri tabanında yer alan müşterilerin medeni hal durumları ‘E’ evli,’B’ bekar ,’O’ dul, ‘D’ boşanmış değerlerine karşılık gelip kişilerin çoğunluğunu 434 kişi ile evliler oluşturmaktadır.



Grafik 3.3. Medeni Hal Değişkeni

Eğitim Durumu Değişkeni

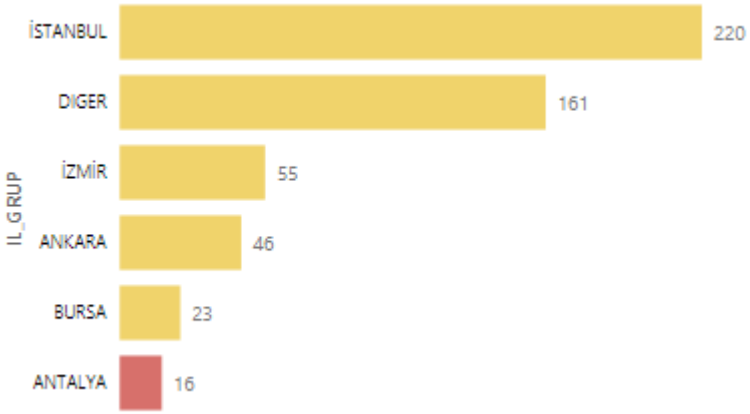
Analize konu olan veri tabanında yer alan müşterilerin eğitim durumu aşağıdaki grafikte gibi olup çoğunluğu 204 kişi ile üniversite mezunu oluşturmaktadır.



Grafik 3.4. Eğitim Durumu Değişkeni Dağılımı

İl Değişkeni

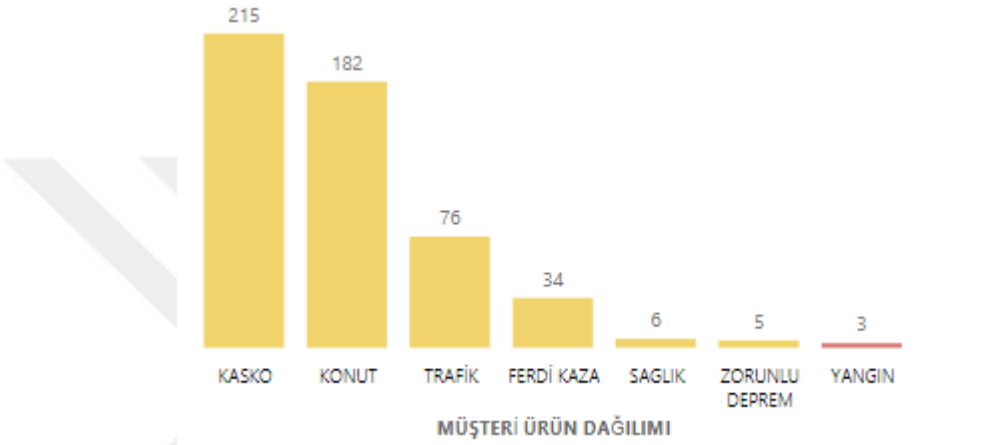
Müşterilerin ikamet ettiği illerin dağılımı grafik 3.5 'te belirtildiği gibi olup il başına düşen kişi sayısının az olduğu iller "DIGER" olarak grublanmıştır.220 kişi İstanbul ili çoğunluğu oluşturmaktadır.



Grafik 3.5. İl Değişkeni Dağılımı

Müşteri Ürün Değişkeni

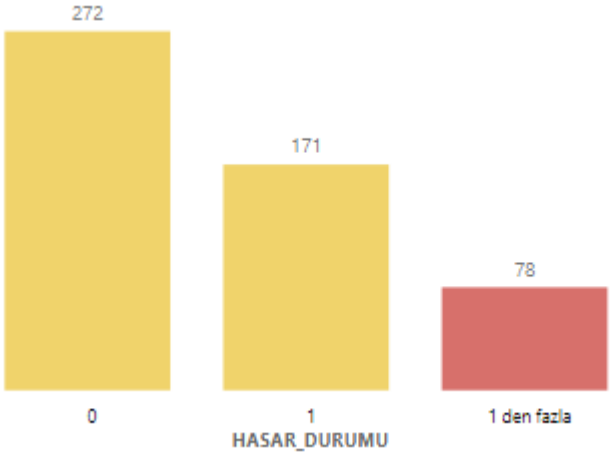
Müşterilerin sigorta şirketinden almış oldukları poliçenin ürünü gösteren bu değişkenin dağılımı grafik 3.6'da gösterilmiştir.Buna göre ağırlıklı olarak 215 kişi ile kasko ürünü çoğunlukta gözükmektedir.



Grafik 3.6. Müşteri Ürün Değişken Dağılımı

Hasar Değişkeni

Müşterilerin sigorta şirketlerinden teminat altına aldıkları mal veya hizmetlerindeki(konut,araç vs.) olası zafiyet hasar olara tanımlanmakta olup müşterilerin ilgili üründe kaç kez hasara uğradığını belirten değişkendir. "0" hiç hasarı olmayan,"1" ise 1 kez hasara uğramış ve 1 den daha fazla hasar adetine sahip olanlar ise "1 den fazla"değerine sahip olarak kategorilenmiştir.Grafik 3.7'de belirtildiği üzere 272 kişi 0 değerinde olup yani hiç hasarı olmayanların çoğunlukta olduğu gösterilmektedir.



Grafik 3.7. Hasar Değişken Dağılımı

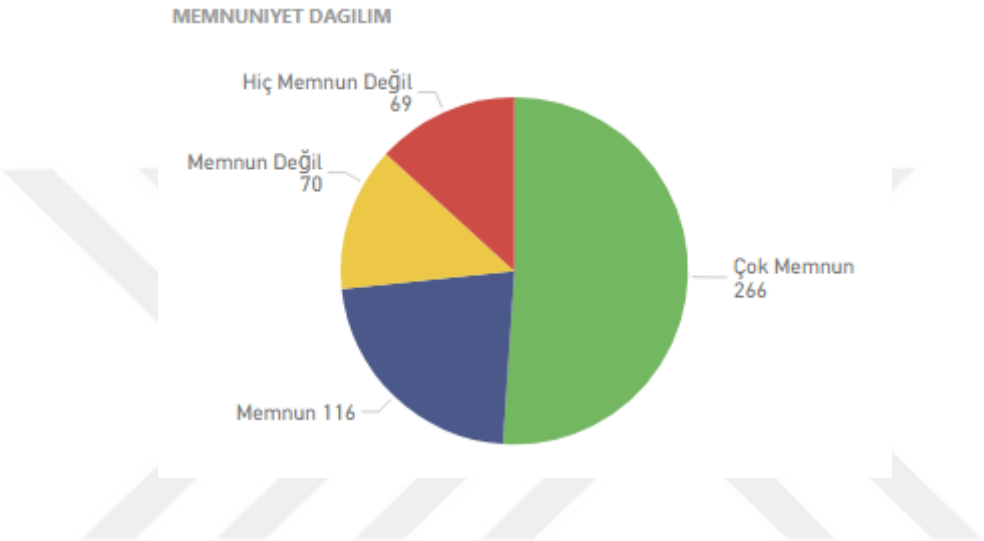
Memnuniyet Puanı

Analize konu olan müşterilerin ankete vermiş oldukları puanlama 1'den 10 'a kadar olup memnuniyet puanları müşteri adedi bazında tablo 3.1 'de verilmiştir.

Memnuniyet Puanı	Müşteri Adeti
0	69
1	7
2	9
3	9
4	13
5	18
6	14
7	20
8	35
9	61
10	266
Toplam:	521

Tablo 3.1 Müşteri Memnuniyet Puan Dağılımı

Puanlamada 0 puanı verenleri "Hiç memnun değil" 1-6 puan arasını verenleri "Memnun Değil" 7-9 puanı arasını verenleri "Memnun" 10 puanı verenleri ise "Çok Memnun" olarak sınıflandırılmıştır.Buna göre müşteri adeti bazında memnuniyet durumu şekil 3.6'da gösterilmiştir.



Şekil 3.6. Memnuniyet Dağılımı

Anket verisinde müşterilerin vermiş oldukları memnuniyetsizlik sebebleri aşağıdaki tablo 3.2 'de verilmiştir.

Memnuniyetsizlik Nedenleri
Bir önerim yok
Çağrı merkezi personelinin bilgi ve tecrübesi artırılmalı
Çağrı merkezine kısa sürede ulaşamama
Çeşitli kanallardan ilettiğim sorularıma daha hızlı cevap verilmeli
Eksperin tutum ve davranışları
Daha fazla ilgi olmalı
İşletmeye istenilen yer veya kanaldan erişim zorluğu

Fiyatlar daha esnek olmalı
Fiyatların daha esnek olmaması
Hasar
Hasar onarım kalitesi
Hasar süreci ile ilgili genel bilgilendirme eksikliği
Hasarın sonuçlandırılma hızı
Hasarın ödenmemesi/eksik ödenmesi
İşlerim daha hızlı sonuçlandırılmalı
Müşteri temsilcisinin tutum ve davranışları
Müşteri temsilcisinin ürün ve teminatları hakkında yeterli bilgilendirme sağlamaması
Onarım şirketi personelinin tutum ve davranışları
Süreç ve işlemlerimle ilgili daha fazla bilgilendirme yapılmalı
Talebim hakkında doğru ve açık yanıtların verilmemesi
Talebin kısa sürede çözülmemesi
Yenileme döneminden önce aranmama
Yüksek fiyat

Tablo 3.2 Müşteri Memnuniyetsizlik Nedenleri

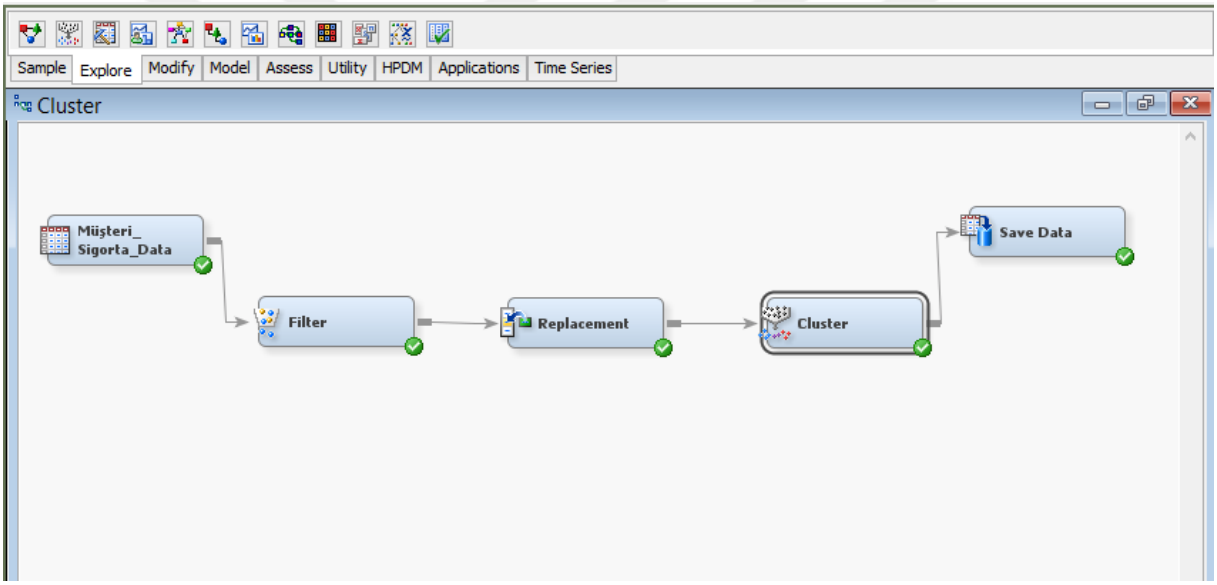
521 müşterinin mevcutda yer alan tüm bilgileri analiz edilmiştir.Memnuniyetsiz olan müşterileri tek bir grupmuş gibi görmek yerine, kaç değişik türde farklı özellikleri taşıyan müşteri grubu olduğunu belirlemek amacıyla K-Means algoritmasıyla kümeleme analizi yapılmıştır. Bu nedenle Şekil 3.6'da belirtilen memnuniyet dağılımında 69 kişi (hiç memnun değil) ve 70 kişi (memnun değil) olan kayıtlar analize sokulmuştur.Dolayısıyla Toplam 521 müşteri kitlesinden memnuniyetsiz olan 139 kişi için aşağıda belirtilen değişkenlere göre SAS Enterprise Miner yazılım programı üzerinde K-Means algoritması kullanılarak kümeleme işlemi yapılmıştır.

SAS Uygulamasında analize giren değişken :

- Memnuniyet Puanı

Kümelerde incelenen değişkenler şunlardır:

- Cinsiyet
- Yaş
- İl
- Medeni Hal
- Eğitim Durumu
- Hasar Adedi
- Ürün
- Ortalama Memnuniyet Puanı



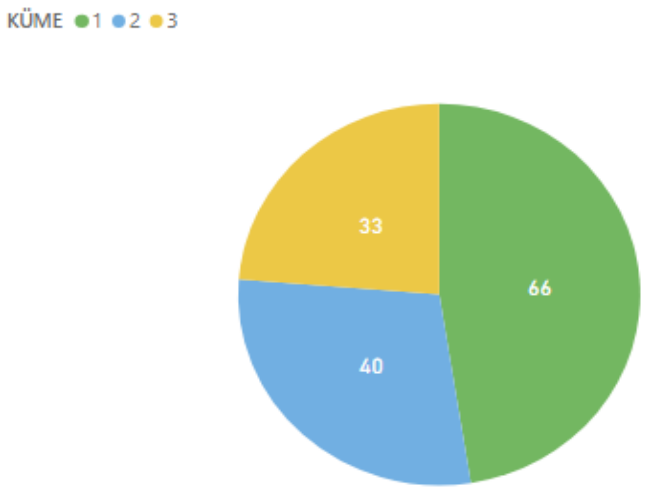
Şekil 3.7. Sas Enterprise Miner Ortamında Uygulama

Şekil 3.7'de gösterildiği gibi 139 kişinin analize giren verileri "Cluster" adımın içinde hazır olarak bulunan K-Means algoritması uygulanır.Uygulamanın sonucunda 3 farklı küme oluşur ve veri madenciliği süreci tamamlanır.Bundan sonraki aşama ise bu 3 farklı kümenin analizi olarak bilgi sunumu bölümde detaylı anlatılmıştır.

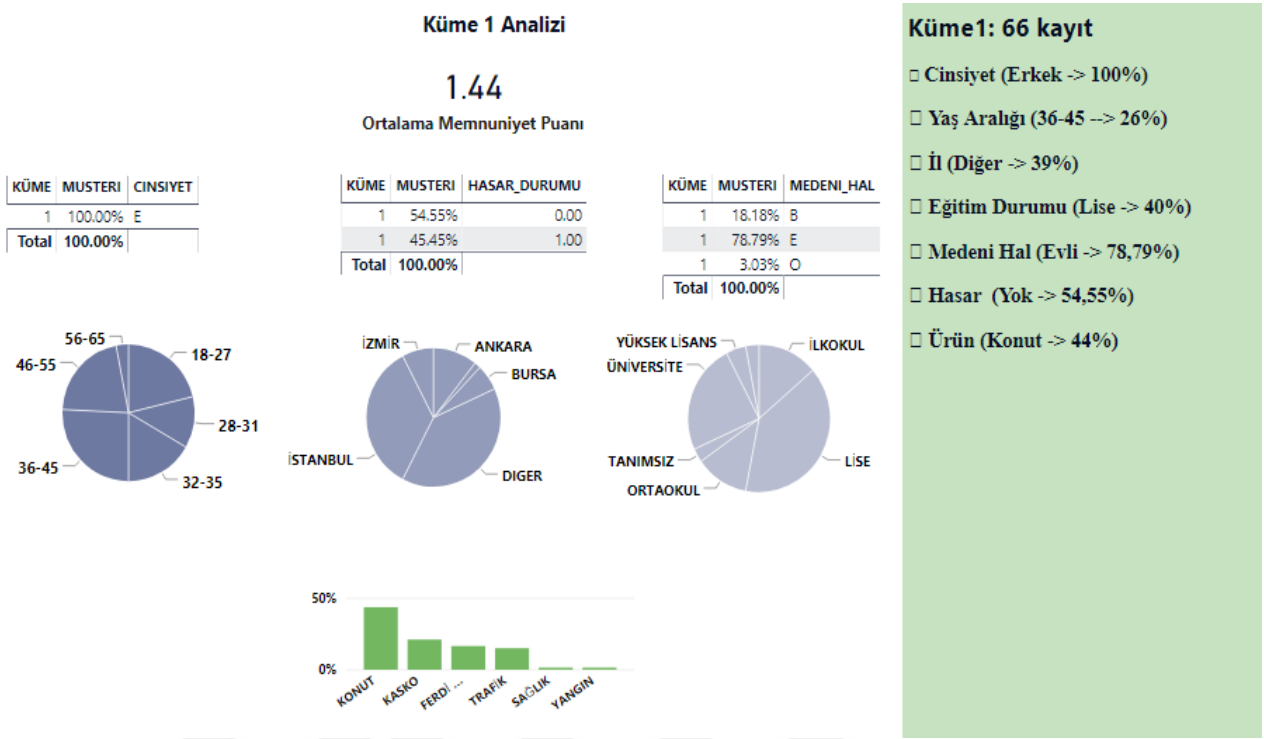
3.5.4. Bilgi Sunumu

K-Means algoritması uygulandıktan sonra, elde edilen kümeleme analizi sonucunda, üç ayrı kümenin ortaya çıktığı görülmüştür.

- Küme1: 66 kayıt
- Küme2: 40 kayıt
- Küme3: 33 kayıt

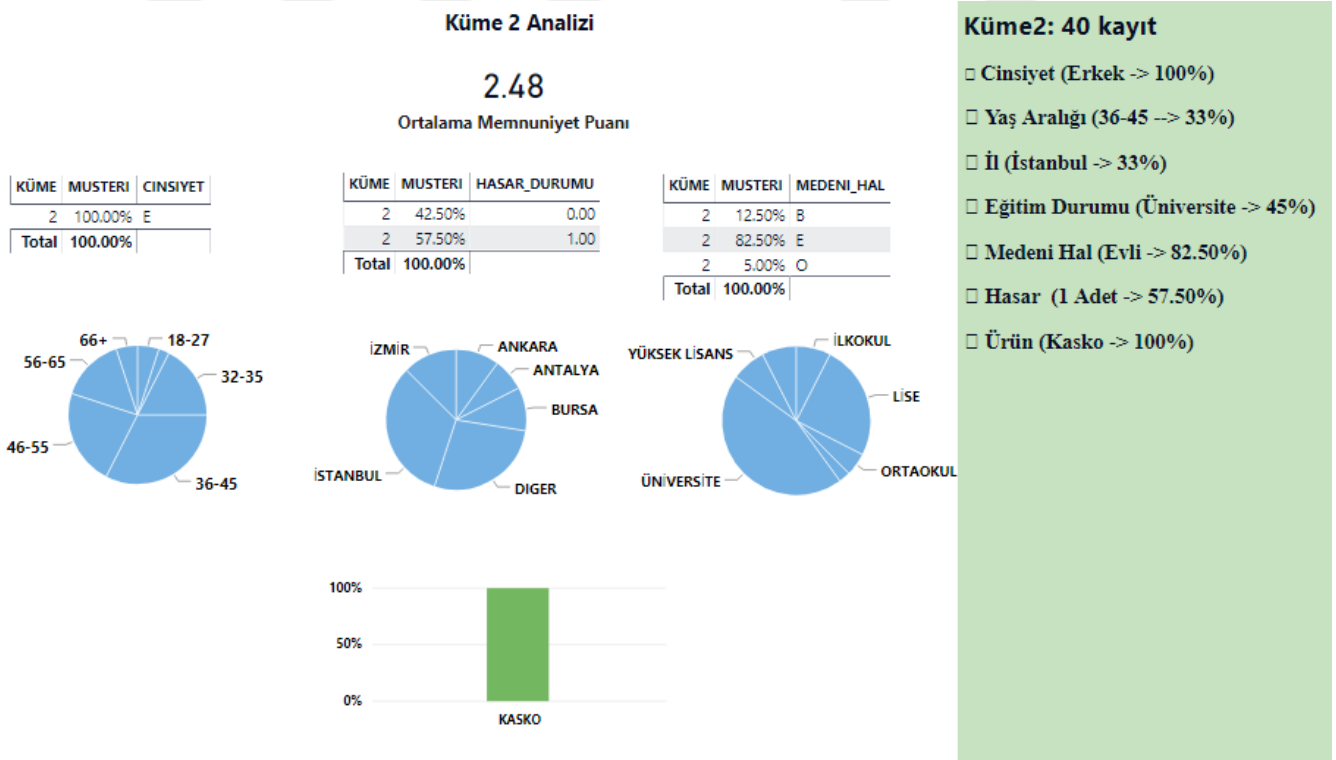


Şekil 3.8. Küme Dağılımları



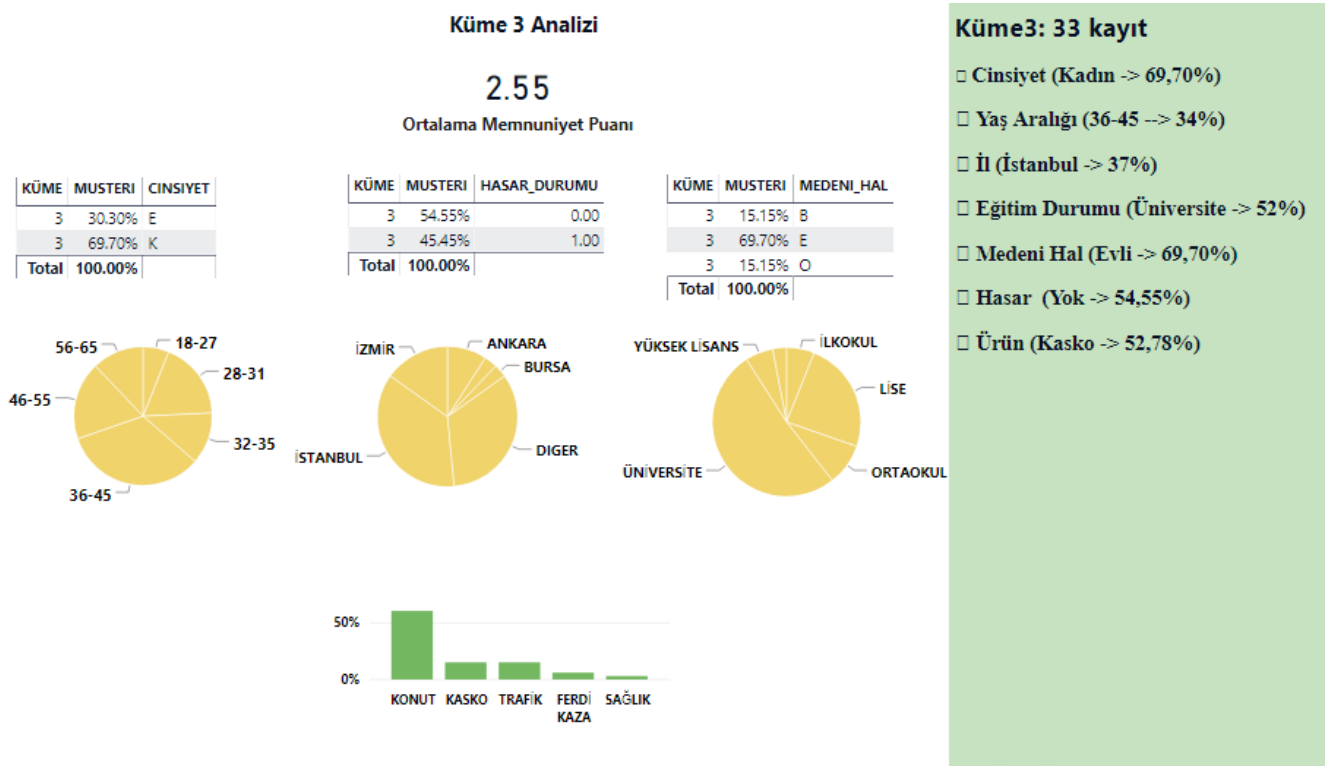
Şekil 3.9. Küme 1 Analizi

*Birinci kümede çoğunluk %30.21 oran ile memnuniyetsizlik sebebini "yüksek fiyat" olarak belirtmiştir.



Şekil 3.10. Küme 2 Analizi

*İkinci kümede çoğunluk %25.42 oran ile memnuniyetsizlik sebebini "eksperin tutum ve davranışları"olarak belirtmiştir." olarak belirtmiştir.



Şekil 3.11. Küme 3 Analizi

*Üçüncü kümede çoğunluk %30.45 oran ile memnuniyetsizlik sebebini "yenileme döneminden önce aranmama "olarak belirtmiştir.

Elde edilen verilerin K-means algoritması ile kümelenmesi sonucunda birbirine benzer müşteri gruplarının oluşturulması sağlanmıştır. Böylece sigorta şirketinin müşterileri daha yakından tanıması ve verilerin anlamlandırılması sağlanmıştır.

Değişkenlerin kendi içinde kümeler üzerine etkisi incelendiğinde aşağıda belirtilen sonuçlara ulaşılmıştır.

Birinci kümenin özelliği, %44 oranla konut poliçesi alan diğer demografik bilgileri Şekil 3.9.Küme 1 Analizinde gösterilen erkekler memnuniyetsizlik sebebini %30.21 oran ile "yüksek fiyat" olarak bildirmişler.Bu küme yine incelediğinde müşterilerin mal veya varlıklarında herhangi bir hasar meydana gelmemiş olup memnuniyetsizlik sebebi sigorta şirketinin hizmetinin kötü olmasından değil ürün fiyatının yüksek olmasından olduğu

gözlenmektedir.Bu kümenin ortalama memnuniyet puanı 1.44 olup bu kümedeki özelliklere sahip müşteri kitlesine ilgili poliçe ürünlerinden kampanya yapılarak ürün fiyatında indirim işlemi yapılabilir.Bu tip kampanyalar sayesinde hasar adedi olmayan yani işletmeyi zarara uğratmadan yüksek getiri sağlayan küme 1 özelliğine sahip müşteri grubunun memnuniyet puanında artış meydana gelebilir.

İkinci kümenin özelliği, %100 oranla kasko poliçesi alan diğer demografik bilgileri Şekil 3.10 Küme 2 Analizinde gösterilen erkekler memnuniyetsizlik sebebini %25.42 oran ile "eksperin tutum ve davranışları" olarak bildirmişler" olarak bildirmişler.Eksper, sigorta konusu risklerin gerçekleşmesi sonucunda ortaya çıkan kayıp ve hasarların miktarını, nedenlerini ve niteliklerini belirleyen ve mutabakatlı kıymet tespiti, ön ekspertiz ve hasar gözetimi gibi işleri meslek olarak yapan tarafsız ve bağımsız kişidir.Bu küme yine incelediğinde hasar adedi 1 olup ağırlıklı olarak 36-45 yaş aralığındaki üniversite mezunu evli erkekler oluşturmaktadır. Hasarı 1 adet olan yani sigorta şirketi ile ilk intibanın oluşma durumunda müşteri memnuniyetsizliğin ortaya çıktığı gözlenmektedir

Üçüncü kümenin özelliği, %52,78 oranla kasko poliçesi alan diğer demografik bilgileri Şekil 3.11 Küme 3 Analizinde gösterilen çoğunluktaki kadınlar memnuniyetsizlik sebebini %30.45 oran ile "yenileme döneminden önce aranmama" olarak bildirmişler.Hasarı 1 adet olan yani sigorta şirketi ile ilk intibanın oluşma durumunda müşteri memnuniyetsizliğin ortaya çıktığı gözlenmektedir.Ortalama memnuniyet puanı 2.55 olan bu kümedeki bireyler ağılrlıklı olarak İstanbul ilinde ikamet etmektedir.

DEĞERLENDİRME VE SONUÇ

Günümüz bilişim teknolojileri vasıtasıyla gerçekleşen yoğun veri üretimi birçok sektör için, üretilen veriyi doğru bir şekilde kullanma ve veriden anlamlı bilgiler elde etme ihtiyacını doğurmaktadır. Doğru yöntemlerle analiz edilen veriden elde edilen bilgi sayesinde karar vericiler karar aşamalarında stratejik hareket etme şansını da yakalayabilirler. Ancak üretilen veriyi klasik yöntemlerle analiz etmek zor ve zahmetli olabilmektedir. Bu aşamada veri madenciliğinin gelişmiş teknikleri ile problemleri çözmek, ilişkileri kurmak, aykırılıkları tespit etmek, gelecekte oluşabilecek durumları tahmin etmek mümkündür.

Yaygın kullanım alanına sahip olan veri madenciliği teknikleri veri tabanlarının içinde bulunan önemli bilgileri kullanıcılara sunmaktadır. Ayrıca veriler arasındaki görülemeyen ilişkileri de ortaya çıkaran bir tekniktir. Veri madenciliği veri tabanlarındaki yararlı bilgiyi ortaya çıkarırken birçok istatistiksel ve matematiksel yöntemleri kullanmaktadır. Veri madenciliği birçok veriyle işlevini yerine getirirken karar ağaçları ve çok değişkenli istatistiksel yöntemlerden de faydalanmaktadır. Bu yöntemlerin başında kümeleme analizi gelmektedir.

Veri madenciliği araçlarıyla ortaya çıkarılan bilgi birçok farklı alanda yaygın olarak kullanılmaktadır. Pazarlama, finans, mühendislik, lojistik, uzay bilimleri, tıp ve psikoloji gibi alanlar veri madenciliğinin kullanım alanlarından bazılarıdır. Bu alanların yanı sıra veri madenciliği tekniklerinin kullanımı yeni alanlarda da denenmektedir. Bu tez çalışmasında bahsedilen alanlardan finans alanlarından sigorta sektöründe veri madenciliği teknikleri uygulanmıştır.

Bu çalışmada Türkiye'de faaliyet gösteren bir sigorta şirketinin müşterilerine farklı tarih aralıklarında satışı gerçekleştirilmiş olan poliçe verileri K-Means algoritması ile analiz edilmiştir.2019.01.01-2019.01.31 tarihleri arasında işletme tarafından yapılan ankette 521 müşterinin memnuniyet verileri alınıp veri tabanına işlenmiştir. Bu müşteri grubunun memnuniyet verisi ile birlikte müşterinin poliçe ürünü, hasar adedi, demografik özellikleri ile ilişkilendirilerek memnuniyetsizliğin altında yatan nedenler araştırılmıştır. Çalışmada özellikle bilgi keşif sürecinin ve adımlarının önemi vurgulanarak, veriyi analiz etmeden önce yapılan ön işleme-dönüştürme işlemleri ile veri analiz edilmeye hazır hale getirilmiştir.

Uygulamada öncelikle her bir değişkenin genel istatistik değerleri belirlenmiştir. Söz konusu değişkenlere ait dağılım grafikleri, ortalamalar, adetler, tanım türleri tablo ve şekiller ile gösterilmiştir. Bu analiz ile elde edilen sonuçlar yardımıyla, şirketin benzer müşterilerinin

özelliklerini tespit etmesi ve onlara uygun yeni pazarlama stratejileri geliştirebilmesi hedeflenmektedir. Son yıllarda önemli ölçüde artan rekabet ortamında, müşterilerin kolayca alternatif hizmetlere yönelebilmelerinden dolayı şirketlerin stratejilerinde ayrılacak müşteriyi önceden tahmin etmek önem kazanmıştır. Ayrılacağı önceden tahmin edilen müşterilere promosyon, indirim, hediye veya avantajlar sağlanması müşterinin ayrılmasını engelleyebilecek ve böylece şirketler uzun vadede daha fazla kâr elde edebileceklerdir.

Elde edilen verilerin K-means algoritması ile kümelenmesi sonucunda birbirine benzer müşteri gruplarının oluşturulması sağlanmıştır.Böylece sigorta şirketinin müşterileri daha yakından tanıması ve verileri anlamlandırması sağlanmıştır.Değişkenlerin kendi içinde kümeler üzerine etkisi incelendiğinde aşağıda belirtilen sonuçlara ulaşılmıştır.

Birinci küme çoğunlukla konut poliçesi alan diğer demografik bilgileri Şekil 3.9.Küme 1 Analizinde gösterilenin müşterilerin tamamını erkekler oluşturmaktadır. Bu kümedeki müşteriler memnuniyetsizlik sebebini %30.21 oran ile "yüksek fiyat" olarak bildirmişler.Bu küme yine incelediğinde müşterilerin mal veya varlıklarında herhangi bir hasar meydana gelmemiş olup memnuniyetsizlik sebebi sigorta şirketinin hizmetinin kötü olmasından değil ürün fiyatının yüksek olmasından olduğu gözlenmektedir.

İkinci kümenin tamamı kasko poliçesi alan diğer demografik bilgileri Şekil 3.10. Küme 2 Analizinde gösterilen müşteriler memnuniyetsizlik sebebini %25.42 oran ile "eksperin tutum ve davranışları" olarak bildirmişler" olarak bildirmişler.Bu küme yine incelediğinde hasar adedi 1 olup ağırlıklı olarak 36-45 yaş aralığındaki üniversite mezunu evli erkekler oluşturmaktadır. Hasarı 1 adet olan yani sigorta şirketi ile ilk intibanın oluşma durumunda müşteri memnuniyetsizliğin ortaya çıktığı gözlenmektedir.Sigorta şirketi eksper alım veya görevlendirme işlemlerinde eksperlere yönelik performans uygulaması ile hem eksperlerin tutuma ve davranışlarında iyileştirme yapılabilir buna paralel olarak hem de müşteri memnuniyetinde artış sağlayabilir.

Üçüncü küme incelendiğinde çoğunlukla kasko poliçesi alan diğer demografik bilgileri Şekil 3.11 Küme 3 Analizinde gösterilen çoğunluktaki kadınlar memnuniyetsizlik sebebini %30.45 oran ile "yenileme döneminden önce aranmama" olarak bildirmişler.Poliçeler 1 yıl süre ile yenilenmekte olup vade bitimine yakın yapılacak sms,e-mail vb.iletişim araçlarıyla olan bilgilendirme işlemleriyle bu müşteri kitlesindeki memmnuniyetsizlik azaltılabilir.

Elde edilen üç küme içinde müşteri cinsiyeti önemli değişkendir.Şekil 3.8'de görülen değerler ilgili kümede kaç adet kayıt olduğunu göstermektedir.Şekil 3.9,şekil 3.10.'da görüldüğü gibi iki kümenin tamamını erkekler oluşturmakla birlikte,şekil 3.11'de ki üçüncü

kümenin çoğunluğu kadınlardan oluşmaktadır.Cinsiyet göz önüne alınarak kadınlara bir satış kampanyası yapılacağı zaman üçüncü kümedeki müşterilerin demografik özelliklerinden faydalanılabilir.

Grafik 3.7. Hasar değişken dağılımına baktığımızda hasar adeti 1 den fazla olan müşterilerin memnnuniyetsiz grubu kapsayan üç küme içerisinde yer almadığı görülmektedir.Bu sonuç hasar adeti olmayan veya hasar adeti 1 olup müşteri ile işletme arasındaki ilk intiba gerçekleştiği durumlarda memnuniyetsizliğin oluştuğu bilgisine varılmaktadır.

Üç küme yine incelendiğinde hasar adetinin önemli bir değişken olduğu görülmektedir.Hasar adet değerinin çoğunlukla yok olduğu ve yaş aralığının 36-45 arasında olduğu şekil 3.9'da müşterilerin genel olarak memnuniyetsizliğin nedeni mal veya hizmetin ya da ürün fiyatının yüksek olması iken şekil 3.11'de ise yenileme döneminden önce aranmama olarak belirtilmiştir.Bu iki kümenin ürün ve demografik özelliklerinden faydalanılarak bu müşteri kitlelerine yapılacak hasarsızlık indirimi ile memnuniyet puanlarında artış sağlanabilir.

Elde edilen sonuçlara göre sigortacılık sektöründe veri madenciliği analizleri veri yapısına göre incelenerek uygulanabilmektedir. Bunun yanı sıra şirketlerin veritabanlarında tutulan verilerden yola çıkılarak, bilgi keşfi yapılmış olup,veritabanında gizli kalmış verilerden bilgi açığa çıkartılmaktadır.Bunun sonucunda değerli ve değersiz müşterilerin farkına varılmaktadır. Bu yolla, müşterilerin kuruma karşı sadakati sağlanırken değerli müşterileri özellikleri de keşfedilerek hangi özelliklere sahip bireylere satış kampanyalarının düzenlenmesi gerektiği hakkında bilgi sahibi olunmuştur.

KAYNAKÇA

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

Akpınar ,Haldun ,Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği

Atbaş, Azize Celile Günay :"Kümeleme Analizinde küme sayısının belirlenmesi üzerine bir çalışma"Ankara Üniversitesi Fen Bilimleri Enstitüsü Yayınlanmamış Yüksek Lisans Tez, 2008,s.17.

Aydın, S. ve Özkul, A.E. (2015). Veri Madenciliği ve Anadolu Üniversitesi Açıköğretim Sisteminde Bir Uygulama. *Eğitim ve Öğretim Araştırmaları Dergisi,* 4(3), 36-44.

Aytaç, M. B. (2013), Doğrudan Pazarlama Aracı Olarak Tele Pazarlama için Veri Madenciliği Çözümleri: Banka Müşterileri Üzerine Bir Uygulama. Yüksek Lisans Tezi, Gazi Üniversitesi, Bilişim Enstitüsü, Ankara.

Bayram, E. 2001. Customer Segmentation and Churn Modeling In Wireless Communications. Yüksek Lisans Tezi, Boğaziçi Üniveristesi Fen Bilimleri Enstitüsü, İstanbul: 1-5.

Coşlu,Eda :Veri Madenciliği,Akademik Bilişim 2013-XV.Akademik Bilişim Konferansı,Bildirileri 23-25 Ocak 2013-Akdeniz Üniversitesi,Antalya

Çakmak,Zeki,Uzgören,Nevin ve Keçek,Gülnur:"Kümeleme Analizi Teknikleri İle İllerin Kültürel Yapılarına Göre Sınıflandırılması ve Değişimlerinin İncelenmesi",Dumlupınar

Dilly Ruth:Data Mining An Introduction,Student Notes 1995.

Dinçer,Esra :"Veri Madenciliğinde K-Means Algoritması ve Tıp Alanında Uygulanması",Kocaeli Üniversitesi Fen Bilimleri Enstitüsü

Dolgun,Özgür M.,Özdemir,Tülin ve Oğuz,Doruk: "Veri Madenciliğinde Yapısal Olmayan Verinin Analizi:Metin ve Web Madenciliği",istatistikçiler Dergisi:İstatistik Ve Aktüerya,2.2.,2009,s.50.

Dunham,Margaret H.:Data Mining – Introductory and Advanced Topics,Person Education,2003.

Everitt,Brian:Cluster Analysis,London,Heinemann Educational Book,1974

Fırat,Seniye Ümit : ”Kümeleme Analizi:İstihdamın Sektörel Yapısı Açısından Avrupa Ülkelerinin Karşılaştırılması”,İstanbul Üniversitesi Soyal Bilimler Dergisi,3.2.,1995,s.46.

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

Goebel,Michael ve Gruenwald,Le :”A Survey of Data Mining and Knowledge Discovery Software Tools”,ACM SIGKDD Explorations Newsletter 1.1.,1999,p.21.

Grabmeier,Johannes ve Rudolph,Andreas:”Techniques of Cluster Algorithms in Data Mining”,Data Mining and Knowledge Discovery 6.4.,2002,p.350.

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

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

Han, J., Pei, J., Kamber, M. 2011. Data mining: concepts and techniques: Elsevier.

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

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

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

Köktürk, Mehtap S.- Ağırbal, Melahat-Aksoy, Bernu –Ardıç, Osman.K. –Cantutan,Deniz-Cenanoğlu,Merve,C.-Demir, Ender-Ergörün,Pınar-Yiğiter,Burçin-Yüksel,Pelin, Bir Eğitim Metodu, Dört Pazarlama Kavramı –Müşteri, Müşteri Memnuniyeti, Memnuniyet/Memnuniyetsizlik Mülakat, Kısım 2, Bölüm 4, (Basılmamış Eser)

Köktürk, Mehtap S., Memnuniyet-Memnuniyetsizlik Hikayeleri Sürdürülebilir Bir Çalışma (2007/2010), Sürat Daktilo Kollektif Şti., İstanbul, 2010

Larose,Daniel T. :Data Mining Methods and Models,A John Wıley & Sons,Inc Publication,Canada 2005.

Nisbet, R., Miner, G. & Yale, K. (2018). Handbook of Statistical Analysis and Data Mining Applications. Elsevier,

Özel, C. ve Topsakal, A. (2014). Veri Madenciliği Kullanarak Beton Basınç Dayanımının Belirlenmesi. *Cumhuriyet Üniversitesi Fen Fakültesi Fen Bilimleri Dergisi,* 35(1), 43-57.

Özbay, Ö. ve Ersoy, H. (2017) Öğrenme Yönetim Sistemi Üzerindeki Öğrenci Hareketliliğinin Veri Madenciliği Yöntemleriyle Analizi. *Gazi Üniversitesi Gazi Eğitim Fakültesi Dergisi,* 37(2), 523-558.

Özbay, Ö. (2015) Veri Madenciliği Kavramı ve Eğitimde Veri Madenciliği Uygulamaları. Inesjournal Uluslar Arası Eğitim Bilimleri Dergisi, 2(5), 262-272.

Özdemir, A., Aslay, F. Y. ve Çam, H. (2009). Veri Tabanında Bilgi Keşfi Süreci: Gümüşhane Devlet Hastanesi Uygulaması. *SÜ İİBF Sosyal Ekonomik Araştırmalar Dergisi,* 1(20), 347-365.

Özkan,Yalçın, Veri Madenciliği Yöntemleri, İstanbul, Papatya Yayıncılık,

Pehlivanoğlu, M. K. ve Duru, N. (2015). Veri Madenciliği Teknikleri Kullanılarak Ortaokul Öğrencilerinin Sosyal Ağ Kullanım Analizi: Kocaeli Ġli Örneği. *Düzce Üniversitesi Bilim ve Teknoloji Dergisi,* 3, 508-517.

Romero, C. & Ventura, S. (2013). Data Mining in Education. Overview, 3, 12-27.

Saharkhiz Aresh,(3 Jan 2009), K-Means Clustering Used in Intention Based Scoring Projects

Sarıman, G. (2011). Veri Madenciliğinde Kümeleme Teknikleri Üzerine Bir Çalışma: K-Means ve K-Medoids Kümeleme Algoritmalarının Karşılaştırılması. *Süleyman Demirel Üniversitesi, Fen Bilimleri Enstitüsü Dergisi,* 15(3), 192-202.

Savaş S., Topaloğlu, N. ve Yılmaz, M. (2012). Veri Madenciliği ve Türkiye‟deki Uygulama Örnekleri. *İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi,* 11(21), 1-23.

Seker, S.E. (2015a). Sosyal Ağlarda Veri Madenciliği (Data Mining on Social Networks). *YBS Ansiklopedi*, 2(2), 30-39.

Selvi, H.Z. ve Çağlar, B. (2017). Çok Değişkenli Haritalama için Kümeleme Yöntemlerinin Kullanılması. *Ömer Halisdemir Üniversitesi Mühendislik Bilimleri Dergisi,* 6(2), 415-429.

Sevindik, T., Kayışlı, K. ve Ünlükahraman, O. (2012). Web Tabanlı Eğitimde Veri Madenciliği. *Turkish Journal of Computer and Mathematics Education,* 3(3), 183-193.

Seyrek, G. H. ve Ata, H. A. (2010). Veri Zarflama Analizi ve Veri Madenciliği ile Mevduat Bankalarında Etkinlik

Silahtaroğlu, G., Veri Madenciliği Kavram ve Algoritmaları. 2013. S-163, 208, 215.

Silahtaroğlu Gökhan, Kavram ve Algoritmalarıyla Temel Veri Madenciliği, , İstanbul, Papatya Yayıncılık, 2008

Sin, K. & Muthu, L. (2015). Application of Big Data In Education Data Mining and Learning Analytics--A Literature Review. *Journal On Soft Computing*, 5(4). 1035-1049.

Şentürk, Aysan, Veri Madenciliği Kavram ve Teknikler, Bursa, Ekin Basım Yayın Dağıtım, 2006

Tekerek, A. (2011). Veri Madenciliği Süreçleri ve Açık Kaynak Kodlu Veri Madenciliği Araçları. *Akademik Bilişim,* 11, 2-4.

Tozak, E. Y., Ersöz, T. ve Ersöz, F. (2017). Demir–Çelik Sektöründe Meydana Gelen iş Kazalarının Veri Madenciliği Kullanılarak Analizi. *3. Ulusal Sigorta ve Aktüerya Kongresi,*Vol.3, 137-144.

Vural, Y. ve Sağıroğlu, ş. (2010). Veritabanı Yönetim Sistemleri Güvenliği: Tehditler ve Korunma Yöntemleri. *Politeknik Dergisi*, 13(2), 71-81.

Yıldız, M. ve şeker, ş. E. (2016). Veri Madenciliği Araçları (Data Mining Tools). *YBS Ansiklopedi,* 3, 10-19.

EKLER

EK-1 MÜŞTERİ TABLOSU	
KOLON	**AÇIKLAMA**
Müşteri_Id	Müşterinin Numarası
Cinsiyet	Müşterinin Cinsiyeti
Meslek	Müşterinin Mesleği
Yaş	Müşterinin Yaşı
İl	Müşterinin İkamet Ettiği il
Eğitim_Durumu	Müşterinin Eğitim Durumu
Medeni_Hal	Müşterinin Medeni Hal
Ad_Soyad	Müşterinin Adı Soyadı
Nüfus_Bilgi	Müşterinin Nüfus Bilgileri
Doğum_Tarih	Müşterinin Doğum Tarihi
Uyruk	Müşterinin Uyruk Bilgisi
Mernis_No	Müşterinin Kimlik Bilgisi

EK-2 MEMNUNIYET TABLOSU	
KOLON	**AÇIKLAMA**
Müşteri_Id	Müşterinin Numarası
Anket_Tarihi	Müşteriye Yapılan Anket Tarihi
Puan	Müşterinin Ankete Vermiş Olduğu Puan
Açıklama	Müşterinin Memnuniyet Açıklaması

EK-3 HASAR TABLOSU	
KOLON	**AÇIKLAMA**
Müşteri_Id	Müşterinin Numarası
Hasar_Adedi	Müşterinin Almış Olduğu Hasarların Toplam Adedi

Ürün	Müşterinin Ürün Bilgisi
Hasar_Tarihi	Hasarın Gerçekleştiği Tarih
İhbar_Tarihi	Hasarın İşletmeye Beyan edildiği Tarih