



T.C.

SÜLEYMAN DEMİREL ÜNİVERSİTESİ

SAĞLIK BİLİMLERİ ENSTİTÜSÜ

BİYOİSTATİSTİK VE TIBBİ BİLİŞİM ANABİLİM DALI

**DATA ANALİTİĞİ İLE DİŞ HEKİMLİĞİ İMMÜNOLOJİ
BÜYÜK VERİSİNDE KÜME YAPILARININ BELİRLENMESİ**

Seren AYHAN

Yüksek Lisans Tezi

DANIŞMAN

Dr. Öğr. Üyesi Adnan KARAIBRAHİMOĞLU

ISPARTA - 2022

KABUL ve ONAY

Sağlık Bilimleri Enstitü Müdürlüğü'ne;

Süleyman Demirel Üniversitesi Sağlık Bilimleri Enstitüsü **Biyoistatistik ve Tıbbi Bilişim Anabilim Dalı Yüksek Lisans Programı** çerçevesinde yürütölmüş olan bu çalışma, aşağıdaki jüri tarafından **Yüksek Lisans Tezi** olarak kabul edilmiştir.

Tez Savunma Tarihi: 04/01/2022

Tez Danışmanı: Dr. Öğr. Üyesi Adnan KARAİBRAHİMOĞLU

Tez Savunma Jürisi

Üye : Prof. Dr. Hikmet ORHAN

Üye : Doç. Dr. Fatih ÜÇKARDEŞ

Üye : Dr. Öğr. Üyesi Adnan KARAİBRAHİMOĞLU

ONAY: Bu uzmanlık tezi, fakölte yönetim kurulunca belirlenen yukarıdaki jüri üyeleri tarafından uygun görölmüş ve fakölte yönetim kurulu kararıyla kabul edilmiştir.

Prof. Dr. Ebru ÇUBUK DEMİRALAY

Enstitü Müdürü

BEYAN

“Data Analitiđi ile Diř Hekimliđi İmm noloji B y k Verisinde K me Yapılarının Belirlenmesi” adlı Y ksek Lisans tezi S leyman Demirel  niversitesi Sađlık Bilimleri Enstit s  Lisans st  Tez  nerisi ve Tez Yazma Y nergesi’ne uygun olarak hazırlanmıřtır.

Bu tez  alıřmasının kendi  alıřmam olduđunu, tezin planlanmasından yazımına kadar b t n safhalarda etik dıřı davranıřımın olmadıđını, bu tezdeki b t n bilgileri akademik ve etik kurallar i inde elde ettiđimi, bu tez  alıřmasıyla elde edilmeyen b t n bilgi ve yorumlara kaynak g sterdiđimi ve bu kaynakları da kaynaklar listesine aldıđımı, yine de bu tezin  alıřılması ve yazımı sırasında patent ve telif haklarını ihlal edici bir davranıřımın olmadıđını beyan ederim.

Hazırlayan

Seren AYHAN

Danıřman

Dr.  đr.  yesi Adnan KARAİBRAHİMOđLU

TEŞEKKÜR

Tez çalışmam süresi boyunca çalışmalarımı yönlendiren, yardımlarını, bilgi birikimini ve önerilerini esirgemeyen saygıdeğer hocam Dr. Öğr. Üyesi Adnan Karaibrahimoğlu'na, Biyoistatistik ve Tıbbi Bilişim ABD Başkanı Hikmet Orhan'a ve öğretim üyesi Dr. Öğretim Üyesi Osman Gürdal'a,

Hayatımın her aşamasında maddi manevi destek ve dualarını esirgemeyen annem, babam ve kardeşime en içten duygularıyla teşekkür ediyorum.

Bu dönem boyunca özellikle manevi olmakla beraber kaynak araştırması ve çeviri konusunda destek ve yardımlarını esirgemeyen sevgili arkadaşlarım Ali Önal, Büşra Bozkurt Çil, Gülcemal Yaşar ve Hatice Yıldız'a teşekkürlerimi sunarım.

Seren AYHAN

Isparta - 2021

ÖNSÖZ

Son yıllarda Veri Madenciliği ve çalışmalarının sağlık alanında kullanımı yaygınlaşmaktadır. Sağlık alanında otomasyon sistemlerinin yoğunlaşması ile hasta ve hastanelerden elde edilen bilgiler her geçen gün artmaktadır. Genel olarak kanser araştırmaları, görüntüleme teknikleri ve en çok kullanımın görüldüğü alan olan sağlık tesisi yönetimlerinde maliyet ile ilgili konularda veri madenciliği ve kümeleme yöntemlerinin oldukça fazla kullanıldığını görmekteyiz. Hastalardan elde edilen kan ölçümleri ile hormon ve immun sistem üzerinde tedavileri öngörmek için sağlık alanında çalışan hekim, hemşire ve diğer tüm sağlık profesyonelleri için büyük veri üretme açısından oldukça önemli olan sağlık sektöründe kümeleme analizi kullanılması önerilmektedir. Bu çalışmanın konu ile ilgili kullanıcılar için faydalı olmasını temenni ederim.

Seren AYHAN
ISPARTA, 2021

ÖZET

Data Analitiği ile Diş Hekimliği İmmünoloji Büyük Verisinde Küme Yapılarının Belirlenmesi

Kümeleme yöntemleri, benzer özellikler taşıyan değişkenlerin bir gruba toplanabilmesini sağlayan önemli analizlerden biridir. Sağlık, ekonomi, işletme veya elektronik gibi oldukça geniş bir kullanım alanına sahip olan bu yöntem sağlık alanında özellikle tanı gruplarının oluşturulmasında daha yaygın olarak uygulama alanı bulmuştur. Bu tez çalışmasının amacı, kümeleme analizi hakkında teorik arka planı anlatmak, farklı kümeleme algoritmalarını hakkında bilgi vermek ve bu algoritmaların örnek büyük veri kümesinde uygulamalarını açıklamaktır. Çalışmamızda, Diş Hekimliği Fakültesi Ağız, Diş ve Çene Cerrahisine başvuran hastalardan elde edilmiş hemogram, biyokimya ve immünolojik ölçümler ile elde edilen veri kümesi kullanılmıştır. Kümeleme algoritmaları içerisinde Fuzzy C-Means, Two-step Clustering ve K-means algoritmaları seçilerek, analiz sonuçları karşılaştırılmıştır. En iyi sonucu veren k-means algoritması seçilerek Python’da analiz edilmiştir. Toplam 5148 hastaya ait kayıtlara göre yaş ortalaması $21,19 \pm 19,23$ yıl olup, yaş aralığı oldukça geniş ve 4-89 yıl olarak hesaplanmıştır. Algoritmalarda sabit 2 küme (hasta-sağlam) oluşturmaları sağlandı. K-means yöntemiyle oluşan immünolojik kümelerde Silhouette değeri 0,830 olarak elde edilmiştir. Açıklayıcılık katsayısı $R^2=0,801$ bulunmuştur. Aynı değişkenler, Fuzzy C-means kümeleme algoritması ile analiz edildiğinde Silhouette skorunun 0,700 olduğu görülmüştür. Açıklayıcılık katsayısı ise $R^2=0,281$ olarak hesaplanmıştır. TwoStep Cluster ise “Good” olarak Silhouette skorları vermiştir. K-means algoritması ile immünoloji, hemogram ve biyokimya değişkenleri ile sağlıklı ve hasta grupları daha yüksek başarı ile elde edilmiştir. Python’da ise Silhouette skoru ise 0,827 olarak elde edildi. Kullanılan ölçümlerin referans aralıklarına göre kümelerdeki bireylerin sağlık durumları ile ilgili tedavi önermeleri yapılmıştır. Herhangi bir sınıfa ait olduğu bilinmeyen bireylerin yapılan ölçümlere göre gruplandırılabilmesi için kümeleme algoritmaları oldukça yararlı bilgiler vermektedir ve kullanılan veri kümesine göre K-means yönteminin daha uygun sonuçlar verdiği söylenebilir.

Anahtar Kelimeler: Data Analitiği, Büyük Veri, Kümeleme Analizi, Python, İmmünoloji

ABSTRACT

Determination of Cluster Structures in Denistry Immunology Big Data with Data Analytics

Clustering methods are one of the important analyzes that enable variables with similar characteristics to be collected into a group. This method, which has a wide area of use such as health, economy, business or electronics, has found more widespread application in the field of health, especially in the creation of diagnostic groups. The aim of this thesis is to explain the theoretical background about clustering analysis, to give information about different clustering algorithms and to explain the applications of these algorithms in a sample big data set. In our study, the dataset obtained by hemogram (blood test), biochemistry and immunological measurements obtained from patients who applied to the Faculty of Dentistry Oral, Dental and Maxillofacial Surgery was used. Among the clustering algorithms, Fuzzy C-Means, Two-step Clustering and K-means algorithms were selected, and the analysis results were compared. The k-means which was analyzed in Python was selected as the best result giver algorithm. According to the records of a total of 5148 patients, the mean age was 21.19 ± 19.23 years, and the age range was calculated as 4-89 years. In the algorithms, they were provided to form 2 fixed clusters (patient-healthy). Silhouette score was obtained as 0.830 in the immunology clusters formed by the K-means method. The explanatory coefficient was $R^2=0.801$. When the same variables were analyzed with the Fuzzy C-means clustering algorithm, the Silhouette score was found to be 0.700. The explanatory coefficient was calculated as $R^2=0,281$. TwoStep Cluster gave Silhouette scores in the “Good” dimension. With the K-means algorithm, immunology, hemogram and biochemistry variables and healthy and patient groups were obtained with higher success. In Python, Silhouette scores was found to be 0,827. The According to the reference intervals of the measurements used, treatment recommendations were made regarding the health status of the individuals in the clusters. Clustering algorithms provide very useful information in order to group individuals who are not known to belong to any class according to the measurements made, and it can be said that the K-means method gives more accurate results according to the data set used.

Key Words: Data Analytics, Big Data, Cluster Analysis, Python, Immunology

İÇİNDEKİLER

	<u>Sayfa</u>
BEYAN.....	i
TEŞEKKÜR	ii
ÖZET.....	iv
ABSTRACT	v
İÇİNDEKİLER	vi
TABLOLAR DİZİNİ	ix
ŞEKİLLER DİZİNİ	x
SİMGELER ve KISALTMALAR DİZİNİ	xi
1. GİRİŞ	1
2. GENEL BİLGİLER.....	2
2.1. Veri Bilimi	2
2.2. Data Analitiği	2
2.3. Veri Tabanları	4
2.4. Veri Ambarları	5
2.5. Veri Madenciliği Süreci	7
2.5.1. Verilerin Hazırlanması	9
2.5.2. Veri Madenciliği Yönteminin Uygulanması	11
2.5.3. Sonuçların Değerlendirilmesi ve Sunulması	13
2.5.4. Veri Madenciliği Uygulama Alanları ve Karşılaşılan Sorunlar	14
2.6. Data Analitiği Uygulama Programları	15
2.7. Büyük Veri	16
2.7.1. Büyük Data Analitiği Yaklaşımı	19
2.7.2. Data Analitiği Döngüsü	19
2.7.2.1. Veriyi Hazırlama	21
2.7.2.2. Modelleme	23
2.7.2.3. Görselleştirme	24
2.7.2.4. Değerlendirme	24
2.8. Data Analitiğinde Yöntemler	25
2.8.1. Birliktelik Kuralı	25
2.8.2. Kümeleme	27
2.8.3. Sınıflandırma	27

2.9. Data Analitiğinde İstatistik	29
2.9.1. İstatistiksel Yöntemler	30
2.9.2. Bayesci İstatistik	30
2.9.3. Temel Bileşenler Analizi	32
2.9.4. Regresyon.....	34
2.10. Data Analitiğinde Çapraz Geçerlilik	35
2.10.1. Model Başarım Ölçütleri.....	36
2.10.1.1. Karmaşıklık Matrisi (Risk Matrix).....	36
2.10.1.2. Doğruluk (Accuracy)	37
2.10.1.3. Kesinlik (Precision).....	37
2.10.1.4. Duyarlılık (Recall)	38
2.10.1.5. F-Ölçütü	38
2.10.2. Klasik Kümeleme Analizinde Küme Geçerlilik İndeksleri	38
2.10.2.1. Silhouette İndeksi.....	38
2.10.2.2. Akaike (AIC) ve Bayesian (BIC) Bilgi Kriterleri	39
2.11. Sağlık Alanında Veri Bilimi.....	40
2.11.1. Kullanım Alanları	40
2.12. Biyokimya ve İmmünoloji	41
2.12.1. Diş ve Çene Cerrahisi	43
3. GEREÇ VE YÖNTEM	44
3.1. Araştırmanın Evreni	44
3.2. Araştırmanın Örneklemi.....	44
3.3. Araştırmanın Türü	44
3.4. Araştırmanın Etiği	44
3.5. Araştırmanın Yöntemi.....	45
3.6. Verilerin Hazırlanması	46
3.7. Kümeleme Analizi	49
3.7.1. Kümelemenin Amacı	49
3.7.2. Kümelemenin Analizinin Temel Mantığı	50
3.7.3. Benzerlik ve Uzaklık.....	51
3.8. Kümeleme Analizinin Temel Yöntemleri	52
3.8.1. Hiyerarşik Kümeleme	52
3.8.1.1. Tek Bağlantılı Kümele Tekniği (Single Linkage - SLINK).....	53
3.8.1.2. Tam Bağlantılı Kümele Tekniği (Complete Linkage - CLINK).....	53

3.8.1.3. Ortalama Bağlantılı Küme Tekniği	53
3.8.1.4. Ward Tekniği	54
3.8.1.5. Medyan Tekniği	54
3.8.1.6. Centroid Tekniği	54
3.8.1.7. DBSCAN Algoritması	54
3.8.2. Bölünmeli Hiyerarşik Yöntemler	55
3.8.2.1. k-Medoid Tekniği	55
3.8.2.2. CLARA ve CLARANS Algoritmaları	56
3.8.3. Hiyerarşik Olmayan Kümeleme.....	57
3.8.3.1. k-Ortalamlar Kümeleme Tekniği (k-means)	57
3.8.3.2. Fuzzy C-Means Kümeleme Tekniği	58
3.8.3.3. İki Aşamalı Kümeleme (TwoStep Cluster) Tekniği	59
3.8.3.4. Kohonen Ağlar	60
3.8.3.2. En Çok Olabilirlik Tekniği.....	61
4. BULGULAR	62
4.1. Tanımlayıcı Analizler.....	62
4.2. Analizler	62
4.3. Sonuçların Karşılaştırılması	84
5. TARTIŞMA	86
SONUÇLAR	92
KAYNAKÇA	94
ÖZGEÇMİŞ.....	103

TABLÖLAR DİZİNİ

Sayfa

Tablo 2.1. Karmaşıklık Matrisi	37
Tablo 2.2. Karmaşıklık Matrisi Harflendirme	37
Tablo 3.1. Veri Kümesini Oluşturan Değişkenler	47
Tablo 3.2. Veri Kümesini Oluşturan Değişkenlerin Referans Değerleri.....	48
Tablo 4.1. Mikrobiyolojik Viral Testlere Ait Tanımlayıcı İstatistikler.....	62
Tablo 4.2. JASP k-means Biyomkiya Değişkenlerinin Atandığı Küme Merkez Değerleri.....	64
Tablo 4.3. JASP Fuzzy C-means Biyomkiya Değişkenlerinin Atandığı Küme Merkez Değerleri.....	65
Tablo 4.4. JASP k-means Hemogram Değişkenlerinin Atandığı Küme Merkez Değerleri.....	66
Tablo 4.5. JASP Fuzzy C-means Hemogram Değişkenlerinin Atandığı Küme Merkez Değerleri.....	68
Tablo 4.6. JASP k-means İmmünolojik Değişkenlerinin Atandığı Kümenin Merkez Değerleri.....	69
Tablo 4.7. JASP Fuzzy C-means İmmünolojik Değişkenlerinin Atandığı Kümenin Merkez Değerleri.....	70
Tablo 4.8. JASP k-means ve Fuzzy C-means Başarım Ölçüt Değerleri	70
Tablo 4.9. SPSS TwoStep Cluster Biyokmya, Hemogram ve İmmünolojik Ölçümlerin Kümeleme Sonuçları	74
Tablo 4.10. SPSS k-means Biyomkiya Değişkenlerinin Atandığı Küme Merkez Değerleri.....	75
Tablo 4.11. SPSS k-means Biyokimya Ölçümleri Anova Sonuçları	76
Tablo 4.12. SPSS k-means Hemogram Değişkenlerinin Atandığı Küme Merkez Değerleri.....	77
Tablo 4.13. SPSS k-means Hemogram Ölçümleri Anova Sonuçları	78
Tablo 4.14. SPSS k-means İmmünolojik Değişkenlerinin Atandığı Kümenin Merkez Değerleri	79
Tablo 4.15. SPSS k-means İmmünoloji Ölçümleri Anova Sonuçları	80

ŞEKİLLER DİZİNİ

Sayfa

Şekil 1.1. Bilgi Piramidi	1
Şekil 2.2. Bilgi Keşfi Sürecinde Veri Madenciliği	7
Şekil 2.3. CRISP-DM Süreci	12
Şekil 2.4. Büyük Veri Yaşam Döngüsü	20
Şekil 3.1. Küme İçi ve Kümeler Arası Uzaklıklar	51
Şekil 4.1. JASP Biyokimya Viral Testlere ait K-means Küme Yapısı.....	64
Şekil 4.2. JASP Biyokimya Viral Testlere Ait Fuzzy C-means Küme Yapısı	65
Şekil 4.3. JASP Hemogram Viral Testlere Ait k-means Küme Yapısı	66
Şekil 4.4. JASP Hemogram Viral Testlere Ait Fuzzy C-means Küme Yapısı	67
Şekil 4.5. JASP Mikrobiyolojik Viral Testlere Ait K-means Küme Yapısı	69
Şekil 4.6. JASP Mikrobiyolojik Viral Testlere Ait Fuzzy C-means Küme Yapısı ...	70
Şekil 4.7. SPSS TwoStep Cluster Biyokimya Ölçümleri Silhouette Değeri	71
Şekil 4.8. SPSS TwoStep Cluster Biyokimya Ölçümleri Kümeye Dâhil Olan Birimlerin Oranları	72
Şekil 4.9. Spss TwoStep Cluster Biyokimya Ölçümleri İlgili Kümelere Ayrılırken Önemli Bulunan Değişkenler	72
Şekil 4.10. SPSS TwoStep Cluster Biyomkiya Değişkenlerinin Atandığı Küme Merkez Değerleri.....	73
Şekil 4.11. SPSS k-means Biyokimya Değişkenlerinin Kümeler Arası Dağılımı.....	76
Şekil 4.12. SPSS k-means Hemogram Değişkenlerinin Kümeler Arası Dağılımı	78
Şekil 4.13. SPSS k-means İmmünolojik Değişkenlerinin Kümeler Arası Dağılımı .	79
Şekil 4.14. Python Biyokimya Viral Testlere Ait Küme Yapısı.....	83
Şekil 4.15. Python Hemogram Viral Testlere Ait Küme Yapısı	83
Şekil 4.16. Python İmmünoloji Viral Testlere Ait Küme Yapısı.....	84

SİMGELER ve KISALTMALAR DİZİNİ

AIC	: Akaike Information Criteria
AIS	: Apriori Algorithm (Agrawal, Imielinski ve Swami)
ANCOVA	: Analysis of Covariance
ANOVA	: Analysis of Variance
BIC	: Bayesian Information Criteria
BMI	: Body Mass Index
CLARA	: Clustering Large Applications
CLARANS	: Clustering Algorithm Based on Randomized Search
CLINK	: Complete Linkage
CRISDM	: Cross Industry Standard Process for Data Mining
DAG	: Directed Acyclic Graph
DBMS	: Database Management
DBSCAN	: Density Based Spatial Clustering of Applications with Noise
DIKW	: Data Information Knowledge Wisdom
GB	: Gigabyte
HDFS	: Hadoop Distributed File System
IBM	: International Business Machines
IDC	: International Data Corporation
JSON	: JavaScript Object Notation
KNIME	: Konstanz Information Miner
MPP	: Massively Parallel Processing
NIST	: National Institute of Standards and Technology
OLAP	: On-Line Analytical Processing
PAM	: Partitioning Around Medoids
PB	: Petabyte
RARM	: Rapid Association Rule Mining
RDBMS	: Relational Database Management
SAS	: Statistical Analysis Software
SLINK	: Single Linkage
SPSS	: Statistical Package for the Social Sciences

SQL	: Structured Query Language
TB	: Terabyte
VM	: Veri Madenciliği
VYS	: Veri Tabanı Sistemleri
WEKA	: Waikato Environment for Knowledge Analysis
XML	: Extensible Markup Language

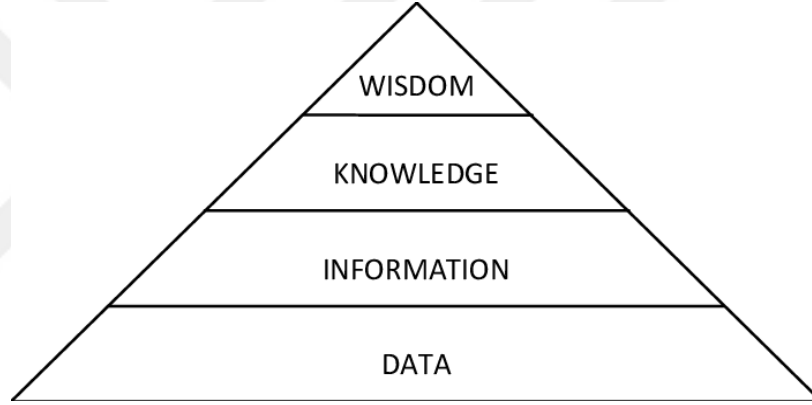


1. GİRİŞ

Veri Nedir? Veri Türleri Nelerdir?

Bilgi (*information*) kavramı veri (*data*) kavramı yerine oldukça fazla kullanılmakta olan bir kavramdır. Fakat veri, nitel (*qualitative*) ve nicel (*quantitative*) kümelerden meydana gelen bilgi parçalarıdır. Bilgi ise bu verilerin işlenmesi ile ortaya çıkan nitel veya nicel ifadelerdir (1).

Verileri işleyerek bilgi, bilgileri toparlayarak ilim (*knowledge*), ilimden de irfan/hikmet (*wisdom*) kavramlarına ulaşılmaktadır. Bu üç kavram arasındaki ilişki Data-Information-Knowledge-Wisdom (*DIKW*) olarak kavramsal model olarak 1989'da literatürde yer almıştır (2).



Şekil 1.1. Bilgi Piramidi

Bilgi piramidindeki basamaklardan ilim/bilim ne kadar geniş olursa olsun, sadece geçmiş ile ilgili bilgi sunabildiği için geleceğe dair tahminler yapmak istiyorsak irfan (*wisdom*) sahibi olmak gerekmektedir. Bir bilgi sisteminde bu ifadenin karşılığı model oluşturmaktır. Veri madenciliğinin temelini oluşturan öğrenme (*learning*), model oluşturma süreci ve karar destek sistemlerinin (*DSS*) temelini oluşturmaktadır (1).

Veri türleri, veri tabanları, analiz ve bilgisayar programlama vb. birçok alanda karşılaşılan bir konudur. Veriler genel olarak sayısal, nesne ve karakter olarak üç temel gruba ayrılırlar ve boyutlarıyla ilgili kısıtlar belirlenir. Yapacağımız işlemler açısından veri türleri, bu işlemleri kısıtlayabilmektedir. Bu nedenle yapılacak işleme göre veriler düzenlenmeli ve gerekli dönüşümler sağlanmalıdır (1).

2. GENEL BİLGİLER

2.1. Veri Bilimi

İlk defa 1960'lı yıllarda "bilgisayar bilimi"nin yerine kullanılan “veri bilimi”, uygulamalı istatistiki bilimler kadar eski bir tarihe sahiptir. Fakat bu bilimde yeni olan şey veriyi düzenlemede kullanılmakta olan araçlar, veriyi toplama ve depolama kapasiteleri ve bilgi işlem gücüdür (3). 2001 yılında William Cleveland tarafından ele alınan “Veri Bilimi: İstatistik Alanının Teknik Alanlarını Genişletmek İçin Bir Eylem Planı” isimli makalede “Veri bilimi” terimi ilk defa kullanılmıştır. Cleveland makalesinde veri biliminden istatistiğin temel alanlarını genişletip geliştirmek için bir plandan bahsetmektedir. Bu plan ise temel bir değişim göstererek yeni bir alan oluşturacaktır. Böylece 2001 yılında yeni oluşan ve bağımsız bir disiplin olarak tanıtılan alan veri bilimidir (4).

2.2. Data Analitiği

Belirlenmiş bir hedefe ulaşmak için alt hedefleri belirlenen projeler hedeflenmiş amaca ulaşma doğrultusunda başlar. Kuruluşlar proje amaç ve hedeflerine ulaşmak için var olan bilgi sistemlerinin çalıştıramayacağı kadar genişlikte ve karmaşık verilerle karşılaşmaktadırlar. Uygun veriler elde edildikten sonra bu verileri işleyip, ilişkilendirme ve karar verme sürecinde bir data analitiğine ihtiyaç olacaktır. Teknolojinin ilerlemesi ve verinin sürekli artmasıyla yapısal verilere, yarı yapısal ve yapısal olmayan veriler de eklenmiştir. Böylece data analitiği karar verme aşamasında anlamlı bilgiyi elde edebilmek için büyük veriler üzerinde kullanılan bir veri analiz tekniğidir (5). Veri analitiği özetle karışık ve büyük verileri, istatistik ve diğer bazı modern hesaplamaları kullanarak veriyi işlenecek hale getirme işlemidir (6).

Büyük data analitiği, farklı birçok türde anlam içeren oldukça geniş çaplı ve farklı kayıtlar için oluşturulmuş analitik yöntemlerin kullanımınıdır. Belirli araçlar içeren büyük veri analitiği bilinen veri tabanı teknikleri kullanılarak işlenmesi güç olan, sürekli değişen ve gelişim gösteren, çok sayıdaki yapısal, yarı yapısal ve

yapısal olmayan verinin bir bütün olarak analizi ile verilerden değer içeren ve anlamlı bilgilere ulaşılmasını amaçlamaktadır (7).

Büyük veri analitiğinde karşılaşılabilecek sorunlara çözüm bulma, verinin tamamının depolanıp depolanmayacağı ve analiz edilip edilmeyeceği, en anlamlı sonuca ulaşmak için büyük verinin nasıl kullanılacağı cevap bulunması gereken başlıca sorunlardır. Bu sorunlar analiz aşamasına geçildiğinde beraberinde büyük zorlukları getirebilmektedir. Büyük veri, yapısal, yapısal olmayan ve yarı yapısal veri çeşitlerinden meydana geldiği için daha gelişmiş ileri tekniklere ihtiyaç vardır. Ayrıca veriye uygulanacak analiz türü, elde edilecek sonuçlara da bağlı olmaktadır. Bu nedenle analiz aşamasına geçildiğinde ya tüm büyük data unsurları birleştirilir, ya da hangi büyük data elemanının elde edilecek sonuçla bağlantılı olduğu belirlenir (8).

Yapısal verilere oranla yapısal olmayan verinin büyüklüğünün ve büyüme hızının oldukça fazla olması, yapısal olmayan verinin veri tabanlarında fazla depolama alanı işgal etmesi önemli zorluklardan biridir. Ayrıca bir diğer zorluk ise yapısal olmayan verinin yapısal veri benzeri ilişkisel veri tabanlarında belirli bir nizamda depolanamaması ve beraber çalışabilirliğin olmamasıdır. Bu zorluklar sebebiyle veri depolama sistemleri ve bu sistemler üzerinde çalışan sorgu dillerinde yenilikler geliştirilmiştir. Bu nedenle, büyük DATA analitiğinde ilişkisel veri tabanları yanı sıra dağıtık dosya sistemleri, SQL (Structured Query Language) gibi ilişkisel veri tabanı sorgu dilleri yerine ise büyük dataya has açık kaynak kodlu yazılımlar geliştirilmiştir. Büyük data analitiğinde bulunan yazılımlar arasında, öncelikle Eşle-İndirge teknolojisini ve HDFS'yi (Hadoop Distributed File System - Hadoop Dağıtık Dosya Sistemi) kullanan Hadoop yazılımıdır. Hadoop şu anda, Amazon, LinkedIn, Twitter, Facebook, Yahoo, IBM ve Adobe gibi önde gelen bilişim şirketinde büyük veri analitiğinde kullanılmaktadır. Ayrıca büyük veri çözümleri sadece Hadoop ile sınırlandırılmamalıdır. Açık kaynak kodlu yazılımlar içerisinde en bilindik Hadoop olsa da büyük data analitiğinde MPP (Massively Parallel Processing), NoSQL, HBase, HCatalog, PIG, Mahout, Cassandra ve In-memory (7) gibi diğer teknolojilerden de faydalanılmaktadır (9).

2.3. Veri Tabanları

Veriler çevremizde gelişen tüm olaylar sonucunda oluşmaktadır. Ancak kullanım hızına kıyasla veriler daha hızlı bir şekilde artmaktadır. Bu hızla artan ve kullanılacak verilerin uygun bir şekilde depolanması gerekmektedir. Düzenlemiş verilerin bir araya getirilip depolanarak oluşturulmuş şekline “veri tabanı” denir. Bu terim ilk yıllarda “kütük (file)” kavramının devamı niteliğinde kullanılmıştır. Bilgisayar bilimlerinde ise veri tabanı kavramı, birbirleri ile anlamlı ve tanımlı ilişkiler bulunabilen bilgiler topluluğudur. Bu bilgilere sistematik bir şekilde erişim sağlanır. Ayrıca yönetilebilir, güncellenebilir ve taşınabilir. Başka bir bilgisayarlarda sistemli bir şekilde depolanmış, programlarca başvuru alan veri yığınıdır (10).

Günümüzde karmaşılaşan dosyalar, çok fazla dosya yapısı ve ilişkileri ayrıca kullanıcıların bu dosya sistemlerinin ulaşım sıkıntıları geleneksel dosyalama sisteminin yetersiz kaldığını göstermiştir. Bu sorunları çözmek, veriye kolay erişim sağlamak ve saklamak konusunda Veri Tabanı Sistemleri (VYS) oldukça işlevseldir. VYS’de veri girişleri, veri depolamak veriye ulaşım sağlayan uygulama programlarına bağlı değildir. Geleneksel dosya sisteminde kayıtlarda veya dosya yapılarında meydana gelebilecek küçük bir değişiklik uygulama programlarının değişmesine sebep olabilmektedir (11). Veri tabanı sistemleri bilgisayar yazılım sistemlerinin önemli bir parçasıdır. Veri toplulukları “*veri tabanı*” olarak adlandırılır. Herhangi bir kuruluşa ait bilgiler bu ortamda yer almaktadır. Bu nedenler veri tabanı sistemleri veri topluluklarının kümeler halinde düzenli bir şekilde bulunduğu ve belirli yazılımlar aracılığı ile yönetildiği yerlerdir (12).

Bir veri tabanının belirli özellikleri içermesi gerekir ve bu özellikler şöyle sıralanabilir:

- Veri tabanı belli bir kuruluşun kullandığı uygulamalarda birbiriyle etkileşimli işletimsel verilerden oluşmalıdır.
- Giriş/çıkış gibi veriler ve geçici veriler veri tabanlarında bulunmaz.
- Veri tabanındaki veriler gereksiz tekrarlardan temizlenmiş, düzenlenmiş şekilde bilgisayar belleklerinde depolanmalıdır.

- Veri tabanında saklanan veriler deęişim gösterebilmektedir. Güncellemeler ile silme, ekleme işlemleriyle veri tabanındaki veriler düzenlenebilmelidir.
- Veri tabanında veriler üzerinde sorgulama, raporlama vb. işlemler yapılabilir.
- Merkezi bir denetim ile veriler üzerinde gerekli işlemler yapılır. Yani uygulama kullanıcıları belirli program dilleri ve komutları kullanarak doğrudan veri tabanında düzenlenmiş verilere erişim sağlayarak bu verileri deęiştirmezler.
- Bir veri tabanını oluşturan, veri saklamayı, çoğaltmayı, güncellemeyi ve yönetmeyi sağlayan programlara Veri Tabanı Yönetme Sistemi (*DBMS – Database Management Systems*) denir. DBMS özellik ve yapısını inceleyen alana Bilgi Bilimi (*Information Science*) denir. Veri Tabanında kayıtların ve bilgi parçalarının tanımlanması en önemli kavramdır (10).

DBMS ile veri tabanının bilgisayar hafızasındaki fiziksel yapılar kullanıcılardan gizlenir ve daha sade mantıksal yapılar tanıtılır. Bağıntısal Veri Tabanı Yönetim Sistemleri (*Relational Database Management Systems - RDBMS*) büyük verilerde güvenli bir şekilde kayıt altına alındığı, bilgilere hızlı ulaşım sağlayan ve birden çok kullanıcıya aynı anda erişime imkânı veren programlardır. MySQL, Oracle13, PostgreSQL, Sybase, BerkeleyDB, Firebird İlişkisel Veri Tabanı sistemleridir. Veri tabanı dillerine SQL, Tel, PL/SQL örnek verilebilir.

Bir evrakın saklanmasına gerek bırakmaması, hızlı güncelleme zaman tasarrufu sağlaması, sadece istenilen bilgiye istenilen zamanda ulaşım sağlaması, tekrarların engellenmesi, hatalı bilgi işlenmesini merkezi bir komuta ile engellemesi, bütünlük sağlaması ve rapor almaya imkan vermesi veri tabanlarının sağlamış olduğu faydalardandır (13).

2.4. Veri Ambarları

Veri ambarı teriminin öncüsü olarak kabul edilen William H. Inmon'un tarafından 1992 yılında ortaya koymuş olduğu "Developing the Data Warehouse" çalışmasında veri ambarı, konu merkezli, kara verme sürecini etkin hale getiren,

bütünleşik ve kalıcı olan veri topluluklarıdır. Analizler ve sorgular için kullanılan, bütünleşmiş bilgi topluluklarının depolanma alanına veri ambarı denir (14). Veri ambarları (*data warehouse*), veri tabanlarının bir parçasıdır. Veri tabanlarını zorlamamak için meydana getirilmiş, özelleştirilmiş, daha hızlı çalışabilen ve bir veri tabanına nazaran daha az veri muhafaza eden yapılardır. Kullanılacak veriyi basit, hızlı ve doğru biçimde analiz etmek için gerekli işlemleri yapabilmektedir (15).

Veri ambarları bazı temel özelliklere sahiptir bu nedenle de veri tabanlarından ayrılmaktadırlar.

- Veri tabanlarının esas amacı tüm verileri depolamakken, veri ambarı özelleştirilmiş ve kullanım için işlenmiş verileri tutar.
- Veri tabanları ile veri ambarları farklı kaynaklardan beslenirler ve bu en önemli farklarıdır. Bir kuruluştaki elde edilen farklı verilerin tümü veri tabanında saklanırken, farklı verilerin bir amaca yönelik birleştirilmesi ve raporlanmaya hazır hale getirilmesi için veri ambarı kullanılır.
- Veri ambarları konu odaklıdır, böylece yapılacak işleme göre özelleştirilebilirler.
- Veri ambarları kullanılan algoritmalarla, trend ve değişimleri takip ederek kendini güncelleyebilir. Veri tabanlarından yapılan güncellemeler ile veri ambarları da güncellenir. Yani veri tabanlarından veri ambarlarına tek yönlü veri akışı sağlanabilmektedir.

Veri ambarları yönetim aşamasında ve özellikle karar verme sürecinde kullanılan sistemlerdir ve veri ambarları bir problemin bütününe yönelik bir bakış açısı sağlar (16).

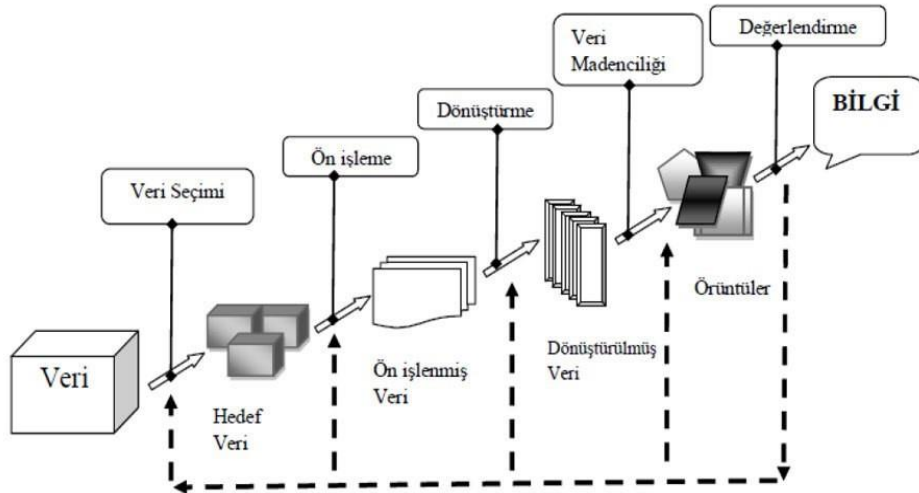
OLAP (On-Line Analytical Processing), veri ambarlarında çeşitli ve farklı izlemlere ait karar vermek için yapılan analizdir. Bu analizler birer sorgulama işlemleridir. Günlük sorgulamalardan farkı ise olasılık içermesidir. Haftalık alış-satış oranı veri ambarı bilgisi içerirken; süt ürünleri satışlarının 10 bini aşması olasılık içerdiği için OLAP'tır. OLAP ile veri madenciliği farklı kavramlardır. OLAP ile analiz yapacak kişi hipotezlere ilişkin örüntü ve anlamları gözeterek bu hipotez ve örüntüleri veri kümesi üzerinde çalışarak kanıtlamaya çalışır. Veri madenciliği ile

çalışacak analizci ise hipotezlere dayanarak hareket etmez. Veri kümesi üzerinden örüntü ve ilişki arar. Fakat OLAP ve VM birbirlerini tamamlamaktadır (17).

2.5. Veri Madenciliği Süreci

Veri madenciliği, bir süreçtir. Veri yığınları arasında, soyut kazı çalışmaları yaparak veriyi ortaya çıkarmaya çalışır. Bunun yanında bilgi keşfi sürecinde örüntüleri ayrıştırarak süzmek ve bir sonraki adıma hazır hale getirmek de bu sürecin bir parçasıdır. Bu süreç Şekil 2.2.'de gösterilmiştir. Üzerinde inceleme yapılan işin ve verilerin özelliklerinin bilinmemesi durumunda ne kadar etkin olursa olsun hiç bir veri madenciliği algoritmasının fayda sağlaması mümkün değildir. Bu sebeple, veri madenciliği sürecine girilmeden önce, başarının ilk şartı, iş ve veri özelliklerinin detaylı analiz edilmesidir. Veri madenciliği sürecinde izlenen adımlar genellikle aşağıdaki şekildedir (18):

1. Problemin tanımlanması,
2. Verilerin hazırlanması,
3. Modelin kurulması ve değerlendirilmesi,
4. Modelin kullanılması,
5. Modelin izlenmesi.



Şekil 2.2. Bilgi Keşfi Sürecinde Veri Madenciliği

Veri tabanlarında öz bilgi keşfi süreci 5 temel aşama içermektedir. Bu aşamalar sırasıyla; (19).

1. Problemin tanımlanması: Öncelikle yapılacak uygulamanın hangi işletme ve amacı için yapılacağının açık ve net bir biçimde ifade edilmelidir. İlgili işletmenin amacı probleme odaklanmaktır ve bu problem açık bir dil ile ifade edilmelidir.

2. Verilerin hazırlanması: (Ön İşleme): Her bir veri analizi iş ve işlemleri öncelikle yeni bir veri setinin toplanması ile başlar. Daha sonra bu veri seti temizlenir ve betimlenir. Bu süreçler tamamlandıktan sonra ancak - veriler analiz edilebilir ve sonuçlara ulaşılmış olur (20). Veri kalitesi veri madenciliğinde anahtar bir konudur. Veri madenciliğinde güvenilirliğin artırılması için, veri ön işleme yapılmalıdır (21). Verilerin hazırlanması aşaması ise şöyledir:

- Verilerin Toplanması: Problem tanımlandıktan sonra o problem için gerekli olan veriler ile verilere ulaşılacak kaynakların belirlenmesi işlemidir (19).
- Verilerin Temizlenmesi: Bu aşamada eksik veriler belirlenmektedir. Aykırı değerler teşhis edilir ve gürültülü veriler temizlenir. Böylece verilerdeki tutarsızlıklar giderilmiş olur.
- Verilerin Birleştirilmesi: Bu aşamada yapılan işlem farklı veri tabanlarındaki verilerin tek bir bütün olması yani (veri ambarında) birleştirilmesi işlemidir. Veriyi elde etme ve sahip olma, verilerin işlenerek iletilmesi ve depolama kapasitesindeki gelişimler işletmelerin sahip oldukları çeşitli ve farklı veri tabanlarının, veri ambarlarında birleştirilmesine yardımcı olmuştur. Özetle, veri ambarları farklı kaynaklarda tutulmakta olan birçok verilerin ortak bir çatı altında birleştirilmesidir (22).
- Verilerin Dönüştürülmesi: Veri madenciliği uygulanacak olan verilerin uygun bir formata getirilmesi işlemi veri dönüştürme aşamasında yapılmaktadır. Bu aşamada veri dönüştürme, verileri düzeltme ya da birleştirme ve bunların yanı sıra genelleştirme-normalleştirme gibi bazı işlemlerde yapılmaktadır.

- Verilerin İndirgenmesi: Veri madenciliği oldukça büyük verileri kapsamaktadır. Bu aşamada veri madenciliği yapılacak bazı veriler oldukça büyük hacimli veri kümeleri olabilir ve bu kümeden daha küçük hacimli veri kümeleri elde edilmesi işlemi yapılmaktadır.

3. Modelin Kurulması ve Değerlendirilmesi: Veri Madenciliğinde öncelikle problem tanımı yapıldığından bahsedilmiştir. Tanımlanan bu problem için en uygun modele ulaşılabilmesi açısından birçok sayıda modelin denenmesi gerekebilmektedir. Bu sebeple, veriyi hazırlama ve model kurma aşamaları, en iyi ve en doğru sonucu verdiği düşünülen model elde edilinceye kadar süreç yinelenabilmektedir.

4. Modelin Kullanılması: Kurulan ve geçerliliği kabul edilmiş bir model ile doğrudan uygulama yapılabilir. Fakat bu model başka uygulamanın alt bir parçası olarak da kullanılabilir (21). Data Mart olarak tanımlanmış olan işlev ile özel veri tabanlarına veri aktarımı yapılabilir. Böylece veri madenciliği sürecine başlanması mümkün olur (19). Veri ambarları hacim açısından Data Mart'lara göre oldukça büyüktür. Data Martlar küçük boyutlu olup (1–10 GB) alana sahip ve bölümsel veri ambarlardır (22).

5. Modelin İzlenmesi: Zaman içerisinde teknolojinin ilerlemesi, verilerin artması ya da değişmesi ile tüm sistemlerin özelliklerinde ve buna bağlı olarak dolayısıyla ürettikleri verilerde de değişiklikler olabilmektedir. Bu nedenle kurulan modeli sürekli olarak izlenmelidir. Böylece yeniden düzenlemeye gereksinim duyulduğunda işlemler yapılabilir (23).

2.5.1. Verilerin Hazırlanması

Veri madenciliğinin aşamasında veri içerisindeki farklı şekildeki, kayıtların tek bir formata dönüştürülmesi ve standartlaştırılması gerekmektedir. Bu aşamanın sonunda aşağıdaki başlıklar sıralanır (24):

- Devam edecek diğer tüm aşamalar için ham verilerden asıl veri kümesine ulaşmak,
- Analiz edilecek verinin ve analiz için uygun olan durum ve değişkenleri belirleyerek seçmek,

- Tekrar eden değişkenleri dönüştürmek,
- Modellemeye uygun olmayan ve katılması istenmeyen verilerin temizlenmesi

Tüm veri madenciliği çalışmaları, verinin sahip olduğu özellikler ve sunumu üzerinedir. Eğer veri işlenecek durumda değilse kullanıcı veriyi uygun duruma getirmek için bir takım önlemler almalıdır. Eksik veya gürültülü, tutarsız ya da işlemeye uygun olmayan gereksiz verilere işlem uygulamak sürecin bir aşamasıdır. Veri madenciliği kullanıcıları kayıtları kullanılabilir olacak şekilde uygun hale getirmelidirler. Genel olarak ikili veya sıralı değişkenler şeklinde veriler hazırlanır. Verilerin bazen standart bir formata dönüştürülemediği ya da indirgenemediği durumlar oluşabilir. Benzer bir şekilde veri madenciliği, sıralı ve nümerik olan değerleri analiz etme amacıyla tasarlanmıştır, kategorize aynı şekilde değilse analiz etmek kolay değildir (25).

Veri madenciliğine başlamadan önce veriler dikkatli şekilde gözden geçirilmelidir. Veri tabanında eksik olabileceği gibi aşırı uç olarak elde edilen değerler ya da yanlış verilerde olabilirler. Bunlara gürültü terimi (*noisy data*) denir. Veri hazırlanırken şu aşamalar gerçekleştirilmelidir (26).

1) Verilerin temizlenmesi: Veri tabanında eksik veriler varsa;

- a. Eksik (kayıp) verileri çıkarmak: Analizci tarafından belirlenen eksik verilerin veri kümesinden çıkarılması gerekmektedir. Fakat burada çıkartma işlemi yaparken veri sayısına oldukça dikkat edilmelidir. Çok sayıda kaydın çıkarılması analizin sonuçlarını olumsuz yönde etkileyebilir.
- b. Eksik ya da boş verileri manuel teker teker doldurmak: Değişken üzerinde eksik verilerin bulunuyorsa bu değişkenin dağılımının belirlenerek ve bu dağılım üzerinden üretilerek elde edilecek rastgele (*random*) sayılarla eksik olan boş gözeler doldurulur.
- c. Bulunan eksik ya da boş gözelerle aynı bilgiyi girmek: Analizci tarafından belirlenecek sabit bir sayı ile işlem gerçekleştirilir.
- d. Eksik ya da boş gözelerle bulunan tüm verilerin ortalama değerini vermek: Var olan verilerin öncelikli olarak ortalaması hesaplanır ve bulunan değer

ortalama deęer girilir. Bylelikle verilerin de sahip olmuř olduęu genel ortalama etkilenmemiř olur.

- e. Regresyon yntemi ile eksik verileri tahmin etmek: Eksik verilere ait olan deęiřkenler regresyona uygun ise elde olan veriler ile gerekli varsayımlar saęlanarak regresyon uygulaması yapılır. Elde edilen denkleme gre eksik deęerler doldurulur.

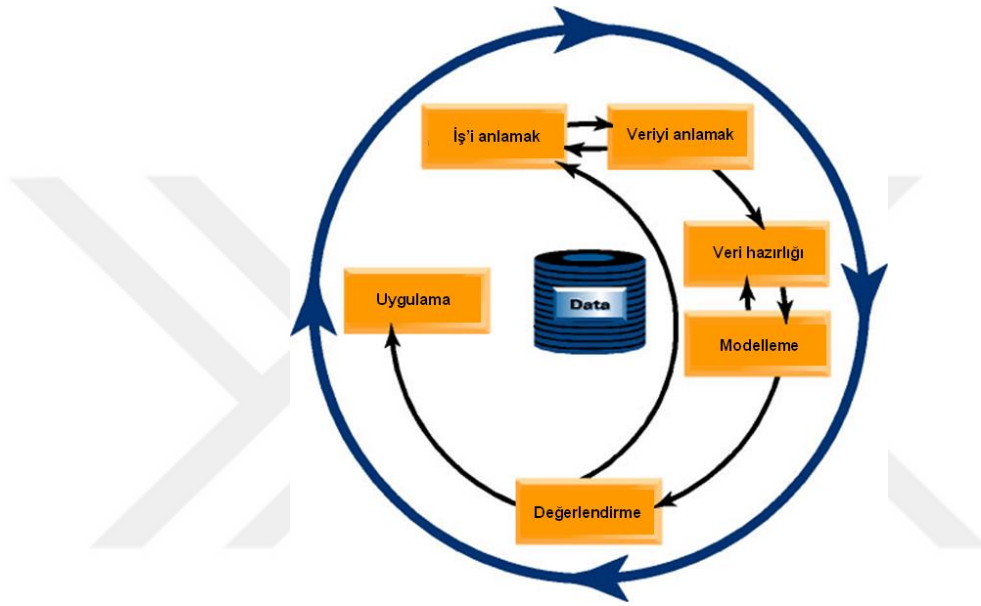
2) Verilerin yeniden yapılandırılması:

- a. Yanlıř olarak gruplandırılmıř veriler: Veri setindeki aynı gruplarda bulunduęu halde aslında farklı gruplarda olmaları gereken veriler bulunabilir. Eęer byle veriler varsa bunların tesit edilerek uygun gruplara dâhil edilmesi gerekir.
- b. Aykırı veya uę deęerlerin tespit edilmesi: Aykırı deęerler verilerin belirli sınırları dışında kalmıř olan ya da verinin ynelimine zıt ynde ilerleyen deęerlerdir. Kesinlikle tespit edilerek dzeltilmelidirler. nk analiz sonuları bu nedenle hatalı olabilir. İstatistiksel yntemlerden olan histogramlar, saılım veya kutu grafikleri aykırı deęerleri tespit etmek iin kullanılabilecek yntemlerdendir.
- c. Veri dnřtrme: Deęiřkenlerin iinde bulunması gereken aralıktan farklı bir aralıkta bulunmasıdır. Deęiřkenin belli bir deęer aralıęında olması gerekiyor ve veri bu deęer aralıęında bulunmuyor ise verilerin belirlenen kullanılacak algoritmaya uygun duruma getirmek iin verilerin yeniden dzenlenmesi gerekir.

2.5.2. Veri Madencilięi Ynteminin Uygulanması

Birok byk kuruluřun (kamu veya zel) ok byk verileri otomatik olarak toplayabilmesiyle veri madencilięi hızla bir řekilde geliřmektedir. Bu sebeple elde edilen verilerin zerinden fazla zaman gemeden rekabete ortamı nedeniyle daha nce analiz edilerek bilgiyi meydana getirmesi gerekmektedir. Veri madencilięine karřı geliřigzel bir yaklařım sergilenmesi nedeniyle endstriler arası standart bir kavram geliřtirilmesi gerekli hale gelmiřtir. Daimler-Chrysler tarafından 1996 yılında, tarafsız sanayi, malzeme ve uygulama propagandasıyla ortaya ıkmıř olan

Cross Industry Standard Process for Data Mining (CRISP-DM) kavramı ilk olarak SPSS ve NCR firmalarının analizcileri aracılığı ile geliştirilmiştir (27). CRISP-DM, veri madenciliği için test edilmiş güvenilir bir süreç modelidir. Bu süreç, iş süreci resmiyete geçmeden önce, verileri benimsemeye başlayıp çözümü elde edene kadar geçen süreçleri ve gerekli adımları kapsar. Şekilde 2.3.'te belirtildiği gibi süreç altı aşamadan oluşmaktadır. Her aşamanın birbiri ile ilişkisini oklar ile takip edebiliriz (28).



Şekil 2.3. CRISP-DM Süreci

- 1) **Amacın belirlenmesi (Business Understanding):** Çalışma yapılacak konu ve konuya ait amaç ve ulaşılabilecek olan hedefler belirlenir. Problemin doğru bir biçimde belirlenebilmesi önemli bir adımdır. Gerekli çözüm yöntemleri için taslak oluşturur.
- 2) **Veriyi Anlama (Data Understanding):** Verilerin oluşturulan plana göre hazırlanması ve veri ile ilgili özelliklerin sunulduğu aşamadır. Veri türleri, kalitesi, veriye ait alt kümelerin özellikleri, verilerdeki hata ve eksikler tespit edilir.
- 3) **Veriyi Hazırlama (Data Preperation):** Verilerin analizden önceki sürecinin tamamını kapsayan en önemli aşamasıdır. Çalışmaya uygun olmayan veriler doğru sonuçlar vermeyecektir. Şekil 2.3. CRISP-DM döngüsünde görülen Araştırmaya ait amacın belirlenmesi, Verinin anlaşılması, Verinin hazırlanarak

modellenmesi, uygulama ve değerlendirme aşamaları için bu verilerin ön analizleri yapılır ve tanımlayıcı istatistikler hesaplanır. Bu sayede gürültü terimleri belirlenir. Eksik verilerin tespiti, hata içeren verilerin düzeltilmesi ve indirgenmesi, dönüştürülecek verilerin dönüşüm işlemlerinin sağlanması vb. veri temizleme kısımları bu aşama ile yapılmaktadır.

- 4) **Modelleme (Modeling):** Bir önceki aşamalardan geçen temizlenmiş veriye hangi modelleme yapılacağına karar verildiği aşamadır. Veri kümesi üzerinde uygun bulunan modelleme uygulanır ve sonuçlar alınır. Bu aşamada birçok teknik kullanılabilir. Bu nedenle bu safhada geriye dönülerek yeniden uygulama yapılabilir. Bu nedenle bu safhada geriye dönülerek yeniden uygulama yapılabilir.
- 5) **Değerlendirme (Evaluation):** Seçilen modellerin uygunluk ve geçerliliğinin genel değerlendirmesinin yapıldığı aşamadır. Elde edilen sonuçlar yorumlanır. Başlangıçta belirlenmiş olan hedef ve amaca ulaşım kontrol edilir. Eksiklik var ise yeniden çalışma yapılması için yol gösterici aşamadır.
- 6) **Sonuçları kullanma (Deployment):** En son elde edilen tüm bulguların değerlendirmesi yapılır. Sonuçların nasıl uygulanacağına karar verilir ve sonuç faaliyet raporu şeklinde sunulur. Bazı durumlarda veri madenciliği sürecini yeniden başlatma gereksinimi duyulabilir. Bu aşama özellikle hedeflere ulaşım açısından hem kurum (işletme) hem de müşterilere (kullanıcı) bağlıdır. (29).

2.5.3. Sonuçların Değerlendirilmesi ve Sunulması

Veri madenciliğinde çalışmanın yarısı örüntüleri bulmaktır. Buradaki esas amaç ise değerlendirme yapabilmek için modeli kullanmaktır. Veri madenciliğinde *Scoring* olarak adlandırılan bu kısım, değerlendirmeyi yapabilmek için seçilen model ve yeni durumları kapsayan veri setini içerir. Eğitilen model ile yeni karşılaşılabilecek durumlar için tahminleme yapılabilir (30).

Raporlama, veri madenciliğinde elde edilen bulguların sunulması için önemli bir dağıtım kanalıdır. Veri kümesine uygulanan modellemelerden sonra belirlenmiş hedeflere ulaşmak için bir değerlendirme raporu hazırlanmaktadır. Bu rapor, hedefleri ve hedeflere ulaşma oranını belirlediği gibi yorumları da içerir. Metin şeklindeki raporlar belli bir kitleye hitap edeceğinden genel kitleye hitap edebilecek

görsel ve işitsel sunum yöntemleri daha etkili olacaktır. Sonuçların tablolar ve grafikler halinde raporlanması karışıklığı önleme, anlamayı kolaylaştırma ve dikkati toplama açısından anlamayı kolaylaştırır. Bu nedenle veriye setine uygun grafik seçimine dikkat edilmelidir. Raporlanan bilginin objektif ve standart bir şekilde grafiklerinin sunulması, yorumlarının ise açık bir biçimde ifade edilmesi gerekmektedir. Temel istatistiksel sonuçlar için Histogram, dal&yaprak, dilim ve kutu grafikleri oldukça kullanışlıdır. Fakat çok değişkenli analizler için bu tür grafikler açıklayıcı olmayabilir. Verinin görselleştirilmesi konusunda veri madenciliği, bilgisayar teknolojisinin ilerlemesi ile iki ve üç boyutlu görsellik sunumları meydana getirmiştir (31).

2.5.4. Veri Madenciliği Uygulama Alanları ve Karşılaşılan Sorunlar

Veri yığınlarının tek başlarına bir anlamı yoktur. Bu yığınlar belirli bir amaç ile işlendiğinde anlamlı sonuçlar alınabilir.

Veri madenciliğinin en çok kullandığı alanlar şu şekilde sıralanabilir (32):

- Endüstri
- Eğitim
- Sağlık
- Bilim
- Mühendislik
- Pazarlama
- Bankacılık
- Sigorta
- e-Ticaret
- Borsa
- Suç Soruşturması
- Telekomünikasyon
- Lojistik
- Perakende

Veri madenciliğinin kullanıldığı alanlarda birçok işlevi bulunmaktadır. Böylece işletmelere veya kurumlara oldukça yarar sağlamaktadır. Müşteri segmentlerini ve tercihlerini anlayarak memnuniyeti artırmak, müşteri kaybını önleme ve bağlılığını artırmak, yeni müşteriler edinmek, çapraz satışı ve ek satışı geliştirmek, operasyonel performansın takip edilmesi, bir ürün veya hizmetle ilgili kampanya programı, hedef kitlenin seçimi, sunum yapılacak kanala kadar tüm sürecin belirlenmesinde veri madenciliğinden faydalanılabilir. Teknik kaynakların en elverişli biçimde kullanılmasını sağlamak, olası risklerin tespit edilmesi, dolandırıcılık olaylarının tespitinde kullanılabilir.

Sağlık alanında ise hasta düzeyinde sağlık hizmeti sunumunun iyileştirilmesi, sağlık kurumlarında yöneticilik alanında iyileştirme, kaynakların etkin kullanımı ile sağlık politikalarının oluşturulması, bazı hastalık tiplerini belirlemek ve sağlıklı-riskli teşhisini koyabilmek için kullanılmaktadır (33).

Veri madenciliğinde ham verilerin kullanılması ile kayıtlar ve veri tabanlarının yapısındaki çok sayıdaki veriler sebebiyle birçok sorun ile karşılaşılabilir. Gelişen teknoloji sayesinde geliştirilen algoritmalara rağmen büyük veriler veri tabanları ile uyum sorunu oluşturabilmektedir. Çünkü elde edilen verilerde ilişki belirlemek için aykırı veriler olabileceği gibi bu durum analiz yapmayı da oldukça zorlaştıracaktır. Bunun yanında eksik verilerin varlığı ve eksik gözlemlerin giderilmesi için yöntemler geliştirilse de eksik veri fazlalığı önemli bir sorundur. Kullanılan analiz programlarını birçoğu verilerin tamamı ile analiz yapmaya uygun olduğundan, eksik veri istatistikî sonuçların geçerliliğini azaltabilmektedir. Bunların dışında ise veri tabanı sorunları, uygun analizlerin seçilememesi, kullanılacak paket programın uyumluluk sorunları, kullanılacak algoritma ile veri setinin uyuşmaması da karşılaşılabilecek sorunlardır. Sonuçların ortaya kullanılmasında kullanılacak ek uygulama ve programlarda ekonomik açıdan sorun oluşturabilmektedir (13).

2.6. Data Analitiği Uygulama Programları

Veri Madenciliği ile analiz yapabilmek için hem ticari hem de açık kaynak programları geliştirilmiştir. Bu programların içeriğinde ise birçok algoritma vardır.

Bu algoritmalar kullanarak kullanılacak verilerden anlamlı bilgiler elde edilebilmektedir. Veri Madenciliği uygulamaları için bilgisayar programı kullanılmaktadır ve bununla birlikte birçok yazılım ve yazılım dilleri geliştirilmiştir (34).

Açık kaynaklı programlara; SPSS Clementine, Excel, SPSS, SAS, Angoss, KXEN, SQL Server, MATLAB örnek verilebilir. Açık kaynaklı kodlamalı programlardan başlıcaları ise Python, WEKA, KNIME, Orange, RapidMiner, jHepWork, Scriptella ETL, ELKI', Rattle, Oracle, Tanagra ve R'dır (30).

2.7. Büyük Veri

Büyük veri (big data) kavramı; verilerin muhafaza edilmesi, analizi ve yönetilmesinde klasik veri tabanı sistemleri yetersiz kaldığında karşımıza çıkmaktadır (35). Büyük veri soyut bir kavramdır. Veri kütlelerinin yanı sıra, kendisi ile "büyük veri" veya "çok büyük veri" arasındaki farkı belirleyen başka bazı özelliklere de sahiptir." Şu anda, büyük verilerin önemi genel olarak kabul edilmiş olsa da, insanların tanımı hakkında hala farklı görüşleri var. Genel olarak, büyük veri, geleneksel IT ve yazılım / donanım araçları tarafından tolere edilebilir bir süre içinde algılanamayan, edinilemeyen, yönetilemeyen ve işlenemeyen veri kümeleri anlamına gelecektir. Farklı kaygılar nedeniyle, bilimsel ve teknolojik işletmeler, araştırma bilimcileri, veri analistleri ve teknik uygulayıcılar farklı büyük veri tanımlarına sahiptir. Aşağıdaki tanımlar, büyük verinin derin sosyal, ekonomik ve teknolojik çağrışımları hakkında daha iyi bir anlayışa sahip olmamıza yardımcı olabilir (36). 2010 Yılında Apache Hadoop, büyük veriyi "genel bilgisayarlar tarafından kabul edilebilir bir kapsamda yakalanamayan, yönetilemeyen ve işlenemeyen veri kümeleri" olarak tanımlamıştır. Bu tanıma dayanarak, Mayıs 2011'de küresel bir danışmanlık ajansı olan McKinsey & Company, inovasyon, rekabet ve üretkenlik için bir sonraki sınır olarak 'Büyük Verileri' açıkladı. Büyük veri, klasik veri tabanı yazılımı tarafından edinilemeyen, depolanamayan ve yönetilemeyen veri kümeleri anlamına gelir. Bu tanım iki çağrışım içerir: Birincisi, veri kümelerinin büyük veri standardına uyan hacimleri değişebilir ve zaman içinde veya teknolojik gelişmelerle büyüyebilir; İkincisi, farklı uygulamalardaki büyük veri standardına uyan veri kümelerinin hacimleri birbirinden farklıdır. Şu anda, büyük

veriler genellikle birkaç TB ile birkaç PB arasında değişmektedir. McKinsey & Company'nin tanımından, bir veri kümesinin hacminin büyük veriler için tek kriter olmadığı görülebilir. Giderek artan veri ölçeği ve geleneksel veri tabanı teknolojileri tarafından ele alınamayan yönetimi sonraki iki temel özelliktir (37).

Nitekim büyük veri 2001 gibi erken bir tarihte tanımlanmıştır. META analisti Doug Laney (Gartner), bir araştırma raporunda 3VS modeliyle artan verilerin getirdiği zorlukları ve fırsatları, yani Hacim, Hız ve Çeşitliliğin artmasını tanımladı (38). Böyle bir model başlangıçta büyük verileri tanımlamak için kullanılsa da, Gartner ve IBM (39) ve Microsoft'un bazı araştırma departmanları (40) dâhil olmak üzere diğer birçok işletme, önümüzdeki on yıl içinde büyük verileri tanımlamak için hala "3VS" modelini kullanıyor (41). "3VS" modelinde Hacim, veri kütlelerinin üretilmesi ve toplanmasıyla veri ölçeğinin giderek daha büyük hale gelmesi anlamına gelir; Hız, büyük verilerin, özellikle veri toplama ve analizinin vb. Güncelliği anlamına gelir. Büyük verinin ticari değerini en üst düzeye çıkarmak için hızlı ve zamanında yürütülmesi gerekir; Çeşitlilik, ses, video, web sayfası ve metin gibi yarı yapılandırılmış ve yapılandırılmamış verilerin yanı sıra geleneksel yapılandırılmış verileri içeren çeşitli veri türlerini gösterir.

3V kavramı günümüzde ilerlemeler ile yerini 5V kavramına bırakmıştır. Büyüklük kavramı veride bir tek hacimsel (Volume) olarak ifade edilmediği gibi, verilerin zaman içerisinde çok hızlı bir hareket göstermesi, işlenmesi (Velocity), verinin aynı zamanda çok oluşu ya da oldukça farklı çeşitlilik göstermesi (Variety) (yapılandırılmış/yapılandırılmamış/yarı yapılandırılmış) olması da büyük veri olarak ifade edilmesinin nedenidir. Bu kavramlara ek olarak ise verilerin anlamlı değerler vermesi (Value) ve doğrulanabilmesi (Verification) büyük veri olarak adlandırılmasına nedendir (42).

Bununla birlikte, diğerlerinin büyük veri ve araştırma alanlarındaki en etkili liderlerden biri olan IDC de dâhil olmak üzere farklı görüşleri vardır. 2011 Yılında, bir IDC raporu büyük verileri "büyük veri teknolojileri, yüksek hızlı yakalama, keşif ve / veya analizi mümkün kılarak, çok çeşitli verilerin çok büyük hacimlerinden ekonomik olarak değer elde etmek için tasarlanmış yeni nesil teknolojileri ve mimarileri tanımlar." (43). Bu tanımla, büyük verilerin özellikleri dört vs., yani

Hacim (büyük hacim), Çeşitlilik (çeşitli modaliteler), Hız (hızlı üretim) ve Değer (büyük değer ancak çok düşük yoğunluk) olarak özetlenebilir. Bu tür 4VS tanımı, büyük verilerin anlamını ve gerekliliğini vurguladığı, yani büyük gizli değerleri araştırdığı için yaygın olarak kabul görmüştür. Bu tanım, büyük verideki en kritik sorunu gösterir; bu, büyük ölçekli, çeşitli türlerde ve hızlı nesil veri kümelerinden değerlerin nasıl keşfedileceğidir. Facebook'un Başmühendis Yardımcısı Jay Parikh'in dediği gibi, "Toplanan verileri kullanmazsanız, yalnızca büyük veriden başka bir veri grubuna sahip olabilirsiniz." (44).

Buna ek olarak NIST, büyük veriyi "Büyük veri, veri hacminin, edinme hızının veya veri gösteriminin, etkin analiz yapmak için geleneksel ilişkisel yöntemleri kullanma kapasitesini sınırladığı veya önemli yatay yakınlaştırma teknolojileriyle etkin bir şekilde işlenebilecek verileri ifade edecektir" olarak tanımlamaktadır ve bu da büyük verinin teknolojik yönüne odaklanmaktadır. Hem endüstriden hem de akademiden büyük verilerin tanımı konusunda önemli tartışmalar olmuştur (45). Uygun bir tanım geliştirmenin yanı sıra, büyük veri araştırması, değerinin nasıl çıkarılacağına, verilerin nasıl kullanılacağına ve "bir grup veriyi" büyük veriye nasıl dönüştüreceğine de odaklanmalıdır (36).

Büyük veri, klasik veri tabanı sistemleri ile işlenmesi oldukça zor ya da mümkün olmayan, farklı hacimlerdeki heterojen veriyi açıklayan yeni oluşturulmuş bir kavramdır ve içeriğini çeşitli dijital içerikler meydana getirmektedir (7):

1. Yapısal veri: Yapılandırılmış veriler, modellenen, girilebilir, depolanan, sorgulanan, işlenen ve görselleştirilebilir tüm veri türüdür. Genellikle, ilişkisel veri tabanlarında veya elektronik tablolarda yönetilen, belirli tür ve boyutlarda önceden tanımlanmış alanlar olarak sunulur. Katı yapıya sahip veriler, işleme yüksek performans yetenekleri veya paralel teknikler gerektirmediğinden yararlı bilgilere ulaşımı kolaylaştırır.
2. Yarı yapısal veri: Yarı yapısal veri, yapısal bir veri türünü belirtmekle birlikte sadece katı bir sistemi içermemektedir. Yani yapıyı tanımlayan modele ek olarak, belirli öğeleri tanımlamak ve veriler içindeki farklı alanların hiyerarşik bir temsilini tanımlamak için kullanılan etiketler ve işaretleyiciler gibi bir tür meta model içerir. Bu veriye verilecek en bilinen

örnekler arasında XML (Extensible Markup Language) ve JSON (JavaScript Object Notation) programlama dilleridir.

3. Yapısal olmayan veri: Yapılandırılmamış veriler, önceden tanımlanmış format olmadan sunulan ve saklanan bir kayıt türüdür. Genellikle kitaplar, makaleler, belgeler, e-postalar ve görüntü dosyaları, sesler ve videolar gibi medya dosyaları gibi serbest biçimli metinlerden oluşur. Bu tür verilerin katı bir biçimde temsil edilmesinin zor olması, işlenmesini zorlaştırır ve bu da NOSQL gibi yeni işleme mekanizmalarının ortaya çıkmasına neden olur.

2.7.1. Büyük Data Analitiği Yaklaşımı

Projeler hedeflenmiş bir amaca ulaşmak için belirlenmiş alt hedefleri doğrultusunda başlar. Bu projelerin hedef ve amaçlarına ulaşabilmesi için organizasyonların hali hazırda bulunan bilgi sistemlerinin işleyemeyeceği büyüklükteki karmaşık veri kümelerine ulaşmaları ve uygun veriyi elde etmeleri gerekmektedir. İstenilen veriye ulaşım sağlandıktan sonra elde edilen yapısal, yarı yapısal ve yapısal olmayan veriyi yekpare olarak aralarında ilişkilendirecek, işleyerek karar verme kademesinde bilgiye erişmeyi sağlayacak veri analitiğine ihtiyaç duyulacaktır. Yapısal verilerin yanı sıra yarı yapısal ve yapısal olmayan veriler de teknolojik gelişmeler sayesinde kullanılmaktadır (5). Farklı ve geniş verileri işlemek için geliştirilen büyük data analitiği, analitik tekniklerin kullanılmasıdır. Sürekli ve hızla değişen ve çok fazla miktarlardaki verileri, veri tabanı teknikleri uygulayarak işlemek ve değerli bilgiler elde etmek amaçlamaktadır (7).

2.7.2. Data Analitiği Döngüsü

Büyük verinin tanımlanmasından sonra yaşam döngüsüne değinilmesi gerekmektedir. Tüm varlıkların bir yaşam döngüsüne sahip olduğu gibi veri kavramı da bir yaşam döngüsü içindedir (Şekil 2.4.). Hiçbir varlık tek başına bir kere ve sonsuza kadar var olamaz. Bu durum veriler içinde geçerlidir. Gelişen teknoloji ile kavram ve değerler değişmektedir. Büyük data her ne olursa olsun, örneğin; sosyal platformlarda toplanan veriler, elektronik postaların tamamı, bir ülkedeki

vatandaşlara ait tüm nüfus kayıtları veya düşünülebilecek herhangi bir büyük veriyi ele aldığımızda, aslında bir yaşam döngüsüne sahip olduğunu görürüz (46).

Essex Üniversitesi tarafından literatüre kazandırılan bu yaklaşım verinin üretimi, işlenmesi, analiz edilmesi ve bozulmalara karşı saklanması, güvenlik problemlerine karşı korunması ya da verilere erişecek kişilere karar verilmesiyle birlikte verinin tekrar kullanılması gibi dönülerden oluşur (47).



Şekil 2.4. Büyük Veri Yaşam Döngüsü

Özetle, verinin üretilmesi (*Creating*), üretim sürecinin nasıl yönetilmesi gerektiği ve verinin ulaşılacağı kaynaklar esas problemlerdir. Datanın işlenme kısmında (*process*) ise ön bir işlem yapılır, eksik veriler toplanır, anlamsız veriler temizlenir. Bu aşamalardan sonra ise analiz aşaması gelir. Analiz aşamasını ise verilerin korunması izlemektedir. Bu aşama, verilerin yedeklenmesi, verinin saklanacağı uygun formatın seçilmesi, paylaşımı arşivlenmesi ve güncel tutulması gibi kararların alınıp problemlerin gözetildiği aşamadır. Verilerin sürekli yenilenmesi nedeniyle güncellenmesi gerekir. Bu sebeple sonraki aşamalarda erişim haklarının belirlenmesi, güvenlik kontrollerin yapılması gerekmektedir. Erişim haklarının paylaşılması, yasal olup olmadığı, verilerin dağıtılması, erişimlerin kimler tarafından ne sıklıkla yapıldığı sorularına cevaplar aranmaktadır (48).

2.7.2.1. Veriyi Hazırlama

Veri madenciliği yapılmadan önce veriler öncelikle dikkatli bir şekilde gözden geçirilmelidir. Çünkü veri tabanlarındaki kayıtlı bulunan ham verilen büyük bir kısmı işlenmemiş, eksik ve gürültülü veriler içerecektir. Bu nedenle verilen düzenlenmesi gerekir. Yanlış veya aşırı uç olarak kaydedilmiş veriler veri tabanında bulunabilmektedir. Bu tanımdaki veriler gürültü terimi (*noisy data*) olarak adlandırılmaktadır. Düzenleme aşamasında uygun olmayan veriler temizlenirken bazı verilerin ise birleştirilmeleri gerekebilir. Aşamaları şu şekilde sıralayabilir (26):

1) Verilerin temizlenmesi: Çalışma yapılacak veri tabanında eksik kayıtlar bulunuyorsa;

- a. Eksik (kayıp) verilerin bulunduğu kaydı çıkarmak: Eksik kayıtlar analizi yapacak ekipler veya kişi tarafından temizlenir. Fakat bu işlem yapılırken çıkarılacak veri sayısına analiz sonuçlarının etkilenmemesi adına dikkat edilmelidir.
- b. Eksik verileri elle teker teker doldurmak: Veri setindeki verilerin sahip olduğu dağılım tespit edilerek boş gözeler bu dağılıma uygun olacak şekilde rastgele (*random*) veri üretilerek doldurulur.
- c. Tüm eksik verilere aynı bilgiyi girmek: Tüm boş gözelerin analizi yapacak ekipler veya kişi ile belirlenmiş ortak bir sayı (değer) ile doldurulmasıdır.
- d. Eksik verilere tüm verilerin ortalama değerini vermek: Genel ortalamayı etkilememek için eldeki verilerin ortalaması hesaplanır. Eksik gözeler belirlenmiş bu ortalama ile doldurulur.
- e. Regresyon yöntemi ile eksik verileri tahmin etmek: Eğer eldeki veri regresyon uygulamaya uygunsa gerekli varsayımlar göz önünde bulundurularak regresyon uygulanır. Ulaşılan denklem kullanılarak eksik değerler tamamlanır.

2) Verilerin yeniden yapılandırılması:

- a. Yanlış gruplandırılmış veriler: Çalışma yapılacak veri setindeki değişkenlerden farklı gruplarda bulunduğu halde aynı grupta olmaları gereken veriler tespit edilerek doğru bir şekilde kategorilenmelidir.
- b. Aykırı değerlerin tespit edilmesi: Bu değerler verideki belirlenen sınırlar dışında alt ve ya üst sınırlarda kalarak verinin trendinin ters yönde değiştiren değerlerdir. Analiz sonuçlarında hataya sebebiyet vermeleri nedeni ile kesimlikle düzenlenmeleri gerekir. Bu aşamada çoğunlukla istatistiksel yöntemler yardımcı olacaktır. Böylece aykırı değerlerin kolayca tespit edilmesinde öncelikli yöntemler olarak histogram, saçılım veya kutu grafikleri kullanılabilmektedir.
- c. Veri dönüştürme: Veri setinde bulunan verilerin bulunduğu bir belli aralık ile ilgileniyorsak ve bazı verilerin bu aralık dışında olması analiz sonuçlarını etkileyecektir. Böyle bir durumla karşılaşıldığında verileri yeniden yapılandırmamız ve uygun algoritmalara uygun hale getirmemiz gerekmektedir. Bunun için geliştirilmiş bazı yöntemler bulunmaktadır:

Min-Maks normalizasyon: Her bir değer minimum ile farkının Min-Maks aralığından ne kadar büyük olduğuna göre hesap yapılmaktadır.

$X = \{x_1, x_2, x_3 \dots x_n\}$ Değer dizisinde Z_i , i .normalleştirilen değer olsun. (0,1) aralığındaki normal değerler için bu yöntem:

$$Z = \frac{X_i - \min(X)}{\max(X) - \min(X)} \quad (2.1)$$

Herhangi bir (a,b) aralığı için Min-Maks yöntemi:

$a = yeni_min$ ve $b = yeni_maks$ olmak üzere;

$$Z = \frac{X_i - \min(X)}{\max(X) - \min(X)} \cdot (yeni_min - yeni_maks) + yeni_min \quad (2.2)$$

Sıfır ortalamalı standart normalizasyon: Her bir değer ortalamaları ile farkının standart sapması değerine bölünmesiyle hesaplanır. $\bar{X}$, X dizisinin ortalama değeri ve σ , dizinin standart sapması olmak üzere;

$$Z_i = \frac{X_i - \bar{X}}{\sigma} \quad (2.3)$$

- d. Veri boyutunu indirgeme: Bu aşamada genel olarak kullanılan iki yöntem bulunmaktadır. Dalga Dönüşümü (*Wavelet Transform*) ve Temel bileşenler analizi bu yöntemlerdendir (49).

2.7.2.2. Modelleme

Bu aşamada çeşitli modelleme tekniklerinin ve algoritmalarının seçilmesi, parametrelerin seçilmesi ve uygulama işlemleri gerçekleştirilir. Veri madenciliği, istatistik, makine öğrenmesi, veri tabanı teknolojileri, yapay sinir ağları, verilerin görselleştirilmesi vb. gibi farklı disiplinlerin birleşimini içermektedir. Genellikle betimsel (*descriptive*) ve çıkarımsal (*predictive*) olarak ikiye ayrılmaktadır (50).

Tahmin edici modeller sonuçları bilinen verilerden bir modele ulaşır, sonuçları bilinmeyen veri kümeleri için sonuç değerlerinin tahminlenmesidir. Sınıflama, Regresyon ve Zaman Serileri Analizi bu kategoride yer almaktadır. Karar verme ve tahminleme aşamasında ise Kümeleme Yöntemleri ve Birliktelik Kuralları kullanılabilecek yöntemlerdir (51).

Betimsel veri madenciliği verilerin genel özelliklerini araştırırken, çıkarımsal veri madenciliği ise verilerden tahmin yapmayı sağlar. Böylece veri madenciliğinde problem tiplerine uygun modellemeler ortaya çıkmaktadır. Bu modeller şöyle sıralanabilir (52):

- Sınıflandırma (Classification): K en yakın komşuluk, sinir ağları, karar ağaçları ve duruma dayalı nedensellik örnekleridir.
- Çıkarım (Prediction): Regresyon, Sinir ağları, k en yakın komşuluk vb. içerir.
- Bağımlılık Analizi (Dependency Analysis): Regresyon ve korelasyon analizleri, Bayesian ağlar ve birliktelik kurallarından oluşur.
- Verilerin Tanımlanması ve Özetlenmesi (Data Description and Summarization): İstatistiksel teknikleri içermektedir.

- Bölme veya Kümeleme (*Segmentation or Clustering*): Kümele teknikleri, sinir ağları ve görselleştirme yöntemleri bu gruptadır.

2.7.2.3. Görselleştirme

İnsanın algılama sistemi 3 boyutludur. Bu nedenle daha fazla boyutu kapsayan datalar insanın algılamasını zorlaştırmaktadır. Veri görselleştirme teknikleri çok boyutlu verileri algılayabileceğimiz şekilde 2 ve ya 3 boyuta indirirken veriler arasındaki ilişkiyi koruyabilmelidir. İndirgeme aşamasında bir miktar veri kaybolacaktır. Muhafaza edilemeyen veriyi en düşük düzeyde tutabilmek temel hedeftir. Temel iki hedeften bir diğeri ise fikir ve kavramların daha iyi anlaşılmasına yardımcı olabilmektir. Bunların hepsi birer bilgi içerdiği için bu tür görselleştirmelere “bilgi görselleştirmesi” (*knowledge visualization*) adı verilir. Veri görselleştirmenin amaçları grafikler ve resimler aracılığı ile yeni fikirler oluşturmak, yeni ilişkiler kurmak, hipotezlerin doğruluğunu sınamak, yeni yapılar keşfetmek veya bu yapıları düzenlemektir. Kısaca problemlerin mantıksal çözümleri için insanların görsel algılamasını kullanabilmektir (53). Bu tür görselleştirmelere “veri görselleştirme” (*data visualization*) ya da “görsel bilgi keşfi” (*visual data exploration*) denir (54).

Veri görselleştirme alanında tanınan Kriegel ve Keim tarafından veri görselleştirme araçları altı sınıfa ayrılmış olup bu sınıfları, geometrik izdüşüm teknikleri, ikon tabanlı teknikler, piksel tabanlı teknikler, hiyerarşik teknikler, grafik tabanlı teknikler ve karma teknikler oluşturmaktadır (55).

2.7.2.4. Değerlendirme

Çalışılan veri kümesine modelleme yapıldıktan sonra veri madenciliği yöntemleri kullanılarak amaçlanan hedeflere ulaşma doğrultusunda raporlama yapılmaktadır. Bu rapor, çalışmanın başında belirlenmiş olan hedefleri, bu hedeflere ulaşma oranını ve analiz sonuçları ile ilgili yorumları içerir. Ancak metin şeklindeki raporlar belli bir kısımda bulunan kullanıcılara hitap ederler. Raporlar, metin halinde hazırlanabildiği gibi tüm kitlelere hitap etme amacı ile görsel ve işitsel olarak da hazırlanabilir. Raporda tablolar kullanmak anlatılan konu için kolaylık sağlayacaktır.

Fakat kullanılacak tabloların fazlalığı, karmaşıklığı ve çok sayıda karmaşık sayılar içermesi dikkati dağıtabilmektedir. Bu sebeplerden dolayı tabloların grafik haline getirilmesi anlamayı kolaylaştırmak açısından önemlidir. Fakat grafik sunumlarında da önemli hususlar bulunmaktadır. Özellikle veri tipine uygun grafik seçimi önemlidir. Eğer ölçekler üzerinde değişiklik yapılmış ise seçilen grafik tipi uygun değilse sonuçları etkileyecektir. Bu nedenle açıklanacak bilginin hem standart hem de objektif bir şekilde grafikler ile sunularak yorumların ise açık ve net bir şekilde ifade edilmesi gerekir. Temel istatistiki sonuçları yorumlamak için Histogram, dal&yaprak, dilim veya kutu grafikleri kullanılabilir. Yapılan çok değişkenli analizlerde ise bahsedilen grafikler ile sonuçları yorumlamada yetersiz kalabilir. Bu nedenle veri madenciliğinde, verileri görselleştirme kavramı bilgisayar teknolojilerinin ilerlemesiyle de iki veya üç boyutlu görsel sunumları ortaya çıkarmıştır. Bu şekilde gelişen sunumlar ise yorumlamayı kolay hale getirmektedir (31).

2.8. Data Analitiğinde Yöntemler

Veri madenciliği çalışmalarında anlamlı sonuçlar alabilmenin en temel adımlardan birisi de algoritmanın seçilmesidir. Algoritma seçerken verilerin yapısı ve türünün nasıl olduğu önemlidir. Data analitiğinde yöntemler genel olarak Tanımlayıcı Modeller ve Tahmin Modelleri olarak ele alınmaktadır.

Tanımlayıcı Modeller;

- Birliktelik Analizi
- Kümeleme
- Örüntü Oluşturma başlıklarından oluşmaktadır.

Tahmin Modelleri ise Sınıflandırma ve Regresyon, Lojistik Regresyon, Zaman Serileri Analizi ve Diskriminant Analizi gibi istatistiki tahminleyici tekniklerden meydana gelmektedir (49).

2.8.1. Birliktelik Kuralı

Kategorik değişkenler arasındaki ilişki veya birliktelikleri incelemek amacı ile kullanılmaktadır. Veri setleri üzerinde gizli kalmış ve ilk aşamalarda tespiti

yapılamamış örüntülerin çıkarılmasına yardımcı olur. Ayrıca verilerin özellikleri veya türlerinde herhangi bir ek bilgiye ihtiyaç duymadan birliktelik kuralları ile kurallar oluşturulabilir (56). Birliktelik kuralları ile küçük bir veri setinden dahi birçok kural elde edilebilir. Bu sebeple kural sayısını düşürmek, geçerli ve güvenilir ilişkiler kurmak için iki kavram ortaya çıkmıştır. Bu kavram Güven (*Confidence*) Sınırı ve Destek (*Support*) Sınırı kavramlarıdır. Veri setleri içindeki verilerin kombinasyonlarının oranını Destek Sınırı verirken, elde edilen kombinasyonlar ile birlikte diğer verilen birlikte geçme olasılığını Güven Sınırı göstermektedir (57).

Birliktelik kuralı, geçmiş verilerin analizi ile bu veriler içindeki birlikteliklerin belirlenmesi ve geleceğe yönelik çalışmalar yapılmasına olanak sağlayan bir yaklaşımdır (58). Örneğin; bir alış-verişte müşterilerin süt ve peynir satın alımlarıyla birlikte yoğurt da satın aldıkları incelensin. Bu şekildeki birliktelik örüntüsünün tespitinde, örüntüde yer alacak birden fazla ürünün satın alınma durumunun birlikte bulunması gerekir. Böylece kurum veya satıcılar, elde edilen bu birliktelik bağıntıları ile etkili ve kazançlı pazarlama ve satış yöntemleri geliştirebilmektedir (59).

Birliktelik Kuralı;

$$A_1, A_2, A_3, \dots, A_m \Rightarrow B_1, B_2, B_3, \dots, B_n$$

Bu ifadede yer alan, A_i ve B_j , yapılan iş, nesne veya durumlardır. Kural, genel olarak " $A_1, A_2, A_3, \dots, A_m$ " iş, nesne veya durumları ortaya çıktığında, sık olarak " $B_1, B_2, B_3, \dots, B_n$ " iş, nesne veya Durumlarının da aynı hareket içinde bulunduğunu gösterir (60).

Kullanıcılar tarafından belirlenmiş Birliktelik Kuralı, en küçük değeri belirlenmiş olan destek ve güvenilirlik açısından eşik değerlerini sağlayacak şekilde hesaplanır. Herhangi bir ürün kümesindeki destek, D harfi ile gösterilir. Bu küme belirtilen tüm hareket ve olaylar içinde ilgili ürün kümesini içeren hareketlerin yüzdesidir. A ve B ürün kümelerinin, birliktelik kuralı " $A \Rightarrow B$ " ise destek ve güven aşağıdaki gibidir.

$$\text{Destek } (A \Rightarrow B) = (\text{A ve B'nin bulunduğu satır sayısı}) / (\text{toplam satır sayısı})$$

(2.4)

Buradaki birliktelik kuralının güven değeri ise, A'yı içeren hareketlerin B'yi de içirme yüzdesidir.

$$\text{Güven } (A \Rightarrow B) = (\text{A ve B'nin bulunduğu satır sayısı}) / (\text{A'nın bulunduğu satır sayısı}) \quad (2.5)$$

Güven değerinin % 100 olduğu durumlarda, kural bütün veri analizlerinde doğru kabul edilir. Böylece bu kurala “kesin” denir (61).

Birliktelik kuralına ilişkin birçok algoritma geliştirilmiştir (61). AIS (62), SETM (63), Apriori (64), Partition (65), RARM - Rapid Association Rule Mining (66), CHARM (67). Bu algoritmalarından ilk olarak geliştirilen AIS iken, en çok bilinen ve kullanılan algoritma ise Apriori'dir (59).

2.8.2. Kümeleme

Kümeleme, bir veri setindeki verileri sınıflandırmak ya da kümelere ayırmak için kullanılan yöntemidir (68). Kümeleme analizi data analitiği, veri madenciliği, istatistik, makine öğrenimi ve biyoloji gibi bir çok alanda kullanılmaktadır. Kümelemede, sınıflama modelinde bulunan veri sınıfları bulunmaz. Kümeleme analizinde verilerin hangi sınıflara ait olduğu bilinmez. Sınıflama modelinde ise, verilerin hangi sınıfa ait oldukları bilinmektedir. Kümelemede işleminde, hangi sınıfa ait olduğu bilinmeyen veri gruplar şeklinde kümelenirler. Kümeleme modeli bazı uygulamalarda, sınıflama yönteminin ön işlemi olarak kullanılabilir (69). Devam eden Gereç ve Yöntem kısmında konu ayrıntılı olarak anlatılacaktır.

2.8.3. Sınıflandırma

Sınıflandırma, veri madenciliği yöntemleri arasında en yaygın kullanılan yöntemlerinden biridir. Sınıflandırmada daha önceden belirlenmiş sınıflar ya da kategorilere bölünmüş hedef kategorik değişkeni belirlenir. Örneğin, hastalık şiddeti ağır-hafif, yüksek – düşük gelir, riskli müşteri-risksiz müşteri vb. şeklinde kategorilenmiştir. Sınıflandırma bankacılık, tıp, hukuk, eğitim veya ülke güvenliği gibi birçok alanda yer almaktadır. Burada ilgilenilen küme çok fazla sayıda veri ve değişken içermektedir. Bu küme kayıt kümesi (*training set*) olarak adlandırılır. Amaç bu kümede bulunmayan bir bilgiye ulaşmaktır. Bu olmayan değişkene ise hedef

değişken (target variable) denir. Sınıflamanın temel süreci, bulunan veri kayıtlarını sınıflayarak hedef değişken kategorilerine ulaşabilmektir.

Matematiksel olarak;

$D = \{t_1, t_2, t_3, \dots, t_n\}$ bir veri tabanı t_i bir kayıt,

$C = \{c_1, c_2, c_3, \dots, c_m\}$ m adet sınıf kümesi olmak üzere

$f: D \rightarrow C$ olarak tanımlansın ve $c_j = \{t_i \mid f(t_i) = c_j, 1 \leq i \leq n, t_i \in D\}$

sınıfına ait kayıtlar içerir şeklinde tanımlanmaktadır (49). (2.6)

Sınıflandırma, tespit edilmiş bir veri kümesini analiz etmek için kullanılır. Kümedeki her örneği alarak, bu örnekleri en az hata ile belirli bir sınıfa atayan bir modeldir. Veri kümesinde bulunan önemli veri sınıflarını doğru bir şekilde tanımlayan modeller çıkarmak için sınıflandırma yöntemi kullanılır. Sınıflandırmanın iki adımı bulunmaktadır. İlk olarak, veri setine sınıflandırma algoritması uygulanır ve böylece model oluşturulur. İkinci adımda ise ulaşılan model, önceden eğitilmiş doğruluğu ve performansı ölçmek için daha önce tanımlanmış bir test veri setinde test edilir (70). Sınıflandırma modellerinde başlıca kullanılan teknikler (49);

- Yapay Sinir Ağları (Artificial Neural Networks)
- Karar Ağaçları (Decision Tree)
- Genetik Algoritmalar (Genetic Algorithms)
- Regresyon (Regression)
- Bayesyen Sınıflandırma (Bayesian Classification)
- K-En Yakın Komşu (K-Nearest Neighbor)
- K-Medoid
- Vaka Tabanlı Akıl Yürütme (Case based reasoning)
- Destek Vektör Makinaları (Support Vector Machine)
- Bellek Temelli Nedenleme (Memory Based Reasoning)

2.9. Data Analitiğinde İstatistik

İstatistiksel tekniklerin verilerin indirgenip modellenebilmesi gibi temel verinin ön işleme adımlarında ve sonuçların değerlendirilerek yorum yapma aşamasında çok sayıda faydası vardır. Veri madenciliği ile istatistiğin birçok yönden oldukça yakın bir ilişkisi bulunmaktadır (71). İstatistik ile veri madenciliği ortak ve en önemli özelliği veriden bilgi öğrenilmesi (72) ya da verinin bilgiye dönüştürülmesidir (73). Hem istatistik hem de veri madenciliği yaklaşımlarının amacı verilerin anlamını çözmektir. Her ikisi de belirsizlikleri ortadan kaldırarak gelecekteki olacak olaylar hakkında bilgi vermek içindir. Ayrıca bu yaklaşımların bir diğer ilgi alanı, bir olayı etkileyecek olan en önemli etmenleri belirleyip türetilen modelleri kullanarak gelecekteki olayları en doğru ve uygun şekilde öngörmektir. İki yaklaşımın benzer yanları olduğu kadar farklı yönleri de bulunmaktadır. Burada dört temel farklılıktan özellikle bahsedilebilir (71):

Bun farklar ise; teorinin rolü, hipotez testi, güven düzeyidir ve genellenebilirliktir. İstatistik bilimi teori ile eş bir ilişki içerisindedir. İstatistiksel analizler teorinin ortaya atılması ile başlar. Teorinin kabul veya ret edilmesi ile ilgilenir. Doğası gereği istatistik süreç olarak doğrulayıcıdır. Buna rağmen veri madenciliğinin odak noktası teorilerin doğrulanması üzerine değildir. Bunun anlamı bilgisayarların otomatik olarak örüntüleri oluşturması değildir. Tam tersine, veri madenciliği açık bir şekilde talimatları analistten bekler. İstatistikte değişkenler arasında veri madenciliğine kıyasla ilişkiler hakkında daha fazla varsayımlar bulunmaktadır. İstatistik tüme varım mantığıyla ilerlerken, veri madenciliğinde tümünden gelim vardır. İstatistiksel araştırmalarda genel olarak ilgilenilen ana kitleden çekilmiş bir örneklemdir. Bu örneklem ile kitle hakkında çıkarımlar yapılmaktadır. İstatistik bireyselliğin üstesinden gelememektedir. Özneler arasındaki benzerliklerin bulunabilmesi nedeniyle tasarlanmıştır. Veri madenciliğinin amacı ise çok daha detaylı, özel ve yerel bilgilere ulaşabilmektir. Hipotezi sınamanın veri madenciliği için önemli bir yeri anlamı yoktur, çünkü veri madenciliğine bir hipotez ile veya özel bir ana kitlenin sonuçlarının genelle yayılması üzerine çalışmak için başlanılmaz. Veri madenciliğinde oldukça büyük ve çok sayıda veri kullanılmaktadır. Kısaca genel olarak ana kitle ile çalışma yapılmaktadır. Önemli bazı diğer farklılıklar ise:

- Veriler istatistiksel arařtırmalarda akla gelen belirli sorular için toplanır. Bu sorulara cevap alabilmek için veriler analiz edilir. İstatistiğin alt disiplinlerinden olan deney tasarımı ve alan arařtırması vb. disiplinler en iyi yolla veri toplamak için ipucu sağlamaktadırlar. Öte yandan veri madenciliğinde veri madenciliği uygulamak için deęil dięer hedeflenen bazı amalar için veriler kullanılmaktadır (74).
- İki yöntem arasındaki en temel ve en büyük farklılık veri kümesinin büyüklüğüdür (50). Birkaç yüz ve ya bin veri içeren bir küme istatistikçi için büyük bir veri seti iken, veri madenciliği bu sayılar, milyon veya milyarı bulabilir.
- Veri madenciliğinde veri kalitesi açısından veri kümesinin ön işlemlerden geçmesi oldukça önemlidir. Doğru çıktılarına ulaşmak açısından kaliteli veriye ulaşmak önemlidir. Fakat veri madenciliğinde yapılan analizlerde kümelerin çok sayıda veri içermesi nedeniyle verilerin ön işlemlerden geçirilmesi oldukça zaman alıcıdır. Farklı ve çeşitli istatistiksel arařtırmalar incelendiğinde, birçok durumda veriler ön işlemlerden geçirilmeksizin doğrudan analize aşamasına geçilebilmektedir. Veri madenciliğinin istatistikten ayrılmasındaki önemli özelliklerden biri ise budur.

2.9.1. İstatistiksel Yöntemler

Veri madenciliğinde gözlemlenen çeşitli problem tipleri çözümlenmektedir. Bu problem tipleri bir ya da birkaç istatistiksel teknik kullanılarak çözümlenebilmektedir. Regresyon, Korelasyon, Temel Bileşenler Analizi, Kümeleme teknikleri ve diskriminant analizi, gibi yöntemler veri madenciliğindeki problemleri çözmede çalıştırılan önemli istatistiksel yöntemlerden birkaçıdır (23).

2.9.2. Bayesci İstatistik

Olasılık tanımlarındaki farklılıklar istatistikte Frekansçı ve Bayesci olmak üzere iki farklı yaklaşımın meydana gelmesine sebep olmuştur. Bayesci yaklaşımın temelini oluşturan Bayes teoremi aşağıdaki şekilde ifade edilir (75).

Teorem 1. A ve B olayları örneklem uzayı üzerinde tanımlanmış iki olay ve $p(B) \neq 0$ olsun. Buna göre;

$$p(A/B) = \frac{p(B/A)p(A)}{p(B)}, \text{ dir.} \quad (2.7)$$

Bu ifade;

- $p(A/B)$ terimi B olayı gerçekleştiğinde A olayının gerçekleşme olasılığıdır.
- $p(B/A)$ terimi A olayı gerçekleştiğinde B olayının gerçekleşme olasılığıdır.
- $p(A)$ ve $p(B)$ olayları, A ve B olaylarının gerçekleşme olasılıklarıdır.

Bayesci çıkarımlar 3 madde ile özetlenebilir (76):

1. Olasılıksal modelin oluşturulması: Gözlenebilen ve gözlemlenemeyen tüm nicelikler için birleşik olasılık fonksiyonu (joint probability distribution) oluşturulur. Tanımlanan bu model toplanan veri ve problem ile tutarlı olmalıdır.

2. Sonsal dağılımın hesaplanması: Bayes teoremi sonsal dağılım hesaplaması için kullanılır.

3. Modelin kontrol edilmesi: Tanımlanan modelin toplanan veriye ne kadar iyi derecede uyduğu ve bunun yanında sonuçların kabul edilebilirliği kontrol edilir.

Bayesci modeller doğrusal regresyon, varyans analizi (ANOVA), kovaryans analizi (ANCOVA) gibi yöntemlerinde temelini oluşturur (76).

1990'lı yıllarda kullanılmaya başlanan Bayesci ağlar, çok boyutlu veri kümesindeki rastlantı değişkenleri arasındaki olasılıksal ilişkileri kodlayan grafiksel modellerdir. Bayesci ağlar hem nedensel hem de olasılıksal özelliklere sahiptir. Bu nedenle bu ağlar sayesinde verinin bilgisi ve uzmanın görüşü kolaylıkla birleştirilebilir. Bunların dışında bayesci ağlar ile ilgilenilen bir problemin kesin olmayan tanım bölgesi ile ilgili bilgi sunulabildiği gibi, güçlü çıkarımlar da yapılabilmektedir. Bayesci ağlardan yararlanmak istatistiksel analizlerde birçok üstünlük sağlayacaktır. Bunlardan birkaçı; olasılık kuramı temelli olmaları nedeniyle her zaman tutarlı sonuçlar vermeleri, değişkenler arasındaki nedensel ilişkilerin anlaşılmasını sağlamaları, uzman görüşünü modellemeye katmaları ve veride kayıp

veya eksik gözlem olması durumunda bile güvenilir çıkarsamalar yapmalarındır (77)(78).

Bayesci ağlar, “yönlü dönüşsüz grafik (directed acyclic graph [DAG])” olarak bilinen bir grafik yapısına sahiptir. İstatistik, makine öğrenmesi ve yapay zekâ alanlarında çok kullanılmaktadır. Sezgisel olarak anlaşılabilirler. Bu nedenle bir rastlantı değişkenleri kümesinin çok değişkenli olasılık dağılımının etkili bir sunumunun ve bu sunum üzerinden çeşitli hesaplamaların yapılmasına yardımcı olacaktır (79). Son yıllarda kesin olmayan uzman görüşlerinin sisteme girmesini sağlayan bayesci ağlar, oldukça önemli bir yöntemdir. Bayesci ağlarda tüm parametrelerin anlaşılabilir bir yorumu olabilmesi diğer olasılık grafik modellerinden ayrıldığı noktadır. Bu nedenle Bayesci ağlar, öğrenme sürecinde zaman tasarrufu sağlayarak doğrudan uzman görüşü ile oluşturulabilir. Her zaman uygun uzman görüşüne ulaşmak kolay olmadığından bu ağların oluşturulmasında veri kümeleri gerekmektedir (77)(78).

2.9.3. Temel Bileşenler Analizi

Orijinal p değişkenlerinden oluşmuş bir veri setinin, daha az sayıda değişken ve bulunan bu değişkenlerin doğrusal bileşenlerini gösteren yeni değişkenlerle ifade etme yöntemine temel bileşenler analizi denir. Aralarında korelasyon bulunan p sayıdaki değişkenin açıklamış olduğu yapıyı, aralarında korelasyon bulunmayan ve orijinal değişken sayısından daha az bir sayıda ($k < p$) orijinal değişkenlerin doğrusal bileşenlerini ifade eden k tane değişkenle ifade etme biçimidir (80).

Temel bileşenler analizi, veri setleriyle ifade edilmiş olan bilgileri daha alternatif bir şekilde ifade etmek amacıyla kullanılmaktadır. Bu analizde bağımlı değişkenler bulunmamaktadır. Bulunan veri setindeki ifadelerin tümü aynı amaca hizmet etmektedir. Yani herhangi bir ifade diğerlerini açıklaması amacıyla seçilmemektedir. Bu analiz eldeki veri setini yeniden ifade etmek amacıyla değil daha çok boyut indirgeme olarak kullanılır. Böylece boyut indirgenerek veri seti yeniden ifade edilmiş olacaktır (81). Temel bileşenler analizi, veriyi indirgemek ve şekli sıkıştırıp özellikleri çıkarmak açısından oldukça önemli bir yöntemdir. Bir veri setinin varyans-kovaryans yapısını ve bu değişkenlerin doğrusal birleşimleri ile

açıklanması, boyutların indirgenmesi ayrıca yorumlanmasını sağlayan temel bileşenler analizi çok değişkenli bir istatistik yöntemdir olarak karşımıza çıkmaktadır (82).

Temel bileşenler analizinin amacı tek bir grup şeklinde ve çok sayıda bulunan değişkenlerin boyutlarını indirgeyerek anlamlı, daha kolay açıklanabilecek ve daha az sayıda değişken elde etmektir. Bu analizde diğer bir amaç ise aralarında korelasyon bulunmayan değişken grupları oluşturmaktır. Kısaca, yüksek düzeyde korelasyon düzeyine sahip verilerden daha az sayıda ve korelasyon bulunmayan değişkenler oluşturmak ve bunları boyutlara indirmek için uygulanan bir yöntemdir (83).

Veri bilimi çalışmalarında da çok sayıda değişken ile çalışılması gerektiğinden, bu durum eğitim (training) sürelerinin oldukça uzun olması, aşırı öğrenme (overfitting) ve çoklu doğrusal bağlantı (multicollinearity) gibi sorunlara zemin hazırlamaktadır. Hazırlanan modellerin en kısa (optimum) süre ve performansla çalışması gerekmektedir. Lojistik regresyon ve lineer regresyon gibi istatistiksel yöntemlerde ortaya çıkan çoklu doğrusal bağlantı sorunu da yanıltıcı sonuçlara sebebiyet verebilmektedir.

Bu sorunları aşmak nedeniyle değişken seçimi (feature selection) ve boyut indirgeme (dimensionality reduction) yöntemleri kullanışlı olacaktır. Değişken seçimlerinde veri setlerindeki değişkenler ya tamamen korunur ya da kaldırılır. Fakat boyut indirgeme ile hali hazırda mevcut olarak bulunan değişkenlerin kombinasyonundan oluşan yeni değişkenler elde edilerek değişken sayısı azaltılır. Böylece veri setindeki tüm özellikler korunmuş olur ve aynı zamanda değişken sayısı azaltılmış olur. Analizlerde yaşanan bu problemleri aşmak için en çok kullanılan ve tercih edilen boyut indirgeme yöntemlerinden biri Temel Bileşenler Analizi'dir. Temel Bileşenler Analizi uygulanırken dikkat edilmesi gereken noktalar vardır. Farklı ölçeklerdeki veriler yanıltıcı bileşenlere sebep olabileceğinden öncelikle veriye mutlaka standardizasyon yapılmalıdır. Analiz aykırı gözlemlerden (outlier) oldukça etkileneceğinden öncelikle aykırı gözlemler temizlenmeli ve Randomized PCA, Sparse PCA gibi alternatif yöntemler kullanılarak sorunlara çözüm üretilebilir.

Temel bileşenler analizi yapılırken aşağıdaki sıralama izlenebilir.

- Her bir boyut için ortalama vektör hesaplanır.
- Kovaryans matrisleri hesaplanır.
- Her boyut için öz vektörleri ve karşılık gelen öz değerleri hesaplanır.
- Her bir öz değer için öz değerler toplamına bölünmesi ile temel bileşenlerin toplam varyansı açıklama yüzdeleri elde edilir.

Temel Bileşenler Analizi;

- Boyut indirgeme
- Verileri korelasyondan arındırma
- Yüksek boyutlu verilerin görselleştirilmesi
- Gürültü filtreleme

gibi çalışmalar için fazlaca yararlıdır (84).

2.9.4. Regresyon

İki veya daha fazla değişkenin arasındaki ilişki veya ilişkilerin matematiksel olarak belirlenmesi yöntemine regresyon denir. Tek değişkenli ve En küçük kareler yöntemiyle elde edilen doğrusal bir regresyon modeli aşağıdaki gibidir.

$$y = \alpha + bx + e \text{ Şeklinde olur.} \quad (2.8)$$

α : sabit katsayısı,

b : Doğrunun eğimini ve regresyon katsayısı,

e : Hata terimidir.

Verilere ait denklem şeklinde $y = \alpha + bx$ olmaz. Çünkü bu denklemin deterministik halini ifade eder. Fakat gerçek gözlemlerin daima stokastik bir yönü bulunur. Hata terimleri ise bu stokastik kısmı oluşturmaktadır. Denklem bize x 'te oluşacak bir birimlik değişimin y üzerinde yaratacağı değişikliği göstermektedir.

Veri madenciliğinde regresyon incelendiğinde ise , y sınıfları temsil ederken, y 'nin alacağı değerlere göre x değerlerinin hangi gruba dâhil olacağı tahminlenir. x Değişkenleri y sınıfını temsil ederler. Fakat temsil ederken oluşturacakları model her

zaman doğrusal olmadığı gibi, kuadratik ya da kübik modeller de karşılaşılabilmektedir. Üçüncü dereceden fazla üst dereceli fonksiyonlar kullanılmamaktadır. Çalışılan veri setindeki verilere bağımlı sonuçlar elde edilmektedir. Bu nedenle üst dereceli fonksiyonların kullanılmasına ihtiyaç yoktur. Tek veya çok değişkenli regresyon modelleri nümerik (sayısal) veriler ile uygulanmaktadır. Fakat verilerin ya da sınıf değişkeninin değerleri sıralı veya nominal değişkenlerden meydana geliyor ise lojistik regresyon modeli kullanılmalıdır (49).

2.10. Data Analitiğinde Çapraz Geçerlilik

Bir projede birden fazla model ile karşılaşılabılır. En doğru modeli seçmek ise oldukça güç bir işlemdir. Bu nedenle oluşturulan modelleri test edecek bazı yöntemler geliştirilmiştir. Çapraz Doğrulama, makine öğrenmesi modellemelerinde yapılan test hatasını daha iyi tahminleyebilmek için model seçiminde kullanılmaktadır. Buradaki fikir, eğitim veri setlerinden doğrulama kümelerini elde etmektir. Sınırlı sayıda veri içeren bir küme bulunmasında çapraz geçerlilik yöntemi kullanılabilir. Bazı çapraz geçerlilik yöntemleri şöyle sıralanabilir (85):

- a. Basit Geçerlilik (Simple Validation): Bir modelin test edilmesi aşamasında kullanılan en basit yöntemdir. Verilerin %5 ile %33 arasındaki kısmı, bu yöntemde test verileri olarak ayrılmaktadır. Kalan diğer kısma ise model oluşturmak için algoritmalar uygulanır. Model elde dildikten sonra daha önce ayrılan test verileri ile test işlemleri yapılır. Yanlış olarak sınıflanmış olan olay sayısının bütün olay sayısına oranı ile hata oranı elde dilmiş olurken, doğru olay sayısının bütün olaya oranlanması ile de doğruluk oranı elde dilmektedir. (Doğruluk oranı = $1 - \text{Hata oranı}$)
- b. Holdout yöntemi: Bu yöntemde ise veriler rasgele iki eşit iki parçaya ayrılmaktadır. İlk önce birinci ayrılan parça üzerinden model tespit edilir. İkinci ayrılan parça üzerinden test işlemi yapılır. Daha sonra yine ikinci parça üzerinden model tespiti yapılır ve birinci parça üzerinden tekrar test işlemi yapılarak hata oranlarının ortalaması elde edilerek kullanılır. Bu yöntem tekrar hesaplama gerektirmediğinden avantajlıdır (86).

- c. K katlamalı çapraz doğrulama: Holdout yönteminin geliştirilmiş halidir. Veri kümesi k alt kümeden oluşur (87). Çapraz geçerlilikte uygulanmış olan yöntem k parça için uygulanmaktadır (85). Bu yöntemde her defa, veri kümesinin k alt kümelerinden bir tanesi test kümesi olarak kullanılırken, kalan diğer $k-1$ alt kümeleri ise eğitim kümesi olarak kullanılmaktadır. Daha sonra tüm elde edilen k denemelerinin ortalama hatası hesaplanır. Verilerin nasıl bölündüğünü daha az önemsemesi bu yöntemin avantajıdır. Bu yöntemde her bir veri bir kez test kümesine, $k-1$ kez ise eğitim kümesine girmiş olur. k arttıkça, tahmini varyans azaltılır. Yöntemin dezavantajı ise eğitim algoritmasının en baştan k kere tekrarlanmasının gerekmesidir. Yani değerlendirme aşamasına geçmek için k kere hesap yapmak gerekecektir. Bu yöntemin farklı bir şekli ise verileri rastgele bir şekilde bir test kümesi ve eğitim kümesine ve k farklı zamanlara bölmektir. Böylece her bir test kümesinin ne boyutta büyük olduğu ve ne kadar denemenin bağımsız olduğu seçilebilir (87).
- d. Bootstrapping: Bu yöntem küçük veri kümelerinde hata düzeyini belirlerken kullanılır. Tüm veri kümesi üzerinde çalışılarak model belirlenir. En düşük 200 olmak üzere çok sayıda tekrarlı örneklemelerle öğrenim kümesi veri kümesinden elde edilir. Böylece hata oranı hesaplanmış olur (88).

2.10.1. Model Başarım Ölçütleri

Elde edilen modelin başarısını değerlendirme amaçlı kullanılan temel kavramlar karmaşıklık matrisi ile ifade edildiğinde doğruluk, kesinlik, duyarlılık ve F-ölçütleridir. Oluşturulan matriste doru sınıfa seçilen veri niceliği ve yanlış sınıfa atanan veri niceliği ile modelin başarısı ölçülebilmektedir (89).

2.10.1.1. Karmaşıklık Matrisi (Risk Matrix)

Bu yöntem modelin doğruluk derecesini görmek için kullanılmaktadır. Genellikle sınıflama problemleri için kurulan modellerinde tercih edilir. Bir örnek vermek gerekirse; aşağıda bulunan matrisin sütunları fiili (gerçek hayattaki), satırları ise tahmini değerleri gösterebilir. Örnekte B sınıfına ait olması gereken 46 elemanın,

daha sonra kurulmuş olan model ile 2'sinin A, 38'inin B, 6'sının C olarak sınıflandığı görülmektedir (90).

Tablo 2.1. Karmaşıklık Matrisi

Fiili			
Tahmini	A Sınıfı	B Sınıfı	C Sınıfı
A Sınıfı	45	2	3
B Sınıfı	10	38	2
C Sınıfı	4	6	40

Kısaca test sonuçlandığında elde edilen başarı karmaşıklık matrisi ile gösterildiğinde; matris satırları test kümesine tabi verileri ait sayıları, sütunlar ise modelin tahminlemesini göstermektedir (89).

Tablo 2.2. Karmaşıklık Matrisi Harflendirme

		Öngörülen Sınıf	
		Sınıf 1	Sınıf 2
Doğru Sınıf	Sınıf=1	a	b
	Sınıf=0	c	d

a: TP (True Pozitif) b: FN (False Negatif) c: FP (False Pozitif) d: TN (True Negatif)

2.10.1.2. Doğruluk (Accuracy)

Bir modelin başarısının ölçülmesinde en yaygın ve basit olarak kullanılan yöntem, modelin doğruluk oranıdır. Bu oran doğru biçimde sınıflandırılmış olan örneklerin sayısının (TP +TN), toplam bulunan örnek sayısına (TP+TN+FP+FN) bölünmesi ile elde edilir. Bulunan bu oranı 1'e tamlayan değer ise hata oranıdır (89).

2.10.1.3. Kesinlik (Precision)

True Pozitif örnek sayısının, sınıfı 1 olarak tahminlenmiş ve yine sınıfı 1 olarak tahminlenmiş tüm örnek sayısına bölünmesi ile kesinlik değeri elde edilir (89).

$$Kesinlik = \frac{TP}{TP+FP} \quad (2.9)$$

2.10.1.4. Duyarlılık (Recall)

Doğru olarak sınıflandırılmış olan pozitif örnek sayısının toplam pozitif örnek sayısına bölünmesi ile duyarlılık elde edilmektedir (89).

$$\text{Duyarlılık} = \frac{TP}{TP+FN} \quad (2.10)$$

2.10.1.5. F-Ölçütü

Model başarımlar ölçümlerinden olan kesinlik ve duyarlılık ölçütleri karşılaştırma yapılacağında anlamlı bir sonuç almada tek başına yeterli olmayacaktır. Bu nedenle hem kesinlik hem de duyarlılık ölçütlerinin ikisini birden değerlendirmek daha iyi ve doğru sonuçlar verecektir. F ölçütü bu şekilde tanımlanmıştır. Kesinlik ve duyarlılık ölçütlerinin harmonik ortalaması F-ölçütü değerini vermektedir (89).

$$F\text{-Ölçütü} = \frac{2 \times \text{Duyarlılık} \times \text{Kesinlik}}{\text{Duyarlılık} + \text{Kesinlik}} \quad (2.11)$$

2.10.2. Klasik Kümeleme Analizinde Küme Geçerlilik İndeksleri

Kümeleme analizinde genel olarak uygun ve doğru olan küme sayısı çoğu zaman bilinmemektedir. Küme sayısı belirlendikten sonra uygun ve doğru analizler için kümeleme geçerlilik teknikleri kullanılmalıdır. Analiz yapılırken aynı küme yapısı içindeki elemanların birbirlerine benzerliklerinin yüksek olması istenirken farklı kümelerde ise bu durumun minimum olması beklenmektedir. Kısaca küme içi varyansların minimum farklı kümeler arasındaki varyansın yüksek değerde olması gerekir (91).

2.10.2.1. Silhouette İndeksi

Silhouette indeksi 1987 yılında Rousseeuw tarafından her bir elamanın atandığı kümeye uygun olup olmadığını test etmek amacı ile oluşturulmuştur.

a_i ; i. elemanın bulunduğu kümedeki tüm noktalara bulunan ortalama benzerliği (uzaklığı)

b_i ; i. elemanın diğer bulunan kümelerdeki tüm noktalara olan ortalama benzerliğinin (uzaklığını) minimumunu göstermek üzere,

Buradan i. birim Silhouette indeksi için;

$$sil_c = \frac{b_i - a_i}{\max((a_i), (b_i))} \quad (2.12)$$

Olarak hesaplanır. sil_i Değeri 1'e yaklaştığında i. birimin atandığı kümeye daha uygun olduğu, eğer sil_i değeri 0'e yaklaşıyor ya da negatif değer alıyorsa i. birimin atandığı kümeye uygun olmadığı söylenebilir (92). Negatif değer oluşuyorsa sadece bir birim en uygun kümesine atanmamıştır. Tüm kümelemenin geçerliliği (kalitesi) için ölçü olan ortalama Silhouette değeri;

$$sil_c = \frac{1}{n} \sum_{s_j \in S} sil_c \text{ şeklindedir.} \quad (2.13)$$

Bu formüle göre uygun küme sayısı maksimum ortalama Silhouette değerine ulaşılandır (93).

2.10.2.2. Akaike (AIC) ve Bayesian (BIC) Bilgi Kriterleri

AIC Akaike Bilgi Kriterleri ve BIC Bayesian Bilgi Kriterleri anlamına gelir. Bu iki terim model seçiminde kullanılmaktadır. İki yaklaşım arasındaki farklar bulunmaktadır. Bu farklar;

- AIC ve BIC karşılaştırıldığında, ek parametreler için risk BIC'de AIC'ye göre daha fazladır.
- AIC genellikle yüksek boyutsal gerçekliğe sahip bilinmeyen bir modele ulaşmayı sağlar. Diğer yandan BIC sadece gerçek modellerde ulaşılan bir değerdir.
- AIC'ye göre BIC daha tutarlı sonuçlar verir.
- AIC, asimptotik olarak çapraz geçerliliğe eşdeğer hale getirmek için iyidir. Tam tersi olarak BIC tutarlı tahmin için iyidir.
- BIC, AIC ile karşılaştırıldığında daha uygun görülmesine rağmen, sayı arttığında daha düşük düzeyde tolerans gösterir.

2.11. Sağlık Alanında Veri Bilimi

Sağlık hizmetlerinde oldukça büyük miktarlarda veri üretilmekte ve entegrasyonu sağlanmaktadır. Bu gelişme özellikle hastalar başta olmak üzere sağlık sistemleri üzerinde büyük gelişmelere yardımcı olmaktadır (94). Fakat sağlık verilerinin çok büyük miktarlara ulaşması veri işleme konusunda klasik yöntemleri yetersiz kılmaktadır. Böylece büyük veri kavramı sağlık hizmetleri alanına girmiştir (95). Sağlık alanı ve sağlık sistemlerinin başarısını artırmak sebebiyle elektronik olarak elde edilen, saklanan otomatik veya rutin olarak kayıt altında tutulan büyük veri kümeleri sağlıkta büyük veri kavramını tanımlar. Ayrıca büyük veriler çok amaçlıdır. Yeniden kullanılabilirler ve birleştirilmiş veri tabanları ya da bağlantıları içerirler (96). Gözlemlenebilir bir olgu hakkında bilgi elde edilmesi aşamasını büyük sağlık verileri kolaylaştırır ve bu veriler benzeri olmayan özelliklere sahip karmaşık veri setleridir. Büyük Sağlık Verilerinde tipik olarak değişik kaynaklardan türetilmiş eksik, heterojen, çoklu-spektral, ve kesin olmayan gözlemlere (örn. hastalıklar, yaralanma, fiziksel veya zihinsel bozukluklar, bu hastalıklara konulan tanılar, tedaviler, hastalıkların önlenmesi) dair veriler vardır (97). Sağlık sektöründe büyük veri büyük imkânlar sunmaktadır. Özellikle ilaç endüstrisi, tıbbi sigorta ve reçeteler konusunda bilgi sahibi olma imkânı vermektedir. Bunların dışında hasta tarafından elde edilmiş veriler, fitness izleme cihazları, kilo tartım terazileri gibi IoT (Internet of Things- Nesnelerin İnterneti) cihazları, kan basıncı monitörleri sayesinde bir kişinin özellikleri ya da günlük yaşamı hakkında bilgi sahibi olunabilir. İnsanların yaşam standartlarını iyileştirmek, ya da kronik hastalıklar, ilaç reaksiyonları ile hastaneye yatış vaka sayılarını düşürmek adına EMR verileri, laboratuvar, ilaç bilgileri, yaşamsal veriler, doktorların notları ve hastalar ile ilgili günlük veya taburculuk belgeleri büyük veriler olarak çok önemli bir alana sahiptir (98).

2.11.1. Kullanım Alanları

Büyük verinin sağlık alanında uygulanması birçok ve büyük avantajlar sağlamaktadır. Bu avantajlar;

- Hastalık tanılarının erken evrelerde teşhis edilebilmesi ile tedavilerin şansı artırılabilir.

- Bireylerin sađlık durumları izlenilerek kontrol sađlanabilir. Bylece toplum ve aile sađlıđı merkezleri aktif bir řekilde kullanılabilir.
- Hastalık gemiřleri iin hastaların verileri kullanılır. Bu sayede bir hastalık meydana geldiđinde hangi tedavi uygulanabilir ya da uygulanan bu tedavinin hangi hastalık ya da hasta iin daha uygun olduđu belirlenebilir.
- Daha hızlı ve kolay bir řekilde maliyet analizleri yapılabilir.
- Bulunan hasta kayıtları ve klinik deney verilerinin analizi ile rnler pazara srlmeden nce oluřabilecek endikasyon ve yan etkiler daha iyi bir řekilde tanımlanabilir.
- Klinik alıřmalarda hasta toplama ařamasında daha kontroll alıřmak ve deney tasarımı daha iyi bir řekilde yapmak iin deneysel hataları azaltarak pazara verilecek rnn bu srecini hızlandırılabilir.
- Gelecekte meydana gelebilecek hastalık tahmin edilebilir. Ayrıca bu tahminlere gre salgın ya da bulařıcı hastalıkları nleyici ařı, ila gibi tedavi yntemlerinin erken tedbir olarak alınmasını sađlayabilir.
- Klinik, operasyonel ve genomik veriler bir araya getirilerek hastalık riski belirlenebilir. Bylece bu risklere karřın en uygun tedavi yntemleri iin alıřmalar yapılır.
- Reete vb. diđer hizmetlerin kt amalı kullanımının nne geilmesi iin faydalanılabilir (99).

2.12. Biyokimya ve İmmnoloji

Canlı dokuları meydana getiren maddeler ve bu maddelerin fonksiyonlarını inceleyen bilim dalına “Biyokimya” denir. En basitinden geliřmiřine her bir canlının ve trlerinin bir kimyası vardır. Bu kimya ise biyokimyanın konusudur. Canlıların dokularından ayrıştırılarak elde edilen her bir maddenin eřitli grevli vardır. Her bir maddenin iřlevini yerine getirmesiyle canlı bir denge iindedir. Bu dengede oluřabilecek en kk bir bozulma ardından gelebilecek pek ok sorunu dođurabilir. Bunun yanı sıra bazen herhangi bir sorundan da bu denge bozulabilir. Biyokimya

tahlilleri bir kan tahlili çeşididir. Vücutta bulunan madde dengesizliğini, bu dengesizliğin neler ifade edebileceğini ortaya koymak için geliştirilmiştir.

Vücutta meydana gelen her türlü işlevin yürütülmesinden kan dokusu sorumludur. Her türlü kimyayı içerisinde bulundurduğundan vücudun durumu hakkında önemli ipuçları verebilmektedir. Bu nedenle en önemli kan tahlillerinden biri biyokimya tahlilleridir. Bu testler sayesinde yaygın olarak görülen birçok hastalık ortaya çıkmaktadır. Genel olarak hormonal (endokrin) ve metabolik rahatsızlıkların değerlendirilmeye alınmasında başvurulur. Ayrıca bu yaşamsal maddelerin eksikliği ya da fazlalığı da vücudumuzdaki organların ne şekilde ve nasıl çalıştığı hakkında hekimlere fikir sunarak yardımcı olmaktadır. Böylece farklı organların kan kimyası üzerindeki etkiler görülebilmektedir. Biyokimya testleri biyokimya laboratuvarında yapılır ve bu testler; rutin biyokimya, idrar, tam kan sayımı, ilaç düzeyleri, hormon ve tümör belirteçleri ile ilgili testlerden oluşur (100).

İmmünoloji, vücudun bağışıklık sistemini inceleyen tıp dalıdır. Bağışıklık sistemi bazı şekillerde savunma yaparak bireyleri enfeksiyonlardan korumaktadır. Bağışıklık sisteminde önemli sorunlar oluştuğunda ya da bağışıklık sisteminde konak hücre dokuları korumak yerine zarar verecek durumlar oluşturduğunda birçok hastalık meydana gelebilir.

İmmünoloji tıbbın oldukça geniş bir dalıdır. Ayrıca immünoterapi, immünojenetik ve evrimsel immünoloji gibi birçok alt dala sahiptir. Farklı disiplinlerde de immünojenetik yönler bulunabilir ya da immünojenetik bulgular kullanılabilir. Eğer bağışıklık sistemi olması gerektiği şekilde çalışmıyorsa, bireylerde otoimmünite, alerji ve kanser gibi hastalıklar ortaya çıkabilir. Bağışıklık sisteminin bozulması bunların dışında, Alzheimer gibi metabolik, kardiyovasküler ve nörodejeneratif koşullar ve immünojenetik kategorisine dâhil edilmeyen birçok rahatsızlığın gelişimine etki edebilir. Astım, Kronik Öksürük, Besin, böcek ya da ilaç alerjileri, immün sistem hastalıklarına verilebilecek örneklerdendir (101).

2.12.1. Diş ve Çene Cerrahisi

Diş hekimliğinin ağız ve çene bölgesinde yumuşak ve sert dokularda fonksiyonel ve estetik olarak gözlenen hastalık, yaralanma veya bozuklukların tedavilerinin tümünü kapsayan uzmanlık alanı Ağız, Diş ve Çene Cerrahisidir (102).

- Normal ve cerrahi diş çekimleri,
- İmplant (Dental ve zigomatik) uygulamaları,
- Çene tümör veya kistlerinin tanısını koyulması ve tedavileri,
- Preprotetik cerrahi operasyonları,
- Sinüs tabanı yükseltme tedavisi işlemleri,
- Kemik ogmentasyonları,
- Odontojenik enfeksiyonların tedavisi,
- Tükürük bezi rahatsızlıklarının tedavileri,
- Travmatik diş ve destek doku yaralanmalarının tanılarının konularak tedavi edilmesi ve uzun dönemde takip edilmesi,
- Çenene meydana gelen kırıkların konservatif ve cerrahi tedavileri,
- Ortognatik cerrahi operasyonlar,
- Çene ekleminde oluşan hastalıklarının tedavileri (artroskopi, fizik tedavi, botoks uygulamaları ve artrosentez) gibi işlemler diş ve çene cerrahisinin tedavi uygulamalarını kapsar. Bu medikal tedaviler genel ve lokal anestezi ile gerçekleştirilir. Bunların dışında sedasyon uygulamaları ile belirli hasta gruplarına diğer disiplinlerin işbirliği ile tüm dental tedaviler sadece cerrahi olarak değil rutin olarak da yerine getirilebilmektedir (103).

3. GEREÇ VE YÖNTEM

3.1. Araştırmanın Evreni

Çalışmamızın evrenini Süleyman Demirel Üniversitesi Diş Hekimliği Fakültesi Ağız, Diş ve Çene Cerrahisi bölümüne başvuran 5148 hasta oluşturmaktadır. Hastalar cerrahi öncesi rutin kan verme işlemini yerine getirmektedirler. Operasyon sırasında uygulanacak olan anestezi ve girişimsel işlemler için hemogram ve biyokimya değerlerinin bilinmesi gerekmektedir. Hastalara ait veriler Diş Hekimliği Laboratuvarında yapılan tahlil sonuçlarına ait olup, 21.10.2016 ile 12.06.2020 tarihleri arasında yaklaşık 5 yıllık sürede toplanmıştır.

3.2. Araştırmanın Örneklemi

Çalışma konumuz büyük veri kümelerinde veri bilimi algoritmalarının uygulanması olduğundan herhangi bir güç analizi gerçekleştirilmemiş ve örneklem değeri belirlenmemiştir. Büyük veri analizi, doğası gereği büyük veri kümelerinde ve herhangi bir örneklem çalışması yapılmadan tasarlanır. Yalnızca, analiz içerisinde test ve eğitim veri grupları oluşturularak tutarlılık değerleri elde edilebilir.

3.3. Araştırmanın Türü

Tez çalışması, retrospektif olarak alınan bilgilerden oluşturulan veri kümesi üzerinde büyük veri analizi tarzında hazırlanmıştır.

3.4. Araştırmanın Etiği

Çalışma için Süleyman Demirel Üniversitesi Diş Hekimliği Fakültesinden yazılı izin onayı ve Tıp Fakültesi İlaç ve Tıbbi Cihaz Dışı Araştırmalara Etik Kurulundan etik onayı alınmıştır. Etik onayına ait tarih 15.04.2020 ve sayı numarası 72867572.050.01.04-52568 olarak kayıtlıdır. Çalışma, Helsinki Deklarasyonuna uygun olarak hastalara ait kişisel verilere yer verilmeden tamamlanmıştır.

3.5. Araştırmanın Yöntemi

Büyük veri analitiği, apriori olarak hipoteze sahip olmayan ve klasik istatistiksel yöntemlerden farklı olarak veri kümesinden faydalı bilgilerin ortaya çıkarılması esasına dayanmaktadır. Büyük veri analitiği makine öğrenmesi algoritmaları kullanılarak gerçekleştirilir. Genel olarak denetimli, denetimsiz ve karma yöntemlere yönelik algoritmalar kullanılmaktadır. Bu çalışmada, denetimsiz öğrenme algoritmalarından en yaygın kullanılan kümeleme algoritmaları kullanılmıştır. Farklı algoritmalar kullanılarak en uygun küme yapısı hakkında öneriler yapılacaktır. Kümeleme yöntemleri denetimsiz öğrenme ile kümeleri oluşturur. Ancak algoritmalar arasında kullanılan iterasyon ve kullanılan ölçüler açısından fark vardır. Kümeleme yöntemlerinde uzaklık ölçüleri kullanılarak küme merkezine olan uzaklıklar her bir değişken için optimum düzeye getirilir ve algoritma sonlandırılır. Algoritmaların ilk kullanıldığı zamanlarda küme sayısı kullanıcı tarafından belirlenirdi. Bu durum, kümeleme yöntemlerinin negatif yönü olarak ifade edilirdi. Ancak, yeni geliştirilen yöntemler ile optimal küme sayısı belirleme aşaması kullanılan yazılıma bırakılmaktadır. Çalışmamızda JASP, SPSS ve Python uygulamaları ile üç farklı kümeleme tekniğine yer verilmiştir. Bunlar K-ortalamlar (K-Means), Bulanık C-ortalamlar (Fuzzy C-Means) ve TwoStep Cluster teknikleridir. Çalışmaya alınan veri kümesi biyokimya, hemogram ve immünolojik olmak üzere üç farklı alt grupta düzenlenerek her bir veri grubu için farklı yöntemler ile kümeler oluşturulmuştur. Küme geçerlilik indeksi olarak Silhouette ölçütü, AIC ve BIC değerleri kullanılırken, açıklayıcılık katsayısı olarak ise R^2 ölçütü ile küme yapıları incelenmiştir. Uzaklık ölçüsü olarak ise tüm yöntemlerde Öklit uzaklığı kullanılmıştır. SPSS Uygulaması TwoStep Cluster ve k-means tekniklerinde oluşturulacak küme sayısının manuel belirlenmesi doğrultusunda AIC başarımlarını ölçütünü kullansa dahi değerini vermemektedir. Ancak belirli bir iterasyon sayısı ile küme sayısının otomatik olarak oluşturulmasına izin verildiğinde uygulama AIC değerlerini oluşan küme sayısı bazında vermektedir. JASP Uygulaması ile Fuzzy C-means ve k-means, SPSS ile TwoStep Cluster ve k-means teknikleri kullanılarak sonuçlar değerlendirilmiştir. Genel olarak k-means tekniğinin daha iyi sonuçlar vermesi nedeniyle k-means yöntemi seçilerek Python Uygulamasında Matplotlib ve Sckit Learn kütüphaneleri aracılığıyla k-means tekniği ile analizler yapılmıştır.

3.6. Verilerin Hazırlanması

Çalışmamızda kullanılacak veri kümesi Diş Hekimliği Fakültesi Ağız, Diş ve Çene Cerrahisi bölümüne başvuran hastaların kan tahlili sonuçlarından elde edilen değerleri içermektedir. Operasyon öncesi alınan kan örnekleri laboratuvarında değerlendirilir. Bu bilgiler, veri tabanı yönetim sistemi tarafından işlenir ve sorguya hazır hale getirilir. Veri tabanı taramasında geriye dönük 21.10.2016 tarihinde girişi yapılan kayda ulaşılmıştır. Bu tarihten sonraki beş yıllık dönemi kapsayan tüm kayıtlar sistemden alınarak MS Office Excel ortamına aktarılmıştır. Veri kümesi demografik özellikler, hemogram ölçümleri, rutin biyokimya ve mikrobiyolojik viral değerler şeklinde gruplandırılmıştır. Demografik özellik olarak hastaların adı-soyadı, dosya numarası, doğum tarihi bilgileri mevcuttu. Cinsiyet bilgisi mevcut değildi, ancak veri sayısı fazla olduğundan isimlerden yola çıkarak cinsiyet bilgisi oluşturma imkânı olmadığından cinsiyet bilgisi veri kümesinde yer alamadı. Hastaların başvuru tarihlerinden başvuru yapılan ay bilgisi hesaplandı ancak kümeleme analizleri içerisinde değerlendirilmedi. Kan tahlili sonuçlarına dayanan bilgiler ise üç alt grupta değerlendirildi: hemogram (kan değerleri), biyokimya ve mikrobiyolojik viral testler. Bu alt grup tahlil sonuçlarının içerdiği değişkenler listesi Tablo 3.1.'de sunulmuştur. Değişkenlerin tamamı sürekli sayısal veri tiplerinden oluşmaktadır.

Tablo 3.1. Veri Kümesini Oluşturan Değişkenler

Hemogram		Biyokimya	
BASO#	Bazofil Sayı	HDL	Yüksek Yoğunluklu Yağ
BASO%	Bazofil Yüzde Oranı	LDL	Düşük Yoğunluklu Yağ
EO#	Eozonofil Sayı	TRİGLİSERİD	Trigliserid
EO%	Eozonofil Yüzde Oranı	KOLESTROL	Toplam Kolesterol
HCT	Hematokrit	AST SGOT	Aspartat Aminotransferaz
HGB	Hemoglobin	ALT SGPT	Alanin Aminotransferaz
LYMPH#	Lenfosit Sayı	SODYUM	Sodyum, Na
LYMPH%	Lenfosit Yüzde Oranı	GLUKOZ	Şeker
MCH	Ortalama Eritrosit Hemoglobin	KLOR	Klor, Cl
MCHC	Ortalama Eritrosit Hemoglobin Konsantrasyonu	KALSİYUM	Kalsiyum, Ca
MCV	Ortalama Eritrosit Hacmi	POTASYUM	Potasyum, K
MONO#	Monosit Sayı	ÜRE	Üre
MONO%	Monosit Yüzde Oranı	BUN	Kan Üre Azotu
MPV	Ortalama Trombosit Hacmi	KREATİNİN	Kreatinin
NEUT#	Nötrofil Sayı	PT	Protrombin Zamanı
NEUT%	Nötrofil Yüzde Oranı	A-PTT	Aktif Parsiyel Tromboplastin Zamanı
PLCR	Büyük Hücreli Trombosit Oranı	TSH	Tiroid Uyarıcı Hormon
PCT	Prokalsitonin	SERBEST T3	Tiroid 3
PDW	Trombosit Dağılım Genişliği	SERBEST T4	Tiroid 4
PLT	Platelet	Mikrobiyoloji	
RBC	Alyuvar	Anti HBS	Anti Hepatit B Virüs
RDWCV	Alyuvar Dağılım Genişliği Varyasyon Katsayısı	ANTI HCV	Anti Hepatit C Virüs
RDWSD	Alyuvar Dağılım Genişliği Standart Sapma	ANTI HIV	Anti İnsan Bağışıklık Virüsü
WBC	Akyuvar	HBSAG	Hepatit B Yüzey Antijeni

Tablo 3.2. Veri Kümesini Oluşturan Değişkenlerin Referans Değerleri

Hemogram	Referans Değer		Biyokimya	Referans Değer
BASO#	100-200 mm ³ µm		HDL	Kadın:> 40 mg/Dl Erkek: > 50 mg/Dl
BASO%	% 0.4 - % 1		LDL	<130 mg/dl
EO#	<500 µm		TRİGLİSERİD	150 - 199 mg/dL
EO%	% 2.8 0 - %7		KOLESTROL	200 - 239 mg/dl
HCT	%35 - %45		AST SGOT	Kadın: 9 – 32 birim/L Erkek: 10 – 40 birim/L
HGB	Yenidoğan: 13,5–24 gr/dL 0-3 Yaş: 11–13 gr/dL 3-11 Yaş: 12–14 gr/dL Yetişkin Erkek: 13,5–17,5 gr/dL Yetişkin Kadın: 12,5–15,5 gr/dL		ALT SGPT	Kadın: 7-35 U/L Erkek:10-40 U/L
LYMPH#	Çocuk: 3000-9500 mcl Yetişkin birey: 1000-4800 mcl		SODYUM	135-145 mEq/L
LYMPH%	% 10.0 - % 48.0		GLUKOZ	Açlık: 70-100 mg/ dl Tokluk: < 140 mg/ dl
MCH	27,5 – 33,2 pg		KLOR	Yenidoğan: 96-113 mmol/ L Yetişkin Birey: 97-107 mmol/ L
MCHC	Yenidoğan: 30 - 36 g/dl 0-3Yaş: 31 - 37 g/dl Yetişkin birey: 33 - 36 g/dl		KALSİYUM	8.5 - 1.03 mg/dL
MCV	80-100 fl		POTASYUM	3 - 5 mEq/L
MONO#	%0 - %9		ÜRE	10 - 40 mg/dL
MONO%	%3.0 - %12.0		BUN	5 - 20 mg/dL
MPV	7-11 fl		KREATİNİN	Kadın: 0.50 - 1.30 mg/Dl Erkek: 0.50 - 1.40 mg/dL
NEUT#	Çocuk: 1000 - 8000 Yetişkin Birey: 1500 - 8000 mcl		PT	10 – 14 Inr, 08 – 1.1 saniye
NEUT%	%45.0 - %78.0		A-PTT	26 - 35 saniye
PLCR	%44 - %56		TSH	0.35 - 4.5 mU/ml
PCT	< 0.05 ng/ml		SERBEST T3	2.5 - 5 pg/ml
PDW	%44 - %56		SERBEST T4	0.6-1 ng/dl
PLT	150 bin - 450 bin mkl		Mikrobiyoloji	Referans Değer
RBC	1 mikro litre başına 4,2 ila 5,4 milyon adet		Anti HBS	Aşısız: < 5 mIU/ml (negatif) Aşılı: ≥12 mIU/ml (pozitif) 100-1000 değerler virüse karşı bağışıklık gösterir
RDWCV	% 11.6 - %14.6		ANTİ HCV	0 - 1 IU/ml
RDWSD	29-46 fL		ANTİ HIV	0,9 ila 1
WBC	%3.8 - %10.0		HBSAG	<1

Veri tabanı sisteminde bulunan testlerin her birisi için ayrı dosya indirildi. Her test her hastaya uygulanmış değildi. Bazı hastaların birden fazla kere tahlil sonuçları alınmıştı. Hemogram testi en yaygın yapılan tahlil olduğundan en fazla sayıda kayıt değeri bu teste aitti. Mikrobiyolojik viral testlerin uygulandığı hasta sayıları daha azdı. Toplam 6487 hastaya ait kayıt oluşturuldu. Aynı hastalara ait tekrarlı kayıtlar temizlenerek toplam 5200 kayıt elde edildi. Veri tabanında toplam 46 adet test tipi mevcuttu. Tüm test dosyaları hastaların dosya numarası ortak bilgisi kullanılarak birleştirilerek toplam 5148 kayıt elde edildi. Tüm analizler bu kayıtlar üzerinden tamamlandı. Mikrobiyolojik ve biyokimya ölçümlerinden eksik olan gözlemler için doğrusal interpolasyon yöntemi ile elde edilen değerlerin imputasyonu gerçekleştirildi. Böylece tüm değişkenler 5148 kayıt içerir hale getirildi.

3.7. Kümeleme Analizi

Kümeleme analizi çok değişkenli istatistik tekniklerinden biridir. Gruplandırılmamış ve grup sayısı bilinmeyen verilen benzer özelliklerine göre sınıflandırma işlemi için kullanılmaktadır. Bu yöntem verilerin birimleri ya da birbirlerine olan benzerliklerine göre ayrı kümelerde toplanmasını sağlayan bir yöntemdir. Kümeleme analizi benzer verilerin bir gruba atanması yönünden diskriminant analizi ile birbirine benzer değişkenlerin aynı gruplarda yer alması ile de faktör analizi ile benzerlik göstermektedir. Ayrıca veri indirgeme özelliğine de sahiptir (104). Diğer birçok çok değişkenli istatistik analizlerinde verilerin normalliği oldukça önemliyken kümeleme analizinde durum farklıdır. Burada uzaklık değerlerinin normalliğine bakılır (105). Belirlenen bir amaca göre kümeleme işlemi, iki gözlem ya da değişkenin yakınlık veya uzaklık ölçülerine bakılarak oluşturulur. Bu analizin başlıca varsayımları, verilerin matrislerinin analizden önce tahmin ve kriter değişkenleri matrislerine bölünmesidir. Ayrıca bu verilerin homojen ya da kısmen heterojen olmasıdır (106).

3.7.1. Kümelemenin Amacı

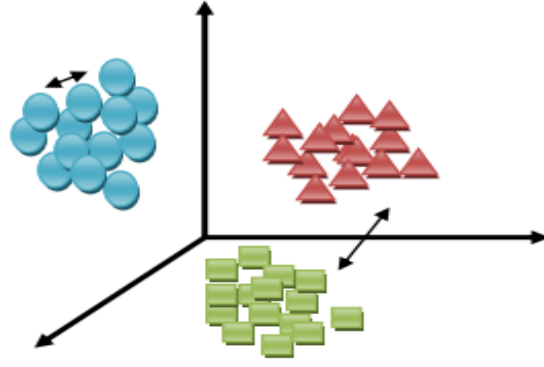
Kümele analizinin amaçları kullanım alanı ve kullanıcının amaçlarına göre değişir. Buna göre kümeleme analizinin amaçları şu şekilde sıralanabilir (107):

- Doğru tiplerin belirlenmesi
- Model oluşturmak
- Gruplara dayalı tahmin yapma
- Hipotez oluşturma ve test etme
- Veri araştırması ve inceleme
- Verileri indirgeme işlemi

Veri madenciliğinde daha gruplar oluşturulmadan pek çok alanda kontrol edilemeyecek büyüklükte büyük hacimli veriler gözlemlenebilir. Bu tür verileri indirgemedi kümeleme analizi oldukça yararlı olacaktır. Pazarlama araştırmaları üzerinden bir örnek verecek olursak; birçok şehir Pazar testi için kullanılabilir. Ekonomik faktörlere bağlı olarak bu şehir sayılarını azaltmak gerekebilir. Bu nedenle bu şehirler arasından birbirlerine en çok benzeyenler gruplara ayrılarak kümelenebilir. Böylece her gruptaki bir şehir test Pazar olarak seçilerek işlem yapılır (108). Verilerin yapısı, bu yapılar ile ilgili hipotezler oluşturma ve daha önceden oluşturulmuş hipotezleri analiz etme açısından kümeleme analizi faydalı olabilir. (104).

3.7.2. Kümelemenin Analizinin Temel Mantığı

Kümeleme analizi bir birim ve bu birimleri ilgilendiren değişkenlerin sınıflandırmalarına ait kesin bilgiler elde edilemediğinde, bir popülasyondan alınmış n tane birimin, p tane değişkene ilişkin gözlemlerinin sonuçlarıyla ilgilenir. Kümeleme de homojen olan ve birbirine benzeyen nesneler gruplar altına toplanarak heterojen gruplar elde edilir. Böylece birimler hiyerarşik bir şekilde düzenlenir. Veriler bu şekilde sınıflandırıldığında gözlem sonuçlarında daha az kayıpla veriler bir araya toplanacaktır. Kümeleme analizi, veri ya da gözlemleri küme içerisinde çok benzer fakat kümeler arasında farklı olacak şekilde sınıflandırır. Kümeleme başarılı bir şekilde yapıldığında, geometrik olarak ifade edilirse gözlemler küme içinde birbirine çok yakın kümeler ise uzak olacaktır (109).



Şekil 3.1. Küme İçi ve Kümeler Arası Uzaklıklar

3.7.3. Benzerlik ve Uzaklık

Verilerin nasıl ve hangi kümeye ayrılacağı, kaç farklı küme oluşturulacağı bu verilerin birbirine olan benzerliğine ve mesafelerine göre belirlenmektedir. Burada verilerin benzerliği aslında verilerin aralarındaki uzaklığı ifade etmektedir. Bu veri tabanında bulunmakta olan her bir verinin diğer bir veri ile olan benzerliği veya diğer verilere uzaklığı olduğu gibi oluşturulan gerçek ve aday kümeler arasındaki benzerlikleri ve uzaklığını içermektedir. Uzaklık Ölçüleri arasında en çok kullanılanlar (110);

- Öklit,

$$d_2(x_i, x_{ij}) = \left[\sum_{k=1}^p |x_{ik} - x_{jk}|^2 \right]^{1/2} \quad (3.1)$$

- Minkowski,

$$d_\lambda(x_i, x_{ij}) = \left[\sum_{k=1}^p |x_{ik} - x_{jk}|^\lambda \right]^{1/\lambda} ; \lambda \geq 1 \text{ için} \quad (3.2)$$

- Manhattan;

$$d_\lambda(x_i, x_{ij}) = \sum_{k=1}^p |x_{ik} - x_{jk}| \quad (3.3)$$

Benzerlik ölçütleri ise;

- Jaccard,
- Dice,
- Cosine,
- Overlap'dir.

Değişken tipleri iki kategoride kesikli ve sürekli olarak sınıflandırılmıştır. Kesikli değişken; bir değişkenin aldığı değerlerin sayısının sayılabilir yani sonlu olmasıdır. Eğer değişken birden fazla aralıkta her değeri alabiliyor ise sürekli değişken olarak ifade edilir. Küme analizinde bu değişkenlerin ölçek türleri önemlidir. Ölçüm düzeyleri isimsel, oransal, aralıklı ve sıralı olmak üzere Stevens tarafından dörde ayrılmıştır (111). Bir değişken oransal ya da aralıklı ölçekler ile tespit edilmiş değerler ise uzaklık ya da ilişki türü ölçümlerden faydalanılır. Gözlemler ikili olarak ölçülmüşlerse bu gözlemler arasındaki benzerlik ve farklılık ölçülerinden faydalanılır. Benzerlik düzeyleri (similarity, sim) matrisi ile gösterilir. sim matrisinin elemanları sim_{ij} şeklinde gösterilmektedir ve şu şekilde hesaplanır:

$$sim_{ij} = 100(1 - d_{ij}/\max(d_{ij})) \quad (3.4)$$

Birimlerin birbirlerine göre olan farklılıkları ise (dissimilarity, diss) sim matrisinden faydalanılarak hesaplanır. Diss matrisinin elemanları $diss_{ij}$ ile ifade edilir ve şekilde hesaplanır (112).

$$diss_{ij} = 100 - sim_{ij} \quad (3.5)$$

3.8. Kümeleme Analizinin Temel Yöntemleri

Kümeleme yöntemleri iki ana gruba ayrılmaktadır. Bu yöntemler Hiyerarşik Kümele Yöntemleri ve Hiyerarşik Olmayan Kümeleme Yöntemleridir. Yöntemlerden en çok kullanılan ise hiyerarşik yöntemlerdir (113).

3.8.1. Hiyerarşik Kümeleme

Hiyerarşik kümeleme yöntemleri birimleri değişik aşamalarda birbirleri ile bir araya getirir. Daha sonra ise ardışık şekilde kümeler belirlenir (114). Kümelere dâhil olacak elemanların hangi benzerlik düzeyinde kümenin elemanı olduğunu belirler (115). Hiyerarşik kümeleme yığılmalı ve bölünmeli hiyerarşik kümeleme olarak iki gruba ayrılmaktadır. Veride bulunan her bir gözlemin küme olarak düşünülüp ve birleştirme işlemleri yapılarak bir küme elde edinceye kadar devam eden yöntem yığılmalı hiyerarşik kümeleme yöntemidir (116). Bölünmeli hiyerarşik kümeleme de ise başlangıçta bütün veriler bir küme olarak ele alınır ve birimler aşamalı bir şekilde farklı kümelere ayrılır. Başlangıçta tüm birimlerin bir küme oluşturduğu kabul

edilerek, birimleri aşamalı olarak kümelere ayırır (112). Hiyerarşik kümeleme analizlerinde en etkili görselleştirme teknikleri dendrogramlardır. Dendrogram, hiyerarşik kümeleme yöntemleri kullanılarak elde edilecek sonuçların kolay bir biçimde anlaşılmasını sağlamaktadır (105). Hiyerarşik Kümeleme de yığılmalı yöntemler ise şunlardır:

3.8.1.1. Tek Bağlantılı Kümele Tekniği (Single Linkage - SLINK)

Farklı iki kümede en yakın gözlemlere göre belirlenen iki küme arası mesafeyi ifade eder. Aynı zamanda bu yöntem en yakın komşu olarak da adlandırılmıştır (117). En kısa uzaklık esasına dayanmaktadır. Uzaklıklar matrisi kullanılarak birbirlerine en yakın kümeler birleştirilir. Bu işlem art arda tekrar edilir (115).

Tek bağlantı kümeleme yöntemi gitgide uzayan kümeler oluşturmaya yatkındır. Sapmış olan aykırı değerlerden çok fazla etkilenir. Bu nedenle Birbirlerinden çok fazla farklı özellikler gösteren iki küme birleştirilebilir. Tek Bağlantı yönteminden iyi sonuçlar alabilmek için kümelerin birbirlerinden oldukça ayrık olması gerekmektedir (118).

3.8.1.2. Tam Bağlantılı Kümele Tekniği (Complete Linkage - CLINK)

Bu teknik en uzak komşuluk olarak bilinmektedir. Tek bağlantı yöntemine çok benzemektedir. Fakat burada iki küme arsında uzaklık olarak alınan mesafe her kümede bulunan eleman çiftlerinin arasındaki en büyük uzaklıktır (116). Bu yöntem özellikle belirgin olmayan, yoğun ve küçük küme yapılarında bulunan verilerde etkili olan bir yöntemdir (107).

3.8.1.3. Ortalama Bağlantılı Küme Tekniği

Ortalama bağlantı yöntemi ile ayrı ayrı gruplarda bulunan eleman çiftleri arasındaki ortalama uzaklık olarak iki küme arasındaki uzaklık alınır (118). Ağırlıksız grup yöntemi olarak da adlandırılmaktadır (117).

Ortalama bağlantı yöntemi dendrogramlar çizme yönünden tam bağlantı yöntemine benzemektedir. Fakat her bir yöntemde uzaklıklar farklı

tanımlanmaktadır. Bu sebeple birleştirmeler farklı seviyelerde meydana gelebilmektedir (119).

3.8.1.4. Ward Tekniği

Bir kümenin ortasında kalan gözlemin, aynı küme içinde bulunan diğer gözlemlerden ortalama uzaklığı hesaplanarak elde edilir (115). Bu yöntem artırımlı (incremental) hata kareler toplamı yöntemi olarak da adlandırılmaktadır (120). Ward'ın geliştirdiği bu yönteminde asıl amaç artan sınıf içi hata kareler toplamını minimum yapmaktır (118), (121).

3.8.1.5. Medyan Tekniği

Medyan bağlantı yönteminde öncelikle iki kümenin arasındaki mesafe; iki kümenin merkezleri arasındaki uzaklık eşit ağırlıklı olarak hesaplanır (122). Eğer Birleştirilecek iki kümenin boyutları birbirinden çok farklı ise, yeni kümenin ağırlık merkezi daha büyük olan kümenin ağırlık merkezine daha yakın olacaktır. Centroid yönteminin bu nedenden dolayı dezavantajlı olduğundan Gower, 1967'de medyan yöntemini geliştirmiştir. Hem benzerlik hem de uzaklık açısından bu yöntem daha uygun olacaktır (122).

3.8.1.6. Centroid Tekniği

Centroid yönteminde iki küme arasındaki mesafe ölçüsü olarak Kareli Euclid (Squared Euclidean) uzaklığı kullanılır (123). Her bir küme, o andaki kümenin ağırlık noktası ile gösterilir. İki küme birleştiğinde ise, ağırlık noktalarının birbirlerinden minimal mesafede olması yeterli olacaktır (124). Bu yöntemin avantajı ise farklı nitelik taşıyan gözlemlerden çok nadir etkilenmesidir (84).

3.8.1.7. DBSCAN Algoritması

Yoğunluğa dayalı kümeleme yöntemlerinde işlem kümeleri yüksek yoğunluklu olabilirler. Bunlar veri kümesinden geri kalan kısımlardır. Burada gürültü terimleri ya da sınır noktaları olarak kabul edilen kısımlar belirlenmiş yoğunluk alanlarının aralarında kalmışlardır. En bilinen yoğunluk tabanlı algoritma DBSCAN

algoritmasıdır. 1996 yılında Martin Esler ve arkadaşları Hans-Peter Kriegel, Jörg Sander ve Xiaowei Xu DBSCAN (Density Based Spatial Clustering of Applications with Noise), geliştirilmiştir. “Ulaşılabilir yoğunluk” esasına dayanarak ϵ komşuluk ve minimum küme sayısı olarak ifade edilmekte olan *minPts* parametrelerini kullanarak çalışır.

İlk adımda, her bir noktanın kendi çevresini kapsayan etki fonksiyonu hesaplanır. Bu fonksiyon Gauss, parabolik ya da kare dalga fonksiyonu olabilir. Veri tabanındaki tüm noktalara fonksiyon uygulanır. Daha sonra ikinci adıma gelindiğinde ise bu fonksiyonların toplamı hesaplanır. Hesaplanan bu fonksiyon; veri tabanının genel yoğunluğunu ifade edecektir. Son adımda ise kümeler matematiksel olarak yoğunluk-çekicileri (density-attractors) kullanılarak tespit edilir. Yoğunluk-çekicilerinden ifadesi genel yoğunluk fonksiyonunun yerel maksimumlarını belirtmektedir. Yoğunluk çekicileri yoğunluk fonksiyonunun gradyenti kullanılarak bulunur (49).

3.8.2. Bölünmeli Hiyerarşik Yöntemler

Bölünmeli Kümeleme Yöntemi, başlangıç olarak bütün birimlerin bir tek küme ifade ettiğini kabul eder. Daha sonra birimleri hiyerarşik n birimi sırasıyla 1, 2, 3,..., n-1, n kümeye parçalamayı amaçlamaktadır. Bu yöntem toplanmış hiyerarşik kümeleme yönteminin tersini ifade eder. Toplanmış hiyerarşik kümeleme yöntemlerine ilişkin sonuçlardan bölen yöntemi gösteren sonuçlara ulaşılabilir (125).

3.8.2.1. k-Medoid Tekniği

Medoid, herhangi bir kümede bulunan tüm elemanlara olan ortalama uzaklığı (benzemezlilik ölçüsü) minimum olan küme elemanıdır. Kaufman ve Rousseeuw tarafından 1987 geliştirilen bu yöntemde amaç, bir grup nesneyi k tane kümeye bölmek ve birbirine fazlaca benzeyen nesnelerin bir arada bulunduğu kümeler elde etmektir. Farklı kümelerdeki gözlemler ise birbirinden benzersiz olmalıdır. k adet temsilci nesne seçilir ve bunlar seçildikten sonra her biri en yakın olduğu temsilciye atanır. Böylece k tane küme elde edilmektedir. Bu yöntem, veri kümesindeki k-medoidi elde ettikten sonra hesaplanan bu medoide olan gözlemlerin uzaklıklarına

göre toplamdaki n noktayı kümelemeye çalışır. Daha sonraki adımda ise temsilci olarak seçilen nesneler temsilci olamayan nesne ile değiştirilerek kümenin niteliği yükseltilmeye çalışılır. Nitelik, nesne ile o nesnenin ait olduğu kümenin temsilci nesnesi arasında bulunan ortalama benzersizlik maliyet fonksiyonu (cost function) kullanılarak ölçülür. K-medoid yöntemi, özellikle aykırı ve tutarsız gözlemlerde iyi sonuçlar vermektedir. Uç değerler bulunduğunda küme merkezine kıyasla daha az etkilenen bir yöntem olmakla birlikte medoid Etrafında Bölümleme (Partitioning Around Medoids-PAM) en çok bilinen algoritmadır. PAM küçük veri tabanlarında çok iyi sonuçlar vermektedir. Fakat hesaplama karmaşıklığı yüksek olması nedeni fazla eleman bulunan veri tabanlarında performansı zayıftır. Büyük nesneler içeren bu veri tabanları için ise CLARA ve CLARANS algoritmaları geliştirilmiştir (126).

3.8.2.2. CLARA ve CLARANS Algoritmaları

PAM, K-medoids algoritmalarının başarısını kanıtlamıştır. Fakat büyük veri tabanlarında aynı başarı elde edilememiştir. Bu nedenle 1990 yılında Kaufman ve Rousseeuw tarafından CLARA (Clustering Large Applications) ortaya atılmıştır (127).

CLARA, veri tabanının tümü üzerinde çalışmak yerine tüm veriden küçük bir örneklem (sample) kümesini temsilcisi seçer. Seçilen bu örnek üzerine PAM algoritmasını uygular. Böylece seçilen örneklem üzerine uygulanan algoritmalarla en iyi sonuç vereni çıktılar. CLARA'nın daha büyük verilerde uygulanabilmesi avantajıdır. Fakat performansı örneklem boyutuna göre değişebilmektedir. Eğer örneklem yeterli düzeyde bağımsız seçilmemiş ise seçilmiş olan örneklem veri tabanını istenilen düzeyde temsil edemez. Bu durum yanlış sonuçlar doğurabileceğinden dolayı ise dezavantajıdır. CLARA algoritmasının karmaşıklığı, s örneklem boyutu olarak verildiğinde $O(ks^2 + k(n-k))$ şeklindedir.

CLARA'nın sonuçlarını örneklem seçimine bağlı olması nedeniyle CLARANS (Clustering Algorithm based on Randomized Search) algoritması Ng ve J.Han tarafından 1994 yılında geliştirilmiştir. CLARANS örneklem seçimindeki sorunlar nedeni ile sabit bir örneklem seçme durumunu ortadan kaldırmıştır. Bu yöntemde göre her aşamada değişen örneklem mevcuttur. Burada rastgele seçilen

noktaların çevreleri dikkate alınmaktadır. K-medoids algoritması ile elde edilen merkez noktalar temsil noktası olarak seçilir. Daha sonra bir başka rastgele nokta elde edilir. Bu bulunan noktanın da çevresine k-medoids uygulanarak yeni merkez noktalar elde edilir. Böylece veri tabanından tarafsız bir örneklem elde edilir.

CLARANS algoritmasının CLARA 'ya göre çok daha iyi sonuçlar verdiği görülmüştür. Bu yöntem istisna bölgeleri de bulma yeteneğini barındırır. Fakat hesaplanabilir karmaşıklığı n nesne sayısı olarak alınırsa $O(n^2)$ olduğu için veri sayısı fazlalaştıkça gereken hesaplama gücü de üstel olarak artacaktır. Bunların dışında ise veri tabanı ana bellekte bulunmuyorsa giriş/çıkış (I/O) işlemi oldukça artacağı için algoritmanın genel performansı düşebilmektedir (128).

3.8.3. Hiyerarşik Olmayan Kümeleme

Küme sayıları hakkında araştırmacı bir ön bilgiye sahip ya da anlamlı olacak küme sayısına karar vermiş ise bu gibi durumda hiyerarşik olmayan kümeleme yöntemi kullanılabilir. Bu yöntemde gözlemler rastgele olarak parçalanabilmektedir. İlk olarak birimlerin ayrılacakları küme sayıları belirlenmelidir. Daha sonra birimlerin hangi kümelere gireceğine karar verilmelidir. Böylece son olarak atama işlemleri yapılır. Hiyerarşik olmayan kümeleme yöntemlerine ise verilebilecek örnekler k-ortalamlar ve en çok olabilirlik yöntemleridir (129), (130).

3.8.3.1. k-Ortalamlar Kümeleme Tekniği (k-means)

1975'de Hartigan tarafından geliştirilen bu teknik, 1965 yılında Forgy ve 1967 yılında MacQueen tarafından ortaya çıkarılmıştır. Yöntem hata kareler azaltma ilkesinden daha sonra elde edilmiştir. Bu teknik uzun süreler boyunca en önemli tekniklerden biri olmuştur (131), (132). Kümeleme analizlerinde mesafe hesaplamada genel olarak Öklid uzaklığı ile işlem yapıldığından;

$\|x_{ij} - v_j\|^2$, $1 \leq j \leq k$ hesaplaması yapıldığında bireyler en yakın olan kümeye sınıflandırılır. Burada x_{ij} j. kümedeki i. nesneyi; v_j ise j. küme merkezini ifade etmektedir (133). Kullanılan uygun uzaklık ölçüsü ise;

$$J(x;v) = \sum_{j=1}^k \sum_{x \in c_j} \|x_j^i - v_i\|^2 \text{ şeklinde hesaplanır.} \quad (3.6)$$

K-Ortalamlar yöntemi, ilk olarak n adet nesneden rastgele olarak k adet nesne seçilir. Bu nesnelerin her bir tanesi, bir kümenin merkezini ya da orta noktasını göstermektedir. Geriye kalan nesneler ise kendisine en yakın bulunan küme merkezine göre kümelerle parçalanırlar. Kısaca bir nesne hangi kümeye mesafe olarak yakınsa oraya atanır. Daha sonra her bir küme için ortalamlar hesaplanarak elde edilir. Elde edilen bu değer ise kümenin yeni merkezidir. İşlem tüm gözlemler bir kümeye yerleşene kadar devam eder. K-Ortalamlar Yönteminin kullanılabilmesi için, elde edilecek küme için kümenin ortalamasının tanımlanması gerekmektedir (134). Burada k değerini, kısaca yeni meydana gelecek küme sayısını belirtme zorunluluğu bir dezavantaj olarak kabul edilebilir. Asıl dezavantajlı durum ise aykırı gözlemler olarak bilinen nesnelere karşı oluşacak duyarlılıktır (135).

K-ortalama yönteminin ana problemleri şu şekilde sıralanabilir (136), (137):

- Ölçeklemeye ve öteki dönüşümlere duyarlılık;
- Zayıf bir optimuma yakınsama eğilimi;
- Ölçüm hatalarına ve aykırı değerlere duyarlılık; yanlışlık (yani yanlış parametre değerleri yakınsama)

Tüm bu problemlere rağmen en büyük sorun ise uygun k değerini hesaplamaktır (138). Belirgin üstünlüğü güvenilir olması iken yorumlaması zordur (139).

3.8.3.2. Fuzzy C-Means Kümeleme Tekniği

Fuzzy C-means kümeleme yani bulanık c-ortalama kümeleme algoritması K-ortalama yöntemi kadar katı bir yöntem değildir. Bir başka deyişle k-ortalama yöntemi gibi birimler kesin olarak bir kümeye dahil olmamaktadır. Bu algorithmada ancak 0 ile 1 arasında değişen bir şekilde kümelerle üyelikler sağlanır. Böylece birimlerin birden fazla kümeye atanması söz konusudur. Bu durum bulanıklığa neden olmaktadır. Klasik yaklaşımlarda bir elemanın bir kümeye ait olup olmama durumu nettir. Kümeye ait birimler 1 değerini alırken, kümeye ait olmayanlar ise 0 değeri almaktadırlar. Fakat bu yöntemde elemanlar $[0,1]$ aralığında sonsuz sayıda değer alabilmektedir. Örnek vermek gerekirse klasik kümeleme yöntemlerinde net bir şekilde hasta-sağlıklı ifadeleri kullanılabilirken, bulanık kümelemede biraz sağlıklı –

daha sağlıklı şeklinde kullanılmaktadır. Klasik kümeleme yöntemlerinde hepsi ya da hiçbiri gibi kesinlik söz konusu iken, bulanık kümeleme yöntemlerinde kısmi ya da belirli derecede mantığı bulunmaktadır. Bu nedenle bu algoritmalarda bir aralıkta değer alma durumu söz konusudur ve elemanlar 0 ile 1 arasında en yüksek olasılığa sahip olan kümelere atanmaktadır.

Eğer her bir birim kümele de eşit üyeliğe sahip ise, kümeleme tamamen bulanık olacaktır. Bulanıklık olup olmadığını tespit edebilmek için Dunn'un parçalama katsayısı kullanılmaktadır. Bu katsayı hesabı ise;

$$F_{(u)} = \sum n_i = 1 \sum k_v = 1 u_{iv}^2 / n \text{ şeklindedir.} \quad (2.14)$$

Bu formülde K küme sayısını, n gözlem sayısını belirtir. u_{iv} değeri i nesnesinin v kümesine olan bilinmeyen üyeliğini belirtmektedir. Bu katsayı ise verilen $[1/K, 1]$ arasında değer alabilmektedir. Tamamen bulanık kümeleme durumu var ise tüm

$u_{iv} = 1/K$ olduğundan $F(u) = 1/K$ değerini alacaktır. Eğer tam anlamıyla katı kümeleme durumu oluşuyor ise tüm $u_{iv} = 0$ ve $F(u) = 1$ olur. Değeri 0 ile 1 aralığında değişmekte olan normalleştirilmiş Dunn katsayısı ise;

$F_{n(u)} = kF - 1$, $k-1$ eşitliğinden elde edilir. Bu katsayı sıfır değerini almış ise tamamen bulanıklık var demektir. Katı kümelenmeyi göstermesi ise 1 değerini alması ile olur. Bu katsayıya diğer bir deyişle bulanıksızlık endeks (nonfuzziness index) de denmektedir (140).

3.8.3.3. İki Aşamalı Kümeleme (TwoStep Cluster) Tekniği

İki Aşamalı Kümeleme (Two step cluster) tekniği hiyerarşik olmayan kümeleme tekniklerinden “k Ortalamalar” ile hiyerarşik tekniklerinden “Ward’ın En Küçük Varyans” tekniklerinin birleşiminden oluşmuş hibrid kümelemedir. Klasik kümeleme teknikleri ile karşılaştırıldığında bu teknik daha öznetelikli kategoriler ortaya koymaktadır. Bu nedenle farklı birçok disiplinden uygulama alanına sahip olmuştur. Bu teknikte öncelikli amaç büyük verileri analiz edebilmektir. İki aşamalı kümeleme tekniği kümelerdeki nesneleri ayırıcı özellik yaklaşımı ile gruplara

ayırmaktadır. Devamlı kategoriler sağlayan yöntem ayrıca en uygun küme sayısını da kendisi belirlemektedir (112).

İki aşamalı kümeleme tekniği, ön-kümeleme, tipik veri türlerini çözümüleme ve kümeleme aşamalarından meydana gelmektedir. Ön kümeleme yapılırken gözlemlerin her biri incelenerek elde edilmiş bir önceki kümeye atanabilir mi yoksa tekrar başka bir küme yeniden mi oluşturulmalı diye karar verilmektedir. Karar verilerin uzaklığına göre Öklid uzaklığı ve log-olasılık uzaklığı olmak üzere sadece iki uzaklık ölçütünü kullanılarak verilmektedir. Tipik veri çözümüleme aşamasında ise, elde edilen kümelerden hiçbirine dahil olamayan veriler değerlendirme aşamasına girer. Burada kümelere girebilecek veriler eklenir fakat hala veri bir kümeye dahil olamıyor ise dış veri olarak ayrılmaktadır. Bu kümeleme tekniğinde bir ağaç yapısı oluşmaktadır. Elde bulunan tüm gözlemle kökten yapraklara doğru atanmaya başlar. Her bir gözlem kendisine en yakın dala tutunur. Eğer verinin dahil olduğu grup sayısı optimum grup üyesi sayısına varmışsa, kullanılan uzaklık ölçütüne göre en yakın başka bir dalda başka bir kümeye atanır. Uygun küme sayısı otomatik olarak belirlenmektedir. İki aşamalı kümeleme tekniği bunun için ise AIC (Akaike's Information Criterion) ya da BIC (Schwarz's Bayesian Information Criterion) yöntemlerini kullanmaktadır (141).

3.8.3.4. Kohonen Ağlar

Self Organizing Map (SOM) olarak da adlandırılan Kohonen Ağlar, kümelemede kullanılan bir yöntemdir. Bu algorithmada önemli olan nokta “nöron”lardır. Nöronlar iki tabakadan oluşmaktadır ve “girdi” ve “çıkı” şeklinde isimlendirilirler. Girdi nöronlarının hepsi çıkı nöronları ile bağlantılıdır. Bu bağlantı ifade etmek için “güç” olarak adlandırılan ölçüler kullanılmaktadır. Algoritmanın uygulanması ile çıkı nöronları en çok veriyi kendilerine bağlamak için çalışmaya başlar. İlişki vektör tabakası iki boyutlu olmak üzere ve ilişkisiz grid yapısı denilen bir yapı oluşturur. İlk olarak tüm girdi nöronları hepsi için her bir çıkı nöronlarına ağırlık ataması yapılır. En güçlü ağırlığa göre çıkı nöronlarına veriler gönderilmektedir. Böylece analiz sonuçlanır ve benzer özelliklere sahip veriler grid üzerinde aynı kümede, farklı özelliğe sahip olanlar ise uzak noktalarda kümelenir (26).

3.8.3.2. En Çok Olabilirlik Tekniđi

Her bir nesne en büyük olabilirlik değeri verecek şekilde daha önceden belirlenmiş olan kümelere atanmaktadır. Diskriminant analizinde kullanılan en çok olabilirlik yöntemine benzemektedir (142). Bu analizde her bir gözlem (her bir x_i gözlem vektörünün bağımsız ve aynı olasılık yoğunluk fonksiyonuna sahip olduğu varsayımı ile) büyük olabilirlik değerine ulaşılacak şekilde daha önceden belirlenmiş kümelere atanır. Kuramsal dayanağı oldukça güçlü bir yöntemdir. Yaygın olarak kullanılması ise özellikle uzun makina zamanı gerektirmesi nedeniyledir (129).



4. BULGULAR

4.1. Tanımlayıcı Analizler

Çalışmada toplam Ağız, Çene ve Diş Cerrahisi bölümüne operasyon için başvuran 5148 hastaya ait kan tahlili sonuçları değerlendirildi. Hastaların kan tahlillerine göre biyokimya, hemogram ve immünolojik değerleri belirlendi. Kan tahlili sonucunda elde edilen bu değerler, referans değerlere göre kıyaslandı. Operasyon öncesi düşük veya yüksek bulunan değerlere göre “Hasta” ve “Sağlıklı” olarak küme grupları oluşturuldu. Bu hastaların yaş ortalaması $21,19 \pm 19,23$ yıl idi ve yaş aralığı oldukça geniş olup en küçük ve en büyük yaş değerleri 4 ve 89 olarak hesaplandı. Hastaların başvuru tarihleri birbirine benzemekle birlikte en yüksek oranda Mart (%11,34), Kasım (%10,93) ve Şubat (%10,27) aylarında başvuru yapıldığı anlaşıldı. En düşük oranda (%4,78) başvurunun ise Ağustos ayında yapıldığı gözlemlendi.

4.2. Analizler

Öncelikle yaş değişkeni ile birlikte viral testler için biyokimya, hemogram ve immünolojik (mikrobiyolojik) değişkenler için tanımlayıcı istatistikler elde edilmiştir (Tablo 4.1.)

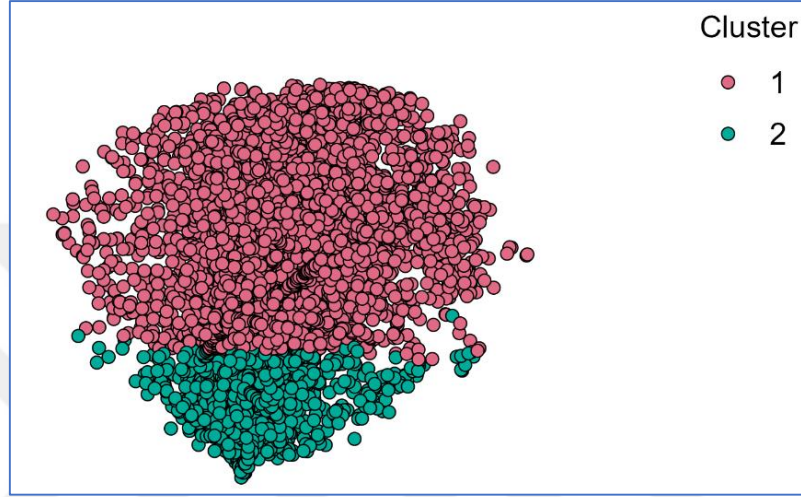
Tablo 4.1. Mikrobiyolojik Viral Testlere Ait Tanımlayıcı İstatistikler

	Değişkenler	Ortalama$\pm$SS	Min-Maks
Biyokimya	Yaş	21,19 $\pm$ 16,23	4 - 89
	Kolesterol	189,10 $\pm$ 19,13	80 - 375
	Kreatinin	0,79 $\pm$ 0,19	0.18 - 2.05
	HDL	58,95 $\pm$ 6,47	15.9 - 137.5
	Potasyum	4,06 $\pm$ 0,37	2.57 - 8.88
	LDL	113,88 $\pm$ 26,82	16 - 244.9
	Sodyum	136,54 $\pm$ 2,72	116.11 - 170.65
	ASTSGOT	23,08 $\pm$ 8,88	2.7 - 154.2
	ALTSGPT	16,79 $\pm$ 11,18	2.8 - 204.4
	Üre	25,75 $\pm$ 7,62	6 - 91.6
	Trigliserid	161,76 $\pm$ 85,77	23.1 - 903.9
	APTT	23,36 $\pm$ 4,03	1.1 - 138
	PT	96,45 $\pm$ 16,16	-103 - 763
	Glukoz	101,60 $\pm$ 23,04	43.8 - 443
	BUN Üre	12,08 $\pm$ 3,73	4 - 91.7

	TSH	2,62±2,03	0.006 - 93.7
	Serbest T3	4,20±0,80	1.02 - 19.1
	Serbest T4	1,45±1,18	0.235 - 21.17
	Kalsiyum	9,53±0,59	6.4 - 14.6
	Klor	105,28±3,19	80.2 - 121.5
Hemogram	Yaş	21,19±16,23	4 - 89
	BASO#	0,02±0,01	0 - 0,6
	BASO%	0,34±0,21	0 - 2,8
	EO#	0,19±0,18	0 - 2,37
	EO%	2,38±2,05	0 - 28
	HCT	38,11±4,30	22,9 - 53,9
	HGB	13,07±1,61	6,5 - 19,4
	LYMPH#	3,06±1,33	0,42 - 12,07
	LYMPH%	37,89±12,19	5,6 - 78,1
	MCH	26,63±2,68	12,7 - 39,5
	MCHC	34,28±1,31	23 - 38,2
	MCV	77,63±7,12	23,7 - 105,6
	MONO#	0,61±0,22	0,06 - 1,97
	MONO%	7,73±2,11	0,9 - 22,5
	MPV	9,89±1,04	7 - 14,2
	NEUT#	4,26±1,89	0,3 - 16,84
	NEUT%	51,64±12,93	8,8 - 89,9
	PLCR	24,35±8,08	5,1 - 57,9
	PCT	0,30±0,08	0,03 - 0,88
	PDW	11,42±2,11	6,6 - 25,7
	PLT	316,58±96,18	26 - 916
	RBC	4,92±0,48	2,62 - 8,2
	RDWCV	14,07±1,61	11,2 - 29,7
	RDWSD	38,53±3,46	25,6 - 69,8
	WBC	8,16±2,50	2,8 - 22,18
	Yaş	21,19±16,23	4,0 - 89,0
İmmunoloji	Anti-HBs	144,58±250,13	-7,0 - 1000,0
	Anti-HCV	0,149±2,573	-73,0 - 70,55
	Anti-HIV	0,366±0,104	0,00 - 1,93
	HBsAG	1,029±15,120	0,0 - 740,8

İlk olarak JASP 16.0 Uygulaması kullanılarak biyokimya değişkenleri k-means ve Fuzzy C-means teknikleri kullanılarak değerlendirildi. Yaş değişkeni de dâhil olmak üzere toplam 20 değişken ile iki farklı teknik ile analiz edildi. K-means kümeleme tekniği ile analiz edilen biyokimya değişkenlerinde son iterasyonda belirlenen yaş merkez noktaları 21 ve 23 olarak elde edildi. 1. Küme n=3833 ve 2. Küme n=1315 değerden oluşarak kümelenme grafiği elde edildi (Şekil 4.1.). Öklit

uzaklık ölçüsü ile yapılan kümelemede model başarımlar ölçütleri; Silhouette değeri 0,580, AIC ve BIC değerleri sırasıyla 2713,680 ve 2713,530, açıklayıcılık katsayısı ise $R^2=0,498$ olarak elde edildi. Kolesterol, LDL ve Trigliserid değişkenlerinde merkez noktaları arasında yüksek fark rastlandı (Tablo 4.2). Bu değerlerin referans değerler ile kıyaslanması sonucu 2. Küme “Sağlıklı” 1. Küme ise “Hasta” olarak isimlendirildi.



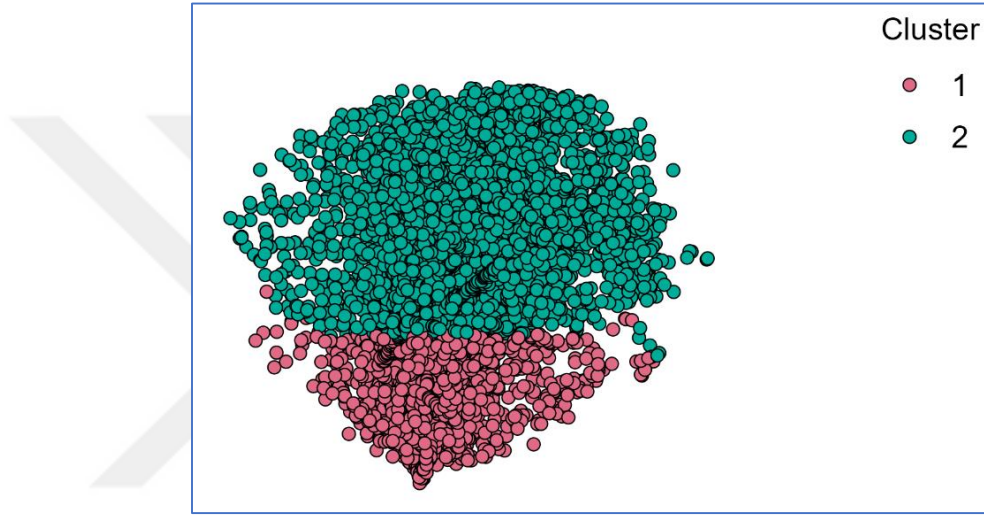
Şekil 4.1. JASP Biyokimya Viral Testlere ait K-means Küme Yapısı

Tablo 4.2. JASP k-means Biyomkiya Değişkenlerinin Atandığı Küme Merkez Değerleri

	Kümeler			Kümeler	
	Küme 1	Küme 2		Küme 1	Küme 2
KOLESTROL	187,335	194,282	APTT	26,382	26,300
KREATİNİN	0,792	0,787	PT	96,748	95,619
HDL	59,387	57,707	GLUKOZ	100,669	104,341
POTASYUM	4,065	4,067	BUNÜRE	12,035	12,240
LDL	108,537	129,485	TSH	2,621	2,644
SODYUM	136,661	136,212	SERBESTT3	4,221	4,152
ASTSGOT	23,218	22,692	SERBESTT4	1,496	1,336
ALTSGPT	16,506	17,642	KALSİYUM	0,10	0,27
ÜRE	25,647	26,081	KLOR	9,538	9,532
TRİGLİSERİD	122,535	276,131			

Aynı değişkenlere uygulanan Fuzzy C-means tekniği ile yapılan analiz sonucunda ise yaş merkez noktaları 16 ve 20 olarak saptandı. 1. Küme n=298 ve

Küme 2 ise n=4850 birey ile oluştu (Şekil 4.2.). Öklit uzaklık ölçüsü ile yapılan kümelemede model başarımları; Silhouette değeri 0,470, AIC ve BIC değerleri sırasıyla 3379,720 ve 3379,570, açıklayıcılık katsayısı ise $R^2=0,465$ olarak elde edildi. LDL, Trigliserid ve Glukoz değerleri arasındaki merkez nokta farklılıkları incelendi (Tablo 4.3.). Fuzzy C-means tekniğinde keskin isimlendirme yapılamayacağından 2. Küme “Daha Sağlıklı” 1. Küme ise “Daha az Sağlıklı” olarak etiketlendi. K-means tekniği ile kıyaslandığında Silhouette değerleri yakın olsa da Fuzzy C-means algoritmasının açıklayıcılık katsayısı daha düşük olarak görüldü.



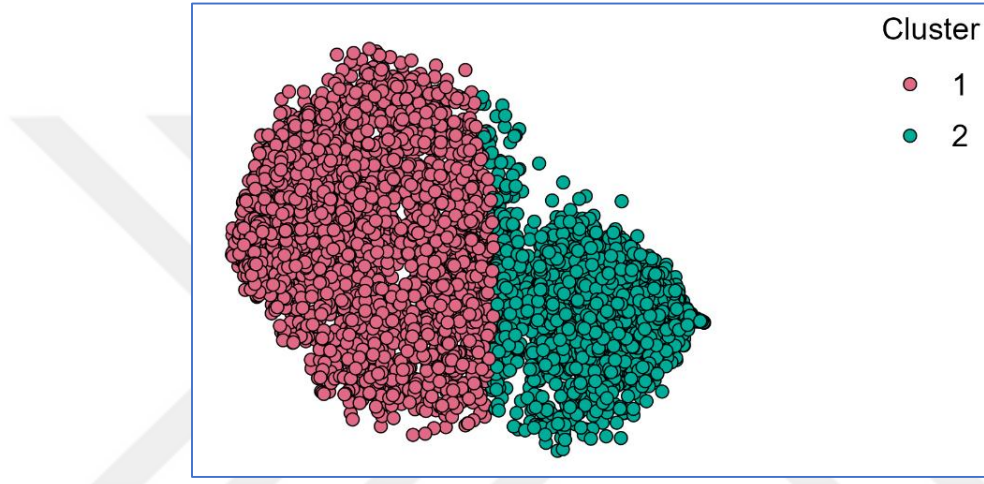
Şekil 4.2. JASP Biyokimya Viral Testlere Ait Fuzzy C-means Küme Yapısı

Tablo 4.3. JASP Fuzzy C-means Biyokimya Değişkenlerinin Atandığı Küme Merkez Değerleri

	Kümeler			Kümeler	
	Küme 1	Küme 2		Küme 1	Küme 2
KOLESTROL	189,781	188,349	APTT	24,235	26,171
KREATİNİN	0,643	0,806	PT	97,780	95,914
HDL	58,726	59,17	GLUKOZ	138,022	96,914
POTASYUM	3,733	4,103	BUNÜRE	11,145	12,015
LDL	148,143	100,797	TSH	2,266	2,327
SODYUM	136,011	136,534	SERBESTT3	4,231	4,168
ASTSGOT	28,565	23,168	SERBESTT4	1,212	1,445
ALTSGPT	15,585	15,232	KALSİYUM	9,688	9,510
ÜRE	24,351	25,591	KLOR	100,914	105,523
TRİGLİSERİD	476,628	134,297			

Hemogram test kümeleri hasta yaşı da değişkenlere dâhil olmak 25 değişken olmak üzere analiz edildi. K-means tekniği ile 2 küme elde edildi. Son iterasyonda

yaş için belirlenen yaş merkez noktaları sırasıyla 26 ve 14 olarak bulundu. Öklit uzaklık ölçüsü ile yapılan kümelemede model başarımlar ölçütleri; Silhouette değeri 0,490, AIC ve BIC değerleri sırasıyla 2207,530 ve 2207,850, açıklayıcılık katsayısı ise $R^2=0,572$ olarak elde edildi. 1. Kümeye $n=3082$ 2. Kümeye ise $n=2066$ birey atandı (Şekil 4.3.). Özellikle Tablo 4.4. LYMPH#, NEUT#, PLCR ve PLT değerleri arasındaki merkez nokta farklılıkları incelendiğinde 1. Küme değerlerinin referans değerlere daha yakın olduğu görülmüştür. Bu nedenle 1. Küme “Sağlıklı” 2. Küme ise “Hasta” olarak etiketlendi.

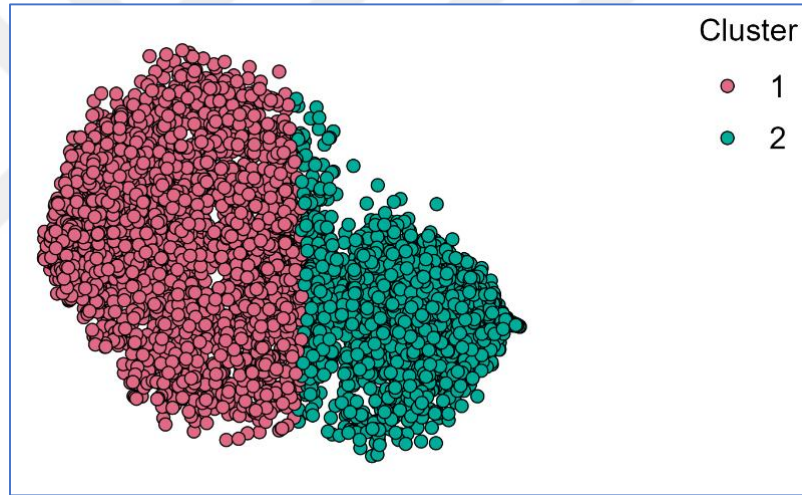


Şekil 4.3. JASP Hemogram Viral Testlere Ait k-means Küme Yapısı

Tablo 4.4. JASP k-means Hemogram Değişkenlerinin Atandığı Küme Merkez Değerleri

	Kümeler			Kümeler	
	Küme 1	Küme 2		Küme 1	Küme 2
BASO#	0,025	0,030	MONO%	7,919	7,453
BASO%	0,344	0,340	MPV	10,231	9,382
EO#	0,173	0,222	NEUT#	4,085	4,529
EO%	2,327	2,470	NEUT%	53,760	48,492
HCT	39,253	36,423	PLCR	27,021	20,370
HGB	13,479	12,481	PCT	0,260	0,381
LYMPH#	2,628	3,705	PDW	12,087	10,445
LYMPH%	35,635	41,255	PLT	254,994	408,470
MCH	27,304	25,643	RBC	4,946	4,891
MCHC	34,308	34,243	RDWCV	13,899	14,344
MCV	79,547	74,784	RDWSD	39,127	37,655
MONO#	0,583	0,671	WBC	7,498	9,155

Aynı değişkenler Fuzzy C-means tekniği ile analiz edildiğinde ise Öklit uzaklık ölçüsü ile yapılan kümelemede model başarımlar ölçütleri; Silhouette değeri 0,480 ile K-means ile elde edilen değere çok yakın olmasına rağmen açıklayıcılık katsayısı ise $R^2=0,348$ K-means tekniğine göre düşük bulundu. AIC ve BIC değerleri sırasıyla 2209,690 ve 2209,010 olarak bulundu. 1. Kümeye $n=3052$ 2. Kümeye ise $n=2096$ birey atanmış oldu (Şekil 4.4.). Merkez yaş ortalamaları ise son iterasyon ile 23 ve 12 olarak elde edildi. İki yöntem aynı veri seti üzerinde kıyaslandığında K-means tekniğinin açıklayıcılık katsayısı olarak daha iyi sonuçlar verdiği görülmektedir. Tablo 4.5. LYMPH#, NEUT#, PLCR ve PLT değeri incelendiğinde 1. Kümenin referans değere daha yakın merkezlenmesi nedeniyle ve Fuzzy C-means tekniğinde keskin isimlendirmeler yapılamadığından dolayı 1. Küme “Daha Sağlıklı”, 2. Küme ise “Daha Az Sağlıklı” olarak isimlendirildi.

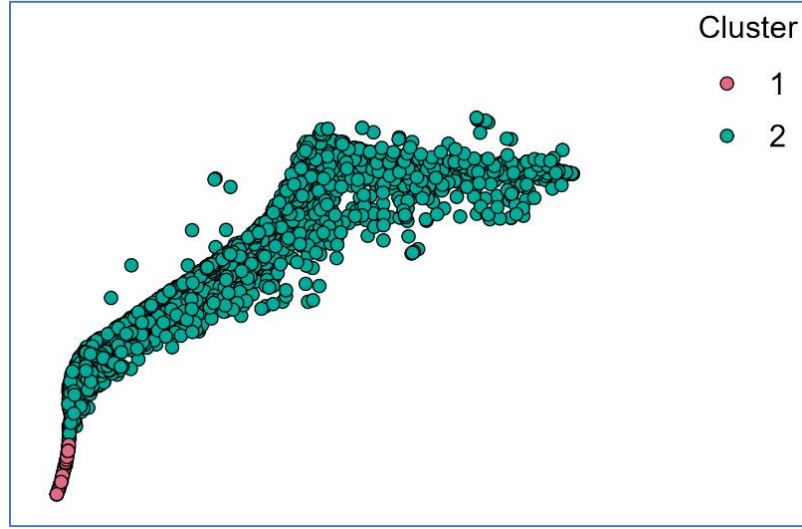


Şekil 4.4. JASP Hemogram Viral Testlere Ait Fuzzy C-means Küme Yapısı

Tablo 4.5. JASP Fuzzy C-means Hemogram Değişkenlerinin Atandığı Küme Merkez Değerleri

	Kümeler			Kümeler	
	Küme 1	Küme 2		Küme 1	Küme 2
BASO#	0,024	0,026	MONO%	7,233	7,612
BASO%	0,310	0,289	MPV	10,007	9,368
EO#	0,173	0,165	NEUT#	4,588	4,023
EO%	2,183	1,189	NEUT%	56,188	45,266
HCT	38,891	36,823	PLCR	25,181	20,407
HGB	13,402	12,666	PCT	0,280	0,351
LYMPH#	2,693	3,899	PDW	11,631	10,340
LYMPH%	34,075	44,947	PLT	281,803	377,208
MCH	26,812	25,798	RBC	5,008	4,922
MCHC	34,435	34,383	RDWCV	13,944	14,173
MCV	77,852	74,989	RDWSD	38,507	37,551
MONO#	0,564	0,659	WBC	8,046	8,770

Son olarak mikrobiyolojik testlerin kümeleri oluşturuldu. Analiz için hastaların yaşı, Anti-HBS, Anti-HCV, Anti-HIV ve HBS-AG sonuçları kullanıldı. K-means tekniğine göre iki küme oluşturuldu ve Öklit uzaklık ölçüsü ile yapılan kümelemede model başarımları ölçümleri; Silhouette değeri 0,830, AIC ve BIC değerleri sırasıyla 6475,970 ve 6475,970 olarak elde edildi. Açıklayıcılık katsayısı $R^2=0,801$ bulundu. Kümelerin büyük olanında 4607 ve ikinci kümede 541 hasta yer aldı (Şekil 4.5.). Hastaların yaşı için belirlenen küme merkez noktaları 24 ve 78 iken son iterasyonda 21 ve 23 olarak belirlendi. Anti-HBS için başlangıçta belirlenen merkez noktaları 2 ve 1000 mIU/ml iken son iterasyonda 67,59 ve 800,20 mIU/ml değerlerine ulaşmıştır. Anti-HBS değerleri arasındaki merkez nokta farklılıkları incelendiğinde 1. Küme değerlerinin referans değerlere daha yakın olduğu görülmüştür (Tablo 4.6.). Bu nedenle 1. Küme “Sağlıklı” 2. Küme ise “Hasta” olarak etiketlendi.

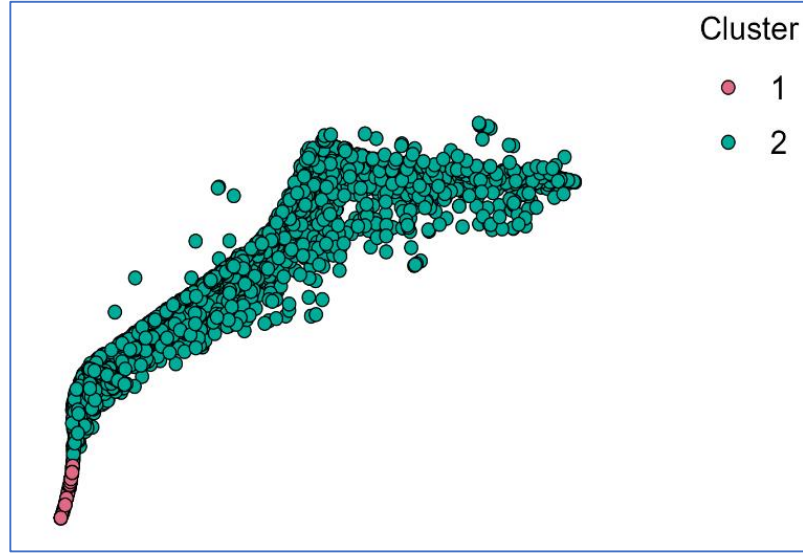


Şekil 4.5. JASP Mikrobiyolojik Viral Testlere Ait K-means Küme Yapısı

Tablo 4.6. JASP k-means İmmünolojik Değişkenlerinin Atandığı Kümenin Merkez Değerleri

	Kümeler	
	Küme 1	Küme 2
ANTIHB	800,200	67,592
ANTIHCV	0,184	0,144
ANTIHIV	0,357	0,367
HBSAG	0,586	1,080

Aynı değişkenler, Fuzzy C-means kümeleme tekniği ile analiz edildiğinde Öklit uzaklık ölçüsü ile yapılan kümelemede model başarımları ölçütleri; Silhouette skorunun 0,700, AIC ve BIC değerlerinin ise sırasıyla 1218,990 ve 1218,450 olduğu görüldü. Açıklayıcılık katsayısı ise $R^2=0,281$ olarak hesaplandı. Kümelerde yer alan hasta sayıları 1146 ve 4002 olarak hesaplandı (Şekil 4.6.). Hastaların yaşı için belirlenen küme merkez noktaları 16 ve 20 olarak belirlendi. Tablo 4.7. Anti-HBS değeri incelendiğinde 1. Kümenin referans değere daha yakın merkezlenmesi nedeniyle ve Fuzzy C-means tekniğinde keskin isimlendirmeler yapılamadığından dolayı 1. Küme “Daha Sağlıklı”, 2. Küme ise “Daha Az Sağlıklı” olarak isimlendirildi.



Şekil 4.6. JASP Mikrobiyolojik Viral Testlere Ait Fuzzy C-means Küme Yapısı

Tablo 4.7. JASP Fuzzy C-means İmmünolojik Değişkenlerinin Atandığı Kümenin Merkez Değerleri

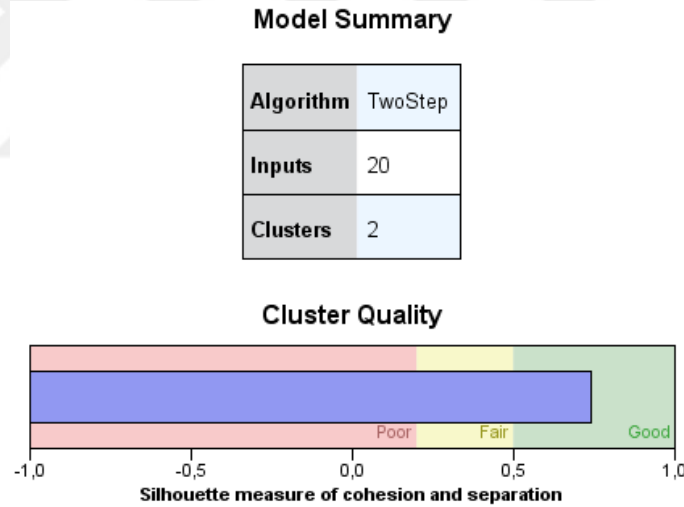
	Kümeler	
	Küme 1	Küme 2
ANTIHB	275,054	43,786
ANTIHC	0,123	0,147
ANTIHIV	0,300	0,366
HBSAG	0,898	0,758

K-means kümeleme tekniği ile karşılaştırıldığında, Fuzzy C-Means tekniğinin Silhouette değerinin ve açıklayıcılık katsayısı değerlerinin daha düşük olması, oluşan kümelerdeki kayıt sayısının birbirine yakın olması K-means kümeleme tekniği ile üretilen küme yapılarının daha uygun olduğunu gösterdi (Tablo 4.8).

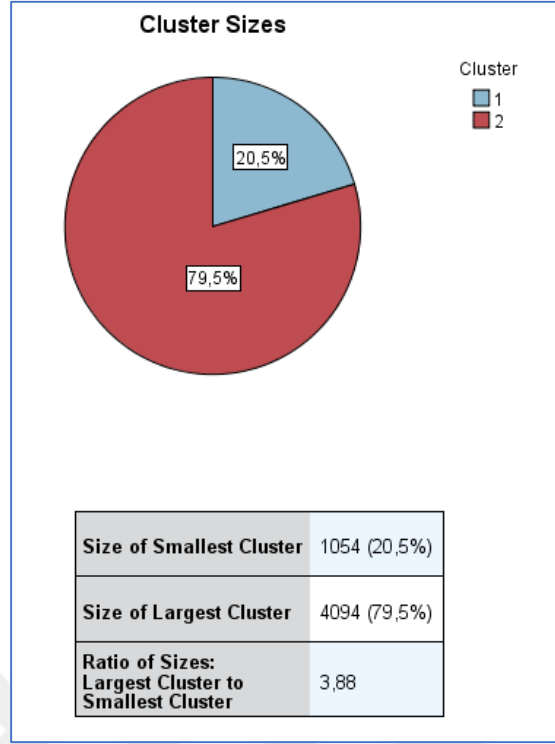
Tablo 4.8. JASP k-means ve Fuzzy C-means Başarım Ölçüt Değerleri

Ölçümler	Teknik	AIC	BIC	Silhouette	R ²	Uzaklık Ölçütü
Biyokimya	k-means	2713,680	2713,530	0,580	0,498	Öklit
	Fuzzy C-means	3379,720	3379,570	0,470	0,465	Öklit
Hemogram	k-means	2207,530	2207,850	0,490	0,572	Öklit
	Fuzzy C-means	2209,690	2209,010	0,480	0,348	Öklit
İmmünoloji	k-means	6475,970	6475,430	0,830	0,801	Öklit
	Fuzzy C-means	1218,990	1218,450	0,700	0,281	Öklit

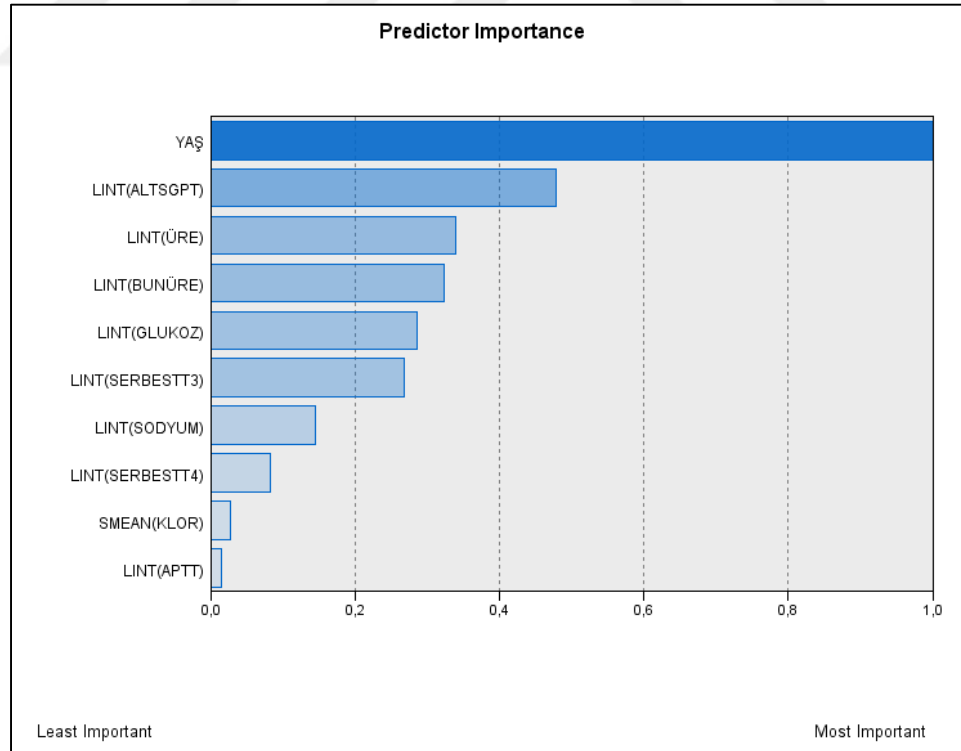
SPSS ile yapılan kümeleme analizinde TwoStep Cluster Tekniği kullanıldı. 20 değişkenden oluşan biyokimya verileri ile Analiz sonuçlarına göre 2 küme elde edilecek şekilde seçim yapıldı. Elde edilen Silhouette değerinin “Fair” ve “Good” a yakın olması grupların kendi içlerinde olabildiğince homojen olduğunu göstermektedir. Öklit uzaklık ölçüsü ile yapılan kümelemede model başarımlı ölçütü Silhouette değeri “Good” olarak bulunmuştur (Şekil 4.7.). Küme 1 (n=1054 %20,5) ve Küme 2 (n=4094 %79,5)’ye atanan değerler elde edildi (Şekil 4.8). Değişkenlerin ilgili kümeye atanırken önem sıraları Şekil 4.9.’da verilmiştir. Buna göre Biyokimya ölçümlerinde en önemli değişkenleri yaş, ALTSSGPT, Üre, Bun, Glikoz ve SerbestT3 değerleri oluştururken, Sodyum, Klor, APTT ve SerbestT4 değerleri düşük düzeyde önem arz etmektedir. Buradaki ilgili değişkenler Şekil 4.10.’da incelendiğinde Küme 2’nin “genç” (Yaş: 15,68) Küme 1’in ise “orta yaş” (Yaş:42,62) bireylerden oluştuğu, referans değerler ile kıyaslandığında Küme 2’nin “Sağlıklı” ve Küme 1’in ise “Hasta” olarak gruplandırılabilceği görüldü.



Şekil 4.7. SPSS TwoStep Cluster Biyokimya Ölçümleri Silhouette Değeri



Şekil 4.8. SPSS TwoStep Cluster Biyokimya Ölçümleri Kümeye Dâhil Olan Birimlerin Oranları



Şekil 4.9. Spss TwoStep Cluster Biyokimya Ölçümleri İlgili Kümelere Ayrılırken Önemli Bulunan Değişkenler



Şekil 4.10. SPSS TwoStep Cluster Biyomkiya Değişkenlerinin Atandığı Küme Merkez Değerleri

Analizler hemogram ve immünolojik ölçümlere de aynı şekilde uygulandı. Her iki ölçüm içinde Öklit uzaklık ölçüsü ile yapılan kümelemede model başarımlar ölçütü Silhouette değeri “Good” düzeyinde elde edildi. Hemogram ölçümleri Küme 1 (n=2244 %43,6), Küme 2 (n=2904 %56,4) ve İmmünolojik ölçümler Küme 1 (n=23 %0,4), Küme 2 (n=5125 %99,6) olarak bulundu (Tablo 4.9.). Hemogram ölçüm değerlerinde en önemli değişkenler sırasıyla PLT, NEUT, MCV, MCH, LYMPH, HGB, HCT, Yaş ve PLCR değişkenleridir. Hemogram ölçümlerinde yaş değişkeni incelendiğinde Küme 1 (Yaş: 11,35) “Genç” ve Küme 2 (Yaş:33,93) “Orta Yaş” bireylerden oluştuğu gözlemlendi. Referans değer sınırlarına göre Küme 1 “Sağlıklı” ve Küme 2 “Hasta” olarak isimlendirildi. İmmünolojik verilerde yaş değişkeninin çok düşük düzeyde önem arz ettiği görüldü. Küme 1 (Yaş: 21,16) ve Küme 2 (Yaş: 27,52) olarak yakın değerler gözlemlendi. Referans değerlere göre karşılaştırma yapıldığında ise Küme 1 “Sağlıklı” ve Küme 2 “Hasta” olarak isimlendirildi.

Tablo 4.9. SPSS TwoStep Cluster Biyokimya, Hemogram ve İmmünolojik Ölçümlerin Kümeleme Sonuçları

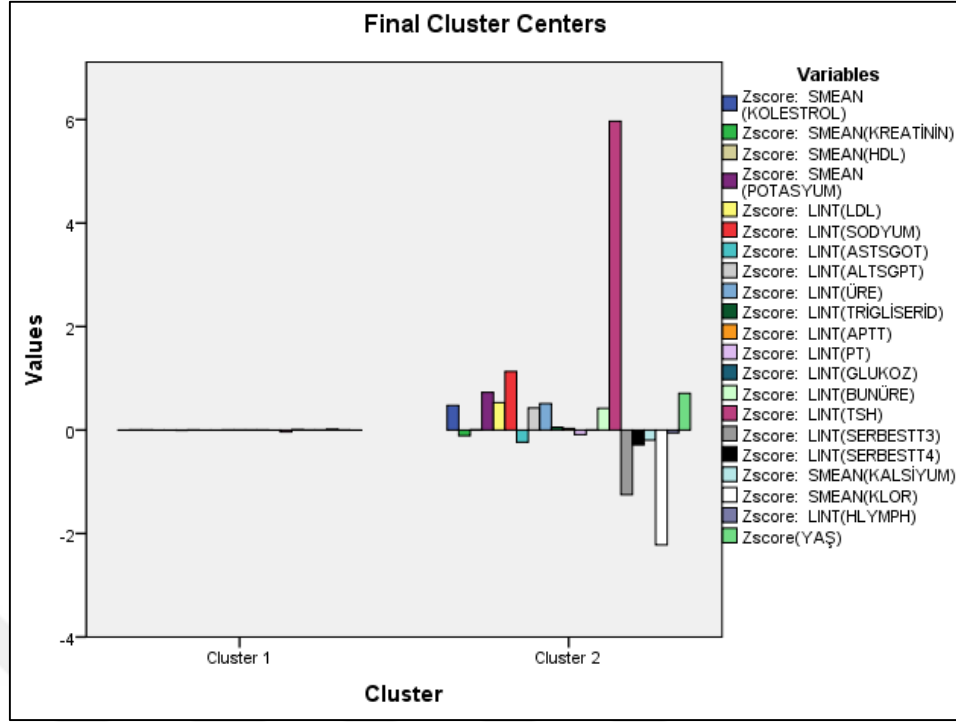
	Silhouette değeri	Kümeye Dahil Değişken	Küme 1 % (n)	Küme 2 % (n)	Kümeye dâhil olan önemli değişkenler
Biyokimya Ölçümleri	Good	20	1054 %20,5	4094 %79,5	Yaş, ALTSSGPT, ÜRE, BUN, GLİKOZ VE SERBESTT3
Hemogram Ölçümleri	Good	25	2244 (%43,6)	2904 (%56,4)	PLT, NEUT, MCV, MCH, LYMPH, HGB, HCT, Yaş, PLCR
İmmünolojik Ölçümler	Good	5	23 (%0,4)	5125 (99,6)	HBSAG, ANTIHIV

Spss ile yapılan kümeleme analizi k-means tekniği ile biyokimya ölçümleri üzerinde 8 tekrarlanma ile 2 küme elde edildi. Ölçümlerin Zscore değerleri alınarak normalleştirildi. Tablo 4.10.’de küme değişkenlerinin merkez noktaları gösterildi. İki küme arasındaki uzaklık 6,805 olarak bulundu. Şekil 4.11.’de ise değişkenlerin iki küme arasındaki dağılım farklılığı bulundu. Buna göre Küme 1 referans değerlere daha yakın olarak “Sağlıklı” Küme 2 ise referans değerlerden fazla sapması nedeni

ile “Hasta” olarak isimlendirildi. Şekilde yaş değişkeninin 2. Kümede 1. Kümeye göre artış göstermesi 1. Kümenin daha genç bireyleri kapsadığını gösterdi. Ayrıca 2. Kümede 1. Kümeye göre Kolesterol, Potasyum, LDL, Sodyum, Üre, BUN Üre ve TSH değerleri oldukça yüksekken SerbestT3, SerbestT4 ve Klor değerleri düşük olarak bulundu. Tablo 4.11. ile ANOVA sonuçları elde edildi. ANOVA sonuçları ile değişkenlerin atandıkları kümeler itibari ile değişkenlik gösterip göstermediği incelendi. Beklenen değişkenlerin atandıkları kümeler itibari ile farklılık göstermesidir. Çünkü kümeleme analizi bu farklılığı kendi oluşturur ve kümeler arası farkı (uzaklığı) maksimum yapar. Kreatinin, ASTSGOT, ALTSGPT, HDL, APTT, PT değerleri kümelere atanma itibari ile farklılık göstermezken diğer tüm değişkenler farklılık gösterdiği gözlemlendi.

Tablo 4.10. SPSS k-means Biyomkiya Değişkenlerinin Atandığı Küme Merkez Değerleri

	Kümeler			Kümeler	
	Küme 1	Küme 2		Küme 1	Küme 2
KOLESTROL	0,00	4,90	APTT	-0,78	2,53
KREATİNİN	0,90	-0,81	PT	41,23	-1,26
HDL	0,00	3,77	GLUKOZ	-0,39	1,14
POTASYUM	0,84	-0,39	BUNÜRE	0,51	1,31
LDL	0,38	2,52	TSH	0,47	44,68
SODYUM	-2,44	-0,34	SERBESTT3	-0,26	-3,95
ASTSGOT	0,71	1,01	SERBESTT4	-0,30	-1,03
ALTSGPT	-0,36	0,51	KALSİYUM	0,10	0,27
ÜRE	0,68	1,31	KLOR	-1,43	0,31
TRİGLİSERİD	0,00	2,01			



Şekil 4.11. SPSS k-means Biyokimya Değişkenlerinin Kümeler Arası Dağılımı

Tablo 4.11. SPSS k-means Biyokimya Ölçümleri Anova Sonuçları

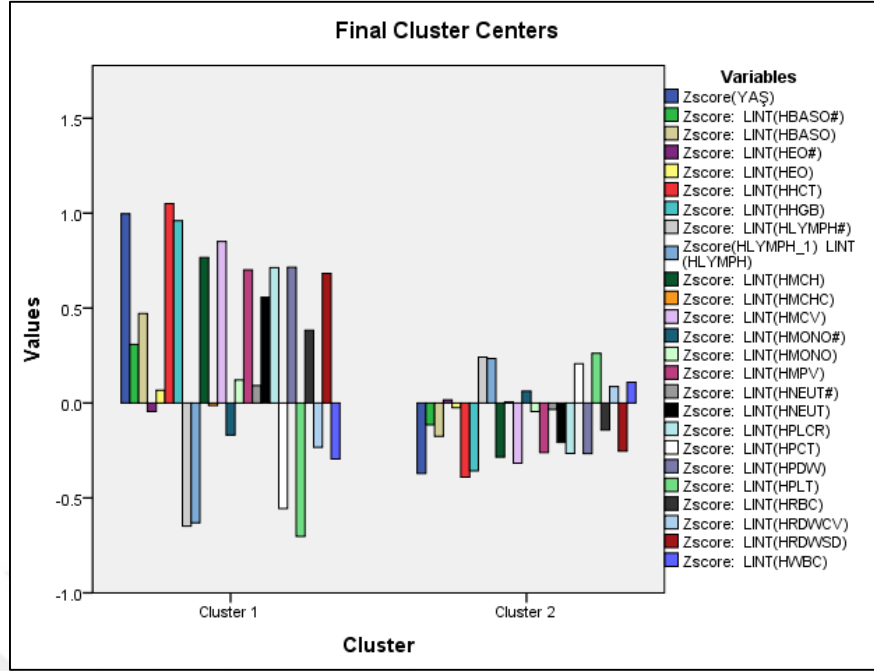
	ANOVA						
	Küme Kareler Ortalaması	sd	p		Küme Kareler Ortalaması	sd	p
KOLESTROL	7,220	1	0,002	APTT	0,026	1	0,872
KREATİNİN	0,393	1	0,531	PT	0,241	1	0,624
HDL	0,001	1	0,981	GLUKOZ	0,341	1	0,001
POTASYUM	17,044	1	<0,001	BUNÜRE	5,685	1	0,001
LDL	9,085	1	0,003	TSH	1146,345	1	<0,001
SODYUM	41,297	1	<0,001	SERBESTT3	50,382	1	<0,001
ASTSGOT	1,778	1	0,182	SERBESTT4	2,670	1	0,001
ALTSGPT	5,874	1	0,015	KALSİYUM	1,202	1	0,002
ÜRE	8,434	1	0,004	KLOR	158,451	1	<0,001
TRİGLİSERİD	0,102	1	<0,001				

Hemogram ölçüm değerleri üzerinde 10 iterasyon ile 2 küme elde edildi. Zscore değerleri alınarak normalize edilen verilerde iki küme arası uzaklık 4,094 olarak bulundu. Kümeye atanan değişkenlerin merkez ölçümleri bulundu. (Tablo

4.12). Şekil 4.12.'de iki küme arasındaki ölçümlerden elde edilen farklılıklar belirtildi. Tüm değerler Küme 1 ve 2 beklenen şekilde birbirlerine zıt olarak kümelendi. Küme 1 de artış gösteren değerler Küme 2 de azalış göstermekteydi. Kısaca Küme 1'de en yüksek değeri gösteren değişken Küme 2'de en düşük değeri göstermektedir. Yaş değişkeni incelendiğinde Küme 1 de ortalama olarak yaş değişkeni yükseldiğinden orta yaş bireyler bu kümeye dâhil iken genç bireyler Küme 2'ye atandı. Küme 1'de yüksek Küme 2'de ise düşük olarak bulunan değerler BASO#, BASO%, HCT, HGB, MCH, MCHC, MCV, MPV, NEUT#, NEUT%, PLCR, PDW, RBC ve RDWSD idi. Küme 1'de düşük iken Küme 2'de yüksek bulunan değerler ise LYMPH#, LYMPH%, PCT, PLT, RDWCV ve WBC idi. Tablo 4.13.'deki ANOVA sonuçlarına göre MCHC değeri dışındaki tüm hemogram ölçümleri kümelere yerleşirken farklılık göstermiştir.

Tablo 4.12. SPSS k-means Hemogram Değişkenlerinin Atandığı Küme Merkez Değerleri

	Kümeler			Kümeler	
	Küme 1	Küme 2		Küme 1	Küme 2
BASO#	30,09	-0,88	MONO%	1,07	-0,72
BASO%	1,17	-1,10	MPV	0,39	0,77
EO#	-0,33	8,09	NEUT#	-0,50	0,23
EO%	0,44	6,18	NEUT%	-0,33	-0,73
HCT	1,34	-0,23	PLCR	0,36	0,93
HGB	1,31	-0,29	PCT	-0,58	-0,46
LYMPH#	-0,26	-0,64	PDW	0,12	0,64
LYMPH%	0,07	-0,11	PLT	-0,64	-0,62
MCH	0,62	-0,75	RBC	0,92	0,42
MCHC	0,24	-8,59	RDWCV	-0,54	-0,29
MCV	0,58	-0,74	RDWSD	0,07	-1,07
MONO#	0,36	0,32	WBC	-0,46	1,20



Şekil 4.12. SPSS k-means Hemogram Değişkenlerinin Kümeler Arası Dağılımı

Tablo 4.13. SPSS k-means Hemogram Ölçümleri Anova Sonuçları

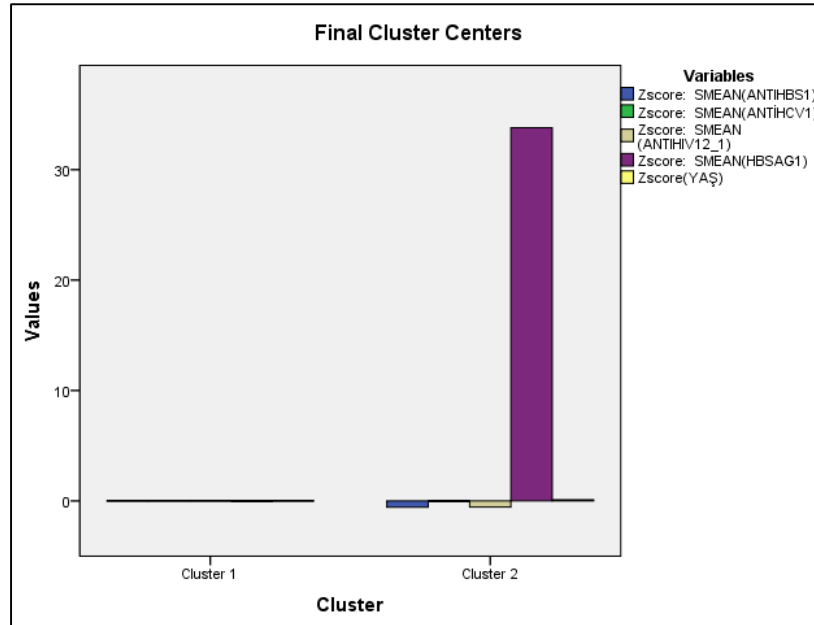
	ANOVA						
	Küme Kareler Ortalaması	sd	p		Küme Kareler Ortalaması	sd	p
BASO#	182,43	1	<0,001	MONO%	28,35	1	<0,001
BASO%	425,63	1	<0,001	MPV	941,69	1	<0,001
EO#	3,79	1	0,005	NEUT#	16,09	1	<0,001
EO%	8,58	1	0,003	NEUT%	593,74	1	<0,001
HCT	2112,88	1	<0,001	PLCR	972,70	1	<0,001
HGB	1766,69	1	<0,001	PCT	592,26	1	<0,001
LYMPH#	806,22	1	<0,001	PDW	979,59	1	<0,001
LYMPH%	762,31	1	<0,001	PLT	946,27	1	<0,001
MCH	1124,32	1	<0,001	RBC	281,14	1	<0,001
MCHC	0,337	1	0,561	RDWCV	104,19	1	<0,001
MCV	1386,63	1	<0,001	RDWSD	891,52	1	<0,001
MONO#	54,80	1	<0,001	WBC	166,69	1	<0,001

İmmünolojik ölçüm değerleri üzerinde 2 iterasyon ile 2 küme elde edildi. Zscore değerleri alınarak normalize edilen verilerin kümelerdeki ortalama değerleri

Tablo 4.14.'de sunuldu. İki küme arası uzaklık 33,837 olarak hesaplandı. Şekil 4.13.'de iki küme arasındaki ölçümlerden elde edilen farklılıklar belirtildi. Tüm değerler Küme 1 ve 2 beklenen şekilde birbirlerine zıt olarak kümelendi. Küme 1 de referans değer gösteren değerlerden ANTIHBS ve ANTIHIV, Küme 2 de azalış, Küme 1 referans değere yakın değer gösteren HBSAG değeri ise Küme 2'de artış göstermekteydi. Ayrıca bu değer kümelemeyi esas belirleyen değişkendi. Yaş değişkeni incelendiğinde Küme 1 ve Küme 2 arasında yaş değişkeni farklılık göstermedi. Tablo 4.15.'deki ANOVA sonuçlarına göre ANTIHCV1 ölçümü dışındaki tüm hemogram ölçümleri kümelere yerleşirken farklılık göstermiştir.

Tablo 4.14. SPSS k-means İmmünolojik Değişkenlerinin Atandığı Kümenin Merkez Değerleri

	Kümeler	
	Küme 1	Küme 2
ANTIHBS	-0,50	-0,57
ANTIHCV	-28,43	-0,02
ANTIHIV	0,00	-0,83
HBSAG	-0,06	48,92



Şekil 4.13. SPSS k-means İmmünolojik Değişkenlerinin Kümeler Arası Dağılımı

Tablo 4.15. SPSS k-means İmmünoloji Ölçümleri Anova Sonuçları

	ANOVA		
	Küme Kareler Ortalaması	sd	p
ANTIHB5	1,31	1	0,004
ANTIHCv	0,003	1	0,959
ANTIHIV	1,23	1	0,004
HBSAG	4573,51	1	<0,001

JASP ve SPSS ile yapılan analizler sonucunda k-means tekniğinin daha iyi sonuç verdiği görüldüğünden Python programlama dilinde k-means kümeleme tekniği analiz edildi. Analiz için Matplotlib ve k-means tekniği için ise Sckit Learn kütüphaneleri kullanıldı. K-means kodları kullanılarak ve grafiğe basılarak alınan sonuçlarda kolay analiz yapabilmek adına veri seti Sarı (Biyokimya), Beyaz (Hemogram) ve Mavi (İmmünoloji) olmak üzere renklere ayrıldı.

Biyokimya Python kodları;

- import numpy as np
- import pandas as pd
- import matplotlib.pyplot as plt
- from sklearn.metrics import silhouette_score
- from sklearn.preprocessing import StandardScaler
- from sklearn.cluster import KMeans
- scaler = StandardScaler()
- fte_colors = {
- 0: "#008fd5",
- 1: "#fc4f30",}
- features = pd.read_csv('diş hekimliği veri.csv')
- features = features.drop (columns=["ADSOYAD", "DOSYANO"])
- yellow = features.drop (features.columns[20:], axis=1)
- y = [[float(x.replace(',','.')) for x in yellow[d]] for d in yellow.columns]

- `yellow = np.array(y).T`
- `y_kmeans = KMeans(n_clusters=2)`
- `y_kmeans.fit(yellow)`
- `y_silhouette = silhouette_score(yellow, y_kmeans.labels_)`
- `fig, ax1 = plt.subplots(figsize=(8, 6), sharex=True, sharey=True)`
- `fig.suptitle(f"BIOKIMYA", fontsize=16)`
- `km_colors = [fte_colors[label] for label in y_kmeans.labels_]`
- `ax1.scatter(yellow[:, 0], yellow[:, 1], c=km_colors)`
- `ax1.set_title(`
- `f"K-MEANS           CLUSTERING\nSilhouette:        {y_silhouette}",`
- `fontdict={"fontsize": 12})`
- `plt.show()`

Hemogram Python Kodları;

- `white = features.drop(features.columns[1:20], axis=1)`
- `white = white.drop(white.columns[25:], axis=1)`
- `w = [[float(x.replace(',','.')) for x in white[d]] for d in white.columns]`
- `white = np.array(w).T`
- `w_kmeans = KMeans(n_clusters=2)`
- `w_kmeans.fit(white)`
- `w_silhouette = silhouette_score(white, w_kmeans.labels_)`
- `fig, ax1 = plt.subplots(figsize=(8, 6), sharex=True, sharey=True)`
- `fig.suptitle(f"HEMOGRAM", fontsize=16)`
- `fte_colors = {`
- `0: "#008fd5",`
- `1: "#fc4f30",}`
- `km_colors = [fte_colors[label] for label in w_kmeans.labels_]`
- `ax1.scatter(white[:, 0], white[:, 1], c=km_colors)`

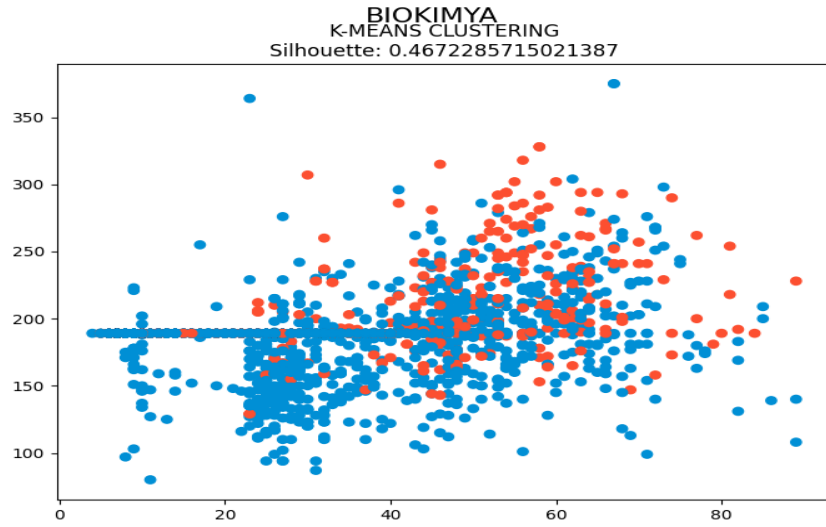
- ax1.set_title(
- f"K-MEANS CLUSTERING\nSilhouette: {w_silhouette}",
fontdict={"fontsize":12})
- plt.show()

İmmünoloji Python kodları;

- blue = features.drop(features.columns[1:44], axis=1)
- a = [[float(x.replace(',','.')) for x in blue[d]] for d in blue.columns]
- blue = np.array(a).T
- b_kmeans = KMeans(n_clusters=2)
- b_kmeans.fit(blue)
- b_silhouette = silhouette_score(blue, b_kmeans.labels_)
- # Plot the data and cluster silhouette comparison
- fig, ax1 = plt.subplots(figsize=(8, 6), sharex=True, sharey=True)
- fig.suptitle(f"IMMUNOLOJI", fontsize=16)
- km_colors = [fte_colors[label] for label in b_kmeans.labels_]
- ax1.scatter(blue[:, 0], blue[:, 1], c=km_colors)
- ax1.set_title(
- f"K-MEANS CLUSTERING\nSilhouette: {b_silhouette}",
fontdict={"fontsize": 12})
- plt.show()

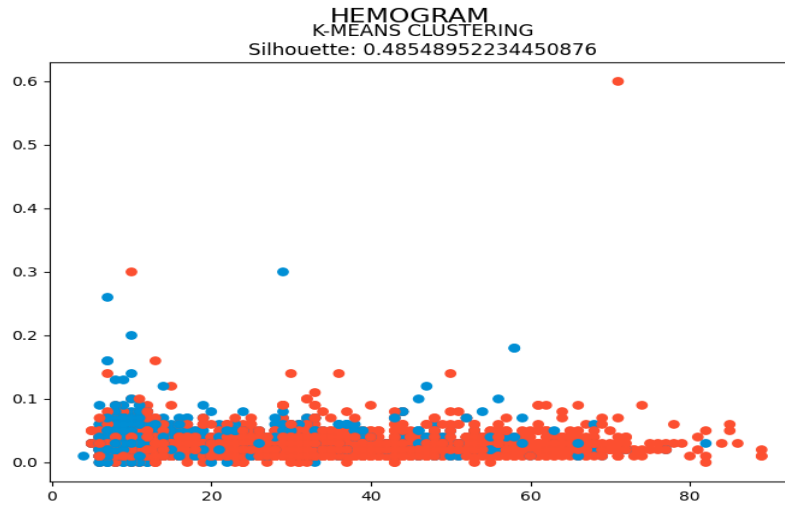
kodları kullanılarak aşağıdaki grafikler elde edildi.

Python ile elde edilen biyokimya ölçümlerine ait küme yapısı Şekil 4.14’te incelendiğinde model başarımlar ölçütü Silhouette değeri 0,467 olarak elde edildi. Grafik incelendiğinde ise alt kısımda kalan mavi noktalardan oluşan Küme 2 skoru düşük grubu oluşturduğundan “Hasta” grup, kırmızı noktalar halinde belirlenen Küme 1 ise “Sağlıklı” olarak etiketlendi. Noktaların keskin bir şekilde ayrılması ile küme içi noktaların merkeze uzaklığının kareler ortalamasının düşük olduğu anlaşılmaktadır.



Şekil 4.14. Python Biyokimya Viral Testlere Ait Küme Yapısı

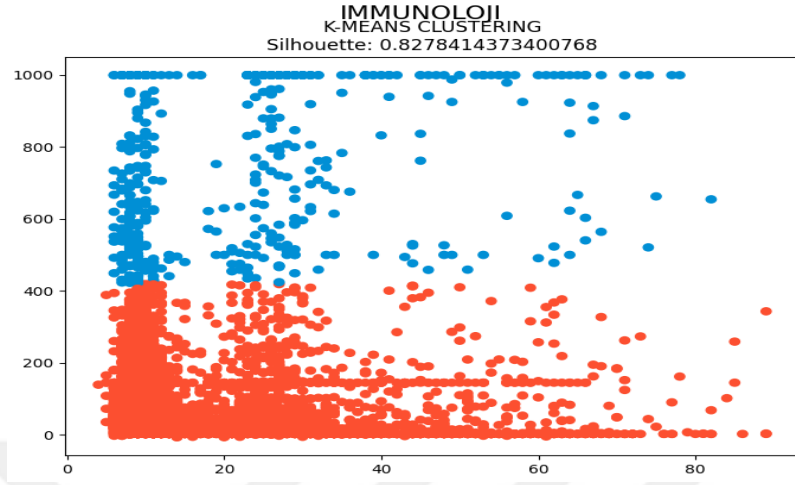
Bir diğer ölçüm olan hemograma ait küme yapısı ise Şekil 4.15'te incelendiğinde model başarımlar ölçütü Silhouette değeri 0,485 olarak elde edildi. Grafik incelendiğinde ise alt kısımda mavi noktalardan oluşan ve skoru daha düşük olması nedeniyle Küme 2 “Hasta” kırmızı noktalardan oluşan Küme 1 “Sağlıklı” olarak etiketlendi.



Şekil 4.15. Python Hemogram Viral Testlere Ait Küme Yapısı

Python ile elde edilen son ölçüm olan immünoloji ölçümlerine ait küme yapısı Şekil 4.16'da incelendiğinde model başarımlar ölçütü Silhouette değeri 0,827 olarak elde edildi. Grafik incelendiğinde ise alt kısımda kalan kırmızı noktalardan oluşan Küme 2 skoru düşük grubu oluşturduğundan “Hasta” grup, mavi noktalar halinde belirlenen Küme 1 ise “Sağlıklı” olarak etiketlendi. Noktaların keskin bir

şekilde ayrılması ile küme içi noktaların merkeze uzaklığının kareler ortalamasının yüksek olduğu ve kümelemenin oldukça iyi bir şekilde olduğu anlaşılmaktadır.



Şekil 4.16. Python İmmünoloji Viral Testlere Ait Küme Yapısı

4.3. Sonuçların Karşılaştırılması

Çalışmada toplam Ağız, Çene ve Diş Cerrahisi bölümüne başvuran 5148 hastaya ait kan tahlili sonuçları Hemogram, biyokimya ve immünolojik (mikrobiyolojik) olarak JASP, SPSS ve Python programlarında analiz edildi. Tüm analizlerde uzaklık ölçütü olarak Öklit uzaklığı seçildi. JASP ile Fuzzy C-Means ve k-means teknikleri karşılaştırıldı. Öncelikle JASP ile yapılan analizlerde sırasıyla biyokimya, hemogram ve immünolojik veriler analiz edildi. Her ölçümde k-means tekniği için keskin şekilde isimlendirilmiş “Hasta” ve “Sağlıklı” olarak 2 küme oluşturulurken, Fuzzy C-means için “Daha az Sağlıklı” ve “Daha Sağlıklı” biçiminde etiketlenmiş 2 adet küme elde edildi. Biyokimya ölçümlerinde verilerde K-means tekniğine göre iki küme oluşturuldu ve Silhouette değeri 0,480 olarak elde edildi. Açıklayıcılık katsayısı $R^2=0,498$ bulundu. Fuzzy C-means ile ise Silhouette değeri 0,470 olarak elde edildi. Açıklayıcılık katsayısı $R^2=0,465$ bulundu.

Hemogram ölçümleri için K-means tekniği Silhouette değerini 0,490 olarak verirken, $R^2= 0,572$ olarak bulundu. Aynı ölçümler Fuzzy C-means tekniği ile tekrarlandığında Silhouette değerini 0,500 olarak verirken, $R^2= 0,366$ olarak bulundu.

İmmünolojik verilerde K-means tekniğine göre iki küme oluşturuldu ve Silhouette değeri 0,830 olarak elde edildi. Açıklayıcılık katsayısı $R^2=0,801$ değeri ile oldukça yüksek bulundu. Aynı değişkenler, Fuzzy C-means kümeleme tekniği ile analiz edildiğinde Silhouette skorunun 0,700 olduğu görüldü. Açıklayıcılık katsayısı ise $R^2=0,281$ olarak hesaplandı. SPSS programı ile ise TwoStep Cluster ve K-means Teknikleri kullanıldı. Her üç ölçüm içinde TwoStep Cluster Tekniği Silhouette değeri “Good” değerindeydi. K-means algoritması ise değerlerin kümelerle atanma itibari ile farklılıklarını ortaya koydu. Bu nedenle K-means kümeleme tekniğinin diğer tekniklere göre daha iyi sonuç verdiği gözlemlendi. Küme sayısını seçme ve verileri etiketlere, konulara ve verilerin içeriğine göre birden fazla kategoride kümeleme şansı sunması nedeniyle çok standart bir sınıflandırma sağlayan K-means kümeleme tekniğinin oldukça uygun olduğu anlaşıldı. Bu nedenle k-means tekniği seçilerek Python’da veriler analiz edildi.

Python programlama dilinde elde edilen sonuçlara göre başarı ölçüm değeri olan Silhouette skoru biyokimya (0,467) ve hemogram (0,485) ölçümlerinde daha düşük elde edilmesine rağmen immünoloji (0,827) oldukça iyi bulundu. Elde edilen grafikler incelendiğinde ise grafik altında kalan küme düşük skoru ifade ettiğinde “Hasta”, grafik üstünde kalan küme ise “Sağlıklı” olarak etiketlendi. Ayrıca immünoloji ölçümlerinde kümelerin oldukça iyi bir şekilde ayrılması küme içi noktaların merkeze uzaklığının kareler ortalamasının yüksek olduğunu gösterdi.

5. TARTIŞMA

Marquezin ve arkadaşlarının 2019'da yapmış olduğu ‘Çocuklarda emme alışkanlıkları, yutulan gıdanın tutarlılığı ve tat duyarlılığı arasındaki ilişkinin kümeleme analizi’ isimli çalışmada emme alışkanlıkları ve yeme davranışı birbiriyle ilişkili olabileceğinden ve aynı zamanda tat, vücut kompozisyonu ve sosyodemografik faktörlerle ilişkili olabileceği belirtilmiştir. Çalışmaya başlangıçta sosyodemografik özellikler, emzirme süresi, geçmiş biberonla besleme ve besleyici olmayan emme alışkanlıkları, tat duyarlılığı, vücut kitle indeksi (BMI) ve alınan gıdaların tutarlılığı hakkında bilgi toplanan 352 çocuk (197 kız; 7-11 yaş arası) dâhil edilmiştir. Tükürük amilaz ve toplam protein konsantrasyonları da belirlenen çalışmada veriler kümeleme analizi (hiyerarşik analiz, K-means ve silhouette plot), tek yönlü ANOVA ve korelasyon testlerine sunulmuştur. Kümeleme analizi, 159 çocuğu içeren üç güvenilir ve anlamlı küme üretmiştir. Küme 1 “iyi tadımcılar”, olarak isimlendirilen kümedeki çocuklar yaş olarak daha büyük ve daha yüksek tat alma duyarlılığına sahiptir. Biberonla daha uzun besleme süresi geçiren küme 2’yi kapsayan çocuklar ise “daha yumuşak gıda tutarlılığı”, “emzirilen” olarak isimlendiren ve anne sütü emme süresi ile daha düşük tutarlılıkla yiyecek yiyen çocuklar ise 3. Kümeyi oluşturmuşlardır (siluet katsayısı = 0.61). Ayrıca küme 3, normal doğum ile dünyaya gelmiş, normal kilolu ve daha düşük oranda geçmişte beslenme dışı emme alışkanlığına sahip çocukları yüksek oranda kapsamaktadır. Tükürük amilazı ile toplam protein konsantrasyonları, tat duyarlılığı ve BMI arasında korelasyon saptanmamıştır. Ancak tat duyarlılığı kız çocuklar arasında daha yüksek bulunmuştur ($p = 0.002$; güç = % 88). Çalışmada, beslenme emme alışkanlıkları, yutulan gıdaların tutarlılığı ve sosyodemografik özellikler arasındaki ilişkiyi vurgulayan üç küme tanımlanmıştır. En önemli sonuç, biberonla besleme ile yumuşak gıda kıvamı tüketimi arasında yakın bir ilişkide gözlemlenmiştir (92).

Dünya çapında anestezi bakımı ve çalışmaları hakkında bir anket yardımı ile profesyonel anestezistlerin görüşlerine başvurulmuş bu çalışmada, bir grup Brezilyalı anestezistin azot oksit / oksijen ve diş hekimleri tarafından yapılan oral minimal / orta sedasyon ile ilgili düşünceleri alınmıştır. K-means kümeleme yöntemi Student t-testi ile analiz edilen çalışmada cinsiyet, ikamet, programının tamamlanmasından bu

yana geçen süre ve diş tedavisi için sedasyon ve genel anestezi sağlama deneyim bilgileri toplanmıştır. Çoğu anestezi uzmanı (% 85,6) diş tedavisi için nadiren veya hiç sedasyon veya genel anestezi sağlamamış ve % 92,8'i diş hekimlerinin diş hekimliği ofisinde orta derecede sedasyon uygulayabileceği ifadesine katılmamıştır. Daha olumlu görüşler (n:21) ve daha az olumlu görüşler (n:90) olarak alınan hekim görüşlerini temsil eden iki küme oluşturulmuştur. “Daha az elverişli” gruptaki anestezistler, “daha elverişli” gruptakilere göre dental sedasyon (P=0.006) ve dental genel anestezi (P=0.008) konusunda daha fazla deneyime sahip bulunmuş, bunun yanında ikamet, program süresinin tamamlanmasından bu yana geçen süre ve cinsiyet anestezistlerin görüşlerini önemli ölçüde etkilememiştir. Orta Batı Brezilya'daki dental sedasyonun bu bölgede hala açıklığa kavuşturulması gereken bir konu olduğu düşünülmekte; hekim ve diş hekimlerinin nüfusa fayda sağlayacak rollerinin belirlenmesi için öneri yapılmaktadır (102).

2020 yılında ‘Filtrek: Fiziksel Aktivite İçin K-Means Kümeleme Yöntemi ile Yeni Bir Aykırı Değer Algılama Yöntemi’ isimli çalışmada aykırı etkiyi azaltarak K-means türevli fiziksel aktivite kümesinin saflığını iyileştirmek için Filtre olarak gösterilen yeni bir algoritma önerilmiştir. Vücuda takılan ivmeölçerlerle elde edilen fiziksel aktivite verilerine K-means kümeleme analizi uygulanmıştır. Local Outlier Factor, Isolation Forests and KNN üç aykırı değer tespit yöntemi ile ortalama küme ve olay saflığı (ACEP) ölçülmüştür. FilterK, ACEP'de (0.581 → 0.596 ile 0.580–0.617'ye kıyasla) karşılaştırılabilir kazançlar sağlarken, diğer yöntemlere göre daha az sayıda aykırı değer çıkarmıştır. Bu yeni aykırı değer tespit yönteminin ana odağı, fiziksel aktivite ivmeölçer verilerinin küme saflıklarını iyileştirmektir. Fakat potansiyel olarak K-means kümelemesi ile yapılacak diğer veri kümesi türlerine de uygulanabileceğini önerilmiştir (131).

Bu çalışmanın amacı, morbid obezite tedavisi arayan 19-68 yaş arası 113 bireyde (92 kadın) ağız sağlığı ve beslenme ile ilgili kalıpları meyve, sebze, şeker ve işlenmiş gıdaların tüketimine ek olarak sosyodemografik değişkenler ve tıbbi kayıt değişkenleri ile belirlemektir. Bunların yanı sıra vücut kitle indeksi, boyun, bel ve kalça ölçüleri, çürük deneyimi (DMFT indeksi), Toplum Periodontal Endeks (TÜFE indeksi) ve tükürük fizikokimyasal verilerde toplanmıştır. K-means yöntemi ile elde edilen 3 küme arası değişkenler arası ilişkiler ANOVA ile belirlenmiştir. 1. Küme

yeme,/gün, gün önce şekerli içecekler ve işlenmiş gıda tüketen öğün sayısı yüksek beden kitle indeksi ile genç katılımcılar tarafından karakterize edilmiştir ve “Genç” ismi ile etiketlenmiştir. Küme 2 “Diyabetik bireyler” etiketli, en yüksek orana sahip yaşlı katılımcılar tarafından kategorize edilmiş olup, diyabetik katılımcıların (% 100'ü diyabetik;% 73 insülin kullanıcısı), daha düşük BMI, daha yüksek DMFT indeksi ve OHIP-14 ve xerostomia skorlarına sahiptirler ve bir gün önce meyve ve sebze tükettiklerini bildirmişlerdir. Son olarak, Küme 3 “Zayıf periodontal sağlık” etiketli, daha kötü periodontal durumu (daha yüksek TÜFE) ve daha düşük tükürük akışı, pH ve tampon kapasitesi olan katılımcılar ile karakterize edilmiştir (132).

K-araç kümeleme algoritmasındaki bu kusuru çözmek için çeşitli hibrit yöntemler tanıtılmıştır. Bu yöntemleri çeşitli açılardan karşılaştıran hiçbir karşılaştırmalı çalışma bulunmadığından, bu makalede genetik algoritmalar, minimum yayılan ağaç ve hiyerarşik kümeleme yöntemi kavramlarını kullanan K-ortalamlar kümeleme algoritması ile üç hibrit yöntemi karşılaştırmıştır. Bu üç melez yöntem önceki araştırmalarda daha fazla dikkat çekmiş olsa da, sonuçlarını karşılaştıran daha az çalışma vardır. Bu nedenle, bu çalışmada örneklem büyüklüğü, sayı özellikleri ve sınıf sayısı açısından farklı özelliklere sahip yedi nicel veri seti kullanılmıştır. Veriler, hibrit yöntemlerin, nihai çözümü elde etmede sıradan K-ortalamlar yönteminden daha yüksek yakınsama oranı ile sonuçlandığını göstermiştir. Ayrıca hiyerarşik kümeleme algoritmasına sahip hibrit yöntemler, diğer iki hibrit yöntemle göre daha az iterasyonla optimal çözüme yakınsar. Hiyerarşik kümeleme algoritmasına sahip hibrit yöntemler, diğer iki hibrit yöntemle göre daha az iterasyonla optimal çözüme yakınsar. Bu nedenle, hesaplama karmaşıklığına rağmen, bu üç hibrit yöntem, K-ortalamlar yönteminde çok fazla gelişmeye yol açmamıştır. Ancak, yöntemleri karşılaştırmak ve sonucu tamamlamak için bir simülasyon çalışmasına ihtiyaç vardır (114).

Diş çürükleri diş minesinde ve sonra dentinde demineralizasyona sebep olan en yaygın ağız hastalığıdır. Demineralizasyon sürecindeki tedavi planında yardımcı olmada “depth-wise” değerlendirme kullanılan bu çalışmada, Raman probu geliştirmek için geniş kompozisyon haritalardan dental dokulardaki mineral bileşim karakterize ederek ve Raman mikrospektroskopi kullanarak temel bilgi sağlamak amaçlanmıştır. Farklı evrelerde diş çürükleri olan 10 farklı yirmilik diş incelenmiş

olup, tüm dişler diş çürükleri bölgelerine göre yarıya kesilmiştir ve Raman görüntüleme mikroskopuyla incelenmiştir. K-means kümeleme analizi ile ön-süreç spektral haritalar tek bir data matrix'e dönüştürülmüştür ve K-means kümeleme analizi ile diş minesini, dentin ve diş çürüklerinin spektrumları belirlenmiştir. Sonuç olarak, K-means kümelemesi diş çürüklerinin denetlenmemiş tanımlama yapılmasının görülebileceğini göstermiştir. Çalışma, aktif diş çürükleri ve lezyonlarına erken tanı koymak ve tespit etmek amaçlı temel metodolojik gelişimleri sunmaktadır (134).

Ulusal Düzeyde HIMS raporları temel alınarak yapılan bu çalışmada Gesicho ve arkadaşları "HIV Göstergesi Verilerinin İzleme Tesisi Raporlamasında Kümelendirme: Kenya Örneği" adlı çalışmada Kenya HIV ulusal raporlama datasının örneği için kullanılan k-means kümeleme algoritma yaklaşımı tesis raporlamasını ve veri kaynaklı karar vermeyi desteklemektedir. Sonuçlar, genel tesis performansının değerlendirmesine ve bu performansın derecelendirmesinde kullanılabilecek 4'lü homojen kümeleme kategorileri ortaya koymaktadır. Bu çalışma yönetim tarafından sağlıklı tesis raporlama performansının ilerlemesini izlemek için ve M&E takımları entegre edilmiş data kaynaklarından büyük miktarda data kullanmaları için iyi bir ana hatır (138).

Doğanın biyolojik veriler üzerindeki etkisi üzerine yapılan çalışmada önerilen Matlab uyumlu SIML Kümeleme Algoritması ile mikro dizi ve SNP data setleri ile test edilen çalışmada klasik kümeleme algoritması ile karşılaştırılmıştır. MLL data setleri üzerinde SML en yüksek kümeleme doğruluğunu başarmış ve rand puanı 4/7 bulunmuştur. Benzer olarak SRBT data setleri üzerinde 3/5, bütün alt türlerin üzerinde 5/7 ve en yüksek skor 4/7 olarak elde edilmiştir. Ayrıca SNP Data kullanılarak 3 kümeleme probleminin SML genel kümeleme üzerinde doğruluk oranı sırası ile 97,3 94,7 ve 100'dür. Maximum likelihood kümeleme yaklaşımı kullanımını öneren bu metodu kullanmanın amacı sadece mesafe bilgisi değil aynı zamanda kümeleme için varyans bilgilerini de dâhil etmektir. Bu metod diğerlerine göre daha verimli ve kullanılabilir görülmüştür (142).

"İstatistiksel aykırı değer tespit yöntemi kullanılarak karar grafiğinden küme centroidlerinin otomatik olarak tanımlanması" isimli 2019'da Yann ve diğerleri

tarafından hazırlanan yayında karar grafiğinden küme merkezlerinin tanımlanması için hala otomatik yöntemler eksikliğinin varlığı vurgulamış ve bu doğrultuda çalışmıştır. Önerilen yöntemde, karar grafiğindeki belirli yoğunluk değerlerindeki tek boyutlu olasılık yoğunluk fonksiyonları, iki boyutlu tahmin çekirdek fonksiyonları kullanılarak tahmin edilmektedir. Daha sonra küme merkezleri belirli yoğunluk değerlerinde hesaplanan beklenti değerleri ve standart sapmalar kullanılarak karar grafiğinde aykırı değerler olarak otomatik olarak tanımlanır. Çeşitli sentetik ve gerçek dünya veri kümeleri üzerinde yapılan deney, çeşitli dağılımlar ve boyutlar ile veri kümesinden ağırlık merkezi tanımlamasında önerilen yöntemin üstünlüğünü göstermektedir. Ayrıca önerilen yöntemin görüntü bölümlmeye etkin bir şekilde uygulanabileceği de gösterilmiştir (123).

Demircioğlu ve Eşiyok tarafından 2019’da yapılan “Covid-19 Salgını İle Mücadelede Kümeleme Analizi İle Ülkelerin Sınıflandırılması” çalışmada amaç, 2019 yılının son dönemlerinde gündeminin en önemli konusu haline gelen COVID-19 salgının ortaya çıkardığı sonuçları ülkeler indeksinde incelemektir. Araştırmaya OECD ve AB üyesi olan 36 ülkenin salgın dönemi sağlık verileri dâhil edilmiş ve benzerlik gösteren ülkeler tespit edilmiştir. Türkiye’nin de çalışmaya dâhil olan ülkeler arasındaki yeri belirlenmeye çalışılmıştır. K-ortalamlar ile yapılan veri analizi WEKA yardımıyla çözümlenmiş ülkeler ikili, üçlü ve dörtlü kümelere ayırmıştır (137).

Türkiye’deki gelişmişlik düzeyleri farklı olan illerden illeri gelişmişlik düzeyine sahip illerin sağlık hizmetleri seviyeleri belirlenerek aralarındaki farklılıkların incelendiği çalışmada Gençöglü, kümeleme analizi yöntemlerinden Ward tekniğini kullanmıştır. 31 adet sağlık değişkeninin alındığı veriler T.C. Sağlık Bakanlığı’nın hazırladığı ‘Sağlık İstatistikleri Yıllığı’ndan temin edilmiştir. Sonuç olarak sosyoekonomik gelişmişlik düzeyi yüksek iller ile sağlık hizmetleri arasında pozitif bir ilişki bulunmuştur (120).

2016’da Çam ve Bircan tarafından hazırlanan çalışmada amaç, hastaların verilerinin bulunduğu çok boyutlu bir veriden kümeleme analizi kullanarak hasta başvuru davranışlarının belirlenmesidir. K-Means ve Yoğunluk Tabanlı Kümeleme tekniği kullanılarak ve ayrıca sayısal verilerin yanında hastaların demografik verileri

de işleme dâhil edilerek, hastalar hakkında daha somut yargılara ulaşılması amaçlanmıştır. Yapılan analizde K-Means Kümeleme yöntemine karşın Yoğunluğa Dayalı olan kümeleme yönteminin daha uygun sonuçlar verdiği görülmüştür (139).

Turhan ve arkadaşlarının 2018 yılında “Potasyum (K) ve Hemoliz İndeksi (HI) Arasındaki İlişkinin Kümeleme Analizi Yöntemi ile İrdelenmesi” çalışmasında hastaların ölçümlerinden elde edilen potasyum (K) ve hemoliz indeksi (HI) arasındaki ilişki Kümeleme Analizi ile incelenmiştir. Buna göre, K ve HI farklarının mutlak değerleri kullanımı, bulunan kümelerde mutlak K ve HI farkları arasındaki korelasyonlara bakılmıştır. K-Means algoritması ve Euclidian mesafe ölçüm tekniği kullanılarak analiz sonuçları elde edilmiştir. Sonuçlara göre HI farkının 295 mg/dL’ den yüksek görülmüş ve kümelerde de oldukça korelasyona rastlanmıştır. Hemolize örneklerde testlerin tekrarına engel teşkil edecek düzeltme faktörü çalışmalarında kullanılması amacı ile kümeleme analizinin bir ön çalışma olarak kullanılabileceği önerilmiştir (124).

SONUÇLAR

Veri madenciliği günümüz analiz yöntemlerinin oldukça önemli bir kısmını oluşturmaktadır. Birçok branşı içine alan bu yöntem multidisipliner bir yöntem olmakla beraber gün geçtikçe artan ve analizi zorlaşan büyük ve karmaşık veriler için oldukça uygundur. Her geçen gün veri madenciliği ile ilgili çalışmalar artarak sürmektedir. Özellikle işletmeler ve finansal yönde kullanılan veri madenciliği son dönemlerde sağlık alanında da oldukça yaygınlaşmaktadır. Yöntemin uygulanacağı birçok yazılım, program ve açık kaynak kodlu uygulamaların olması da kullanımını kolaylaştırmaktadır. Veri madenciliği yöntemlerinin en önemlilerinden birisi de Kümeleme Algoritmalarıdır. Büyük veri kümeleri uğraş alanı olan bu yöntem verileri kolayca ilgili kümelere uygun teknikler kullanarak atamakta ve birçok alanda doğru ve kullanışlı sonuçlar elde etmeyi sağlamaktadır. Klasik istatistiksel yöntemlerde düşük sayıda veri ve değişken ile işlem yapmak gerekir. Bunun yanı sıra klasik yöntemler, birçok varsayım da gerektirmektedir. Veri madenciliğinin oldukça büyük veri kümeleri ile çalışabilmesi, çok ve çeşitli değişkenlerle işlem yapması ve varsayım gerektirmemesi gibi özellikleri tercih nedenidir. Ek olarak hipotezlere ihtiyaç duyulamaması da önemli bir özelliğidir. Bunların yanı sıra veri madenciliği uzun ve zorlu bir süreçtir. Veri sayılarının oldukça geniş ve büyük olması özellikle veriyi analize hazırlama aşamasında zaman açısından sorun yaratabilmektedir. Analiz aşamasında uygun model seçimi, raporlama aşamasında ise kolay anlaşılır grafik ve tabloların yapılması açısından da süreci zorlaştırmaktadır.

Sağlık alanında otomasyon sistemlerinin yoğunlaşması ile hasta ve hastanelerden elde edilen bilgiler her geçen gün artmaktadır. Sağlık sektörü büyük veri üretme açısından oldukça önemlidir. Genel olarak hastalık tedavileri, kanser araştırmaları, görüntüleme teknikleri ve en çok kullanımın görüldüğü alan olan hastane yönetiminde maliyet ile ilgili konularda veri madenciliği ve kümeleme yöntemlerinin oldukça fazla kullanıldığını görmekteyiz. Bu tez çalışmasında Ağız, Çene ve Diş Cerrahisi bölümüne başvuran 5148 hastaya ait kan tahlili sonuçları değerlendirildi. Öncelikle veri madenciliği süreçlerine uygun olarak outliner uygulaması ile aykırı ve gürültülü veriler temizlendi. Daha sonra ise regresyon tahmin yöntemi ile boş gözeler doldurularak veriler analize uygun hale getirildi.

Veriler imm nolojik, hemogram ve biyokimya deęiřkenleri olarak k melendi.    farklı program ile analiz edilen verilerde JASP, SPSS ve Python uygulamaları kullanıldı. Fuzzy C-means, TwoStep ve K-means k meleme teknikleri kullanılarak karřılařtırmalar yapıldı. Alınan sonu lara g re yař deęiřkenine g re gen  ve orta yař bireyler, kullanılan  l  m deęiřkenlerine g re ise saęlıklı ve hasta řeklinde etiketlenen gruplar belirlendi. Referans deęerler ile kıyaslanan  l  mlerde hangi deęerlerin k melere g re artıř veya azalıř g sterdięi belirtildi. Yapılan analizler sonucunda k-means teknięinin daha iyi sonu lar verdięi g r ld . Bu  alıřma, bug ne kadar aęız, diř ve  ene cerrahisinde  alıřılmamıř imm nolojik, biyokimya ve hemogram  l  mleri kullanılarak ‘‘Hasta’’ ve ‘‘Saęlıklı’’ gruplar elde etme amacı ile yapılan bir  alıřmadır. Aęız, diř ve  ene cerrahisi alanında genel olarak tedavi y ntemleri, anestezi uygulamaları ya da hekim g r řleri alınarak yapılan  alıřmalar mevcuttur. Bu  alıřmaların bir oęu ise yurt dıřında yapılan  alıřmalardır. Bu nedenle, b y k veri k mesine sahip olan arařtırmacılara k meleme analizini  nermekteyiz. Veri madencilięi verinin hazırlanması adımları dikkatli bir řekilde izlendikten sonra b y k verilerde k mele analizi ve k-means teknięinin kullanımının faydalı olduęunu vurgulamaktayız.  zellikle iyileřme a ısından zorlu bir s re  ve cerrahi operasyonlar gerektiren aęız, diř ve  ene cerrahisi iřlemlerinde, k meleme analizinin, kandan elde edilen deęerlerle diř tedavisi  ncesi imm nolojik ve hormonal rahatsızlıkların belirlenmesini kolaylařtıran bir y ntem olduęunu d ř nmekteyiz.  alıřmada kullandığımız řekilde  ok sayıda deęiřken ile iřlem yapabilmesi nedeniyle k meleme analizi, kolay kullanım, iterasyon ve k me sayısı se im kolaylıęı saęlayan k-means teknięi saęlık alanında elde edilen veri k meleri i in olduk a uygun y ntemler olarak g r lmektedir.

KAYNAKÇA

1. Köse İ. Veri Madenciliği Teori Uygulama Felsefesi. 1st ed. Silahtaroglu G, editor. İstanbul, Türkiye: Papatya Bilim Üniversite Yayıncılığı; 2018. 51 p.
2. Ackoff R. From Data to Wisdom. J Appl Syst Anal [Internet]. 1989;16(1):3–9. Available from: [https://www.scirp.org/\(S\(351jmbntvnsjt1aadkposzje\)\)/reference/ReferencesPapers.aspx?ReferenceID=713373](https://www.scirp.org/(S(351jmbntvnsjt1aadkposzje))/reference/ReferencesPapers.aspx?ReferenceID=713373)
3. Frankfield J, Banton C. Investopedia [Internet]. Data Science. Available from: <https://www.investopedia.com/terms/d/data-science.asp>
4. Cleveland WS, Cleveland WS. Data Science: an Action Plan for Expanding the Technical Areas of the Field of Statistics. Int Stat Rev. 2001;69(1):21–6.
5. Gandomi A, Haider M. Beyond the hype: Big Data Concepts, Methods, and analytics. Int J Inf Manage [Internet]. 2015;35(2):137–44. Available from: <http://dx.doi.org/10.1016/j.ijinfomgt.2014.10.007>
6. Maptriks [Internet]. 2020. p. 1. Available from: <https://maptriks.com/tr/>
7. Gahi Y, Guennoun M, Mouftah HT. Big Data Analytics: Security and Privacy Challenges. 2016 IEEE Symp Comput Commun (ISCC. 2016;50–3.
8. Katal A, Wazid M, Goudar RH. Big Data: Issues, Challenges, Tools and Good Practices. 2013 Sixth Int Conf Contemp Comput. 2013;404–9.
9. Mantha B. Five Guiding Principles for Realizing the Promise of Big Data. Bus Intell J. 2014;19(1):8–11.
10. Usgurlu B, Top ÖÖ, Ungan E, Demirörs O. A Clustering Based Functional Similarity Measurement Approach. Proc - 36th EUROMICRO Conf Softw Eng Adv Appl SEAA 2010. 2010;371–5.
11. Özkan Y. Veri Tabanı Sistemleri. 4. baskı. İstanbul: Papatya Bilim Üniversite Yayıncılığı; 2015. 252 p.
12. Özkan Y. Veri Madenciliği Yöntemleri. 3. Baskı. Çölkesen R, editor. İstanbul, Türkiye: Papatya Bilim Üniversite Yayıncılığı; 2016. 240 p.
13. Çıngır H. Veri Madenciliğine Giriş. 2007.
14. Inmon WH. Developing the Data Warehouse. 3. Baskı. New York, NY, United States: John Wiley & Sons Publication; 1992. 312 p.
15. Hoffer JA, Prescott MB, McFadden FR. Modern Database Management. In: 8. Edition, editor. Prentice Hall; 2001.
16. Bonifati A, Cattaneo F, Ceri S, Fuggetta A, Paraboschi S. Designing Data Marts for Data Warehouses. ACM Trans Softw Eng Methodol. 2001;10(4):452–83.
17. Küçüksille EU. Veri Madenciliği ve Uygulama Programları [Internet]. Isparta, Türkiye; 2010. Available from: VeriTabaniYonetimSistemleriDersNotlari.org.tr/ab06/ozet/61.html, [Ziyaret Tarihi:29.05.2021]

18. Shearer C, Watson HJ, Grecich DG, Moss L, Adelman S, Hammer K, et al. The CRISP-DM model: The New Blueprint for Data Mining. J Data Warehous [Internet]. 2000;5(4):13–22. Available from: www.spss.com/Cnwww.dw-institute.com
19. Akpınar H. Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği. İstanbul Üniversitesi İşletme Fakültesi Derg. 2000;29(1):1–22.
20. Dasu T, Jahnson T. Exploratory Data Mining and Data Cleaning. New Jersey, USA: John Wiley & Sons Publication; 2003.
21. Oğuzlar A. Veri Ön İşleme. Erciyes Üniversitesi İktisadi ve İdari Bilim Fakültesi Derg. 2003;(21):67–76.
22. Şentürk A. Veri Madenciliği Kavram ve Teknikler. Bursa: Etkin Kitabevi; 2006.
23. Tüzüntürk S. Veri madenciliği ve istatistik. Uludağ Üniversitesi İktisadi ve İdari Bilim Fakültesi Derg. 2010;29(1):65–90.
24. Larose DT, Larose CD. Discovering Knowledge in Data: An Introduction to Data. 2. Edition. New York, NY, United States: John Wiley & Sons Publication; 2014. 336 p.
25. Benoît G. Data mining. Annu Rev Inf Sci Technol. 2002;36:265–310.
26. Larose DT. Discovering Knowledge in Data: An Introduction to Data. In: Discovering Knowledge in Data: An Introduction to Data. 2. Edition. USA; 2005. p. 30–6.
27. Cespivová H, Rauch J, Svatek V, Kejkula M, Tomeckova M. In Knowledge Discovery and Ontologies (KDO) at ECML/PKDD. ECMLPKDD04 Work Knowl Discov Ontol KDO'04 Pisa [Internet]. 2004;1–12. Available from: <http://nb.vse.cz/~svatek/kdo04cespivova.pdf>
28. Amanet H. CRISP-DM Nedir? [Internet]. 2020. Available from: <https://hasan-amanet.medium.com/crisp-dm-nedir-767b697caeae>
29. Çınar H, Arslan G. Proceedings of 17th Statistics Research Symposium. 3245th ed. Çankaya, ANKARA: TÜİK; 2008. 304–314 p.
30. Tekerek A. Veri Madenciliği Süreçleri ve Açık Kaynak Kodlu Veri Madenciliği Araçları. Ankara, Türkiye; 2011.
31. Berardi M, Appice A, Loglisci C, Leo P. Supporting visual exploration of discovered association rules through multi-dimensional scaling. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics). 2006;4203 LNAI:369–78.
32. <https://www.mysoft.com.tr/>. Mysoft. Mysoft Dijital Dönüşüm A.Ş. p. 1–3.
33. Koyuncu AS, Özgülba N. Veri Madenciliği: Tıp ve Sağlık Hizmetlerinde Kullanımı ve Uygulamaları. Int J Informatics Technol. 2009;2(2):21–32.
34. Dener M, Dörterler M, Orman A. Açık Kaynak Kodlu Veri Madenciliği Programları : WEKA ' da Örnek Uygulama. Akademik Bilişim'09 - XI. Akademik Bilişim Konferansı Bildirileri. Şanlıurfa, Türkiye; 2009.

35. Sağiroğlu Ş, Bensghir TK, Doğdu E, Kılıç K, Saran AN, Köse İ. Büyük Veri ve Açık Veri Analitiği: Yöntemler ve Uygulamalar. 1. Baskı. Şeref S, Koç O, editors. Ankara,Türkiye: Grafiker Yayınları; 2017. 344 p.
36. Chen M, Mao S, Liu Y. Big data: A survey. Mob Networks Appl. 2014;19(2):171–209.
37. Manyika J, Chui Brown M, B. J. B, Dobbs R, Roxburgh C, Hung Byers A. Big data: The next frontier for innovation, competition and productivity. McKinsey Glob Inst [Internet]. 2011;(June):156. Available from: https://bigdatawg.nist.gov/pdf/MGI_big_data_full_report.pdf
38. Laney D. 3D Data Management: Controlling Data Volume, Velocity, and Variety. Meta Gr Appl Deliv Strateg. 2001;19:171–209.
39. Zikopoulos P, Eaton C, Deroos D, Deutsch T, Lapis G. What Is Big Data? Hint: You’re a Part of It Every Day. Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data. 2012. 5–13 p.
40. Meijer E. The world according to LINQ. Commun ACM. 2011;54(10):45–51.
41. Beyer M. Gartner says solving big data challenge involves more than just managing volumes of data [Internet]. Gartner. 2011. Available from: <http://www.gartner.com/it/page.jsp>
42. Gobble MM. Big Data : The Next Big Thing in Innovation Big Data : The Next Big Thing in Innovation. Res Manag. 2013;56(1):64–7.
43. Gantz J, Reinsel D. Extracting Value from Chaos. IDC Anal Futur. 2011;3(6):1–12.
44. V M-S, Cukier K. Big data: a revolution that will transform how we live, work, and think. Dolan E, Houghton MH, editors. 2013.
45. Team OR. Big data now: current perspectives from OReilly Radar. OReilly Media; 2011.
46. Seker SE. Sosyal Ağlarda Akan Veri Madenciliği. YBS Ansiklopedisi. 2014;1(3):21–5.
47. Eynden VV, Corti L, Bishop L. Managing and Sharing Research Data: A Guide to Good Practice. Essex SAGE; 2014.
48. Seker SE. Büyük Veri ve Büyük Veri Yaşam Döngüleri. YB. 2015;2(3):10–7.
49. Karaibrahimoğlu A. Veri Madenciliğinden Birliktelik Kuralı İle Onkoloji Verilerin Analiz Edilmesi: Meram Tıp Fakültesi Onkoloji Örneği. Selçuk Üniversitesi, KONYA; 2014.
50. Oğuzlar A. Veri Madenciliğinde Birliktelik Kuralları. Uludağ Üniversitesi İİBF Derg. 2004;315–21.
51. Alaybeg F. Veri Madenciliği Giriş, Yöntemleri ve Metodolojileri [Internet]. Big Data Enthusiast | Machine Learning | Data Science. 2019. p. 1–5. Available from: <https://furkanalaybeg.medium.com/veri-madenciligi-ve-yontemleri-d0e2fd238e44>

52. Tutorial D. Problem Understanding [Internet]. 2004. p. 1. Available from: http://dms1.irb.hr/tutorial/tut_prob_understand.php
53. Faloutsos C. Searching multimedia databases by content. Boston, USA: Kluwer Academic Publishers; 1996.
54. Inselberg A, Dimsdale B. Parallel Coordinates: A Tool for Visualizing Multidimensional Geometry. *Proc IEEE Vis.* 1996;361–75.
55. Keim DA, Kriegel HP. VisDB: Database Exploration Using Multidimensional Visualization. *IEEE Comput Graph Appl.* 1994;14(5):40–9.
56. Nisbet R, Elder J, Miner G. Handbook of Statistical Analysis and Data Mining Applications. Academic Press; 2009.
57. Hatipoğlu B. Derhane Eğitiminin Üniversiteye Yerleşmedeki Etkisinin Veri Madenciliği ile İrdelenmesi. İstanbul Aydın Üniversitesi; 2013.
58. Frawley WJ, Piatetsky-Sahpiro G, Matheus CJ. Knowledge Discovery Databases: An Overview, in *Knowledge Discovery in Databases*. AAAI 58 AI Mag. 1991;1–27.
59. Agrawal R, Srikant R. Mining Sequential Patterns. 11th Int Conf Data Eng Taipei, Taiwan. 1995;3–14.
60. Zhu H. On-Line Analytical Mining of Association Rules. Simon Fraser University, Ottawa, Canada; 1995.
61. Özçakır FC, Çamurcu AY. Birlikte kuralı Yöntemi için Bir Veri Madenciliği Yazılımı Tasarımı ve Uygulaması. İstanbul Ticaret Üniversitesi Fen Bilim Derg. 2007;6(12):21–37.
62. Agrawal R, Imielinski T, Swami A. Mining Association Rules between Sets of Items in Large Databases. *Proc ACM SIGMOD Int Conf Manag Data (ACMSIGMOD '93)*, Washington, USA. 1993;207–16.
63. Houtsma M, Swami A. Set-Oriented Mining for Association Rules in Relational Database. *Proc 11th IEEE Int Conf Data Eng Taipei, Taiwan.* 1995;25–34.
64. Agrawal R, Srikant R. Fast Algorithms for Mining Association. *Proc 20th Int Conf Very Large Databases (VLDB '94)*, Santiago, Chile. 1994;487–99.
65. Savasere A, Omiecinski E, Navathe S. An Efficient Algorithm for Mining Databases Association Rules in Large. *Proc 20th Int Conf VLDB.* 1995;432–44.
66. Das A, Ng WK, Woon YK. Rapid Association Rule Mining. *Proc Tenth Int Conf Inf Knowl Manag ACM Press Atlanta, GA, USA.* 2001;487–99.
67. Zaki MJ, Hsiao CJ. CHARM: An Efficient Algorithm for Closed Itemset Mining. 2nd SIAM Int Conf Data Min (SDM'02), Eds Grossman, R L, Han, J, Kumar, V, Mannila, H ve Motwani, R, Siam, Arlington, VA, USA. 2002;457–73.
68. Karypis G, Han E-HS, Kumar V. CHAMELEON: A Hierarchical Clustering Algorithm Using Dynamic Modeling. *Computer (Long Beach Calif).* 1999;68–75.

69. Ramkumar GD, Swami A, America H. Clustering Data Without Distance Functions. IEEE Data Eng Bull. 1998;9–14.
70. Nikam SS. A Comparative Study of Classification Techniques in Data Mining Algorithms. Orient J Comp Sci Technol. 2015;8(1):13–9.
71. Chung-Mei Z, Luan J. Data Mining: Going Beyond Traditional Statistics. New Dir Institutional Res. 2006;131:7–16.
72. Ganesh S. Data Mining: Should it be included in the ‘Statistics’ curriculum? Sixt Int Conf Teach Stat Cape Town, South Africa. 2002;7–12.
73. Kuonen D. Data Mining and Statistics: What is the Connection [Internet]. Data Mining and Statistics: What is the Connection. 2004. Available from: <https://tdan.com/data-mining-and-statistics-what-is-the-connection/5226>
74. Oğuzlar A. Veri Ön İşleme. Erciyes Üniversitesi İktisadi ve İdari Bilim Fakültesi Derg. 2003;21:67–76.
75. Koptur M. Bayesci İstatistiğe Giriş. 2020;(February).
76. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. Bayesian Data Analysis. Third Edit. 2013.
77. Ersel D. An Original Combined Interestingness Measure in Association Analysis. Hacettepe University, Ankara, Türkiye; 2012.
78. Heckerman D. Bayesian Networks for Data Mining. Data Min Knowl Discov. 1995;79–119.
79. Ben-Gal. Bayesian Networks, Encyclopedia of Statistics in Quality & Reliability. Ruggeri F, Faltin R, Kenett R (eds), editors. Wiley & Sons.; 2007.
80. Johnson RA, Wichern DW. Applied Multivariate Statistical Analysis. Fifth Edit. New Jersey, USA: Prentice-Hall Inc.; 2002. 425 p.
81. Jackson BB. Multivariate Data Analysis An Introduction. First Edit. Richard d Irwin; 2004. 225 p.
82. Yayıcı AÖ. Temel Bileşenler Analizi için Robust Algoritmaları. Gazi Üniversitesi; 2006.
83. Marascuilo LA, Levin JR. Multivariate Statistics In The Social Sciences: A Researcher’s Guide. California, ABD: Brooks/Cole Publishing Company; 1983. 120 p.
84. Muratlar ER. Makine Öğrenmesine Çok Değişkenli İstatistiksel Yaklaşımlar: Temel Bileşenler Analizi [Internet]. VBO. 2019. p. 1–2. Available from: <https://www.veribilimiokulu.com/makine-ogrenmesine-cok-degiskenli-istatistiksel-yaklasimlar-temel-bilesenler-analizi/>
85. Kavurkacı Ş, Aydın Gürkaş Z, Şamlı R. Büyük Ölçekli Veri Tabanlarında Bilgi Keşfi. İstanbul Üniversitesi Bilgisayar Mühendisliği Bölümü. 2017.
86. Lee W, Stolfo SJ, Mok KW, Street W, York N. A Data Mining Framework for Building Intrusion Detection Models *. DARPA NSF. 2007;1–13.

87. Makine Öğrenmesi Çapraz Doğrulama (Cross Validation) [Internet]. Veri Bilimcisi. 2021. p. 1–2. Available from: <https://veribilimcisi.com/2017/07/13/capraz-dogrulama-cross-validation-nedir/>
88. statistics.com [Internet]. The Institute for Statistics Education. 2021. p. 1. Available from: <https://www.statistics.com/resources/glossary/b/b>
89. Coşkun C, Baykal A. Veri Madenciliğinde Sınıflandırma Algoritmalarının Bir Örnek Üzerinde Karşılaştırılması. 2015 p. 1–8.
90. Karar Verme ve Veri Madenciliği [Internet]. bilimselkonular.com. 2010. p. 2. Available from: <https://www.zovovo.com/>
91. Bolshakova N, Azuaje F. Cluster Validation Techniques for Genome Expression Data. Signal Processing. 2003;83(4):825–33.
92. Marquezin CMS, Darlle S. A, Juliana NA, Taís de SB, Fonseca FLA, Gavião, Maria BD, et al. A Cluster Analysis of the Relationship Between Sucking Habits, Consistency of Food Ingested, and Taste Sensitivity in Children. J Texture Stud. 2019;50:248–56.
93. Rousseeuw PJ. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. J Comput Appl Math. 1987;20:53–65.
94. Herland M, Khoshgoftaar TM, Wald R. A Review of Data Mining Using Big Data in Health Informatics. Vol. 1, Journal of Big Data. 2014.
95. Olaronke I, Oluwaseun O. Big Data in Healthcare : Prospects , Challenges and Resolutions. Futur Technol Conf San Fr United States. 2016;1152–7.
96. Habl C, Renner AT, Bobek J, Laschkolnig A. Study on Big Data in Public Health , Telemedicine and Healthcare. 2016.
97. Dinov ID. Volume and Value of Big Healthcare Data Ivo. J Med Stat Inf. 2016;4:1–15.
98. Big Data Value Association. Big Data Technologies in Healthcare. Big Data Technologies in Healthcare Needs, opportunities and challenges. 2016.
99. Beyan T. Sağlıkta Dönüşüm ve Büyük Veri Uygulamaları [Internet]. 2016. p. 1–4. Available from: <https://saglikvebilisim.wordpress.com/2016/02/18/saglikta-donusum-ve-buyuk-veri-uygulamalari/>
100. ALAB. No Title [Internet]. 2019. p. 1. Available from: <http://www.alab.com.tr/>
101. Pembekar. No Title [Internet]. 2018. p. 1–3. Available from: <https://www.milliyet.com.tr/pembekar/immunoloji-nedir-neye-bakar-alerji-ve-bagisiklik-sistemi-hastaliklari-doktoru-hangi-hastaliklara-bakar-6103758>
102. Costa PSS. Dental Sedation by Dentists : A View From Anesthesiologists Working in Central Western Brazil. Int Anesth Res Soc. 2010;110(1):110–4.
103. Durmuş Hİ, Polat ME, Habek O. Ağız, Diş ve Çene Cerrahisi Nedir? [Internet]. Şanlıurfa, Türkiye; 2019. Available from: http://dishekimligi.harran.edu.tr/assets/uploads/other/files/dishekimligi/files/Ağız_diş_ve_çene_cerrahisi__tanıtıcı%283%29.pdf

104. akmak Z. Kmeleme Analizinde Geerlilik Problemi ve Kmeleme Sonularının Değerlendirilmesi. Dumlupınar niversitesi Sos Bilim Derg. 1999;3:187–205.
105. Tatlıdil H. Uygulamalı ok Değışkenli İstatistiksel Analiz. H. . Fen. Ankara, Trkiye: Engin Yayıncılık; 1992. 252 p.
106. Atamer B. Kmeleme Analizi (Cluster Analysis) ve Kmeleme Analizinin İla Sektrne Uygulanması. İstanbul; 1992.
107. Ball GH. Classification Analysis. Menlo Park Calif Standford Res Inst. 1970;
108. Everitt BS, Landau S, Leese M, Stahl D. Cluster Analysis. 5th Editio. David J. Balding, Noel A.C. Cressie, Garrett M. Fitzmaurice, Harvey Goldstein, Geert Molenberghs, David W. Scott, Adrian F.M. Smith, Ruey S. Tsay SW, editor. London, UK: Wiley & Sons.; 2011. 19–348 p.
109. Hair JFJ, Anderson ER, Tatham LR. Multivariate Data Analysis With Readings. 5th Editio. USA: Prentice-Hall Inc.; 1998.
110. Atılğan E. Karayollarında Meydana Gelen Trafik Kazalarının Karar Ağları ve Birliktelik Kuralı ile Analiz Edilmesi. 2011.
111. Servi T. ok Değışkenli Karma Dağılım Modeline Dayalı Kmeleme Analizi. ıkurova niversitesi; 2009.
112. zdamar K. Paket Programlarla İstatistiksel Veri Analizi 2. Eskişehir: Kaan Kitabevi; 1999.
113. Ada M, Altunay F, Civelek M, Kaplan S, Ko P. Kmeleme Analizi. Başkent niversitesi Derg. 2019;1–8.
114. Pourahmad S, Basirat A, Rahimi A, Doostfatemeh M. Does Determination of Initial Cluster Centroids Improve the Performance of K -Means Clustering Algorithm? Comparison of Three Hybrid Methods by Genetic Algorithm, Minimum Spanning Tree, and Hierarchical Clustering in an Applied Study. Comput Math Methods Med. 2020;2020:1–1.
115. Kalaycı ř. SPSS Uygulamalı ok Değışkenli İstatistik Teknikleri. 4. Baskı. Asil Yayın Dağıtım; 2009.
116. Mitchell CM, Kaufman CE, Beals J, Cottier N, Dethlefsen AD, Wilson A, et al. Identifying Diverse HIV Risk Groups Among American Indian Young Adults : The Utility of Cluster Analysis. AIDS Behav. 2004;8(3):263–75.
117. Jain AK, Dubes RC. Algorithms for Clustering Data. New Jersey, USA: Prentice Hall, Englewood Cliffs; 1988.
118. Everitt BS, Landau S, Leese M. Cluster Analysis. New York, NY, United States: Oxford University Press Inc; 2001.
119. Fırat Oktay S. Kmeleme Analizi: İstihdamın Sektrel Yapısı Aısından Avrupa lkelerinin Karşılaştırılması. İ Sos Bilim Derg. 1997;3(2):50–9.
120. Genoğlu P. Trkiye’deki İllerin Gelişmişlik Dzeyi Dikkate Alınarak Saėlık Hizmetlerinin Kmeleme Analizi Aracılığıyla Değerlendirilmesi. Erciyes niversitesi İktisadi ve İdari Bilim Fakltesi Derg. 2018;52:301–24.

121. Ward JJH. Hierarchical Grouping to Optimize an Objective Function. J Am Stat Assoc. 1963;58(301):234–44.
122. Gower JC. A General Coefficient of Similarity and Some of Its Properties. Int Biometric Soc. 1971;27(4):857–71.
123. Yan H, Wang L, Lu Y. Identifying Cluster Centroids from Decision Graph Automatically Using a Statistical Outlier Detection Method. Neurocomputing [Internet]. 2019;329:348–58. Available from: <https://doi.org/10.1016/j.neucom.2018.10.067>
124. Turhan K, EngİN YZ, Kamburoğlu S, Kurt B, Örem A. Potasyum (K) ve Hemoliz İndeksi (HI) Arasındaki İlişkinin Kümeleme Analizi Yöntemi ile İrdelenmesi. Biyoistatistik ve Tıp Bilişimi AD, Karadeniz Teknik Üniversitesi, Trabzon; 2018 p. 106–12.
125. Yıldırım S. Fuzzy Kümeleme Tekniği Kullanarak Türkiye’deki Bir GSM Operatörü Kullanıcılarının Profili. 2007.
126. Karaağaoğlu E. Veri Madenciliği ve Yeni Nesil Yöntemleri. n’den N’ye Gezinti İstatistik Dergisi,. 2012;9:34–8.
127. Kaufman L, Rosseeauw PJ. Finding Groups in Data: An Introduction to Cluster Analysis. New York, USA: John Wiley and Sons Inc.; 1990.
128. Dinçer E. Veri Madenciliğinde k-Means Algoritması ve Tıp Alanında Uygulaması. Kocaeli Üniversitesi; 2006.
129. Tatlıdil H. Uygulamalı Çok Değişkenli İstatistiksel Analiz. Ankara, Türkiye: Cem Web Ofset; 1996.
130. Everitt B, Dunn G. Applied Multivariate Data Analysis. New York, USA: Oxford Uni. Press; 1992. 101 p.
131. Jones PJ, James MK, Davies MJ, Khunti K, Catt M, Yates T, et al. FilterK: A New Outlier Detection Method for k-means Clustering of Physical Activity. J Biomed Inform. 2020;104:1–10.
132. Marquezin CMS, Chaves-júnior SDC, Jr Rasera I, Pacheco ERP, Gavião BMD, Lamy E, et al. Oral Health and Nutritional Characteristics of Adults With Morbid Obesity : A Multivariate Analysis. 2020;7(November):1–10.
133. Cebeci Z, Yıldız F, Kayaalp T. K-Ortalamlar Kümelemesinde Optimum K Değeri Seçilmesi. 2. Ulusal Yönetim Bilişim Sistemleri Kongresi, Bildiriler Kitabı (Ed: Ü. Özen ve ark.). 2015.
134. Gupta S Das, Killenberger M, Tanner T, Rieppo L, Saarakkala S, Heikkilä J, et al. Mineralization of Dental Tissues and Caries Lesions Detailed with Raman Microspectroscopic Imaging. R Soc Chem. 2021;146:1705–13.
135. Han J. Data Mining Concepts and Techniques. San Francisco, ABD: M. K: Publishes; 2000.
136. Estivill-Castro V, Yang J. Fast and Robust General Purpose Clustering Algorithms. Probl Peredachi Informatsii. 2004;8(2):127–50.

137. Demircioğlu M, Eşiyok S. Covid – 19 Salgını ile Mücadelede Kümeleme Analizi ile Ülkelerin Sınıflandırılması. İstanbul Ticaret Üniversitesi Sos Bilim Derg Covid-19. 2019;37:369–89.
138. Gesicho MB, Babic A, Were MC. K- Means Clustering in Monitoring Facility Reporting of HIV Indicator Data : Case of Kenya. Importance Heal Informatics Public Heal Dur a Pandemic J. 2020;143–6.
139. Bircan H, Çam S. Veri Madenciliğinde Kümele Analizi ve Sağlık Sektöründe Bir Uygulaması. Cumhuriyet Üniversitesi İktisadi ve İdari Bilim Derg. 2016;17(2):85–96.
140. Pasin Ö. Sağlık Alanında Yapılan Araştırmalarda Kümeleme Algoritmalarının Kullanımı: Bir Uygulama. Düzce Üniversitesi; 2015.
141. Taşkın Ç, Emel GG. Veri Madenciliğinde Kümeleme Yaklaşımları ve Kohonen Ağları ile Perekendecilik Sektöründe Bir Uygulama. Süleyman Demirel İktisadi ve İdari Bilim Fakültesi Derg. 2010;15(3):395–409.
142. Sharma A, Shigemizu D, Boroevich KA, Lopez Y, Kamatani Y, Kubo M, et al. Stepwise Iterative Maximum Likelihood Clustering Approach. BMC Bioinformatics. 2016;17(319):2–14.