T.C.
İSTANBUL ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

Doktora Tezi

VERİ MADENCİLİĞİ YAKLAŞIMI KULLANILARAK İNTERNET ERİŞİMLİ TELEVİZYON KULLANICI VERİLERİNİN ANALİZİ

Tevfik ÖRKÜN

Enformatik Anabilim Dalı

Enformatik Programı

DANIŞMAN
Doç. Dr. Hulusi GÜLSEÇEN

Aralık, 2022

İSTANBUL

Bu çalışma, 27.12.2022 tarihinde aşağıdaki jüri tarafından Enformatik Anabilim Dalı, Enformatik Programında Doktora tezi olarak kabul edilmiştir.

Tez Jürisi

Doç. Dr. Hulusi GÜLSEÇEN(Danışman)
İstanbul Üniversitesi
Fen Fakültesi

Prof. Dr. Küçük Hüseyin KOÇ
İstanbul Üniversitesi-Cerrahpaşa
Orman Fakültesi

Doç. Dr. Zümrüt ECEVİT SATI
İstanbul Üniversitesi
Siyasal Bilgiler Fakültesi

Doç. Dr. Emre AKADAL
İstanbul Üniversitesi
İktisat Fakültesi

Dr. Öğr. Üyesi Zerrin AYVAZ REİS
İstanbul Üniversitesi-Cerrahpaşa
Hasan Ali Yücel Eğitim Fakültesi

İntihal Programı Beyanı

20.04.2016 tarihli Resmi Gazete'de yayımlanan Lisansüstü Eğitim ve Öğretim Yönetmeliğinin 9/2 ve 22/2 maddeleri gereğince; Bu Lisansüstü teze, İstanbul Üniversitesi'nin abonesi olduğu intihal yazılım programı kullanılarak Fen Bilimleri Enstitüsü'nün belirlemiş olduğu ölçütlere uygun rapor alınmıştır.

Proje Destekleri

Tezden Üretilmiş Yayınların Künye Bilgileri

ÖNSÖZ

Tez çalışması sırasında desteğini her zaman yanımda hissettiğim değerli danışman hocam Doç. Dr. Hulusi GÜLSEÇEN'e yardımlarından dolayı en içten teşekkürlerimi sunarım.

Tez süreci boyunca, tez çalışmalarımın ilerlemesine yönelik yapıcı eleştirileri ve önerileri ile katkı sağlayan kıymetli hocalarım Prof. Dr. Küçük Hüseyin KOÇ'a ve Doç. Dr. Zümrüt Ecevit SATI'ya teşekkür ederim.

Doktora eğitimim boyunca desteklerini esirgemeyen değerli hocam Prof. Dr. Sevinç GÜLSEÇEN'e teşekkürlerimi sunarım.

Çalışmalarımda her zaman motivasyon kaynağı olan mesai arkadaşlarıma ve dostlarıma teşekkür ederim.

Doğrudan veya dolaylı olarak yolculuğumda bana değer katan herkese teşekkürlerimi sunarım.

Babamı, annemi ve ağabeyimi saygı, özlem ve minnetle anarım.

Desteklerini her zaman yanımda hissettiğim değerli aileme sabır ve yardımlarından dolayı teşekkür ederim.

Aralık 2022 Tevfik ÖRKÜN

İÇİNDEKİLER

Sayfa No

ŞEKİL LİSTESİ

Sayfa No

TABLO LİSTESİ

Sayfa No

SİMGE VE KISALTMA LİSTESİ

Simgeler **Açıklama**

γ : Konum parametresi
$\lambda(x)$: Üstel arıza fonksiyonu
μ : Ortalama değer
σ_x : Standart sapma
$R(x)$: Güvenilirlik fonksiyonu
$f(x)$: Olasılık yoğunluk fonksiyonu
β : Biçim parametresi
η : Ölçek parametresi
$F(x)$: Kümülatif dağılım fonksiyonu
$h(x)$: Hata oranı
E_a : Aktivasyon enerjisi
k : Boltzman sabiti
T : Kelvin sıcaklı derecesi (℃ + 273)
HF_A : Hızlandırma faktörü
A : Ürün veya malzeme karakteristiği
τ_0 : 25 ℃ ortam sıcaklığındaki ürünün yaşam ömrü
T_0 : 25 ℃'deki jonksiyon sıcaklığı
τ' : *T'* sıcaklığındaki ürünün yaşam ömrü
n : Nitelik sayısı
d : Uzaklık
E : Hata karelerinin toplamı
c_i : Küme merkezi
n_j : j kümesinin boyutu
m : Entropi toplam küme sayısı
E_j : Entropi
n : Entropi veri sayısı
$I(X,Y)$: X ve Y değişkenleri arasındaki karşılıklı bilgi
$H(X)$: X kümeleme yöntemine göre entropi değerini
VI : Bilgi değişimi indeksi

$\mathbf{R(i,j)}$: Küme geri çağırma değeri

P : Kesinlik değeri

n_{ij} : j kümesi içerisinde i sınıfına ait nesne sayısı

n_j : j kümesi içindeki nesne sayısı

n_i : i kümesi içindeki nesne sayısı

$\mathbf{F(i,j)}$: i kümesi ve j kümesine ait F indeksi

$tr(s_{Bk})$: CH indeksi kümeler arası mesafe karelerinin toplamı

$tr(s_{wk})$: CH indeksi küme için mesafe karelerinin toplamı

C : C indeksi

S : C indeksi küme içi eleman çiftleri arasındaki mesafelerin toplamı

S_{max} : Küme içi elamanlar arasındaki en küçük mesafeyi

S_{min} : Küme içi elemanlar arasındaki en uzak mesafeyi

Kısaltmalar **Açıklama**

ALT : Accelerated Life Test

ASSIST : Alliance for Solid State İllumination System and Technologies

CRISP-DM : Cross-Industry Standard Process for Data Mining

HALT : Highly Accelerated Life Testing

HASA : Highly Accelerated Stress Auditing

HASA : Highly Accelerated Stress Screnning

HDI : Human Development Index

IES : Illuminating Engineering Society

KDD : Knowledge Discovery in Database

LED : Light Emitting Diode

MAR : Missing at Random

MCAR : Missing Completely at Random

MNAR : Missing Not at Random

MTTF : Mean Time to Failure

pdf : Probability density function

SEMMA : Sample, Explore, Modify, Model, Assess

UNDP : United Nations Development Programme

ÖZET

DOKTORA TEZİ

VERİ MADENCİLİĞİ YAKLAŞIMI KULLANILARAK İNTERNET ERİŞİMLİ TELEVİZYON KULLANICI VERİLERİNİN ANALİZİ

Tevfik ÖRKÜN

İstanbul Üniversitesi

Fen Bilimleri Enstitüsü

Enformatik Anabilim Dalı

Danışman : Doç. Dr. Hulusi GÜLSEÇEN

Teknolojik gelişmelere ve ihtiyaçlara bağlı olarak veri miktarı ve çeşidi her geçen gün artmaktadır. Verilerin doğru şekilde toplanarak analiz edilmesi, veriler arasındaki ilişkilerin keşfedilmesi gereklilik haline gelmiştir. Bu ihtiyacı karşılamak için sınıflandırma, kümeleme ve birliktelik kuralı gibi farklı veri madenciliği yaklaşımları yaygın olarak kullanılmaktadır.

Tez çalışmasının amacı kullanıcı verilerinin analizi sonucu kullanıcı dostu, dayanıklı, segmente edilmiş ürünlerin oluşturulmasına yönelik bilgilerin elde edilmesidir. Kümeleme ve birliktelik kuralı yöntemleri kullanılarak, kullanıcıların kullanım alışkanlıklarına göre gruplandırılması hedeflenmiştir. Kümeleme yöntemi vasıtasıyla müşteri segmentasyonu yapılması amaçlanmıştır. Birliktelik kuralı analizi sonucunda ürün tasarımına yönelik hangi kaynakların birlikte kullanıldığının belirlenmesi hedeflenmiştir. Kullanım süresine göre ürün dayanıklılık durumunun değerlendirmesi amaçlanmıştır.

Bu tez çalışması kapsamında tüketici elektroniği sektöründe faaliyet gösteren bir firmanın internet bağlantılı televizyon kullanıcılarının kullanım süresi ve kaynak kullanım verileri ele alınmıştır. Veri madenciliği süreç adımları izlenerek analiz çalışmaları gerçekleştirilmiştir. Verilerin ön işlenmesi, dönüştürülmesi ve analizi adımları sırasında R dili kullanılmıştır. Kullanım alışkanlıklarına yönelik 17 adet nitelik değerlendirme kapsamına alınmıştır. Kullanım alışkanlıklarına yönelik modelleme aşamasında K-ortalamalar kümeleme yönteminden faydalanılmıştır. Kümeleme performansını ölçmek ve optimum küme sayısını belirlemek

amacıyla iç indekslerden yararlanılmıştır. Kümeleme çalışması sonucunda kullanıcıların çoğunlukla harici oynatıcıları ve uydu yayınları başta olmak üzere sayısal kaynakları kullanmayı tercih ettikleri saptanmıştır.

İlgili ürün segmentinde hangi kaynakların sıklıkla kullandığı tespit edilmiştir. Belirlenen minimum destek ve güven değeri kriterlerine göre birliktelik kuralı analizi yapılmıştır. Farklı sayıda grupları içeren birliktelik kuralları oluşturulmuştur. Birbirine özdeş olan kaynakların kullanım oranları değerlendirilmiştir.

Kullanım süresine yönelik güvenilirlik analizi tekniğinden yararlanılmıştır. Yıllar içinde ışık akısındaki değişimler hesaplanmıştır. Arrhenius sıcaklık yaşam modeli parametrelerinden yararlanılarak hesaplanan 82,5 °C jonksiyon sıcaklığındaki ışık akısı değeri interpolasyon yöntemi ile belirlenmiştir. LED ışık akısının %70 seviyesine düşeceği zaman öngörülmüştür. Yaşam-stres analizinde sıcaklığın hızlandırıcı stres etkisi gözlemlenmiştir. Belirlenen yıllık ortalama izleme süresi verisinden yararlanılarak güvenilirlik ve arıza oranı değerleri hesaplanmıştır.

Çalışma sırasında Birleşmiş Milletler Gelişim Programı tarafından yayınlanan İnsani gelişmişlik raporu verilerinden yararlanılmıştır. Ülkeler arasında insani gelişmişlik düzeyinin yıllar içindeki trendi değerlendirilmiştir. Bu amaçla Avrupa kıtasında yer alan 31 ülkenin 1995 ile 2021 yılları arasındaki insani gelişmişlik indeksi değerleri incelenmiştir. Avrupa kıtasında yer alan ülkeler arasında gelişmişlik düzeyi açısından farklılık durumu Anova yöntemi ile analiz edilmiştir. Hangi ülkeler arasında fark olduğunu belirleyebilmek için Games-Howell testi kullanılarak post-hoc analizi yapılmıştır. Televizyon izleme süreleri ile insani gelişmişlik düzeyi arasındaki ilişki korelasyon yöntemi ile incelenmiştir. Korelasyon analizi sonucunda insani gelişmişlik düzeyi ile televizyon izleme oranları arasında zayıf negatif bir ilişki olduğu görülmüştür.

Aralık 2022, 138 sayfa.

Anahtar kelimeler: Veri madenciliği, k-ortalamalar, insani gelişmişlik indeksi, segmentasyon, internet erişimli televizyon, arrhenius yaşam stres modeli

SUMMARY

Ph.D. THESIS

ANALYSIS OF INTERNET ACCESS TELEVISION USER DATA BY USING DATA MINING APPROACH

Tevfik ÖRKÜN

İstanbul University

Institute of Graduate Studies in Sciences

Department of Informatics

Supervisor : Assoc. Prof. Dr. Hulusi GÜLSEÇEN

The volume and variety of data increases day by day depending on technological developments and needs. It has become a necessity to collect and analyze the data accurately and discover the relationships between the data. Several data mining approaches such as classification, clustering and association rule are widely used to satisfy this needs.

The aim of thesis study is to obtain information for creation of user-friendly, robust and segmented products as a result of the analysis of user data. By using clustering method and association rules, it is targetted to group users according to their usage habits. It is aimed to make customer segmentation through the clustering method. As a result of association rule analysis, it is aimed to determine which ports are used together for product design. It is also intended to evalute the reliability of examined product according to usage time.

Within the scope of this thesis, the usage time and port usage data of internet-connected television users of a consumer electronic company were investigated. Analyzes were carried out by following the data mining process steps. The R language was used during the preprocessing, conversion, analysis steps of the data mining. 17 attributes related to usage habits were evaluated. K-means clustering method was used in the modeling phase for usage habits. Internal indexes were used to measure the clustering performance and determine the optimum

number of clusters. As a result of the clustering study, it has been determined that users mostly prefer to use digital ports, especially external ports and satellite broadcasts.

It has been determined which ports are frequently used in the relevant product segment. Association rule analysis was performed according to minimum support and confidence criteria determined. Association rules containing different numbers of groups were created. Utilization rates of ports that are identical to each other were evaluated.

The reliability analysis technique was used for the life time evaluation of product. Changes in luminous flux over the years have been calculated. The luminous flux value at the 82,5 °C junction temperature which was calculated with the Arrhenius temperature life model parameters , was determined by means of the interpolation method. It has been predicted when LED flux will decrease to %70 level. By using the determined annual average watching time data, reliability and failure rate values were calculated.

In this study, the Human Development Report data published by the United Nations Development Program was used. For the years between 1995 and 2021 , trends of human development level among 31 countries in the European continent were evaluated. The difference between the countries in the European continent in terms of the level of development was analyzed by the Anova method. Post-hoc analysis was conducted by using Games Howell test to determine which countries differ. Using the data of the human development report, the relationship between the television watching time and the level of human development was examined by correlation method. As a result of the correlation analysis, it was seen that there is a weak negative relationship between the level of human development and television viewing rates.

December 2022, 138 pages.

Keywords: Data mining, k-means, human development index, segmentation, internet access television, arrhenius life stress model

1. GİRİŞ

Günümüzde veri, buna bağlı olarak da bilgi miktarının artması tüm dünyayı etkileyecek şekilde değişimlere, zihinsel dönüşümlere, iş süreçlerinin evrilmesine yol açmaktadır. Bu değişime ayak uydurabilmenin anahtarı elde edilen bilgiyi doğru kullanmaya dayanmaktadır. Francis Bacon 16. yüzyılda yayınladığı eserinde "bilgi güçtür" ifadesini kullanılmıştır. Bu ifade çağımızda daha fazla anlam kazanmaktadır. Taylor (1911) Bilimsel Yönetimin İlkeleri isimli kitabını yayınlayarak üretkenlik kısıtlarının ortadan kaldırılıp, verimliliğin ön plana alındığı endüstri çağının başlamasına öncülük etti. Kullanıcıya özgü üretimin yapıldığı üretim çağı yerine endüstri çağına geçiş yapılmış oldu. Rekabetin düşük olduğu, müşteriye özgü çözümlerin sunulmadığı endüstri çağı 1980'li yıllara kadar devam ederek, yerini evrensel pazarlar dönemine bıraktı (Waters, 1995). Bu son dönemde, dünya çapında büyük firmalarla aynı pazarda olmak dinamik süreçlerle iş yapış tarzını ve değer katan çözümler sunmayı gereklilik haline getirdi. Bu gereklilikleri sağlayabilmenin yolu da bilgiyi kullanmaktan geçmektedir. Elde edilen bilgiyi analiz ederek kullanmak gerek bireyler için gerekse şirketler için kaçınılmaz hale gelmiştir.

Şirketler açısından müşterilere yönelik yenilikçi, rekabetçi, doğru zamanda ürün ve hizmet sunabilmenin sırrı elde ettikleri verileri enformasyona, enformasyonu bilgiye doğru ve hızlı biçimde dönüştürmeleriyle sağlanmaktadırlar. Şirketler elde ettikleri bilgiler ile müşterilere yönelik kullanıcı dostu çözümler oluşturabilmekte, bilgileri tasarım geri beslemesi olarak kullanarak daha yenilikçi, dayanımı güçlü, uzun ömürlü ürünler sunabilmektedirler. Şirketler analiz sonuçlarını değerlendirerek, karar mekanizmalarında kullanabilmekte, daha yalın süreçler oluşturarak, karlılıklarını artırabilmektedirler.

Müşterilerin kullanım alışkanlıklarının belirlenmesine yönelik sepet analizlerinin yapılması, kullanıcıya uygun ürün, hizmet çözüm sunulabilmesi için segmentasyon çalışmasının yapılması günümüz pazarlama stratejilerinin bir adımı haline gelmiştir.

Müşteri ihtiyaçlarını daha iyi karşılayabilecek yenilikçi çözümler ve yeni özellikler sunabilmek için müşteri kullanım alışkanlıklarının analiz edilmesi tasarım ve pazarlama yönünden kurumun yetkinliklerinin güçlenmesini sağlamaktadır. Üründe kullanılan özellikleri müşterilerin ne kadar faydalı buldukları, kullanım ihtiyaçlarını ne kadar karşıladıklarını anlamak kurumlar

açısından önemli bilgilerdir. Elde edilen bilgiler şirketlere, kullanıcılarına sundukları değerleri daha da iyileştirmelerine imkân verir. Bu durum müşteri ihtiyaçlarının daha iyi karşılanmasını sağlayarak, müşteri memnuniyetinin artmasına imkân vermekte, ürünün veya çözümün potansiyel pazar payının artmasına yol açabilmektedir.

Tez kapsamında beş bölüm yer almaktadır. Tezi oluşturan kısımlardan, birinci kısım "Giriş" bölümünü, ikinci kısım "Genel Kısımlar" bölümünü, üçüncü kısım "Malzeme ve Yöntem" bölümünü, dördüncü kısım "Bulgular" kısmını, sonuncu kısım ise "Tartışma ve Sonuç" bölümünü oluşturmaktadır.

Genel kısımlar bölümünde veri madenciliği kavramı, veri madenciliği süreçleri, kümeleme yöntemi, müşteri segmentasyonuna yönelik birliktelik kuralları ve güvenilirlik kavramı üzerine bilgilere değinilmiş, literatür araştırması sonuçlarına yer verilmiştir.

Malzeme ve yöntem bölümünde tez çalışması sırasında izlenecek olan yöntem, kullanılacak olan modeller paylaşılmıştır.

Bulgular bölümünde müşteri kullanım alışkanlıklarına yönelik yapılan kümeleme çalışmaları modeli, birliktelik çalışmaları modeli, kullanım alışkanlıklarının değerlendirilmesi, ürün ömür hesaplama yöntemlerine yönelik yapılan analiz çalışmaları paylaşılmıştır.

Tartışma ve Sonuç bölümünde ise kullanım alışkanlıkları verilerine yönelik kurulan modellerle ilgili sonuçlara ve gelecekte yapılabilecek çalışmalara değinilmiştir.

2. GENEL KISIMLAR

2.1 VERİ MADENCİLİĞİ KAVRAMI

Günümüzde oluşturulan ve depolanan veri miktarının artmasına bağlı olarak, karmaşıklaşan veriler arasındaki ilişkilerin keşfedilmesi, yorumlanması ihtiyacına bağlı olarak veri madenciliği yöntemleri geliştirilmiştir.

Veri tabanlarından bilgi keşfi veriler içinde yer alan geçerli, farklı, potansiyel olarak faydalı ve anlaşılabilir örüntülerin belirlemesine yönelik gerçekleştirilen basit olmayan bir süreçtir. Veri tabanlarından bilgi keşfi, veriden yararlı bilgilerin keşfedilmesi için gerekli bütün süreci kapsarken, veri madenciliği bilgi keşfi sürecinin belirli bir safhasını tariflemektedir. Veri madenciliği yaklaşımı, veriden örüntülerin çıkartılması için gerekli özgün algoritmaların uygulamasıdır (Fayyad ve diğ., 1996).

Büyük veri yığınları içerisinden örüntülerin keşfedilip, ilgili örüntülerden yararlanılarak bilgi edinilmesi amacıyla kullanılan veri madenciliği yöntemlerinin günümüzde birçok farklı kullanım alanı mevcuttur. Veri madenciliğinin kullanıldığı uygulama alanları arasında sosyal ağ analizi, medikal ve sağlık hizmetleri, pazarlama, bulut bilişim, eğitim, kurumsal bilgi güvenliği, bankacılık sayılabilir (Kumar ve Sharma, 2017; Yoo ve diğ., 2011; Dhar, 1998).

Veri madenciliği farklı disiplinler ile etkileşim halindedir. Veri tabanı teknolojisi, enformatik, görselleştirme, istatistik, makine öğrenmesi ve diğer disiplinler veri madenciliği ile ilişki disiplinler arasında belirtilebilirler (Han ve diğ., 2012).

Veri madenciliği çalışmaları sırasında yaygın olarak kullanılan modeller arasında sınıflandırma, birliktelik kuralları, kümeleme, regresyon, eğri uydurma sayılabilir.

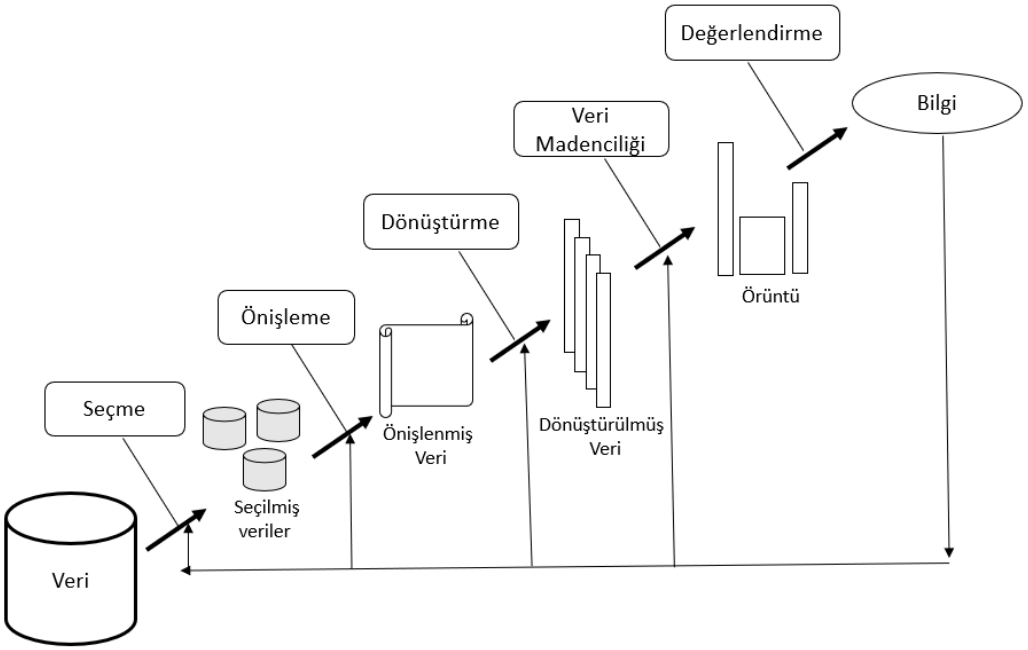
Veri madenciliği yöntemleri kullanılarak veriler arasında ilişkili kurulabilmesi, verilerin sınıflandırılabilmesi ve tahmin yapılabilmesi mümkün olabilmektedir.

2.1.1 Veri Madenciliği Süreçleri

Veri madenciliği çalışmalarında sıklıkla kullanılan modeller farklı süreç adımlarını ve geçişleri içermektedirler.

2.1.1.1 KDD

KDD (Knowledge Discovery in Databases) Veri Tabanından Bilgi Keşif Süreci, Fayyad ve diğ. (1996) tarafından "Veri Madenciliğinden Veri Tabanındaki Bilgi Keşfine" isimli makalede sunulan 5 adımdan oluşmaktadır. Bu adımlar sırasıyla seçme, ön işleme, dönüştürme, veri madenciliği ve değerlendirme aşamalarından oluşmaktadır. KDD modelinde, veri seçimi aşaması öncesinde problemi tanımlama, veri kaynağından veri toplama ve veri anlama safhalarının tamamlandığı varsayılmaktadır.



Şekil 2.1: KDD sürecine genel bakış (Fayyad ve diğ., 1996).

Bilgi keşif sürecinin adımları aşağıdaki şekilde açıklanabilir.

Seçme aşamasında, veri seti incelenerek, tanımlanan probleme uygun, önemli olduğu öngörülen veriler seçilir. Veri madenciliği çalışmaları seçilen bu veriler üzerinde yapılır.

Ön işleme aşamasında, eksik verinin tamamlanması, eksik kayıtlara ait satırların veri kümesinden çıkarılması, eksik verilerin sabit bir değer veya ortalama değer ile doldurulması, gürültü verilerinin temizlenmesi, yeni özniteliklerin oluşturulması gerçekleştirilir. CRISP-DM süreçlerinde yer alan veriyi anlama aşaması, KDD süreçlerinde ön işleme ve dönüştürme adımlarından oluşmaktadır. Ön işleme çalışmalarının kalitesi, veri madenciliği çalışmalarının başarısını arttıracaktır.

Dönüştürme aşamasında, veri madenciliği yöntemlerine uygun olarak verinin, farklı tiplere ve içeriğe dönüştürülmesi gerçekleştirilir.

Veri madenciliği, örüntülerin elde edilebilmesi için veri madenciliği yöntemlerinin kullanıldığı aşamadır.

Değerlendirme, veri madenciliği yöntemleri kullanılarak elde edilen örüntülerin yorumlanarak bilgiye dönüştüğü aşamadır. Yorumlama sonucunda yeterli bilgi elde edilemez ise önceki adımlar tekrar edilir. Aynı süreç adımı CRISP-DM ve SEMMA süreçlerinde de aynı isimle değerlendirme adımı olarak isimlendirilirler.

2.1.1.2 CRISP-DM

CRISP-DM (Cross-Industry Standard Process for Data Mining) Çapraz Endüstri Standard Süreci Modeli, Daimler Chrysler AG (Almanya), NCR System Engineering Copenhagen (A.B.D ve Danimarka), SPSS Inc. (A.B.D) ve OHRA Verzekeringen en Bank Groep B.V.(Hollanda) firmaları tarafından oluşturulmuştur (Chapman ve diğ., 2000). CRISP-DM 2000'li yılların başlarında oluşturulan günümüzde de yaygın olarak kullanılmaya devam eden sektörden bağımsız bir süreçtir. Süreç; iş süreçlerini anlama, veriyi anlama, veriyi hazırlama, modelleme, değerlendirme ve ürünleştirme olarak isimlendirilen 6 adımdan meydana gelmektedir. Aşamalar arasında ileri ve geri yönde geçişler olabilmektedir (Azevedo ve Santos, 2008).

Schröer ve diğ. (2021) tarafından yapılan literatür araştırmasında CRISP-DM süreci kullanılarak çoğunlukla sağlık, eğitim ve araştırma başta olmak üzere mühendislik ve üretim, kamu, yönetim, enformasyon teknolojileri, gıda, finans ve eğlence alanlarına yönelik çalışmalar yapıldığı görülmüştür.

İş süreçlerini anlama aşamasında, mevcut ve gerekli kaynaklar hakkında genel bir bakış elde etmek için iş durumu değerlendirilir. Veri madenciliği hedefinin belirlenmesi bu aşamadaki en önemli unsurlardan biridir. Öncelikle veri madenciliği türü açıklanmalıdır. (örneğin sınıflandırma) ve veri madenciliği başarı kriterleri (kesinlik gibi) belirlenmelidir. Bir proje planı oluşturulmalıdır.

Veriyi anlama aşamasında, gereksinimlere uygun verinin toplanması, toplanan verinin veya mevcutta hazır olan verinin araştırılması, tanımlanması ile ilgili temel görevler yerine getirilir. Toplanan verinin gürültülü veya eksik olup olmadığına yönelik veri kalitesi kontrol edilir.

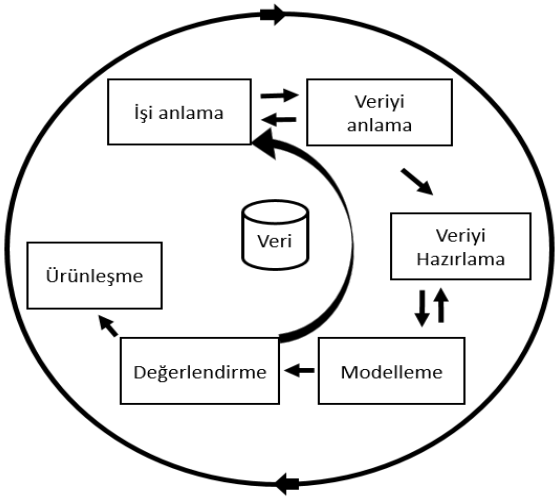
Veriyi hazırlama aşamasında, toplanan veriler ile ilgili olarak hangi çalışmaların yapılacağına karar verilerek, bu çalışmalar sırasında kullanılacak metotlar belirlenir. Toplanan veriler içindeki eksik kısımların tamamlanması veya veri kümesinden hariç tutulması kriterleri belirlenerek, veri seçimi yapılmalıdır. Belirlenen yönteme bağlı olarak öznitelikler oluşturulmalıdır.

Modelleme aşamasında, modelleme tekniğinin seçilmesi, test senaryosunun ve modelinin oluşturulması gerçekleştirilir. Tanımlanan iş problemine ve eldeki verilere bağlı olarak uygulanacak veri madenciliği tekniği seçilir. Seçime bağlı olarak bir makine öğrenmesi ve istatistiksel model geliştirilir. Modele uygun olarak verinin düzenlenmesi, bazı parametrelerin ayarlanması gerekebilir.

Değerlendirme aşamasında sonuçlar, belirlenen iş hedeflerine göre kontrol edilirler. Bu aşamada, şimdiye kadar olan sürecin bütün aşamaları gözden geçirilerek, değerlendirilir (Schröer ve diğ., 2021). Karara göre ürünleşme aşamasına geçilebilir veya ilk adıma geri dönülebilir.

Ürünleşme (Deployment) aşamasında, sonuç raporu oluşturulur veya sistemin çalışan bir uygulaması geliştirilir. Bu aşama ürün haline gelme planı, izleme ve bakım kısımlarını içerir (Schröer ve diğ., 2021).

CRISP-DM modelinin aşamaları ve safhalar arasında meydana gelen döngüler Şekil 2.2‘de gösterilmiştir. CRISP-DM modelinde işin anlaşılması ile verinin anlaşılması aşamaları arasında ve verinin hazırlanması ile modelleme aşamaları arasında döngü oluşabilmektedir. Değerlendirme aşamasında, elde edilen sonuçlara göre ilk adıma dönülerek sistemin bütün aşamaları genel olarak gözden geçirilebilir.



Şekil 2.2: CRISP-DM sürecine genel bakış (Chapman ve diğ., 2000).

2.1.1.3 SEMMA

SEMMA (Sample, Explore, Modify, Model, Assess) , analitik yazılım şirketi SAS tarafından geliştirilen ve örnekleme, inceleme, düzenleme, modelleme, değerlendirme adımlarından oluşan bir süreçtir (Azevedo ve Santos, 2008).

CRISP-DM sürecinin işi anlama safhasında gerçekleştirilen problemin tanımlanması işlemi SEMMA sürecinde yer almaz, bu yönü ile CRISP-DM'den ayrılır. Dolayısı ile değişen problemlere özgü olarak farklı veri işlemleri gerçekleştirilmez. Veri, farklı amaçlarla kullanıma yönelik olarak bir kez işlenir (Şeker, 2018).

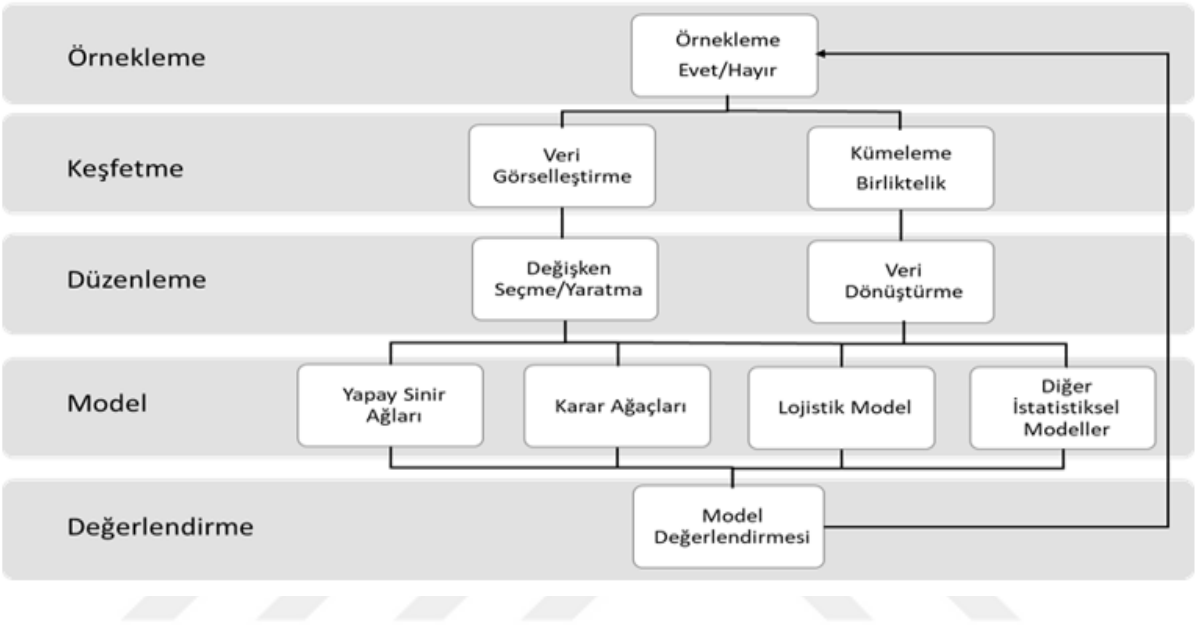
Örnekleme, SEMMA sürecinin ilk aşamasıdır. SEMMA aşamasında, büyük bir veri kümesinin içinden belirli bir kısmı seçilerek alınır. Örnekleme sırasında alınan veri miktarı, işlem hızını artırabilmek için ve kolay analiz yapabilmek için olabildiğince küçük seçilmeli fakat anlamlı bilgi içerecek kadar yeterli büyüklüğe de sahip olmalıdır. Bütün veri kümesi yerine örnekleme sırasında alınan sınırlı boyuttaki veri üzerinde işlem yapıldığından, işlem kapasitesi açısından kısıtlı imkanların olduğu koşullarda kolaylık sağlayacaktır (McCue, 2015).

Keşif aşamasında, beklenilmeyen eğilimler ve anormallikler araştırılarak, verilerin analiz ve keşifleri gerçekleştirilir. Veri içerisindeki örüntü ve ilişkiler araştırılır.

Düzenleme aşamasında, değişkenler üzerinde oluşturma, ilave seçim ve dönüştürme işlemleri uygulanır. Türev değişkenler bu aşamada yaratılabilir.

Modelleme aşamasında, hedeflenen sonuçların doğru tahmin edilebilmesi için düzenlenen veriler üzerinde işlem yapan makine öğrenmesi ve istatistiksel yöntemlerden yararlanılır.

Değerlendirme aşamasında, veri madenciliği süreçleri uygulanarak elde edilen bulgular yararlılık ve güvenirlik açısından gözden geçirilerek, model başarısı değerlendirilir.



Şekil 2.3: SEMMA sürecine genel bakış (Şeker, 2018).

Veri madenciliği projelerinde, SEMMA yönteminin kullanım kolaylığı süreçlere rahat hâkim olunmasını sağlamakta, projelerin koordineli şekilde geliştirilmesine ve sürdürülebilmesine imkân vermektedir (Azevedo ve Santos, 2008).

Tablo 2.1: KDD, SEMMA ve CRISP-DM süreç adımları (Azevedo ve Santos, 2008).

KDD	SEMMA	CRISP-DM
KDD Öncesi	_ _ _	İş Süreçlerini Anlama
Seçme	Örnekleme	Veriyi Anlama
Önişleme	Keşfetme	
Dönüştürme	Düzenleme	Veriyi Hazırlama
Veri Madenciliği	Model	Modelleme
Değerlendirme	Değerlendirme	Değerlendirme
KDD Sonrası	_ _ _	Ürünleşme

2.1.2 Veri Önişleme

Veri kümesindeki verinin durumu göz önüne alınarak veri önişleme çalışmaları sırasında veri eksiklerinin ve belirsizliklerinin temizlemesi, standardizasyon, veri dönüştürme ve boyut indirgeme işlemleri gerçekleştirilebilir.

2.1.2.1 Veri Temizleme

Veri madenciliği sırasında ele alınan veriler her zaman tam ve eksiksiz olmayabilirler. Analiz edilecek verilerde gürültülere, eksikliklere ve belirsizliklere rastlanılmaktadır. Verilerin içlerinde yer alan gürültü ve eksikliklerden arındırılmadan analiz edilmeleri, bilgi keşfi çalışmalarında yanlış örüntülerin oluşmasına neden olarak, hatalı yargıların elde edilmesine yol açarlar. Veri madenciliği sürecinde verideki gürültü, eksiklik, hata ve belirsizliklere yönelik olarak temizleme, aykırı değerlerin belirlenmesi ve veri setinden çıkarılmaları, verinin yeniden işlenmeleri önemli yöntemler olarak kullanılmaktadır (Han ve diğ, 2012). Veri temizleme işlemi, eksik verilerin kaldırılmasını veya farklı metotlarla elde edilen değerlerle doldurulmasını, veri üzerindeki gürültünün ortadan kaldırılmasını, tutarsız verilerin düzeltilmesini ve aykırı gözlem analizini içerir.

Eksik veriler istatistiksel analiz sırasında karşılaşılan yaygın sorunlardandır. Eksik verinin toplam veri içerisindeki oranı analiz yöntemleri ve sonuçlarını etkilemektedir. %1'den az kayıp veri oranları genellikle önemsiz olarak kabul edilirken, %1-5 arasındaki kayıp veri yönetilebilir olarak değerlendirilir. Bununla birlikte %5-15 arasındaki kayıp veriyi işlemek karmaşık yöntemler gerektirmektedir ve %15 üzerinde kayıp veri herhangi bir yorumu ciddi anlamda etkilemektedir (Acuna ve Rodriguez, 2004).

Eksik veri oluşmasının farklı nedenleri vardır. Bazen değerler gerçekten mevcut olmadığı için veri kümesinde bu değerler eksik görünürler. Tasarımın hatalı veya eksik olması, cevap vericinin bilinçli olarak cevap vermemesi, veri toplama cihazlarındaki hatalar eksik veriye yol açabilmektedir.

Rubin (1976) tarafından eksik veriler, Tamamen Rassal Eksiklik (MCAR), Rassal Eksiklik (MAR), Rassal Olmayan Eksiklik (MNAR) olarak üç gruba ayrılmıştır. Bu 3 eksik veri sınıfı eksik veri mekanizmaları olarak da adlandırılırlar (Fox ve diğ., 2015).

Tamamen Rassal Eksiklik, diğer değişkenlerden ya da yapısal bir problemden kaynaklamayan rastgele oluşan gözlemlerdir. Eksiklikler, kişinin yaptığı analizlerdeki değişkenlerle ilgisi olmayan bir süreçten veya tamamen rassal bir süreçten kaynaklanır. Bu durum bir madeni para atma veya zar atmaya benzer.

Rassal Eksiklik, değişkenin kendisi ile ilgisi olmayıp, diğer değişkenlere bağlı olarak oluşabilecek eksiklik türüdür. Bir değişken değerinin eksik olma olasılığı, kısmen veri setinde gözlemlenen diğer değişkenlere bağlıdır, ancak eksik olan değerler değişkenin kendi değerleri ile ilişkili değildir (Newman, 2014).

Rassal Olmayan Eksiklik, göz ardı edilemeyecek olan ve yapısal problemler ile ortaya çıkan eksiklik türüdür. Bir değişken değerinin eksik olma nedeni değişkenin kendisi ile ilgidir.

Han ve diğ. (2012) veri kümesinde eksik veri olması durumunda yapılabilecek işlemleri aşağıdaki şekilde sıralamışlardır:

- Eksik verinin bulunduğu ilgili satırın görmezden gelinerek veri kümesinden çıkarılması,
- Eksik verileri el ile doldurulması,
- Eksik verilerin genel bir sabit ile doldurulması,
- Eksik veriler yerine ortalama bir değer veya ortanca (medyan) değer koyulması,
- Benzer sınıfa ait örnekler için özniteliğin ortalamasının veya ortanca değerinin eksik verilerin yerine koyulması,
- Eksik verilerin tahmin edilen en olası değer ile doldurulması.

Eksikliklerin rassallık durumu ve toplam veri içerisinde kayıp veri miktarına göre silme işlemine karar verilebilir. Eğer kayıplar çok az veride varsa ve rassal oluşmuşlar ise eksik gözlemler veri kümesinden silinebilmektedir. Kayıpların rassal olmadığı durumlarda yapılacak silme işlemine karar verilmeden önce eksik veri oranı ve değişkenin önem durumu kontrol edilmelidir. Bu durumda yapılacak silme işlemi yanlılıklara sebep olabileceğinden silme yerine farklı değer koyma yöntemleri kullanılması uygun olabilir (Tabachnick ve Fidell, 1996).

Silme yöntemleri arasında yer alan liste bazlı silme yöntemi ve çiftler bazında silme yöntemi yaygın şekilde kullanılmaktadır. Kullanılan silme yöntemlerinin güvenilirliğin azalması ve yanlılığa yol açma gibi etkileri olmaktadır (Allison, 2009).

Liste bazında silme yöntemi verilerin MCAR varsayımını sağlaması durumunda kullanılabilir, kestirim sırasında herhangi bir yanlılığa yol açmaz. Verilerin MCAR varsayımını sağlayamayıp, MAR varsayımını sağlaması durumunda ise liste bazlı silme yöntemi yanlılığa yol açabilir. Çiftler bazında silme yöntemi, örneklemin büyük olduğu ve MCAR varsayımının sağlandığı durumda kararlı bir şekilde yanlılık göstermeden çalışmaktadır. Verilerin MCAR varsayımını sağlamayıp, MAR varsayımını sağlaması durumunda ise çiftler bazlı silme yöntemi kestirim sonuçları yanlılık gösterebilir (Allison, 2009).

Kullanılan programlarda eksiklikler NA ya da null olarak belirtilmektedir. Eksik verilerin veri seti içerisinden silinmesi sırasında bilgi kaybı kaynaklı oluşacak zarar dikkate alınmalıdır. Veri kümesi içerisinde yer alan NA değerleri her zaman eksiklik anlamına gelmez. Veri kümesindeki eksikliğin sebebinin yapısal bir nedenden mi kaynaklandığının belirlenmesi önemlidir. Eksik veri seti üzerinde yapılacak işlemler modelleme çalışmasının güvenilirliğini etkileyeceğinden silme işlemlerine yönelik çalışmalarda dikkatli olunmalıdır. Salgado ve diğ. (2016) eksikliğin nedenin MCAR mı, MAR mı yoksa MNAR mı olduğunun mutlaka değerlendirilmesi gerektiğini belirtmekte, verideki ve kayıttaki kayıp oranının dikkate alınması gerektiğini vurgulamaktadırlar.

Eksik veri rassallığının incelemesinde kullanılan yöntemler; görsel teknikler, bağımsız iki örneklem t testi, korelasyon testi ve Little'nin MCAR testidir.

Eksik veri problemini gidermenin diğer yöntemi ise değer atamadır. Değer atama tekli değer atama ve çoklu değer atama olarak iki gruba ayrıldığı (Fox ve diğ., 2015) gibi geleneksel ve modern tipte atama yöntemleri olarak da gruplandırılabilmektedir. Yaygın tekli değer atama yöntemleri arasında ortalama ve regresyon modelleri sayılabilir. Eksik verileri hesaplama ve tahmine dayalı olarak tamamlamak için kullanılan istatistiksel tabanlı geleneksel yöntemler ile model tabanlı modern tipteki atama yöntemleri mevcuttur. Geleneksel yöntemler arasında ortalama değer atama (ortanca, ortalama veya medyan), en benzer değer atama (hot deck), en yakın komşu değeri atama yöntemleri sayılabilir. Tahmin esasına dayanan modelleme bazlı değer atama yöntemleri arasında ise çoklu değer atama, en çok olabilirlik yöntemi, regresyon

ataması yer almaktadır. Regresyon ataması modelinde eksik olan değer, değişkenin regresyon denkleminden yararlanılarak tahmin edilebilir.

Veri setinde bulunan eksik veri oranına ve rassallık durumuna göre seçilecek yönteme karar verilir. Eksikliklerin tamamen rassal dağıldığı ve az oranda kayıp veri olması durumunda basit atama yöntemlerinin, aksi durumlarda gelişmiş atama yöntemlerinin tercih edilmeleri uygun olacaktır (Osborne, 2013; Schafer, 1999).

Eksik veriler ve eksik verileri gidermeye yönelik yapılan işlemler birçok sorunu da beraberinde getirebilmektedir. Veri kümesinde eksik veri bulunması durumda, eksik verilerin çalışma sırasında kullanımları istatiksel model sonrasında yapılacak tahminlerde yanlılık oluşmasına yol açmaktadır. Bilgi kaybı ve istatiksel analiz gücünün zayıflaması da eksik veri durumunda oluşan diğer olumsuzluklar olarak ortaya çıkmaktadırlar. Liste bazlı silme işlemlerinde bazı değerlerin istatistiksel analizden çıkarılması t testi gibi istatistiksel testlerde serbestlik hata derecesinin düşmesine neden olabilmektedir. Yaygın olarak kullanılmakta olan birçok istatistiksel yöntem, eksik veri olması durumda işlem yapmaya uygun değildir veya kullanımları sırasında problemlerle karşılaşılabilinmektedir. Eksik veri olması durumunda bu eksiklilerin temel nedenini bulmaya yönelik sarf edilen efor, zaman ve kaynak kaybına yol açmaktadır (Peng ve diğ., 2007).

Verinin geri kalanından oldukça farklı olan, uygulamadaki örüntülere uymayan gözlemler aykırı değer olarak adlandırılır. Gözlem, ölçüm ve kayıt sırasında oluşan farklı durumlar aykırı değer oluşmasına yol açabilmektedirler. Hatalı ölçüm yapılması, ölçüm doğru yapılsa dahi kayıt sırasında veri girişinde yapılan hatalar veri kümesinde aykırı değerler oluşmasına neden olabilir. Bunların dışında gözlem sırasında ölçümler doğru yapılsa dahi normal değerlerin dışında farklı popülasyondan da veri gelebilir, bu değerler sıklıkla gözlemlenmeyen nadir karşılaşılan değerleri temsil ederler (Cateni ve diğ., 2008).

Aykırı değerlerin bulunmasına yönelik gözlem çalışmaları, aykırı değer tespiti veya anomali tespiti olarak isimlendirilmektedir (Aggarwal, 2017). Aykırı değer tespiti büyük veri kümelerinin temizlenmesi ve kullanıcı için yararlı olan bilgilerin tutulmasına yardımcı olmaktadır (Tang ve Ngan, 2016). Bununla birlikte veri kümelerinin çeşitliliği ve veri boyutundan kaynaklanan kompleksitenin artması aykırı değer analizini zorlaştırmaktadır. (Bhatt ve diğ., 2016). Aykırı değer analizinin kullanıldığı alanlar arasında kredi kartı

dolandırıcılıkları, iklim olayı tespiti, medikal bakım, mekanik arıza tespiti, hata tespiti, terörist tespiti, kötü amaçlı URL tespiti, finansal uygulamalar, coğrafi bilgi sistemi sayılabilir. (Pang ve diğ., 2021; Bhatt ve diğ., 2016) Aykırı değerler tahmine yönelik testlerin etkinliklerini olumsuz etkiler.

Aykırı değerlerin belirlenmesine yönelik yöntemler arasında öğrenme tabanlı yöntemler, istatistiksel yöntemler, yakınlık tabanlı yöntemler ve sıralama tabanlı yöntemler yer almaktadırlar (Tang ve Ngan, 2016).

Aykırı değerlerin geliştirilecek olan model üzerindeki olumsuz etkilerini azaltmak veya yok etmek için farklı veri madenciliği algoritmaları kullanılabilir. Aykırı değerlerin ortadan kaldırılması kritik ve dikkatli olunması gereken bir adımdır. Aykırı değerlere yönelik gerçekleştirilecek tespit ve yok edilme işlemleri sırasında yanlışlıkla önemli gizli bilgilerin silinmemesi hususuna önem verilmelidir (Cateni ve diğ., 2008).

Aykırı değer analizi tek değişkenli aykırı değer analizi ve çok değişkenli aykırı değer analizi olarak iki gruba ayrılabilir.

Tek değişkenli aykırı değer analizinde kullanılan aykırı değer temizlenmesine yönelik tanıma yöntemleri olarak sektör bilgisi, standart sapma yaklaşımı, Z skor yaklaşımı ve kutu grafiği (Interquartile Range- IQR) sayılabilir.

Tukey tarafından geliştirilen kutu grafiği tek değişkenli aykırı değer analizinde çoğunlukla kullanılan yöntemlerden biridir. Kutu grafiği çalışma yöntemi gözlemlerin belirli bir aralık değeri içerisinde olmasına dayanır. Bu amaçla çeyrek değerlerinden faydalanılır. [Q1-1.5IQR, Q3+1.5IQR] aralığının dışında bulunan gözlemler aykırı değerler şeklinde etiketlenirler. İlgili aralığı belirlerken kullanılan Q1 ve Q3 sırasıyla birinci ve üçüncü çeyrekleri belirtirken, IQR ise çeyrek arası aralığı olarak tanımlamaktadır. Yapılan araştırmalarda kutu grafiğinin simetrik verilere daha uygun olduğunu görülmüştür (Junsawang ve diğ., 2021).

Aykırı değer temizleme yöntemleri arasında; aykırı değerleri silme yöntemi, ortalama değer ilave etme yöntemi ve baskılama yöntemi sayılabilir. Ortalama ilave etme yönteminde, belirlenen aykırı değerler ortalama değer ile doldurulur. Baskılama yönteminde (Swamping) aykırı değerler belirlendikten sonra, alt sınır değerinden küçük olan aykırı değerlerin alt sınır

değerine, üst sınır değerinden büyük aykırı değerlerin üst sınır değerine eşitlenir. Baskılama yönteminde değerler sınır değerlere doğru baskılanmaktadır.

Çok değişkenli aykırı değer analizinde kullanılmakta olan lokal aykırı değer faktörü (LOF, Local Outlier Factor) bir noktanın yerel yoğunluğunun, komşu noktaların yoğunluğu ile karşılaştırılmasına dayanır. LOF yöntemi ile belirlenen aykırı değerler, tek değişkenli aykırı değer yönteminde olduğu gibi silme ortalama değer ilave etme, baskılama yöntemleri ile temizlenirler.

2.1.2.2 Normalizasyon/ Standartlaşma

Veri madenciliği sürecinin adımlarından biri olan ön işleme safhasında veri temizleme, veri indirgeme, değişken dönüşümü çalışmalarının yanında veri normalizasyon işlemi de yapılır.

Veri madenciliği çalışmaları sırasında veri kümesinde yer alan verilerin doğrudan kullanımı uygun olmayabilir. Ortalama ve varyans değeri diğerlerinden büyük olan değişkenler görece kendilerinden düşük ortalama ve varyans değerine sahip değişkenleri baskılayacaklardır. (Özkan, 2020). Veri kümesi içinde bulunan değişkenlerin farklı ölçü birimleriyle ölçülmeleri durumunda sonuçlar küçük veya büyük değerlerden fazlasıyla etkileneceklerdir. Büyük değerlerin küçük değerleri baskılamasından ötürü sonuç üzerindeki etkisi artarak, analiz sonuçlarının değişmesine yol açar. Bu durumun önüne geçebilmek için değişkenlerin normalizasyonu veya standardizasyonuna yönelik farklı veri dönüştürme teknikleri kullanılır. Normalizasyon işlemi ile bütün değerler aynı ölçeğe getirilirler. Kullanılan yöntemler arasında Min-max normalleştirmesi, z-skor normalleştirmesi, 10 tabanlı logaritma kullanımı sayılabilir.

Min-max normalizasyon işlemi sırasında öncelikle veri seti içerisinde yer alan en küçük ve en büyük değerler tespit edilir. Veri setindeki değerlerin en küçüğü 0, en büyüğü 1 olacak şekilde veriler 0 ile 1 aralığına dağıtılarak standardizasyon işlemi gerçekleştirilir. Min-max normalizasyonu aşağıdaki formül ile ifade edilir.

$$X_{nor} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{2.1}$$

Z-skor standartlaştırma metodunda ortalama değer ile standart sapma değerlerinden faydalanılarak normalizasyon işlemi gerçekleştirilir. μ ortalama değeri, $\sigma_x$standart sapmayı belirtecek şekilde Z-skor normalizasyonu aşağıdaki formül ile ifade edilir.

$$X_{nor} = \frac{x-\mu}{\sigma_x} \quad (2.2)$$

Z-skor normalizasyonu sonrası değişken ortalaması 0, standart sapması 1 olan veri setine dönüşür.

2.1.2.3 İndirgeme

Günümüzde veri boyutunun büyümesi işlem süresinin artmasına, analiz çalışmalarının zorlaşmasına yol açmaktadır. Bu durumun önüne geçebilmek için veri indirgeme yöntemleri uygulanır. Boyut sayısının azaltılması, örnekleme, genelleme ve veri küpü kullanılan yöntemler arasındadır.

2.1.3 Veri Madenciliği Yöntemleri

Veri madenciliği çalışmaları sırasında kullanılan farklı yöntem ve algoritmalar mevcuttur. Yaygın olarak kullanılmakta olan veri madenciliği yaklaşımları arasında kümele, sınıflandırma ve birliktelik kuralları sayılabilir. Kullanılan öğrenme metoduna göre veri madenciliği modelleri, denetimli tip öğrenme (supervised learning,) ve denetimiz (unsupervised learning) tip öğrenme yöntemleri olarak iki gruba ayrılabilirler. Denetimli tip öğrenme uygulamalarında girdi veri değerlerinin bir kısmı eğitim seti olarak kullanılarak, çıktı değeri tahmin edilmeye çalışılmaktadır. Denetimsiz tip öğrenme uygulamalarında ise girdi değerleri arasındaki örüntüler bulunup, birbirleri arasındaki ilişki keşfedilmeye çalışılır. Denetimsiz tip öğrenme uygulamalarında eğitim seti bulunmaz. Denetimli tip öğrenme yöntemi sırasında eğitim seti oluşturulması ihtiyacı bulunduğu için, denetimsiz tip öğrenme yöntemine göre çalışma süresi açısından denetimsiz tip öğrenme yönteminin dezavantaja sahip olduğu söylenebilir.

Kümeleme çalışmaları denetimsiz tip öğrenme modeline, sınıflandırma çalışmaları ise denetimli tip öğrenme çalışmaları grubuna örnek verilebilir. Yapay Sinir Ağları, Destek Vektör Makineleri ve Karar Ağacı, Lojistik Regresyon yaygın olarak kullanılan denetimli tip öğrenme yöntemleri arasındadır. Kümeleme çalışmaları sırasında temelde bir öngörüden yararlanılmadan, veri seti içerisinde yer alan benzer özellik taşıyan veriler gruplanmaya çalışılır. Birliktelik kuralları çalışmaları da sırasında da doğru kararlar verebilmek için veriler analiz edilerek gizli kalmış örüntüler bulunmaya çalışılır.

Birliktelik kuralları ve özdüzenleyici haritalar denetimsiz tip öğrenme uygulamalarına örnek olarak verilebilirler. Veri madenciliği çalışmalarında istatistik ve yapay zekâ analizleri kullanılmaktadır.

2.2.KÜMELEME

2.2.1 Kümeleme Kavramı

Kümeleme işlemi nesneleri benzerlik özelliklerine kategorize ederek gruplara ayırmaya yarayan istatistiksel bir yöntemdir. Kümeleme işleminin amacı gözlemlenen nesnelerin benzerliklerini (yakınlıklarını) göz önünde bulundurarak, benzer nesneleri aynı kümede toplayarak, ayrık küme grupları oluşturmaktır. Bu sayede farklı kümeler içinde yer alan nesnelerin uzaklıkları maksimum, aynı küme içindeki nesnelerin uzaklıkları minimum seviyeye getirilir (Hastie ve diğ., 2009).

Veri madenciliği yöntemlerinden biri olan kümeleme günümüzde farklı alanlarda yaygın olarak kullanılmaktadır. Biyomedikal (McLachlan, 1992; Dilts ve diğ., 1995; Lu ve diğ., 2019; Suzuki ve Shimodaira, 2006), konuşma tanıma (Ault, 2018), örüntü tanıma (Chen ve diğ., 2019), görüntü işleme, sosyal ağ analizi, uzay araştırmaları, derin öğrenme gibi güncel konularda kümeleme çalışmaları yapılmaktadır.

Kümeleme gözetimsiz öğrenme şekli olduğu için oluşturulacak kümelerin özellikleri ve sayısı önceden bilinmemektedir. Bazı algoritmalar oluşturulacak küme sayısını kullanıcıdan girmesi isterler. Burada girilecek uygun değerdeki küme sayısı, kümeleme kalitesi için kritik öneme sahip olmaktadır. Kümeleme performansını ölçmek ve optimum küme sayısını belirlemek için içsel ve dışsal indeksler kullanılmaktadır. Veri üzerinde ön bilgiye sahip olunmadığı durumda içsel indeksler kullanılır. Veri madenciliği çalışmaları sırasında kullanılan farklı kümeleme yöntemleri bulunmaktadır. Yöntemler özellikleri itibari ile farklı gruptaki yöntemler ile çakışabildiğinden net bir gruplama yapılamamaktadır. Literatürde gruplamalar arasında farklılıklar görülebilmektedir. Aggarwal (2015) tarafından yapılan kümeleme yöntemleri gruplandırması aşağıdaki şekildedir:

- Temsil Tabanlı Yöntemler (Representative – Based Algorithms),
- Hiyerarşik Yöntemler (Hierarchical Clustering Algorithms),
- Olasılık Modeli Tabanlı Yöntemler (Probabilistic Model-Based Algorithms),

- Izgara Tabanlı ve Yoğunluk Tabanlı Yöntemler (Grid-Based & Density Based Algorithms),
- Grafik Tabanlı Yöntemler (Graph-Based Algorithms).

Kümelemeye yönelik algoritmalar ve bu algoritmalar üzerine çalışma yapan kişiler aşağıdaki tabloda belirtilmiştir (Han ve diğ., 2012; Aggarwal, 2015).

Tablo 2.2: Kümeleme algoritmaları ve çalışma sahipleri.

Yöntem	**Algoritma**	**Çalışma Yapan Kişi**
K-Means		Lloyd, MacQueen
K Medoids	PAM, CLARA	Kaufman & Rousseeuw
K-Modes		Huang, Chaturevedi & Green & Carroll
K Prototypes		Huang
Scalable Yöntem	CLARANS	Ng & Han, Ester & Krigel&Xu
K-Based Scalable		Bradley & Fayyad & Reina
Agglomerative Hierarchical		Day & Edelsbrunner
Agglomerative Hierarchical	AGNES	Kaufman & Rousseeuw
Divise Hierarchical	METIS	Karypis & Kumar
Divise Hierarchical	DIANA	Kaufman & Rousseeuw
Scalable Yöntem	BIRCH	Zhang & Ramakrishnan &Livny
Scalable Yöntem	CURE	Guha & Rastogi & Shim
Hierarchical	ROCK	Guha & Rastogi & Shim
Grafik Tabanlı	CHAMELEON	Karypis & Han & Kumar
Probabilistic Hierarchical		Friedman, Heller & Ghahramani
Yoğunluk Tabanlı	DBSCAN	Ester & Kriegel & Sander & Xu
Yoğunluk Tabanlı	OPTICS	Ankerst & Breunig & Kriegel & Sander
	DENCLUE	Hinneburg & Keim
	DENCLUE2.0	Hinneburg&Gabriel
Izgara Tabanlı	STING	Wang&Yang&Muntz
Izgara Tabanlı	WAVECLUSTER	Sheikholeslami&Chatterjee&Zhang
Izgara Tabanlı	MAFIA	Goil &Nagesh&Choudhary
Scalable Yöntem		Gibson&Kleinberg&Raghavan, Guha&Rastogi&Shimi, Ganti&Gehrke&Ramakrishnan
Fuzzy Kümeleme		Kaufman &Rousseeuw, Bezdek&Paul
High Dimensional Clustering	CLIQUE	Agrawal&Gehrke&Gunopulos&Raghavan
High Dimensional Clustering	PROCLUS	Aggarwal&Procopiuc&Wolf&Yu&Park
High Dimensional Clustering	ORCLUS	Aggarwal&Yu

Literatürde kümelemenin Hiyerarşik Kümeleme ve Hiyerarşik Olmayan Kümeleme yöntemleri şeklinde iki ana başlıkta ele alındığı yaklaşımlarda da mevcuttur. Hiyerarşik Olmayan Kümeleme yöntemlerinde küme sayısı ile ilgili olarak ön fikir söz konusu iken, Hiyerarşik Kümele yöntemlerinde ise söz konusu değildir. Hiyerarşik Kümeleme Yönteminde önceden tanımlanan kriterlere göre kümeler kademeli olarak alt kümelere ayrıştırma veya kümelerin birleştirilmesi yaklaşımıyla meydana getirilirler (Lorr, 1983; Kaufman ve Rousseeuw, 1990). Hiyerarşik kümeleme yaklaşımına örnek olarak en yakın komşu algoritması ile en uzak komşu algoritması verilebilirler. Hiyerarşik olmayan kümeleme yaklaşımında gözlemler önceden belirlenen sayıda kümelere dağıtılırlar, bu yönü ile daha az hesaplama yükünün oluşması sağlanır. K-ortalamalar ve K-Medoids algoritmaları hiyerarşik olmayan yaklaşım türündendir (Dilts ve diğ., 1995).

2.2.2 Uzaklık Kavramı ve Çeşitleri

Kümeleme yönteminde verilerin benzerlik veya farklılıklarını saptayabilmek için kullanılan ölçülerden bir tanesi de uzaklıktır. Kümeleme çalışmalarında nümerik nitelikler için kullanılan uzaklık ölçülerinden bazıları Öklid, Manhattan ve Minkowski'dir (Dilts ve diğ., 1995). Nümerik nitelikler için nesneler arasında uzaklık mesafesi azaldıkça benzerlik artmakta, mesafe arttıkça benzerlikler azalmaktadır.

En yaygın olarak kullanılan uzaklık ölçüsü Öklid uzaklığıdır. İki nesne arasındaki mesafenin düz çizgi uzunluğu şeklinde ifade edilmesidir. n adet nümerik değere sahip a ve b nesneleri $a = (x_{a1}, x_{a2}, \ldots, x_{an})$ ve $b = (x_{b1}, x_{b2}, \ldots, x_{bn})$ olarak tanımlanabilir,

İki nesne arasındaki Öklid uzaklığı aşağıdaki şekilde gösterilebilir.

$$\mathrm{d}(a,b) = \sqrt{(x_{a1} - x_{b1})^2 + (x_{a2} - x_{b2})^2 + \cdots + (x_{an} - x_{bn})^2} \tag{2.3}$$

Nümerik nitelik sayısının 2 olması durumunda a ve b nesneleri arasındaki Öklid uzaklığı aşağıdaki şekilde ifade edilir.

$$\mathrm{d}(a,b) = \sqrt{(x_{a1} - x_{b1})^2 + (x_{a2} - x_{b2})^2} \tag{2.4}$$

Minkowski uzaklık fonksiyonu aşağıdaki şekilde gösterilebilir. Minkowski uzaklık fonksiyonu daha geneldir.

$$\mathrm{d}(a,b) = \sqrt[p]{(x_{a1} - x_{b1})^p + (x_{a2} - x_{b2})^p + \cdots + (x_{an} - x_{bn})^p} \quad (2.5)$$

Öklid ve Manhattan uzaklıkları, Minkowski uzaklığının özel bir durumlarıdır. Minkowski formülünde değişken sayısı olan p değerinin 2 olması durumunda Öklid uzaklığı elde edilir. Minkowski formülünde p değerinin 1 olması durumunda ise Manhattan uzaklığı elde edilir.

Nümerik nitelik sayısının n olduğu durumda, a ve b nesneleri arasındaki Manhattan uzaklığı (şehir blok uzaklığı) aşağıdaki şekilde ifade edilir.

$$\mathrm{d}(a,b) = |x_{a1} - x_{b1}| + |x_{a2} - x_{b2}| + \cdots + |x_{an} - x_{bn}| \quad (2.6)$$

Öklid ve Manhattan uzaklıklarından a, b ve c olarak isimlendirilen 3 farklı nesne için aşağıdaki matematiksel özellikleri sağlaması beklenir.

$d(a,b) \geq 0, Uzaklık\ değeri\ negatif\ sayı\ olamaz$. (2.7)

$d(a,a) = 0, Bir\ nesenenin\ kendisine\ olan\ uzaklığı\ sıfırdır.$ (2.8)

$d(a,b) = 0 \leftrightarrow$ a=b, Uzaklık 0 ise iki nesne aynı değere sahiptir. (2.9)

$d(a,b) = d(b,a), Uzaklık\ simetriktir.$ (2.10)

$d(a,c) \leq d(a,b) + d(b,c)$. Üçgen eşitsizliği. İki nesne arasındaki uzaydaki mesafe, dolaylı olarak farklı bir nesneye uğrayarak gidilen mesafeden daha fazla olamaz. (2.11)

Minkowski genel formülünde değişken sayısı p'nin ∞ değeri alması durumunda sup veya dominans olarak isimlendirilen en uzak mesafe elde edilir. Sup uzaklık değerinin bulunması için iki nesne arasındaki en uzak mesafe değerini elde edilmesini sağlayan nitelik seçilir.

Sup uzaklığı aşağıdaki formül ile ifade edilir.

$$\mathrm{d}(a,b) = \lim_{P \to \infty} \left(\sum_{n=1}^{m} |x_{an} - x_{bn}|^{\frac{1}{p}} \right)^p \quad (2.12)$$

2.2.3 K-ortalamalar

K-ortalamalar algoritması kümeleme çalışmalarında yoğun olarak kullanılmakta olan temel algoritmalardan bir tanesidir.

Standart algoritma ilk olarak Lloyd (1957) tarafından darbe kodu (Pulse Code Modulation) modülasyonu için bir teknik olarak önerilmiştir (Lloyd, 1982). Sonrasında MacQueen (1967) tarafından yapılan çalışma ile geliştirilmiştir. Hiyerarşik olmayan kümeleme algoritmalarındandır.

K-ortalamalar algoritması ile bir gözlem sadece bir kümede içinde yer alacak şekilde, veriler ayrık kümelere yerleştirilirler. Bu yöntemde oluşturulacak küme sayısına başlangıçta karar verilir. Öklid uzayında veri seti rastgele olarak başlangıçta belirlenen k adet kümeye dağıtılır. C1, C2… Ck olarak isimlendirilen kümelerin kesişimlerinin boş küme olması gerekir. Bu durum $(1 \leq i, j \leq k)$ için $Ci \cap Cj = \emptyset$ şeklinde ifade edilebilir. (Han ve diğ., 2012). Her kümeye yerleştirilen gözlemlerin ortalamaları alınarak, her kümenin ayrı ayrı merkezleri bulunur. Sonrasında her bir gözlemin küme merkezlerine olan uzaklıkları bulunur. Benzerlik oranının küme içinde maksimum olacağı, kümeler arası benzerlik oranının ise minimum olacağı şekilde gerek görülen gözlemler farklı kümelere yerleştirilir. Yeniden her kümenin ortalama değeri belirlenerek, yeni küme merkezleri saptanır ve her bir gözlemin yeniden küme merkezlerine uzaklıkları hesaplanır. Hata kare değerini minimize edecek şekilde tekrar eden adımlar, değişiklik olmayıncaya kadar devam eder. Sonlanmaya kadar olan olacak tekrar sayısına, yazılım ile başlangıçta da karar verilebilir.

Hata kare hesaplaması ile kümelere ayırma işleminin kalitesi değerlendirilir. Küme içi ayırma kalitesi küme merkezi ile o kümeye ait bütün gözlemlerin Öklid uzaklıklarının karelerin toplamı ile bulunur. Bütün kümelerin kalitesi ise bütün kümelerin hata kare değerlerinin toplanması ile elde edilir. Bütün kümelerin hata kare toplamı (E) aşağıdaki formül ile elde edilir. Formülde pϵCi olacak şekilde, c_i küme merkezini ifade etmektedir.

$$E = \sum_{i=1}^{k} \sum_{P \in C_i} \mathrm{d}(p, c_i)^2 \tag{2.13}$$

K-ortalamalar algoritmasının yaygın olarak tercih edilmesinin nedenleri arasında basit yapısı, uygulama kolaylığı ve deneysel araştırma sonuçlarındaki başarısı sayılabilir (Wang ve diğ., 2017; Steinley, 2006)]. Büyük veri setleri üzerinde işlem yapma açısından etkilidir. Fakat aykırı değerlere karşı duyarlı olması yönüyle dezavantaja sahiptir.

2.2.3.1 K Küme Sayısının Belirlenmesi ve İndeksler:

Kümeleme çalışmasının temel adımlarından bir tanesi de küme sayısının belirlenmesidir. Küme sayısı, objektif olarak belirlenebildiği gibi kümeleme çalışması yapan kişinin, geçmişteki benzeri deneyimlerine ve konuya yönelik bilgisine dayanılarak sübjektif olarak da saptanabilmektedir (Cebeci ve diğ., 2015).

Farklı indeksler kullanılarak kümeleme çalışması sonunda elde edilen değerler geçerlilik ve iyilik yönünden sınanırlar. Kümele işlemi başlatılmadan önce belirlenen küme sayısı ve başlangıç küme merkezleri, kümeleme işleminin sonucu üzerinde etkili olmaktadır. k küme sayısı değiştirilerek K-ortalamalar algoritması çalıştırılır. Çıktı değerleri farklı indekslerle değerlendirilerek en uygun küme sayısı değeri saptanır. K küme sayısını belirlemeye yönelik indeksler iç, dış ve rölatif olmak üzere üç gruba ayrılabilirler (Cebeci ve diğ., 2015; Wang ve diğ., 2017).

Küme sayısının belirlenmesine yönelik farklı yollar mevcuttur. Küme sayısının belirlenmesine yönelik basit yöntemlerden birisi de veri sayısını dikkate alan yaklaşımdır. Bu yönteme göre n noktalı bir veri setinde, küme sayısının yaklaşık $\sqrt{\frac{n}{2}}$ olarak alınır. Bu metoda göre her bir kümenin eleman sayısının $\sqrt{2n}$ nokta olması beklenir (Han ve diğ., 2012).

Kümelemenin temel amacı küme içindeki nesnelerin olabildiğince birbirine benzemesi, farklı kümelerde bulunan nesnelerinde olabildiğince birbirinden uzak olmasını sağlamaktadır. Küme içlerinin yoğunluğu ve kümeler arasının ayrılma değeri kümelemenin kalitesini yansıtır. Kümeleme kalitesini gösteren indekslerden bazılarının değeri artıkça kümeleme kalitesi artarken bazılarının değeri azaldıkça kümeleme kalitesi artar. Daha iyi kümeleme kalitesi için indekslerin davranışı aşağıdaki tabloda yukarı ve aşağı ok yönleri ile özetlenmiştir. Yukarı ok yönü ile belirtilen indekslerin değeri artıkça kümeleme kalitesi artarken, aşağı ok yönü ile belirtilen indekslerin değeri azaldıkça kümeleme kalitesi azalır (Arbelaitz ve diğ., 2013; Kim ve Ramakrishna, 2005).

Tablo 2.3: İndekslerin değer değişimi (José-García ve Gómez-Flores, 2021).

İndeks İsmi	Kısaltma	Daha İyi Kümeleme Kalitesi
Silhouette	Sil	↑
Dunn	DI	↑
Calinski-Harabasz	CH	↑
Gamma	G	↓
C	CI	↓
gD31 (Generalized Dunn)		↑
gD33 (Generalized Dunn)		↑
gD41 (Generalized Dunn)		↑
gD43 (Generalized Dunn)		↑
gD51 (Generalized Dunn)		↑
CS		↓
Davies_Bouldin	DB	↓
Score Fonksiyonu	SF	↑
Sym		↑
SymDB		↓
SymD		↑
Sym33		↑
COP		↓
Negentropy Increment	NI	↓
SV		↑
OS		↑
XB		↓

2.2.3.2 Dışsal İndeksler

Kümeleme çalışmalarıyla elde edilmiş sonuçların, önceden elde edilmiş bilgiler, uzman görüşleri ve kıyas kriterlerine göre karşılaştırılması ile dışsal değerlendirme yapılır. Karşılaştırmada kullanılan dışsal ölçüler arasında Entropi indeksi, Saflık indeksi (Purity), NMI, VI ve F indeksi sayılabilir (Zaki ve diğ., 2014).

2.2.3.4 Entropi İndeksi

Entropi i kümelerin sınıf etiketlerinin saflığını ölçer. Her kümenin entropi değeri aşağıdaki formül ile hesaplanır (Rendón ve diğ., 2011).

$$E_j = \sum_i P_{ij} Iog(P_{ij}) \tag{2.14}$$

Toplam entropi değeri kümelerin entropi değerlerinin ağırlıklı toplamı ile bulunur, aşağıdaki formül ile hesaplanır. m toplam küme sayısını, n veri sayısını, n_j ise j kümesinin boyutunu ifade eder.

$$E = \sum_{j=1}^{m} \frac{n_j}{n} E_j \tag{2.15}$$

2.2.3.5 NMI – Normalize Karşılıklı Bilgi İndeksi

Etiketlenmiş 2 nesnenin Normalize Karşılıklı Bilgi (Normalized Mutual Information) indeksi aşağıdaki formül ile hesaplanır.

$$NMI(X,Y) = \frac{I(X,Y)}{\sqrt{H(X),H(Y)}} \tag{2.16}$$

Formülde $I(X,Y)$; X ve Y değişkenleri arasındaki karşılıklı bilgiyi, $H(X)$; X kümeleme yöntemine göre entropi değerini ve $H(Y)$ ise Y gerçek etiket durumuna göre entropi değerini ifade etmektedir (Rendón ve diğ., 2011).

Kümeleme kalitesine yönelik Normalize Karşılıklı Bilgi İndeksinin aldığı değer [0,1] aralığındır. Bu indeksin aldığı değerin 1'e yaklaşması kümeleme kalitesinin daha iyi olduğunu belirtmektedir (Altunkaynak, 2019).

2.2.3.6 VI indeksi- Bilgi Değişimi İndeksi

Bilgi değişimi indeksi (Variation of information) aşağıda belirtilen formül ile hesaplanır. Karşılıklı bilgi ve entropi değerlerine bağlıdır.

$$VI = H(X) + H(Y) - 2I(X,Y) \tag{2.17}$$

2.2.3.7 F İndeksi

Her sınıf için küme geri çağırma (R, Recall) ve kesinlik (P, Precision) değerleri aşağıdaki formüllerde belirtildiği şekilde hesaplanır. n_{ij}, j kümesi içerisindeki i sınıfına ait nesne sayısını

n_j j kümesi içerisindeki nesne sayısını ve n_i ise i sınıfı içerisindeki nesne sayısını belirtmektedir (Dalli, 2003; Rijsbergen, 1979).

$$R(i,j) = \frac{n_{ij}}{n_i} \tag{2.18}$$

$$P(i,j) = \frac{n_{ij}}{n_j} \tag{2.19}$$

i sınıfı ve j kümesine ait F indeksi aşağıdaki formül ile hesaplanır.

$$F(\mathrm{i},j) = \frac{2R(\mathrm{i},j)\cdot P(i,j)}{R(\mathrm{i},j)+P(\mathrm{i},j)} \tag{2.20}$$

F indeksi [0,1] aralığında değer alır ve değeri arttıkça kümeleme kalitesi artar.

2.2.3.8 Saflık (Purity)

Ci kümesinin saflık değeri aşağıdaki formülde belirtildiği şekilde hesaplanır.

$$purity_i = \frac{1}{n_i} max_{j=1}^{k}\{n_{ij}\} \tag{2.21}$$

Bütün kümeleme çalışmasının saflık değeri tekil kümelerin saflık değerlerinin ağırlıklı toplamıyla elde edilir. C kümeleme yöntemi için saflık değeri aşağıdaki formüle göre hesaplanır.

$$purity = \frac{1}{n}\sum_{i=1}^{r} max_{j=1}^{k}\{n_{ij}\} \tag{2.22}$$

2.2.3.9 İçsel İndeksler

İç İndeksler, aynı küme içindeki elemanların ne kadar yakın olduğunu, kümeler arasındaki elemanların ise ne kadar uzak mesafede olduğunu yani benzeşme değerlerinin düşük olduğunu anlamak için kullanılır. Kümeleme kalitesinin hesaplanması sırasında benzerlikler dikkate alınır. İç indekslere örnek olarak Dunn, Davies_Bouldin, C indeksi, SD indeksi sayılabilir. (Wang ve diğ., 2017). İç indeksler ile yapılan kümeleme değerlendirmelerinde sınıf etiketlerine ihtiyaç duyulmaz.

Rölatif İndeksler, kümeleme sonuçlarını kıyaslamak için farklı parametre değerlerindeki aynı algoritmaları kullanırlar (Wang ve diğ., 2017).

2.2.3.10 Silhouette İndeksi

a(i) değeri, kümenin i. elemanının bulunduğu küme elemanları ile mesafesi veya benzeşmezlik ortalamasıdır. b(i) değeri ise kümenin i. elemanının diğer küme elemanları ile olan mesafesi veya benzeşmezlik ortalamasını belirtecek şekilde Silhouette indeksi aşağıdaki formülle ifade edilir (Rousseeuw, 1987).

$$S(\mathrm{i}) = \frac{b(\mathrm{i}) - a(\mathrm{i})}{maks((a(\mathrm{i}), b(\mathrm{i}))} \tag{2.23}$$

$$S(i) = \begin{cases} 1 - \frac{a(\mathrm{i})}{b(\mathrm{i})}, & \text{eğer} \quad a(\mathrm{i}) > b(\mathrm{i}) \\ 0 \quad , & \text{eğer} \quad a(\mathrm{i}) = b(\mathrm{i}) \\ \frac{b(\mathrm{i})}{a(\mathrm{i})} - 1 \ , & eğer \quad a(\mathrm{i}) < b(\mathrm{i}) \end{cases} \tag{2.24}$$

Yukarıdaki formülden de anlaşılabileceği gibi küme içi ve kümeler arası mesafe değerine göre Silhouette indeksinin aldığı değer -1 ile +1 arasında değişir.

$$-1 \leq S(\mathrm{i}) \leq 1 \tag{2.25}$$

Silhouette indeksinin pozitif değerler alması istenir. İndeksin değeri 1'e yaklaştıkça kümelemenin kalitesi artar. a(i) değerinin olabildiğince 0'a yaklaşması beklenir. a(i)=0 olduğunda, yani kümenin i. elemanının bulunduğu kümenin elemanları ile benzeşmezlik ortalaması 0 olduğunda indeks maksimum değeri olan 1 değerini alır. İndeks 0 değerini alıyorsa iki küme arasındaki sınıra yakın değer aldığı anlaşılır (Mohammed, 2014).

İndeksin negatif değer alması istenmez. Eğer indeks -1 değerini alıyorsa tamamen yanlış bir kümede olduğu sonucuna varılır. Silhouette indeksi hem kümelerin ayrı ayrı kümeleme kalitesini, hem de bütün elemanları kapsayan kümeleme çalışmasının kalitesini hesaplamak için kullanılabilir. Tüm kümenin kümeleme kalitesi, kümedeki elemanların Silhouette indekslerinin ortalaması alınarak bulunur.

n adet elemana sahip bütün kümenin Silhouette indeksi kümeyi oluşturan elemanların S(i) değerlerinin ortalamasına eşittir, aşağıdaki formül ile ifade edilir.

$$Sil = \frac{1}{n}\sum_{i=1}^{n} S(i) \tag{2.26}$$

2.2.3.11 Dunn İndeksi

Dunn (1974) tarafından kümeleme algoritmasının kalitesini değerlendirmek amacıyla oluşturulan Dunn indeksi, farklı kümelerde bulunan noktalar arası en kısa mesafenin, aynı küme içerisinde iki nokta arasındaki en uzak mesafeye oranı olarak tanımlanır. Dunn indeksi, küme elemanları arasında değişimin az olduğu, yoğun kümelerin oluşturulması amacına dayanmaktadır. Dunn indeksinin değeri arttıkça kümelemenin kalitesi artar.

$d(C_i, C_j)$, C_i ve C_j kümeleri arasındaki mesafeyi, $d'(C_k)$ k kümesinin içerisindeki iki nokta arasındaki mesafeyi ve küme çapı olarak isimlendirilir. n küme sayısını belirtecek şekilde Dunn indeksi aşağıdaki şekilde hesaplanır.

$$D = \min_{1\leq i\leq n}\left\{\min_{1\leq j\leq n, i\neq j}\left\{\frac{\mathrm{d}(C_i,C_j)}{\max_{1\leq k\leq n}\mathrm{d}'(C_k)}\right\}\right\} \tag{2.27}$$

Kümeleme çalışmasında değerlendirilen veri seti kompakt ve iyi ayrılmış kümelerden oluşuyorsa, kümeler arası mesafenin uzak, küme çaplarının da küçük olması beklenir. Yoğun ve iyi ayrılmış kümelere sahip veri setine yönelik kümeleme kalitesini gösteren Dunn indeksinin değeri büyüktür. Dunn indeksine yönelik dezavantajlar arasında zamansal kompleksliği ve veri setindeki gürültüye karşı olan hassasiyeti sayılabilir. Verideki gürültüden dolayı küme çapının hatalı algılanması, Dunn indeksinin denklemde değerinin yanlış hesaplanmasına yol açacaktır (Halkidi ve diğ., 2002).

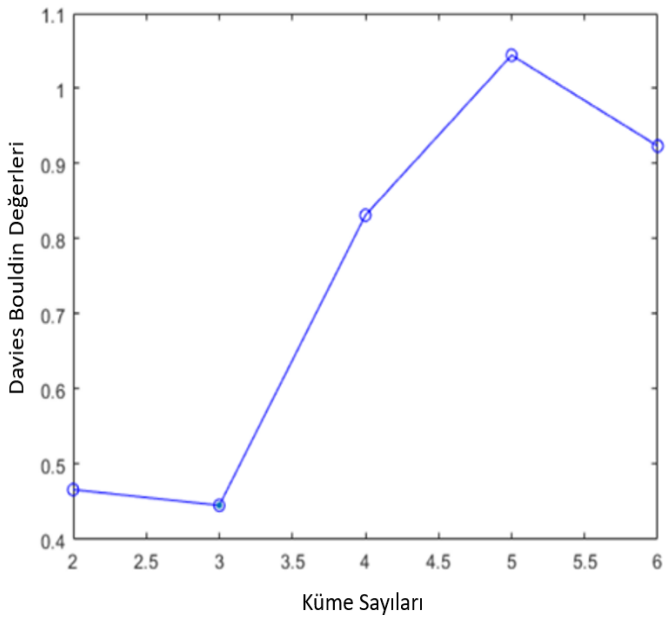
2.2.3.12 Davies_Bouldin İndeksi

Düşük Davies_Bouldin indeksi, kümeler arası mesafenin fazla olduğunu, yani kümelerin iyi ayrıştığını, kümelerin kendi içindeki noktaların küme merkezine mesafelerinin az olduğunu belirtir. Davies_Bouldin indeksinin düşük olması kümeleme kalitesinin iyi olduğunu gösterir (Davies ve Bouldin, 1979).

n elemanlı kümede, $d(c_i, c_j)$ C_i ve C_j kümelerinin merkezleri arasındaki mesafeyi, σ_i ve σ_j ise sırasıyla C_i ve C_j kümesindeki elemanların kendi küme merkezlerine olan mesafelerinin ortalamasını gösterecek şekilde Davies_Bouldin indeksi aşağıdaki formül ile hesaplanabilir.

$$DB = \frac{1}{n}\sum_{i=1}^{n} \underset{i\neq j}{max} \left\{\frac{\sigma_i+\sigma_j}{\mathrm{d}(c_i,c_j)}\right\} \tag{2.28}$$

Kümelerin birbirleri arasındaki benzerliklerin en az seviyede olması hedeflendiğinden, kümeleme işleminde DB değerinin minimum olması amaçlanır. Küme sayısı ile DB indeksi arasında bir eğilim (trend) yoktur, bu nedenle küme sayılarına karşı DB değeri grafiğinde DB değerinin minimum olacak şekilde grafik üzerinde optimum küme sayısı seçilir. (Halkidi ve diğ., 2002) Aşağıdaki küme sayısına karşı Davies_Bouldin değeri grafiği için optimum küme sayısı 3'tür.



Şekil 2.4: Davies_Bouldin indeksine göre optimum küme sayısı.

2.2.3.13 Calinski-Harabasz İndeksi

n elemanlı kümede, k küme sayısını gösterecek şekilde Calinski-Harabasz indeksi aşağıdaki formül ile hesaplanabilir (Calinski ve Harabasz, 1974).

$$CH = \frac{n-k}{k-1} \cdot \frac{tr(s_{Bk})}{tr(s_{wk})} \quad (2.29)$$

Formülde yer alan $tr(s_{Bk})$ ifadesi kümeler arası mesafe karelerinin toplamını, $tr(s_{wk})$ ifadesi ise küme içi mesafe karelerinin toplamını ifade etmektedir. Calinski-Harabasz indeksinin büyük olması kümeleme kalitesinin yüksek olduğunu gösterir. Bu sebeple Calinski-Harabasz değerine karşı küme sayısı grafiğinde Calinkski-Harabasz değerinin en yüksek olduğu k küme sayısı, en uygun küme sayısı değeri olarak saptanır. Yoğun ve iyi dağılmış kümeler elde edebilmek için kümeler arası mesafe karelerinin toplamı $tr(s_{Bk})$ 'nın maksimum, küme içi mesafe karelerinin toplamı $tr(s_{wk})$'nın minimum olması istenir.

2.2.3.14 C İndeksi

n ifadesi küme içindeki eleman çiftlerinin sayısını, S ifadesi aynı küme içinde bulunan eleman çiftlerinin mesafelerinin toplamını, S_{min} ifadesi eleman çiftleri arasındaki en küçük mesafeyi, S_{max} ifadesi eleman çiftleri arasındaki en büyük mesafeyi belirtecek şekilde C indeksinin değeri aşağıdaki formül ile hesaplanabilir.

$$C = \frac{S-S_{min}}{S_{max}-S_{min}} \quad (2.30)$$

Kümeleme kalitesinin değerini ölçmek için hesaplanan C indeksinin aldığı değer ne kadar küçükse kümeleme kalitesi o kadar iyidir. (Pala ve Aksaraylı 2020; Silahtaroğlu, 2020). C indeksi [0,1] aralığında değer alır.

2.2.3.15 Xie Beni İndeksi

Xie Beni indeksi, Bulanık C Ortalamalar algoritması kullanılarak elde edilen kümelerin kalitesini ölçmek için geliştirilmiştir.

Xie Beni indeksi kümelerin kompaktlık değerinin, ayrışma değerine oranına göre hesaplanır. Kümeler birbirinden ne kadar ayrı ve kompakt olursa XB indeksi değeri düşer yani kümeleme kalitesi artar. XB indeksi aşağıdaki formül ile hesaplanabilir (Xie ve Beni,1991).

$$XB = \frac{\sum_{i=1}^{C} \sum_{j=1}^{n} \mu_{ij}^2 \|v_i - x_j\|^2 / n}{\min_{i \neq j} \|v_i - v_j\|^2} \quad (2.31)$$

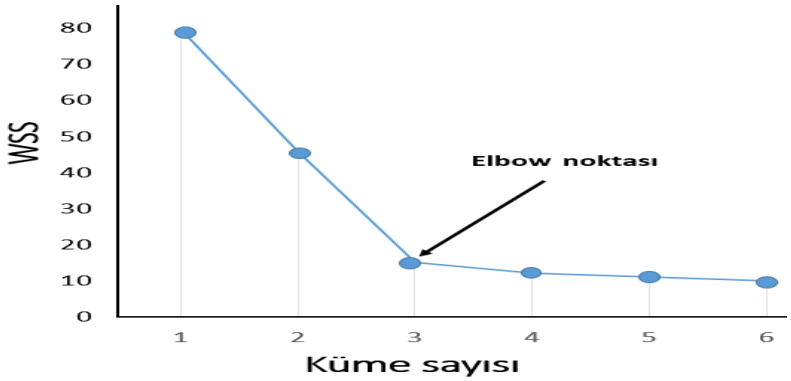
μ_{ij}, üyelik değerini, v_i kümelerin ağırlık merkezlerini, x_j vektörü belirtmek üzere denklemde m=2 değerini alması durumunda, Bulanık C Ortalamalar Algoritmasına yönelik XB değeri aşağıdaki şekilde ifade edilebilir.

$$XB = \frac{J_2}{n \cdot (d_{min})^2} \quad (2.32)$$

d_{min} değeri küme merkezleri arasındaki en küçük mesafeyi belirtmektedir.

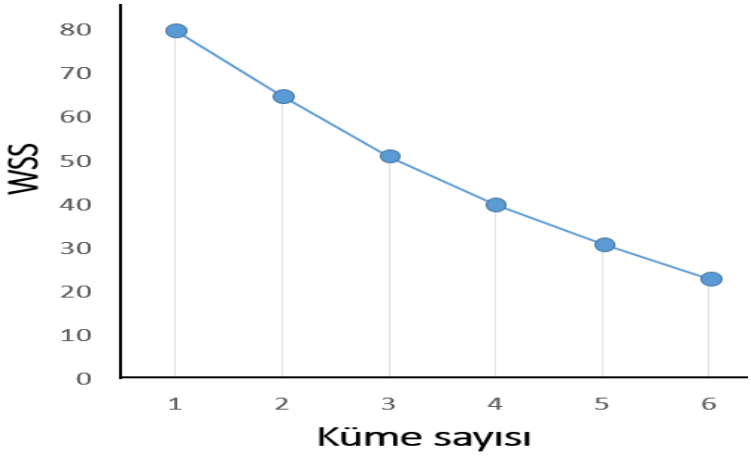
2.2.3.16 Elbow Yöntemi

Gözetimsiz kümeleme yönteminin önemli adımlarında biri olan küme sayısını belirlemek amacıyla kullanılan görsel bir tekniktir. Belirlenen her küme sayısı değeri için noktaların küme merkezlerine olan kareler toplamı bulunur. İyi belirlenen bir kümelemede, toplam küme içi kareler toplamının (within-cluster sum of square, WSS) küçük olması beklenir. Farklı değerdeki küme sayılarına göre toplam küme içi kareler toplamı hesaplanır. Küme sayısına karşılık gelen WSS değerlerine göre grafiği çizilir. Grafikte küme sayısı artıkça kare toplamında değişimin yeterince değişmediği nokta optimal k küme sayısı olarak belirlenir. Optimal küme sayı değerinden sonra, küme sayısı arttırılmaya devam edilmesi durumunda yeterince değişim olmayacak şekilde eğri plato şeklini alacaktır. Değişim miktarı küme sayısını daha fazla arttırmanın maliyeti karşılamayacaktır.



Şekil 2.5: Elbow noktasına göre küme sayısı.

Elbow yöntemi görsel bir tekniktir bu yüzden yukarıdaki grafikteki gibi değişimin net bir şekilde plato halini aldığı dirsek noktası belirgin ise, bu yöntem ile k küme değeri belirlenebilir. Görsel olarak değişimin plato halini almadığı durumlarda Elbow yöntemine göre optimal küme sayısına karar verilemeyecektir. Aşağıdaki şekildeki gibi eğrilerde net bir değişim gözlemlenemediği için Elbow yöntemi yararlı olmayacaktır (Shi ve diğ., 2021).



Şekil 2.6: Elbow yöntemine göre belirsizlik durumu.

2.2.5 R ile Küme Sayısının Belirlenmesi

Küme sayısını belirlemeye yönelik geçerlilik indekslerini kullanan farklı R dili paketleri mevcuttur. Bu paketler arasında cclust, clusterSim, clv, clValid, clusterCrit, NbClust, fviz_nbclust sayılabilir (Cebeci ve diğ., 2015).

clusterCrit, fviz_nbclust ve NbClust fonksiyonları kümeleme sayısı hesaplanmasında yaygın olarak kullanılmaktadırlar. factoextra R paketi içerisinde yer alan fviz_nbclust fonksiyonu elbow, silhouette ve gap istatistik yöntemlerini kullanarak K-ortalamalar ve K-Medoids gibi kümeleme metotlarında küme sayının hesaplanmasında kullanılabilir. NbClust R paketi içerisinde yer alan NbClust fonksiyonu, 30 farklı indeks kullanarak kümele sayısının hesap edilmesini sağlar. Küme sayısı, mesafe ölçüsü ve kümeleme yönteminin değişimine göre kullanıcıya en iyi kümeleme yapısını belirlemesine yönelik seçenekler sunmaktadır. R dili altında çalışan tek bir fonksiyon ile aynı anda bütün indekslere göre hesaplama yaparak küme sayısının belirlenmesini sağlar, kullanıcıya zaman anlamında da avantaj oluşturur (Charrad ve diğ., 2014).

2.3 BİRLİKTELİK KURALI

Veri madenciliğinin temel konularında biri olan birliktelik kuralları, olayların birbirleriyle olan ilişkisini inceleyerek, birlikte gerçekleşme olasılıklarını açıklar. Büyük veri kümeleri içinde birliktelik analizine yönelik ilk araştırma Agrawal ve diğ. (1993) tarafından yılında yapılmıştır.

İlişki analizi, perakende sektöründe müşteri satın alma alışkanlıklarını belirlemek amacıyla sıklıkla kullanılmaktadır. Literatürde pazar sepet analizi ismiyle de adlandırılan birliktelik kuralları müşterilerin satın aldıkları ürünleri analiz ederek, hangi ürünleri birlikte alma eğiliminde olduklarını ortaya çıkarır. Marketler ve mağazalar birliktelik kuralları sonucunda müşteriler tarafından birlikte alınan ürünleri birbirlerine yakın raflarda satışa sunmaktadırlar. Web tabanlı satış yapan alışveriş siteleri, video ve müzik uygulamaları kullanıcı ve müşterilerinin kullanım alışkanlıklarını analiz ederek, benzer ürünleri öneri olarak sunabilmektedirler.

Birliktelik kuralı sıklıkla kullanıldığı market sepeti analizi faaliyetlerinin yanında sayım verilerinin analizinde, CRM (Müşteri ilişkileri yönetimi) kullanıcı alışkanlıklarının analizinde, medikal tanı faaliyetlerinde, yerleşim düzeni tasarım çalışmalarında kullanılmaktadır.

Birliktelik kuralı analizi eş zamanlı gerçekleşen ilişkilerin değerlendirmesinde kullanıldığı gibi birbirini izleyen dönemlerde gerçekleşen ardışık zamanlı ilişkilerin değerlendirilmesinde de kullanılabilmektedir.

Birliktelik kurallarına göre ilişkiler ortaya çıkarılırken destek, güven ve kaldıraç ölçütlerinden yararlanılır. Destek (support) değeri bir seçeneğin tüm gözlem kümesi içerisinde kaç defa tekrarladığını belirtir. Minimum destek değeri önceden belirlenmelidir. n bütün gözlem kümesini ifade etmek üzere A ürün grubu ve B ürün grubunu birlikte içeren gözlemler N(A,B) şeklinde belirtmektedir. A ve B seçenek kümesine ait destek değeri aşağıdaki formül ile ifade edilebilir (Altunkaynak, 2019).

$$Destek = P(A,B) = \frac{N(A,B)}{n} \tag{2.33}$$

Güven ölçütü, A seçeneğini seçenlerin B seçeneğini de seçme olasılığını ifade etmektedir, aşağıdaki formül ile hesaplanabilir.

$$Güven = P(B/A) = \frac{N(A,B)}{N(A)} \tag{2.34}$$

Kaldıraç ölçütü Destek/Güven değerini belirtir. Kaldıraç değeri aşağıda yer alan formül ile hesaplanır.

$$Kaldıraç = \frac{P(A)}{P(B/A)} \tag{2.35}$$

Destek değeri [0,1] aralığında değer alır. Güven değeri de [0,1] aralığında değer alır. Birliktelik kuralları uygulanırken destek ve güven ölçütlerine yönelik eşik değeri belirlenir. Eşik değerinin altında kalan seçenekler elenir.

Tablo 2.4: Birliktelik kuralı algoritmaları (Erpolat, 2022).

Sıralı Algoritmalar	**Paralel ve Dağıtılmış Algoritmalar**
AIS	CD
SETM	PDM
Apriori	DMA
Apriori TID	DD
Apriori Hybrid	IDD
OCD	HPA
Partioning (Bölümleme Tekniği)	PAR
Sampling (Örnekleme Tekniği)	Candidate Distribution (Aday Dağılımı)
DIC	SH
CARMA	HD
FP Growth	

Birliktelik kuralında kullanılan algoritmalar yukarıdaki tabloda gösterildiği gibi sıralı ve paralel tip algoritmalar olmak üzere iki gruba ayrılabilirler. Birliktelik Kuralları analizi için kullanılan bazı algoritmalar arasında AIS, Apriori, Carma, GRI, Eclat, FP-Growth, Recursive Elemination, Tertius sayılabilir.

Birliktelik kuralı algoritmaları, işlem yapılan verinin büyüklüğü, hızı ve doğruluğu kriterleri yönünden aşağıdaki tabloda kıyaslanmıştır (Ghafoor ve Ahmad, 2021). Tabloda giriş veri büyüklüğü için 1 en az, 4 en fazla veriyi ifade edecek şekilde 4 kademede belirtilmiştir. Hız; 1 en yavaş, 4 en hızlı değeri ifade edecek şekilde 4 farklı kademede belirtilmiştir. Algoritmanın doğruluğu 1 en düşük, 4 en yüksek değeri ifade edecek şekilde 4 farklı kademede belirtilmiştir.

Tablo 2.5: Birliktelik kuralı algoritmalarının kıyaslaması (Erpolat, 2022).

Sıralama	Algoritma İsmi	Giriş Veri Büyüklüğü	Hız Faz1	Hız Faz2	Doğruluk
1	AIS	2	1	1	1
2	SETM	2	1	1	2
3	Apriori	1	3	1	2
4	Apriori TID	3	1	3	3
5	Apriori Hybrid	4	3	3	4
6	Eclat	3	3	2	3
7	Recursive Elimination	3	3	1	3
8	TERTIUS	1	1	2	3
9	FP Growth	4	3	3	4
10	Dyn-FP Growth	4	3	3	4
11	PSO	4	4	4	4

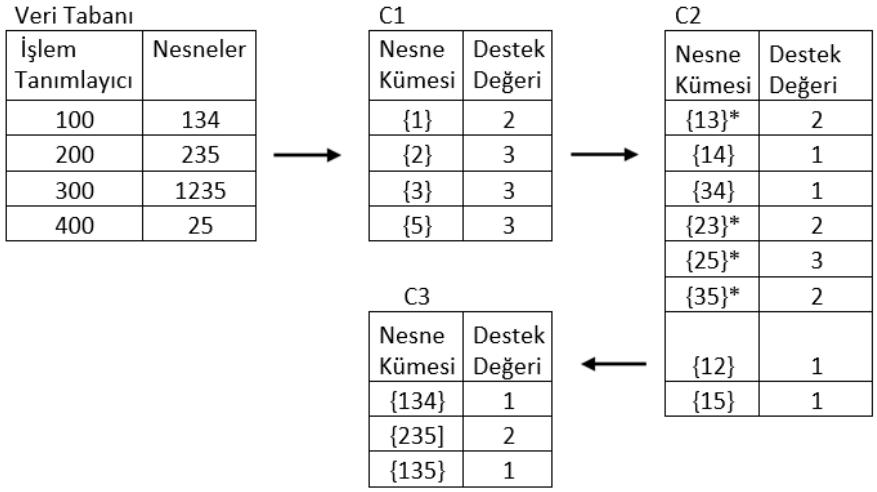
Kıyaslama tablosuna göre işlem yapılan veri büyüklüğü, işlem hızı ve algoritmanın doğruluk sonucu kriterlerinin hepsinde FP Growth algoritması ile Apriori ve Eclat algoritmalarına kıyasla daha iyi sonuçlar elde edilmiştir.

2.3.1 AIS Algoritması

AIS algoritması veri yığınları için birliktelik kuralı alanında geliştirilmiş ilk algoritmadır. AIS Algoritmasına yönelik çalışma 1993 yılında Agrawal, Imielinski ve Swami tarafından yayınlanmıştır. Aday nesne kümeleri, veri tabanı tarandığı esnada oluşturulur ve sayılır. Bir işlem tarandıktan sonra, önceki taramada geniş nesne kümelerinden hangilerinin bu işlemde de yer aldığı belirlenir. Bu işlemdeki nesneler ile yaygın nesne kümeler genişletilerek yeni aday nesne kümeleri oluşturulur. Geniş bir nesne küme olan *l*'nin nesneler ile birleşerek genişleyebilmesi için eklenecek olan nesnenin, büyük olması ve *l* nesne kümesinin bütün nesnelerinden sözlük harf sırası olarak sonra gelmesi gerekir. Budama tekniği kullanılır. Gereksiz olan nesneler atılır. Minimum destek değeri ile kıyaslama yapılır. Minimum destek değerine sahip olan ya da daha büyük destek değerine sahip olan aday nesneler yaygın nesne kümeler olarak belirlenir. Yaygın nesne kümeler sonraki adımdaki aday nesne kümenin elde edilmesinde kullanılır. Bu işlemler yeni yaygın nesne bulunamayıncaya kadar devam eder (Agrawal ve diğ., 1993, Agrawal ve Srikant, 1994).

AIS algoritmasında veri tabanı yaygın nesneleri bulmak için çok defa taranır, bu durum işlem hızının düşmesine yol açar. Oluşturulan aday nesne kümelerinin çoğunun yaygın nesne küme olmaması algoritmanın dezavantajıdır.

AIS algoritması çalışma şekli örneği aşağıdaki şekilde gösterilmiştir.



Şekil 2.7: AIS algoritması örneği (Kumbhare ve Chobe, 2014).

2.3.2 SETM Algoritması (Set Oriented Mining)

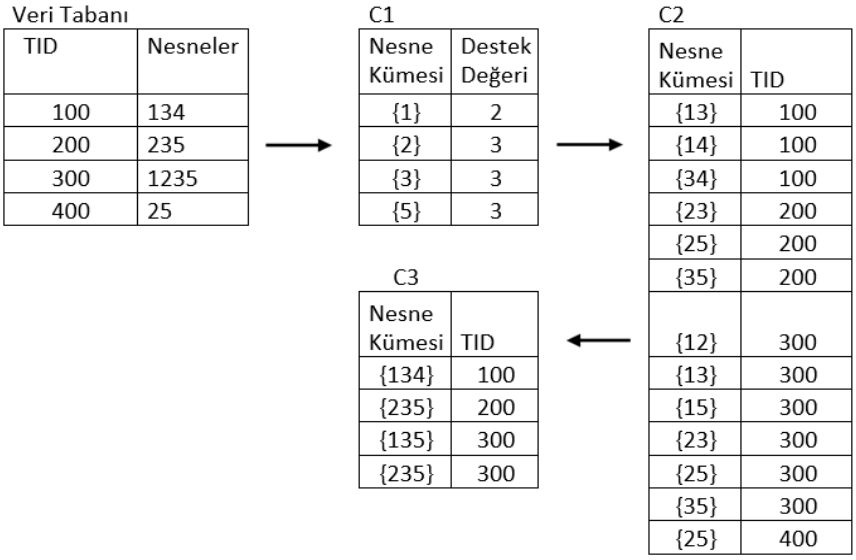
SETM algoritması, SQL kullanılarak geniş nesne kümelerinin hesaplanması için Hatsuma ve Swami (1995) tarafından önerilmiştir.

L_k geniş nesne kümesinin her elemanı, nesne ismi ve nesne işlem numarası olan TID (Transaction Identifier) olmak üzere 2 parametreden oluşur. Benzer şekilde aday nesne kümesi C_k de, TID ve nesne kümesi ismi parametrelerini içerir (Houtsma ve Swami, 1995).

AIS algoritmasında olduğu gibi aday nesne kümeleri veri tabanından okuma işlemi esnasında oluşturulur fakat sayma işlemi geçişin sonunda yapılır. Yeni aday nesne kümeleri AIS Algoritmasındakiyle aynı şekilde oluşturulur ancak oluşturulan işlemin TID'si (Transaction Identifier) aday nesne kümesiyle birlikte sıralı bir yapıda kaydedilir. Geçişin sonunda bu sıralı yapı toplanarak aday nesne kümelerinin destek sayıları belirlenir (Kumbhare ve Chobe, 2014; Agrawal ve Srikant, 1994).

AIS algoritmasında olduğu gibi STEM algoritmasında da veri tabanı birçok defa taranır. Yeni bir geniş nesne kümesi bulunamadığında algoritma sonlandırılır. Aday nesne kümelerinin boyutunun büyüklüğü STEM algoritmasının performansını etkileyen bir faktördür. Aday nesne kümelerinin boyutunun artması dezavantaj oluşturmaktadır. Aday nesne kümelerinin boyutu büyüdükçe, çok sayıda TID'ı depolamak için gerekli bellek ihtiyacı da artmaktadır. Geçişin sonunda aday nesne kümelerinin destek değerlerinin sayılması esnasında, aday nesne kümeleri sıralı olmadığından, tekrar sıralama ihtiyacı ortaya çıkmaktadır. Sayma ve minimum destek eşik değerini sağlamayan aday nesne kümelerinin budanması sonrasında nesne kümelerinin TID değerlerine göre yeniden sıralanması gerekmektedir (Agrawal ve Srikant, 1994). Bu durum iş yükünü artırarak hız kaybına yol açmaktadır.

SETM algoritması çalışma şekli örneği aşağıdaki şekilde gösterilmiştir.



Şekil 2.8: SETM algoritması örneği.

2.3.3 Apriori Algoritması

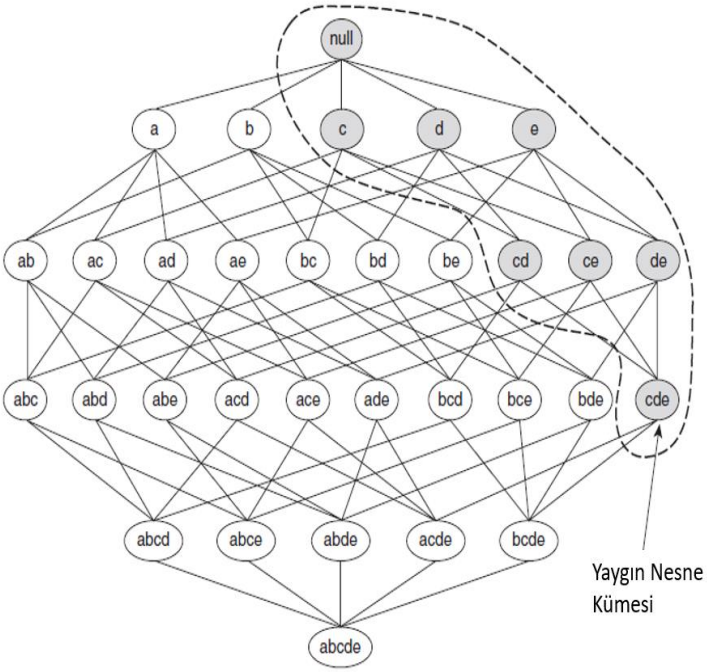
Agrawal ve diğ. (1993) tarafından geliştirilen Apriori algoritması, bilgileri bir önceki adımdan almasından dolayı isimlendirilmesinde önceki (priori) ifadesi kullanılır. Bir nesne küme yaygın ise boş olmayan bütün alt kümeleri de yaygındır. Örneğin {c,d,e} nesne kümesi yaygın ise onun bütün alt kümeleri {c,d}, {c,e}, {d,e}, {c}, {d}, {e} yaygındır (Tan ve diğ., 2019).

Yaygın nesne kümelerini keşfetmek için kullanılan algoritmalar veriler üzerinden birçok defa geçerek, tararlar. İlk geçiş esnasında her bir nesnenin destek değeri hesaplanır ve minimum

destek seviyesine sahip yaygın nesne kümeleri belirlenir. Sonraki geçişlerde ise önceki tarama sırasında yaygın olarak belirlenen nesne kümeleri kullanılarak aday nesne kümeleri olarak isimlendirilen potansiyel nesne kümeleri elde edilir. Veriler taranırken aday nesne kümelerinin gerçek destek değeri hesaplanır. Taramanın sonunda hangi aday nesne kümelerin gerçekten yaygın olup, sonraki taramada kullanılacağına karar verilir. Sonraki taramalarda önceki nesne kümeleri kullanılmaya devam edilir, bu işlem yeni bir yaygın nesne küme bulunamayıncaya kadar devam eder (Agrawal ve Srikant 1994).

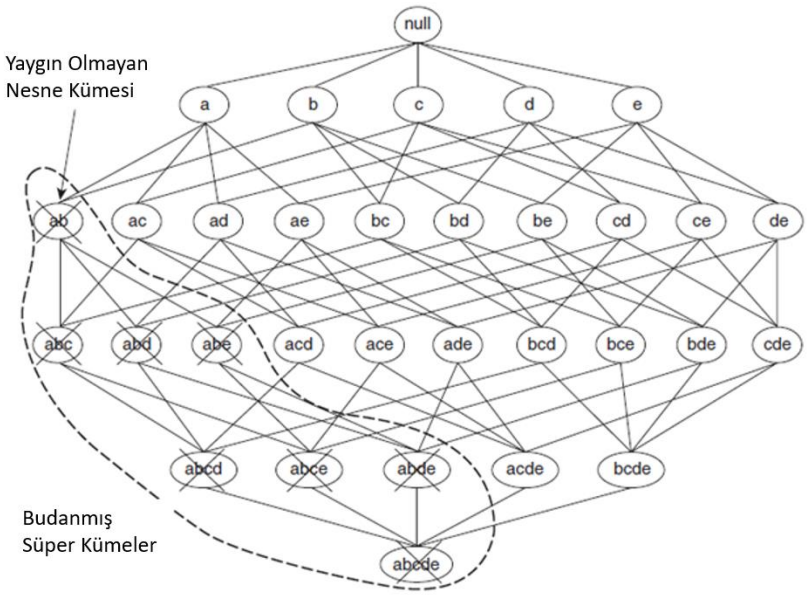
```
// D ,işlem veritabanı
// min_sup, minimum destek eşik değeri
// L, geniş nesne kümeleri
 L1 = Geniş 1-nesne kümeleri (D);
 for (k =2; L_{k-1} ≠ Ø ; k++) {
      Ck = apriori gen(L_{k-1});
      for each transaction t ∈ D {
           Ct = subset (Ck, t) ;          // t'nin alt kümelerindeki adaylar
           for each candidate c ∈ Ct
                c.count++; }
 Lk ={c ∈ Ck|c.count ≥ min sup} }
 return L= Uk Lk;
 procedure apriori_gen(L_{k-1}: geniş (k -1)-nesne kümeleri )
      for each itemset l1 ∈ L_{k-1}{
           for each itemset l2 ∈ L_{k-1}{
                if (l1[1] = l2[1] ^ l1[2] = l2[2])  ^...^ (l1[k -2] =
                l2[k -2])^(l1[k -1] < l2[k -1]/ then {
                   c l1 ⋈ 1 l2; // adayların üretilmesi
                 if has_ infrequent_subset(c, L_{k-1});) then
                        delete c; // budama adımı: gereksiz adayların çıkarılması
                 else add c to Ck;
      } } }
      return Ck;
  procedure has_infrequent_subset(c: candidate k-itemset;
                  L_{k-1}: frequent (k-1)-itemsets); // önceki bilginin kullanılması
       for each (k -1)-subset s of c {
            if s∉ L_{k-1}:  then
                 return TRUE;}
       return FALSE;
```

Şekil 2.9: Apriori algoritması.



Şekil 2.10: Yaygın nesne kümesi (Tan ve diğ., 2019).

Yaygın olmayan bir nesne kümenin bütün süper kümeleri de yaygın değildir. Örneğin {a,b} yaygın değilse, onun bütün süper kümeleri de {a,b,c}, {a,b,d}, {a,b,e}, {a,b,c,d}, {a,b,c,e}, {a,b,d,e}, {a,b,c,d,e}yaygın değildir.



Şekil 2.11: Yaygın olmayan nesne kümesi (Tan ve diğ., 2019).

2.3.4 AprioriTID Algoritması

AprioriTID algoritması, Agrawal ve Srikant (1994) tarafından Apriori algoritması ile birlikte büyük veri tabanındaki nesneler arasındaki birliktelik kurallarının keşfine yönelik olarak AIS ve SETM algoritmalarının yerine önerilmiştir.

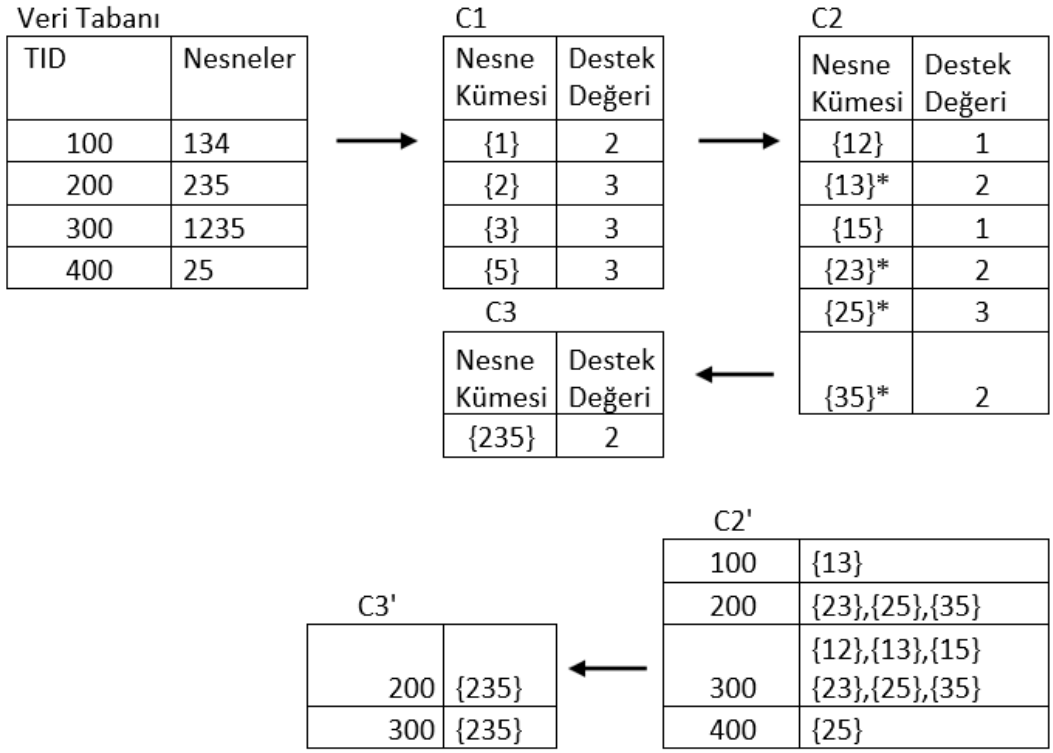
Apriori algoritması her geçişte destek değerlerini hesaplamak için bütün veri tabanını tarar. AprioriTID algoritmasında ise ilk geçişten sonra aday nesne kümelerinin destek değerlerini belirlemek için bütün veri tabanı kullanılmaz. Sonraki adımlardaki destek değerleri, önceki aday kümelerine ait destek değerlerinden elde edilir. Aday nesne kümelerinin belirlenmesi süreci Apriori algoritması ile benzerdir. Aday nesne kümelerinin destek değerlerini hesaplamak için C' olarak isimlendirilen küme kullanılır. AprioriTID algoritmasının sonraki geçişlerde performansı Apriori algoritmasından daha iyidir. Veri kümesinin boyutu küçük iken AprioriTID daha iyi performans gösterirken, veri boyutu büyüdüğünde Apriori algoritmasının performansı AprioriTID algoritmasının performansından daha iyidir (Kumbhare ve Chobe, 2014; Agrawal ve Srikant, 1994).

AprioriTID algoritmasının kaba kaynak kodu aşağıdaki şekildedir (Aggrawal ve Srikant 1994).

```
L1 = {geniş 1-nesne kümeleri};
C1 = D veritabanı;
for (k =2; L_{k-1} ≠ Ø ; k++)  do begin
        Ck = apriori- gen(L_{k-1}); // yeni adaylar
        |Ck| = Ø;
        for all entries t ∈ | Ck-1| do begin
            // Ck  içindeki t.TID numaralı işlemleri içeren aday nesne kümelerini belirle
            Ct = { c ∈ Ck | (c - c[k]) ∈ t.nesnekümleri ^ (c– c [k-1]) ∈ t.nesnekümeleri};
            for all candidates c ∈ Ct  do
                c.count++;
            if (Ct ≠Ø) then | Ck | += < t.TID, Ct  >;
        end
        Lk = {c ∈ Ck | c.count >= minsup}
end
Answer = L= Uk Lk;
```

Şekil 2.12: Apriori TID algoritması.

AprioriTID algoritmasının çalışma şekli örneği aşağıdaki şekilde gösterilmiştir.



Şekil 2.13: Apriori TID algoritması örneği (Kumbhare ve Chobe, 2014).

2.3.5 Apriori Hibrid Algoritması (Frequent Pattern Growth)

İlk geçişlerde Apriori algoritmasının performansı daha iyiyken, sonraki geçişlerde AprioriTID algoritmasının performansı daha üstündür (Aggrawal, 1994). Geçişlerdeki performans durumları gözetilerek Apriori ve AprioriTID algoritmaları kullanılarak Apriori Hibrit algoritması oluşturulmuştur. Apriori Hibrid algoritması, ilk geçişlerde Apriori algoritmasını kullanırken, sonraki geçişlerde AprioriTID algoritmasını kullanmaktadır. Geçişler sonunda Ck değerlerinin hesaplanması gerekir. Hibrid algoritma içinde Apriori'den, AprioriTID'e anahtarlayarak geçişler maliyet artışına yol açmaktadır (KP ve diğ., 2012).

2.3.6 FP Growth Algoritması (Frequent Pattern Growth)

Apriori algoritması, basit yapısına karşılık ölçeklenebilirlik açısından son derece yetersizdir. Algoritmanın diğer bir dezavantajı ise, adayların bulunması için veri tabanının her seviyede yeniden taranmasıdır (Akpınar, 2017).

Apriori algoritması işlem yapılan veri kümesi büyüklüğü az iken başarılı sonuçlar elde ederken, veri büyüklüğü arttıkça algoritmanın performansı düşmektedir.

FP Growth algoritması Han ve Kamber (2000) tarafından geliştirilmiştir. FP Growth algoritmasında öncelikle veri tabanı Apriori algoritmasında olduğu gibi bir kez taranır. Birinci tarama sonucunda nesneküme içerisinde yer alan her bir öğenin frekans değeri belirlenerek, öğeler sıklık sayılarına göre büyükten küçüğe olacak şekilde sıralanırlar.

İkinci tarama sonucunda FP-Tree olarak isimlendirilen sıklık örüntü ağacı oluşturulur.

Tablo 2.6: Öğe sıklıkları ve indirgenmiş işlem veri tabanı (Borgelt, 2005).

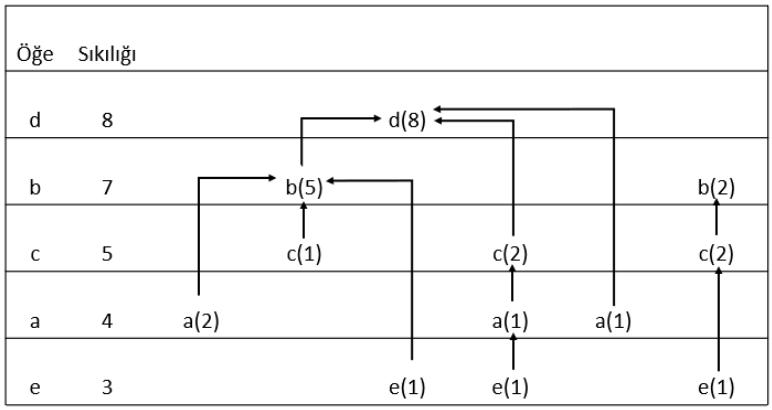
Veri Tabanı
adf
acde
bd
bcd
bc
abd
bde
bceg
cdfg
abdh

Nesne	Sıklık değeri
d	8
b	7
c	5
a	4
e	3
f	2
g	2
h	1

İndirgenmiş Veri Tabanı
da
dca
db
dbc
bc
dba
dbe
bce
dc
dba

Yukarıdaki tabloda solda yer alan işlem veri tabanındaki öğelerin frekans değerleri belirlenerek ortada yer alan öğelerin sıklık değerlerine göre sıralandığı başlık tablosu (header table) oluşturulur. Destek değeri, destek minimum eşik değerine eşit veya büyük olan nesneler listeye eklenir. Destek değeri, eşik değerinden küçük olan öğeler listeden çıkarılır, böylece sık kullanılmayan öğelerin ağaç yapısına dahil olması engellenmiş olur. Analist tarafından belirlenen minimum eşik değerinin 3 olduğu durumda f, g ve h öğeleri veri tabanından çıkarılarak yukarıdaki tabloda sağ tarafta yer alan indirgenmiş işlem veri tabanı elde edilir (Borgelt, 2005).

FP-tree destek değeri yükseldikçe öğelerin köke olan yakınlıkları artacaktır. Çok tekrarlı öğelerin ön ekler olarak birleştirilmesi sonucunda sıkıştırma işlemi gerçekleştirilir. Sıklık veri ağacının dallarındaki her öğenin yanında yer alan sayı o yolun kaç defa kullanıldığını ifade eder. Öncesinde ağaçta yer alan bir öğe ise o düğüme ait destek değeri bir artırılır. FP-tree analizi sıklığı en düşük öğe ile başlar. Yukarıdaki tabloda sıklığı en düşük öğe e olduğu için e öğesini içeren yollar dbe, dcae ve bce olacaktır.



Şekil 2.14: Sıklık veri ağacı (FP_Tree) (Borgelt, 2005).

İkinci tarama sonucunda ise elde edilen sıklık örüntü ağacı yukarıdaki şekilde gösterilmiştir (Borgelt, 2005).

Ağaç oluşturulması sonrasında koşullu örüntü ağacı meydana getirilir. Koşullu örüntü ağacı kullanılarak algoritma özyinelemeli olarak tekrar çalıştırılır. FP-Growth algoritması, böl-yönet metoduna göre yaygın şekilde kullanılan öğeleri kendi içinde alt kümelere ayırır. Her bir veri tabanı parçası üzerinde ayrı madencilik analizi yapılır. Her bir örüntü parçasının ilgili olduğu veri kümesi üzerinde işlem yapılır. Bu nedenle, bu yaklaşım incelenen örüntü boyutu büyümesi ile birlikte aranacak veri kümelerinin boyutunu önemli ölçüde azaltabilir (Han ve diğ., 2012).

2.3.7 Recursive Elemination

Recursive Elemination, yaygın nesne kümelerini bulmaya yönelik olarak FP_Growth algoritmasından oldukça etkilenmiş ve H_mine algoritmasına çok benzeyen bir algoritmadır. Önek ağaçları veya diğer karmaşık veri yapıları olmadan, işlemleri doğrudan işleyerek gerçekleştirir. Güçlü yanı hızı değil basitliği olmasına karşın hız açısından bazı veri setlerinde Apriori ve Eclat'tan daha iyi performans gösterir. Temel olarak tüm işler, nispeten az sayıda kod satırıyla yazılabilen basit bir özyinelemeli (recursive) fonksiyon ile yapılır (Goethals ve diğ., 2005).

Apriori ve FP_Growth algoritmalarından olduğu gibi öncelikle önişlem ile yaygın olan nesne kümeleri tespit edilir. Önişlem sırasında yaygın nesne kümelerinin tespiti için bütün veri tabanı taranır. Nesneler sıklıklarına göre sıralanır. Kullanıcı tarafında belirlenen eşik seviyesi değerine göre nesneler arasında eleme yapılır. Özyinelemeli fonksiyon ile yaygın olmayan nesneler

elenir. Recursive Elemination algoritması yaygın nesneler üzerinde işlem yaparken her düğümün kendisinden sonraki düğümün adres bilgilerini tuttuğu bağlı liste veri yapılarını kullanır (Satyavathi ve diğ., 2019).

Satyavathi ve diğ. (2019) tarafından Recursive Elemination ve FP_Growth algoritmalarının hız performanslarını kıyaslama yönelik T10I4D100K veri seti kullanılarak farklı minimum destek seviyelerindeki işlem zamanları test edilmiştir. Yapılan testin sonucuna göre Recursive elemination algoritmasının, FP_Growth algoritmasına göre daha hızlı olduğu görülmüştür. Elde edilen sonuçlar aşağıdaki Tablo2.7 'de gösterilmiştir. Vinay ve diğ. (2013) tarafından Mushroom veri seti kullanılarak yapılan testlerde de Recursive Elemination algoritmasının, FP_Growth'a göre işlem hızının destek değerine göre eşit veya daha iyi olduğu görülmüştür.

Tablo 2.7: Minimum destek oranı ve işlem zamanı.

Minimum Destek (%)	İşlem Zamanı (Saniye)	
	Recursive Elemination	FP Growth
5	0.465268	0.604799
10	0.200576	0.321492
15	0.131207	0.224248
20	0.103727	0.196747
25	0.0762267	0.173898
30	0.0719919	0.160361
35	0.0537936	0.146814
40	0.0449078	0.142579
45	0.0499542	0.133693
50	0.0410684	0.13411
55	0.0321825	0.134515
60	0.0320987	0.120978

2.3.8 ECLAT (Equivalence Class Transformation)

İşlem veri tabanında sık görülen öğeleri keşfetmek amacıyla Zaki (2000) tarafından önerilen algoritmadır. Yatay formattaki işlem veri kümesini dikey formata dönüştüren bir yöntemdir. İşlemler ile öğeler yer değiştirir. Arama metodu olarak Apriori'den farklı olarak derin öncelikli yapıyı kullanır. Öncelikle ikili öğe setlerine bakılır , sonrasında öğe seti sayısı bir artırılarak devam edilir. Bu durum sık görülen öğe veya aday öğe bulunamayıncaya kadar devam ettirilir (Han ve diğ., 2012).

2.3.9 TERTIUS Algoritması

Flach ve Lachiche (2001) tarafından geliştirilen Tertius algoritması doğrulama değerlerine göre kuralları kullanır. Birinci dereceden mantık gösterimine göre çalışır. Sınıf indeksi, sınıflandırma, onay eşiği, onay değerleri, frekans eşiği, eksik değerler, gürültü eşiği, tekrar kalıpları, sayı kalıpları, ROC analizi ve çıktı değerleri gibi çeşitli seçenekleri veya parametreleri içerir. Kurallar içinde yer alan kalıp sayılarına dayalı yoğun çalışma zamanlarından olumsuz etkilenir. İlk adımda daha fazla zaman harcanırken, sonraki adımlarda daha az zaman ihtiyacı oluşmaktadır. Büyük veri setlerinde saatlerce süren işlem zamanı gereksinimi vardır (Kaur ve Chauhan, 2014; Baher ve Lobo L.M.R., 2012; Sheikh ve diğ., 2021).

2.4 ELEKTRONİK SİSTEMLERDE GÜVENİLİRLİK ANALİZİ

2.4.1 Güvenilirlik Kavramı

İstatistikçiler, matematikçiler ve mühendisler tarafından güvenilirliğe yönelik çok farklı tanımlar yapılabilmektedir. Güvenilirlik, belirlenen şartlar altında bir ekipman parçasının, bir sistemin veya bir ürünün belirli zaman aralığında tatmin edici bir şekilde performansını gösterme olasılığıdır. Mutlak kesinlikle davranışı tahmin etmek mümkün olmadığı için olasılık devreye girmektedir. Girdi, çıktı, sıcaklık ve yük gibi çalışma koşulları ile birlikte tatmin edici performans kriterinin de tanımlanması gerekir (XPPower.com).

2.4.2 Güvenilirlik Kavramı Çalışmasının Yararları

Yeni malzemelerin kullanımı, teknolojinin gelişmesi ile birlikte birçok yeni özelliği barındıran komplike ürünler, entegre sistemler hayatımıza girmeye başlamıştır. Tüketiciler kendilerine sunulan ürünlerin kaliteli, güvenilir ve uzun ömürlü olmalarını beklemektedirler. Bu durum üreticiler arasında rekabetin artmasına yol açmakta, üreticilerin yaşam süresi daha uzun ve güvenilir ürünler üzerinde çalışmalar yapmalarına neden olmaktadır. Olasılık, istatistik ve mühendislik yöntemleri kullanılarak yapılan güvenilirlik analizleri gelişmeye devam etmektedir. Üreticiler gelişen güvenilirlik yöntemlerini ve araçlarını kullanarak arıza oranlarını daha hassas analiz edebilmekte, garanti sürelerine yönelik daha doğru öngörülerde bulunabilmektedir. Üreticiler, güvenilirlik analizleri sonucunda elde ettikleri verilerin sentezine göre tasarımsal değişiklikler yapabilmektedirler. Pazarda artan rekabet ve kişiye özgü ürün talepleri, ürün çeşitliliğini arttırmış, ürünün pazara sunulması için gereken zamanı kısaltmıştır.

Bu durum test sürelerinin kısalmasına yol açmış, üretici tarafında ürünler müşteriye daha ulaşmadan yapılacak ömür hesaplarına yönelik çalışmaları için güvenilirlik araçlarının kullanımını gerekli kılmıştır.

Güvenilirlik analizi metodolojisi (IEC 62308), kavramsal fikir aşaması, tasarım geliştirme, üretim, kurulum, operasyon ve bakım, geri dönüşüm ve imha safhalarını kapsayacak şekilde sürekli değerlendirmeyi içermektedir.

Güvenilirlik testleri sayesinde tasarımsal kısıtlar veya eksiklikler varsa tespit edilmekte, kriterlere uyacak şekilde önlemler alınabilmektedir. Alınan önlemler mekanik tasarım değişikliği, elektronik tasarım değişikliği veya her ikisini de içerecek şekilde olabilmektedirler. Diğer taraftan test sonuçlarına göre kriterler sağlanabiliyorsa farklı alternatif malzemelerin kullanımı da değerlendirilebilmektedir.

Güvenilirlik çalışmaları toplam ürün veya sistem maliyetini etkiler. Ürün ve sistem üzerindeki güvenilirlik arttıkça tasarım, geliştirme, kontrol ve test maliyetleri artacaktır. Kullanılan komponent sayısı değişeceğinden veya daha yüksek kalitede malzeme kullanılacağından dolayı malzeme maliyeti ve üretim maliyeti artacaktır. Diğer taraftan ise bakım ve tamir maliyetleri düşecektir.

Güvenilirlik testleri sonucunda elde edilen hata oranı, geri dönüş oranları tahminleri sayesinde üreticiler tarafından garanti maliyetleri daha doğru hesap edilebilmektedir. Garanti maliyetlerinin doğru hesaplanması ile oluşabilecek finansal yükler önlenebilir. Daha uzun garanti süreleri müşterilere öneri olarak sunulabilir.

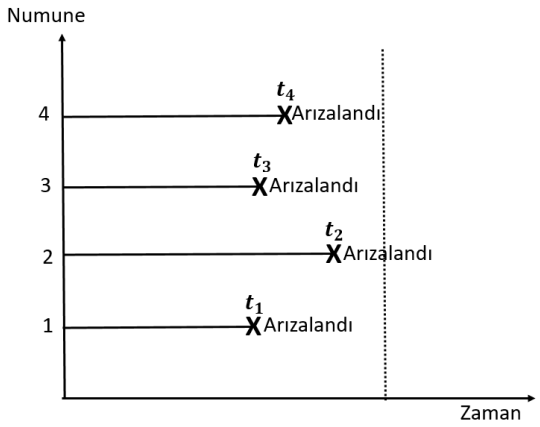
2.4.3 Güvenilirlikte Kullanılan Veri Türleri

Güvenilirlik analizleri sırasında kullanılan istatiksel modeller, tahminlerde bulunmak için verileri kullanırlar. Ürünün yaşamına yönelik ve başarısızlık zamanına yönelik tahminlerin başarısı elde edilen verileri kalitesi, doğruluğu ve eksiksizliği ile doğru orantılıdır (Reliasoft Corporation, 2014).

Ömür testlerine alınan numunelerden, arızalananların bozuluncaya kadar süreleri ile bozulmayanların test sonuna kadar geçen süreleri analizler için kayıt altına alınılırlar. Ömür

analizlerinde kullanılan veriler, numunelerin test süresi boyunca arızalanıp arızalanmamalarına göre tam veri ve sansürlü veri olarak iki gruba ayrılırlar.

Verilerin tam veri olarak adlandırılabilmesi için, test sırasında bütün numunelerin bozuluncaya kadar olan sürelerinin kayıt altına alınmış olmaları gerekir. Yapılan test sırasında bozulmayan numune olursa, kaydedilen veri seti tam veri olarak nitelendirilemez. Tam veri setinde bütün numunelerin arızalandığı an gözlemlenir ve analiz çalışmalarında kullanılmak için kaydedilirler.



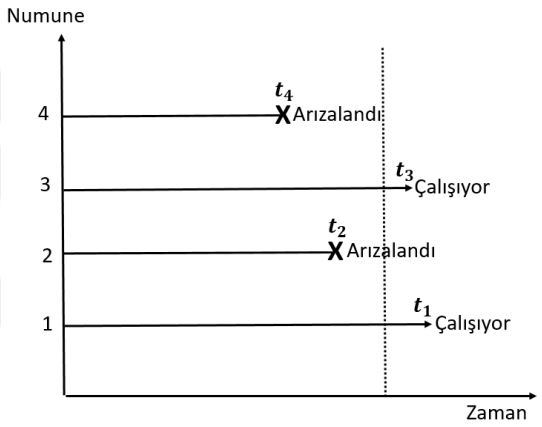
Şekil 2.15: Tam veri.

Analiz çalışmalarında kullanılmak üzere bütün numunelerin arızalandığı an test ekipmanlarının yetersizliği, test ölçüm zaman aralığı veya maliyet nedenleriyle kayıt altına alınamayabilir. Çoğu zaman test numunelerinin hepsi arıza yapmayabilir veya hata yapmalarına kadar geçen süre gözlemlenemez. Bu durumda her numunenin veya bireyin çalışmanın başlangıcından bozuluncaya kadarki gözlem verisi elde edilemez. Bu tip veriler sansürlü veri olarak isimlendirilirler. Sansürlü veriler, hata yapmadan geçen test sürelerini içerirler. Sansürlemenin sağdan sansürleme, aralık sansürlemesi ve soldan sansürleme olmak üzere üç tipi mevcuttur.

Çalışmanın bitirilmesi için belirlenen bir gözlem anında, başlangıçtan o ana kadar bozulma veya başarısızlık olayı gerçekleşmez ise numune veya birey yaşam faaliyetlerini sürdürmeye devam edecektir, bozulmanın oluştuğu an bilinemeyecektir. Numune veya bireyin yaşam süresinin çalışma zamanından fazla olmasından dolayı, bozulma süresinin tespit edilemediği bu tip sansürlemeye sağdan sansürleme adı verilir. Mühendislik, sağlık, ekonomi gibi farklı alanlarda karşılaşılan sağdan sansürlü veriler gözlem veri kümesinden çıkartılabildiği gibi

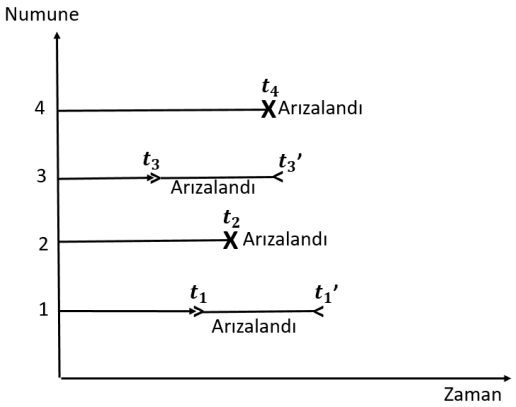
analiz çalışmalarına dahil edilebilmek için Kaplan-Meier ağırlıkları ve KNN yerine koyma yöntemlerinden de faydalanılabilmektedir.

Sağdan sansürlemede ilgilenilen olay, örneğin bozulmaya kadar geçen süre veri aldığımız gözlem anının sağında kalacaktır. Başka bir deyişle veri aldığımız noktanın sağ tarafında ilgili numuneler bir süre daha çalışmaya devam edecektir. Sağdan sansürlü veri yapısı aşağıda gösterilmiştir. Aşağıdaki grafik örneğinde, test edilen 4 numuneden 2 tanesinde bozulma (arızalanma) gözlemlenirken, 2 tanesinde bozulma gözlenmemiştir. Arızalanmayan 2 numune veri noktasının sağ tarafında bozulmadan bir süre daha çalışır. Arızalanmayan 2 numuneden dolayı sağdan sansürlenmiş veri örneğine sahip olunur.



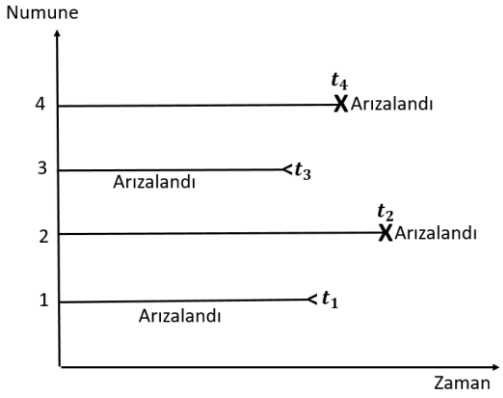
Şekil 2.16: Sağdan sansürlü veri.

Aralıklı sansürlü veriler, bozulmanın tam olarak ne zaman olduğunun bilinmediği, arıza anına yönelik belirsizliğin olduğu durumu yansıtır. Yapılan testlerde gözlemlemenin sürekli yapılamadığı, kesikli ölçümler alınmak zorunda olunduğu durumlarda veya ölçüm ekipmanın örnekleme sıklık yeteneğinden dolayı aralıklı sansürlü veriler elde edilmek zorunda kalınabilir. Bozulma, önceki çalışır durumda olduğu test anı ile son test anı arasındaki zaman aralığında meydana gelebilir. Aşağıdaki aralıklı sansürlü veri grafiği örneğinde birinci ve üçüncü numunelerin tam olarak ne zaman bozulduğu tespit edilememekle birlikte, bozulmanın meydana geldiği aralık belirtilebilmektedir.



Şekil 2.17: Aralık sansürlü veri.

Soldan sensörlü veriler bozulmanın belirli bir zaman aralığından önce olduğunun bilindiği fakat bozulmanın tam olarak ne zaman meydana geldiğinin bilinmediği gözlem sonuçlarını içerir. Soldan sensörlü veriler aralıklı sansürlü verilere benzerler, aralarındaki farklılık ise soldan sensörlü verilerde aralığın başlangıç zamanı sıfırdan başlar. Aşağıdaki grafikte birinci ve üçüncü numunelerin bozulmaları soldan sensörlü veri tipine örnektir.



Şekil 2.18: Soldan sansürlü veri.

2.4.4 Güvenilirlik Sistemlerinde Kullanılan Ömür Dağılımları

Ürün ve sistemlerin güvenilirlik analizlerinde yaygın olarak kullanılan hata zamanı dağılım modelleri arasında üstel, normal, lognormal, Weibull, Burr-XII, gama ve en küçük değer dağılımları sayılabilir. Bu dağılımlara ilave parametreler ekleyerek farklı yeni modeller türetilmesine de sıklıkla rastlanılmaktadır. Marshall ve Olkin (1997) Üstel ve Weibull dağılımlarına yeni parametre eklenmesi, Mudholkar ve Srivastava (1993) 2 parametreli Weibull

dağılımına yeni parametre eklenmesi, Mahdavi ve Kundu (2017) 1 parametreli üstel dağılıma ikinci bir parametre eklenmesi yönünde çalışmalar yapmışlardır.

2.4.4.1 Üstel Dağılım

Üstel dağılım güvenirlik analizlerinde yaygın olarak kullanılan hata dağılımıdır. Üstel dağılım zamanla değişmeyen, sabit arıza oranına sahip bireylerin davranışlarını modellemek amacıyla kullanılır. Weibull dağılımında β'nın 1 değerini alması durumunda üstel dağılım elde edilir. Üstel dağılım, süre uzasa da arıza oranı değişmediğinden hayatta kalma olasılığı aynı olacaktır, bu nedenle üstel dağılım, hafızasız dağılım olarak da nitelendirilir. Üstel dağılımın hafızasının olmaması aşağıdaki olasılık formülü ile ifade edilebilir.

$$P(X > a + b \mid X > a) = P(X > b) \tag{2.36}$$

Üstel dağılıma sahip x rastgele değişkeni için tek parametreli olasılık yoğunluk fonksiyonu *f(x),* ölçek parametresi (1/λ) sıfırdan büyük olacak şekilde aşağıdaki gibi ifade edilir.

$$f(x) = \lambda e^{-\lambda x} \ , x > 0 \ , \lambda > 0 \tag{2.37}$$

Üstel dağılama sahip x rastgele değişkeninin dağılım fonksiyonu F(x) aşağıdaki şekilde belirtilir.

$$F(x) = 1 - e^{-\lambda x} \ , x > 0 \ , \lambda > 0 \tag{2.38}$$

Ortalama değer veya arızalanmaya kadar ortalama zaman (MTTF) E(x) ve varyans var (x) ise sırasıyla aşağıdaki gibi ifade edilir.

$$E(x) = 1/\lambda \tag{2.39}$$

$$var(x) = (1/\lambda)^2 \tag{2.40}$$

Orijinal üstel dağılımın ortalamasının 1/λ eşittir, hafıza özelliği bulunmaması nedeni ile zamanda alınan bir kesitte artık kısmının da ortalama değeri değişmez 1/λ olur.

2 parametreli Üstel dağılım fonksiyonu ölçek parametresi ile birlikte konum parametresini γ de içerir, şekil parametresi içermez. Üstel fonksiyonun tek bir şekli vardır. Üstel yoğunluk dağılım fonksiyonu aşağıdaki şekilde belirtilir.

$$f(x) = \lambda e^{-\lambda(x-\gamma)} \ , f(x) \ \geq, x \geq 0 \ , \lambda > 0 \tag{2.41}$$

Konum parametresi γ sıfırdan büyük ise dağılım orijinin sağına doğru kayar. Üstel kümülatif dağılım fonksiyonu F(x) ve güvenilirlik fonksiyonu R(x) ise sırasıyla aşağıdaki şekilde gösterilir.

$$F(x) = 1 - e^{-\lambda(x-\gamma)} \tag{2.42}$$

$$R(x) = 1 - F(x) = e^{-\lambda(x-\gamma)} \tag{2.43}$$

Üstel arıza oranı fonksiyonu λ(x) aşağıdaki şekilde ifade edilir. Değeri sabittir.

$$\lambda(x) = f(x)/F(x) = \ \lambda \tag{2.44}$$

Üstel dağılımdan yararlanılarak yeni üstel dağılım modelleri de türetilmiştir. Geliştirilen yeni üstel dağılım modelleri arasında üstel üstel dağılım (Gupta ve Kundu, 2001), genelleştirilmiş üstel dağılım (Gupta ve Kundu, 2007), ters üstel dağılım (Keller ve diğ., 1982) sayılabilir.

2.4.4.2 Weibull Dağılımı

Wallodi Weibull tarafından 1951 yılında bulunan ve genellikle ömür süresi hesaplamalarında kullanılan dağılım türüdür. İstatistiksel modellerde ve veri analizlerinde 2 ve 3 parametreli türleri sıklıkla kullanılmakla birlikte 2 parametreli türünün kullanımı daha yaygındır.

2 parametreli Weibull dağılımı, β biçim parametresi ve η ölçek parametresini içerir. 2 parametreli Weibull dağılımının olasılık yoğunluk fonksiyonu pdf (probability density function) aşağıdaki şekilde ifade edilir.

$$f(x) = \frac{\beta}{\eta}\left(\frac{x}{\eta}\right)^{\beta-1} exp\left[-\left(\frac{x}{\eta}\right)^{\beta}\right] \quad , f(x) \ \geq \ 0, x \ \geq \ 0 \tag{2.45}$$

Olasılık yoğunluk fonksiyonunda biçim ve ölçek parametreleri (β > 0, η > 0) sıfırdan büyük değerler alırlar. İki parametreli Weibull kümülatif dağılım fonksiyonu *F(x)* aşağıdaki şekilde ifade edilir.

$$F(x) = 1 - exp\left[-\left(\frac{x}{\eta}\right)^{\beta}\right] \tag{2.46}$$

İki parametreli Weibull dağılımının güvenilirlik fonksiyonu *R(x)* aşağıdaki şekilde belirtilir.

$$R(x) = 1 - F(x) = exp\left[-\left(\frac{x}{\eta}\right)^{\beta}\right] \tag{2.47}$$

Hata oranı *h(x)* aşağıdaki şekilde hesaplanır.

$$h(x) = \frac{f(x)}{R(x)} = \frac{\beta}{\eta}\left(\frac{x}{\eta}\right)^{\beta-1} \tag{2.48}$$

3 parametreli Weibull dağılımı, β biçim parametresi ve η ölçek parametresinin yanında üçüncü parametre olarak γ konum parametresini de içerir. Üç parametreli Weibull olasılık dağılım fonksiyonu *f(x)* aşağıdaki şekilde ifade edilir.

$$f(x) = \frac{\beta}{\eta}\left(\frac{x-\gamma}{\eta}\right)^{\beta-1} exp\left[-\left(\frac{x-\gamma}{\eta}\right)^{\beta}\right], \beta > 0, \eta > 0, x \geq 0 \tag{2.49}$$

Üç parametreli Weibull kümülatif dağılım fonksiyonu *F(x)* aşağıdaki şekilde ifade edilir.

$$F(x) = 1 - exp\left[-\left(\frac{x-\gamma}{\eta}\right)^{\beta}\right] \tag{2.50}$$

Üç parametreli Weibull olasılık dağılımında konum parametresi sıfır değerini aldığında (γ=0), 2 parametreli Weibull dağılımı elde edilir.

Weibull dağılımı biçim parametresi β'nın aldığı değerlere göre farklı dağılımlara dönüşür. Biçim parametresi sıfır ile bir arasında değer aldığında (0 <β < 1) bozulma oranı zaman içinde artar. β=1 değerini aldığında Weibull dağılımı üstel dağılıma eşit olur.

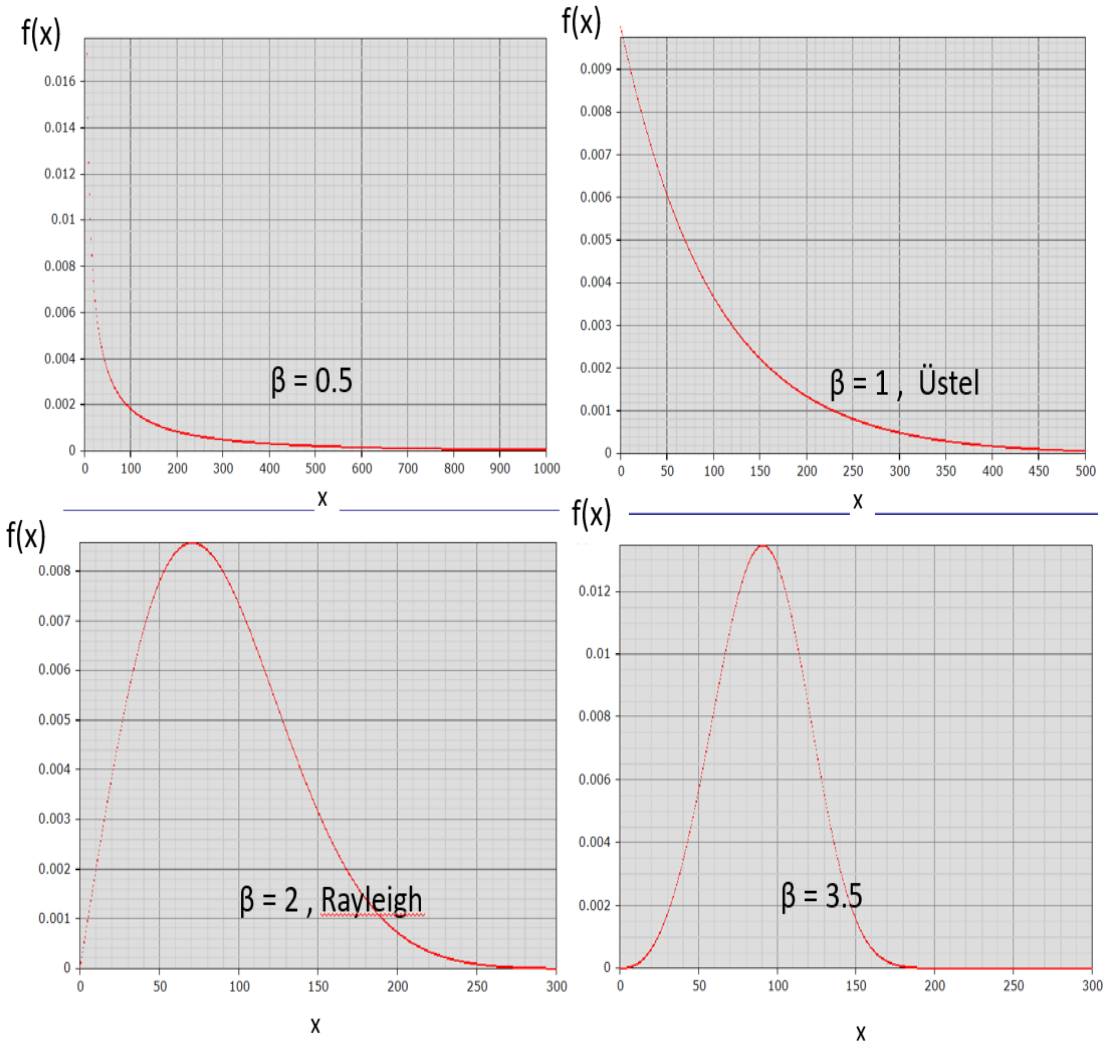
$$f(x) = \frac{\beta}{\eta}\left(\frac{x}{\eta}\right)^{\beta-1} exp\left[-\left(\frac{x}{\eta}\right)^{\beta}\right] = \frac{1}{\eta} exp - \left(\frac{x}{\eta}\right) \quad , \beta = 1 \tag{2.51}$$

β= 2 değerini aldığında Weibull dağılımı Rayleigh dağılımına eşit olur.

$$f(x) = \frac{\beta}{\eta}\left(\frac{x}{\eta}\right)^{\beta-1} exp\left[-\left(\frac{x}{\eta}\right)^{\beta}\right] = \frac{2}{\eta}\left(\frac{x}{\eta}\right) exp\left[-\left(\frac{x}{\eta}\right)^{2}\right], \beta = 2 \tag{2.52}$$

β= 2,5 değerini aldığında Weibull dağılımı lognormal dağılımına, β= 3,5 değerini aldığında ise Weibull dağılımı normal dağılıma yaklaşır.

Biçim parametresinin, β, aldığı farklı değerlere göre (β =0.5, β=1, β=2, β=3.5) Weibull olasılık yoğunluk fonksiyonunun değişimi, pdf, aşağıdaki şekilde gösterilmiştir.



Şekil 2.19: Biçim parametresindeki değişimin Weibull pdf üzerindeki etkisi.

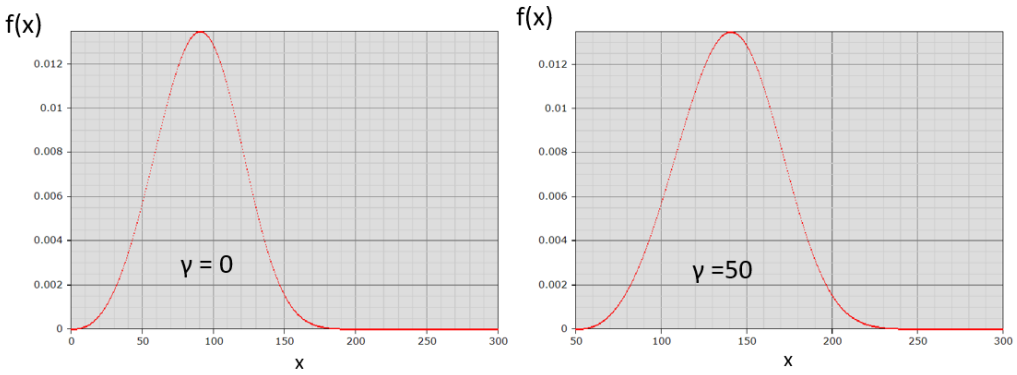
Biçim parametresi, β, 0 <β < 1 aralığında iken f(x) monoton bir biçimde azalır ve a değerinin ötesindeki artışlarda dış bükey hale dönüşür.

$x \to 0$, $f(x) \to \infty$,

$x \to \infty$, $f(x) \to 0$,

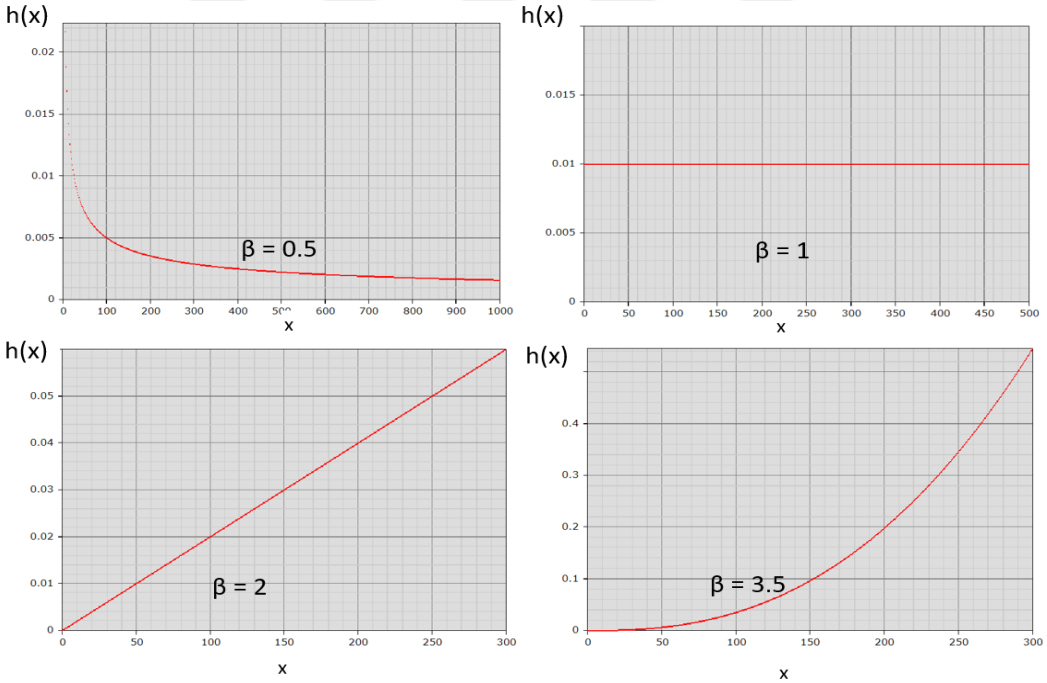
Biçim parametresi, β, β> 1 olduğunda, f(x) = 0, t=0.

Konum parametresinin, γ, aldığı farklı değerlere göre (γ=0, γ=50), Weibull olasılık yoğunluk fonksiyonunun değişimi, pdf, β=3.5 iken aşağıdaki şekilde gösterilmiştir.



Şekil 2.20: Konum parametresindeki değişimin Weibull pdf üzerindeki etkisi.

Konum parametresinin değerine göre pdf eğrisi, zaman ekseninde kayar. Konum parametresinin sıfır değerini, γ=0, aldığı durumda eğri x=0 noktasından yani orijinden başlar. Konum parametresinin sıfırdan büyük olduğu değerlerde, γ>0, dağılım orijinin sağından, sıfırdan küçük olduğu değerlerde ise, γ<0, dağılım orijinin solundan başlar.



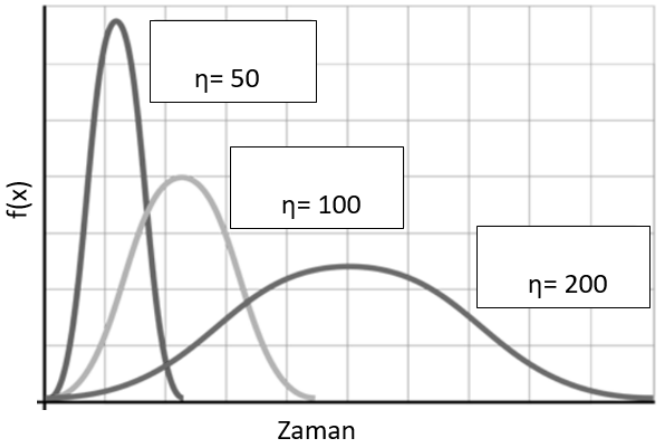
Şekil 2.21: Biçim parametresindeki değişimin arıza oranları üzerindeki etkisi.

Biçim parametresinin, β, aldığı farklı değerlerin (β =0.5, β=1, β=2, β=3.5) arıza oranları h(x), üzerindeki etkisi yukarıdaki şekilde gösterilmiştir.

Biçim parametresi, β, β <1 aralığında iken arıza oranlarınki düşme zamanla azalır. β =1 iken, arıza oranı sabittir. β >1 iken ise arıza oranları zamanla hızlı şekilde artar. β değerinin değişimi ve Weibull dağılımı ile küvet eğrisinin 3 evresi de modellenebilir.

Biçim parametresi, β, 0 <β <1 aralığında iken R(x) monoton bir biçimde, hızlı bir şekilde düşer ve şekli konvekstir. β =1 iken, monoton şekilde düşmeye devam eder, fakat düşüş oranı 0 <β < 1 aralığındaki düşmeye göre daha azdır ve şekli konvekstir. β >1 iken ise bükülme noktasına kadar R(x)'nin düşüş miktarı artar. Bükülme noktasında sonra ise R(x) keskin bir şekilde düşer.

Biçim parametresi, β sabit tutulup, ölçek parametresi η'nün değeri artırılırsa dağılımın yüksekliği azalırken basıklığı artar. Ölçek parametresinin değeri azalırken ise yüksekliği artar, daha dar hale gelir.



Şekil 2.22: Ölçek parametresindeki değişimin Weibull pdf üzerindeki etkisi.

Yukarıdaki şekilde de görüldüğü gibi biçim parametresi, β, sabit olacak şekilde ölçek parametresi büyütüldüğünde pdf zaman ekseninde genişler. Pdf eğrisinin altındaki alan sabit olduğundan, ölçek parametresinde artışa bağlı olarak pdf eğrisinin tepe değeri azalır.

2.4.4.3 Lognormal Dağılımı

Çarpık bir dağılım olan Lognormal dağılımı, dağılımın aralığın sonundaki kuyrukta (sol) yoğunlaştığı halleri modellemek için kullanılabilir. Farklı müşterilerce tüketilen elektrik miktarları, sistemlerin arızalı olarak geçirdiği süreler, tamir süresi, ampullerin ışık yoğunlukları, kimyasal işlem kalıntılarının konsantrasyonu, farklı müşterilerin otomobilleri ile

yaptıkları mesafe değerleri lognormal dağılıma örnek olarak sayılabilirler (Ireson ve diğ., 1996).

Hızlandırılmış yaşam testlerinde (ALT) ürünler daha yüksek gerilim, sıcaklık, basınç gibi stres faktörlerine maruz bırakılarak ürünlerin daha hızlı bozulmaları hedeflenmektedir. Bu sayede yapılan testlerin süreleri ve maliyetleri azaltılmış olur. Artırılmış stres seviyelerinde elde edilen veri analiz testlerinin sonuçları sonrasında İnterpolasyon yapılır. Yapılan çok sayıda araştırmadaki ALT faaliyetleri lognormal ve Weibull modellerine dayandırılarak gerçekleştirilmiştir (Pascual and Montepiedra, 2005). Şekil ve konum parametrelerini içeren 2 parametreli Weibull ve Lognormal modelleri literatürde yaygın olarak ömür hesabı modellerinde kullanılmıştır (Upadhyay ve Sharma, 2018).

Lognormal dağılımına sahip sürekli tipteki t rastgele değişkeninin olasılık yoğunluk fonksiyonu (pdf) aşağıdaki şekildedir. μ, dağılımın konum parametresi, σ ise şekil parametresini ifade etmektedir.

$$f(t;\mu,\sigma) = \frac{1}{t\sigma\sqrt{(2\pi)}} exp\left[-\frac{1}{2}\left(\frac{lnt-\mu}{\sigma}\right)^2\right] \quad t > 0,\ \infty < \mu < -\infty,\ \sigma > 0 \tag{2.53}$$

Kümülatif dağılım fonksiyonu ise aşağıdaki şekilde ifade edilir.

$$F(t;\mu,\sigma) = \frac{1}{\sigma\sqrt{(2\pi)}} \int_0^t \frac{1}{y} exp\left[-\frac{1}{2}\left(\frac{lny-\mu}{\sigma}\right)^2\right] dy \tag{2.54}$$

Lognormal dağılımın güvenilirlik fonksiyonu aşağıdaki şekilde belirtilir.

$$R(t) = 1 - \Phi(\frac{lnt-\mu}{\sigma}) \tag{2.55}$$

2.4.4.4 Burr-XII Dağılımı

1942 yılında Burr tarafından önerilen 2 parametreli Burr-XII dağılımı güvenilirlik analizlerinde, başarısızlık zamanı modellenmesinde ve yazılım güvenirliliği analizlerinde kullanılmıştır (Tekin ve Özel, 2020). Esnek yapısı, arıza süresini modellemedeki kullanım ve hesaplama kolaylığı Burr-XII dağılımının avantajları arasındadır (Zimmer ve diğ., 1998).

X rastgele değişkeninin Burr-XII dağılımına sahip olması durumunda olasılık yoğunluk (*f* (x; λ, β)) ve dağılım (*F*(x; λ,β)) fonksiyonları aşağıdaki şekilde ifade edilir (Kuş ve Akdoğan, 2010).

$$f_{BXII}(x;\lambda,\beta) = \lambda\beta x^{\beta-1}\,(1+x^{\beta})^{-(\lambda+1)}\,,\ x>0,\ \lambda>0,\beta>0 \tag{2.56}$$

$$F_{BXII}(x;\lambda,\beta) = 1-\ (1+x^{\beta})^{-\lambda} \tag{2.57}$$

λ=1 olması durumunda Burr-XII dağılım fonksiyonu log-lojistik yapısına dönüşür. Burr-XII dağılımı için yaşam ve bozulma fonksiyonları sırasıyla aşağıdaki şekilde hesaplanır (Tekin ve Özel, 2020).

$$S_{BXII}(x;\ \lambda,\beta)\ =\ 1\ -\ F_{BXII}(x,\lambda,\beta)\ =\ (1+x^{\beta})^{-\lambda} \tag{2.58}$$

$$h_{BXII}(x;\ \lambda,\beta)\ =\ \frac{\lambda\beta x^{\beta-1}(1+x^{\beta})^{-(\lambda+1)}}{(1+x^{\beta})^{-\lambda}} \tag{2.59}$$

2.4.5 Hızlandırılmış Yaşam Testi Kavramı

Hızlandırılmış yaşam testlerinde, ürünün normal çalışma koşullarda karşı karşıya geldiği seviyesinin üzerinde stres uygulanarak, bozulmalar tespit edilir. Hızlandırılmış yaşam testlerinin hedefi daha az numune, daha az maliyet bütçesiyle stres faktörlerini kullanarak ürünün bozulma zamanını hızlandırmaktır. Testler sonucunda tespit edilen bozulmalara karşı ürün tasarımında değişiklik yapılarak, ürün ömrünü arttırmaya yönelik iyileştirmeler yapılır. Güvenilirlik testlerini stres seviyesini arttırmadan normal çalışma koşullarında yapmak zaman ve maliyet açısından pratikte uygun olmayacaktır.

Hızlandırılmış yaşam testlerinde (Accelarated Life Test) kullanılan farklı stres faktörleri mevcuttur. Stres faktörleri ürün, malzeme veya sistem üzerine tek başına uygulanabileceği gibi birden fazla stres faktörü belirli koşullar altına aynı anda da uygulanabilir. Yaygın olarak ALT testlerinde kullanılan stres faktörleri arasında titreşim, şok, bükülme, sıcaklık, nem, elektriksel alan, gerilim, sayılabilir. Stres faktörü seviyesi test süresi boyunca sabit kalabilir veya şiddeti artırılabilir. Stres faktörü seviyesinin değişmeden, aynı seviyede ürüne, malzemeye veya sisteme uygulandığı test türü sabit stres faktörlü test olarak ifade edilirken, zaman içinde stres faktörü seviyesinin belirli zaman aralıklarında arttırıldığı test türü ise adım stres faktörlü test olarak adlandırılır.

Ürün ve malzemelerin ömürlerine yönelik hızlı şekilde bilgi edinmek üzerine kurgulanan hızlandırılmış yaşam testleri ile ilgili geçmişte farklı çalışmalar yapılmıştır. Mann (1972), doğrusal kestirim yöntemini kullanarak uç değer dağılımının yüzdelik dilimini (percentile) tahmin etmeye yönelik çalışma yapmış, tip II sensör verileri için yaklaşık uygun değer planı

oluşturmuştur. Meeker ve Nelson (1974) sansürlü verilerin uygun test planını ele almışlar, uç değer dağılımı ve Weibull dağılımı için En Çok Olabilirlik Tahminleyicilerini kullanmışlardır. Nelson ve Kielpinski (1976) hızlandırılmış yaşam testine yönelik plan üzerinde çalışmışlardır. Normal veya lognormal dağılıma sahip stres ve ürün ömrü arasındaki basit doğrusal ilişkinin tahmini üzerinde durmuşlardır. Planlarını, sıcaklığı stres faktörü olarak kullandıkları hızlandırılmış yaşam testinde uygulamışlardır. Çalışmalarında Arrhenius modeli kullanmışlardır. Xu ve Tang (2015) hızlandırılmış yaşam testi planı için Bayes tekniklerini kullanmışlardır.

Güvenilirliğe yönelik yapılan hızlandırılmış yaşam testleri, nitel hızlandırılmış yaşam testleri ve nicel hızlandırılmış yaşam testleri olarak iki farklı grupta ele alınmışlardır.

Nicel hızlandırılmış yaşam testleri ile belirlenen seviyelerde arızalanmaya kadar geçen süre veya bozulma dağılımı bilgisi elde edilmeye çalışılır. Elde edilen bilgiler kullanılarak ekstrapolasyon yöntemi ile ürünün normal çalışma koşullarındaki ömür davranışına yönelik kestirimde bulunulur (Nelson, 1982). ALT testleri nicel hızlandırılmış yaşam testleri sınıfındandır.

Ürün geliştirme çalışmaları sırasında tasarım kaynaklı noksanlıkları tespit etmeye yönelik nitel hızlandırılmış yaşam testleri gerçekleştirilir. Bu testler sırasında normal çalışma koşullarının ötesinde ağırlaştırılmış stres şartları uygulanarak tasarımsal eksiklikler ortaya çıkarılır. Belirlenen eksikliklere yönelik, zayıflığın kök neden analizi yapılarak süreç iyileştirme ve tasarım değişikliği çalışmaları gerçekleştirilir. HALT testleri nitel hızlandırılmış yaşam testleri sınıfındadır.

Ürün güvenilirliğine ve kalitesine yönelik uygulanan farklı hızlandırılmış testler mevcuttur. Bu testler arasında ALT (Accelerated Life Testing), HALT (Highly Accelerated Life Testing), HASS (Highly Accelerated Stress Screnning), HASA (Highly Accelerated Stress Auditing) yer almaktadır. ALT testinin temel amacı normal çalışma koşullarının ötesinde ağırlaştırılmış şartlar altındaki test sonuçlarından elde edilen verilerden, ürünün veya sistemin ömür tahmininin yapılabilmesidir. HALT testinin temel amacı tasarım aşamasında iken ürün tasarımındaki temel zayıfların ve problemlerin saptanarak düzeltmelerinin ve iyileştirilmelerinin sağlanmasıdır. HALT testleri sırasında hata daha tasarım safhasının başlangıç evrelerinde tespit edildiğinden, sonraki aşamalarda tespit edilmesine göre maliyet ve

müşteri memnuniyeti açılarından avantaj sağlar. Bu amaçla HALT testleri sırasında tasarım eksiklikleri tespit etmek için belirlenen stres faktörü değerleri, normal çalışma koşullarının ötesinde teknoloji limitlerine kadar artırılılarak ürün ve sistem bozulmaya zorlanır. HALT testleri ağırlaştırılmış stres koşulları altında yapıldığından hata oluşumu, normal koşullar altına yapılan testlere göre çok daha hızlıdır. HASS ve HASA testlerin üretim süreci problemlerinin tespit edilmesi, düzeltilmesi ve ürünler müşteriye sevk edilmeden ayıklanarak seçilmesi amacıyla kullanılırlar. HALT tasarım aşamasında uygulanırken, HASS üretim aşamasında kullanılır.

2.4.6 Güvenilirlik Analizlerinde Kullanılan Paket Programlar

Güvenilirlik analizlerinde kullanılan ticari paket programlar arasında Reliasoft, SuperSMITH, ALD, FReET, Struel, Item Software, Isograph, PTC Windchill sayılabilir (Symynck ve Bal, 2011; Menčík, 2016).

Reliasoft programının Weibull++ paketi ve SuperSMITH Weibull paketi, arıza davranışlarının tahmin edilmesi ve analizinde güveninirlik mühendisleri tarafından kullanılan yazılım araçlardandır. SuperSmith Weibull paketi ile güvenilirlik analizi yapmak için Weibull, Lognormal, Gumbel ve normal dağılımlarının olasılık grafikleri elde edilebilir (Symynck ve Bal, 2011).

Relisoft programının ALTA modulü hızlandırılmış yaşam testleri analizde, DOE modulü deney tasarımında, BlockSim güvenilirlik blok diyagramlarının yaratılmasında, RENO modulü risk karar analizinde, RCM modulü güvenilirlik merkezli bakım çalışmalarında kullanılır.

ALD yazılım paketi, Thales, BAE sistemleri, Locheed Martin, Nasa, MBDA, Sagem gibi sivil ve askeri havacılık, iletişim, uzay ve elektronik firmalarında güvenlik ve güvenilirlik analizlerinde kullanılmaktadır. ALD yazılım paketi güvenilirlik tahmini, kullanılabilirlik, bakım analizi, güvenlik yönetimini içeren araçlardan oluşmaktadır.

FReET, mühendislik uygulamalarında istatislik, hassasiyet ve güvenilirlik analizlerinde kullanılan yazılım paketidir.

Struel, özellikle inşaat mühendisliğinde yapıların güvenilirliği analizinde kullanılan paket programları içerir. Statrel, Comrel ve Sysrel isimli programlarından oluşur. Statrel, güvenilirlik

odakli istatiksel analiz için kullanılır. Comrel, Zamanla değişmeyen ve zamana bağımlı analizler için kullanılır. Sysrel, sistem güvenilirlik analizlerinde kullanılır.

Item Software, farklı analiz ve güvenilirlik tahminlerine yönelik modülleri içeren paket programdır. MIL-HDBK-217 modulü, IEC 62380 Electronik Güvenilirlik Tahmin modulü, NSWC Mekanik Güvenilirlik Tahmin Modulü, China 299B Elektronik Güvenilirlik Tahmin modulü, Telcordia Elektronik Güvenilirlik Tahmin modulü, FMECA hata, etki, şiddeti analizi modulü, FTA hata ağacı analizi modulü ve RBD güvenilirlik blok diyagramları modulü paket program içerisinde yer alır (Menčík, 2016).

Isograph, bakım ve yedek parça optimizasyonu, sistem kullanılabilirliği tahmini, yaşam döngüsü maliyeti tahminleri için kullanılır. Bu amaçlar Weibull analizi modulü ve yaşam döngüsü tahmini modulünü içerir. Entegre görsel Güvenilirlik modulü ile de hata oranı, bakım tahmini, hata ağacı, güvenilirlik blok diagramı analizlerinin yapılmasına imkan sağlar.

PTC Windchill, eski adıyla Relex Software, kalite, güvenilirlik, risk yönetimine yönelik modulleri içeren entegre paket programdır. PTC Windchill ALT modulü hızlandırılmış yaşam testlerine yönelik ürün güvenirliliği analizinde, PTC Windchil Weibull modulü yaşam veri analizinde, PTC Windchill LCC modulü yaşam döngüsü maliyeti analizinde, PTC Windchill Prediction modulü komponentleri arıza oranı tahmininde kullanılır.

3. MALZEME VE YÖNTEM

Bu çalışmada analizler sırasında tüketici elektroniği sektöründe faaliyet gösteren bir firmaya ait kullanıcı veri seti kullanılmıştır. Analiz için veri seti kullanılan firmanın Türkiye'deki pazar payı %20'nin üzerindedir, Avrupa pazarında da ilk 10 firma arasında yer almaktadır. Veri seti internet bağlantısı olan televizyon kullanıcılarının, kullanım sürelerini ve kullanım alışkanlıklarını içermektedir. Çalışma kapsamında farklı ülkelerdeki izleyicilerin kullanım süreleri ve kullanım alışkanlıkları ele alınarak, kümeleme ve birliktelik kuralları ve güvenilirlik analizine yönelik uygulama çalışmaları gerçekleştirilmiştir.

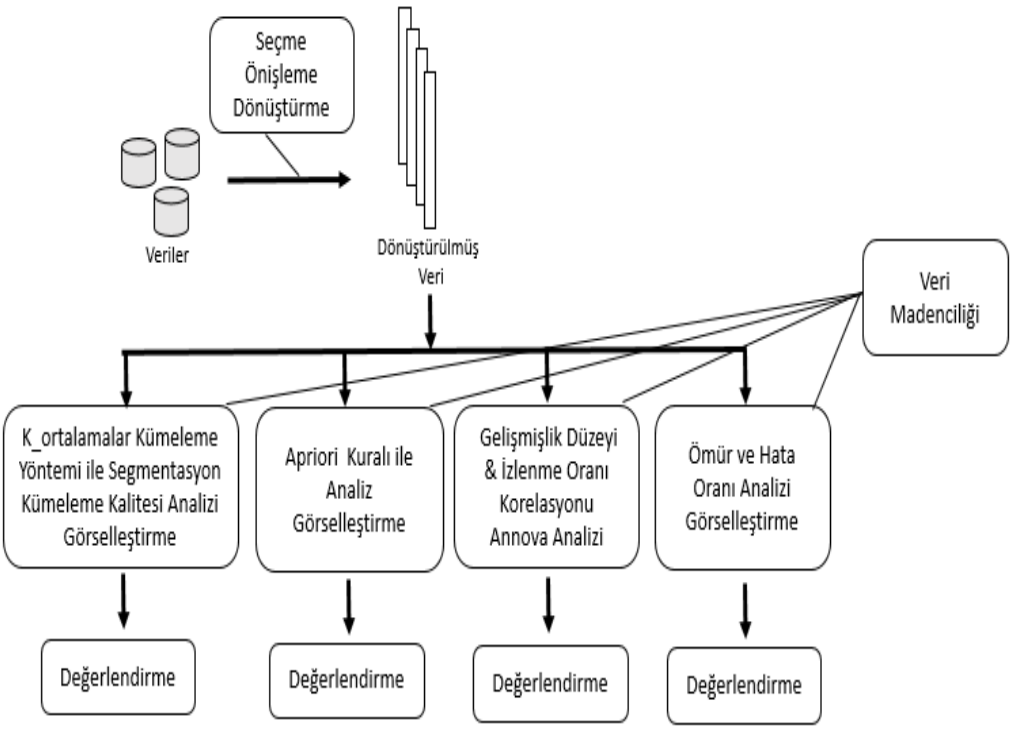
Açık kaynak kodlu olması, veri madenciliği alanında analiz ve görselleştirme uygulamalarında sıklıkla yararlanılması sebebi ile çalışmalar R programlama dilinde, Rstudio ortamında gerçekleştirilmiştir.

Ham veriler, Json formatında ele alınarak işlenip sonrasında csv (Comma Seperated Value) formatına çevrilmiştir. Rstudio programı kullanılarak önişleme ve dönüştürme işlemleri gerçekleştirilmiştir. Csv formatına çevrilen veriler Rstudio programı vasıtası ile uygulamaya hazır hale getirilmişlerdir. Format dönüşümü ve önişleme çalışmaları için rjson, xlsx, dplyr kütüphaneleri kullanılmıştır. Önişleme çalışmaları sırasında değişken dönüşümü işlemleri gerçekleştirilmiştir. Farklı değişken tipindeki kaynak kullanım süreleri tam sayı (integer) veri tipine dönüştürülmüştür. Zamana yönelik analizler yapabilmek için kullanım süreleri saat birimine dönüştürülmüştür. Tarih bilgisi yıl, ay, gün formatına getirilmiştir. Veriler, tarih ve model numaralarına göre filtrelenmiştir. Aynı seri numarasına sahip verilerin oranı %0,1'den düşük olduğu için fazla kopyalar silinerek veri seti içerisinde bir seri numarasından sadece bir kayıt olacak şekilde veri temizleme işlemi gerçekleştirilmiştir. Veri transferinin izlenebilirliği amacıyla gönderilen test verileri, veri kümesinden çıkarılmıştır. Teknik nedenlerden dolayı tarih bilgisinin alınamayarak başlangıç tarihi 1970 UTC zamanı olarak varsayılan veriler ve 10 yıldan daha uzun süreli verilerin oranı %0,5'ten küçük olduğu için bu veriler , veri kümesinden çıkarılmıştır. Veri kümesinde yer alan farklı ölçeklerdeki verilerin diğer nitelikleri baskılamalarının önüne geçebilmek için min-max normalizasyon işlemi gerçekleştirilmiş, bütün veriler aynı ölçeğe getirilmiştir. Her gözlem için, kullanıcı bazlı kaynak izleme süreleri hesaplanarak önişleme süreci tamamlanmıştır.

Veriler içerisinden nitelik seçimi yapılmıştır. Kullanılan kaynak bilgilerini, izleme süresini, ülke bilgisini içeren nitelikler alınarak, 32 nitelik arasından analiz kapsamı dışındaki niteliklerin olduğu veriler çıkarılmıştır. Değerlendirme çalışması 17 nitelik üzerinde yapılmıştır.

Çalışma kapsamında veri madenciliği süreci adımları izlenerek analiz çalışmaları gerçekleştirilmiştir. K-ortalamalar kümeleme yöntemi vasıtasıyla, kullanıcıların kullanım alışkanlıklarına göre segmente edilmesi amaçlanmıştır. Müşteri kullanım alışkanlıkları segmentasyonuna göre farklı ürün konfigürasyonlarının belirlenmesi amaçlanmıştır. Apriori birliktelik kuralları kullanılarak hangi kaynakların birlikte kullanıldığının bulunması hedeflenmiştir. İzleme süreleri ele alınarak ülkeler arası izleme sürelerinin arasında bir fark olup, olmadığı izleme süreleri ile gelişmişlik düzeyi arasında bir ilişki incelenmesi amaçlanmıştır. İzleme sürelerinin ürün ömrü üzerindeki etkisinin değerlendirilmesi hedeflenmiştir.

Çalışma kapsamında izlenen süreç adımlarının akışı ve veri madenciliği aşamasında kullanılan yöntemler aşağıdaki şekilde belirtilmiştir.



Şekil 3.1: Çalışma süreç akışı ve kullanılan yöntemler.

Tez çalışması sırasında veri madenciliği safhası sırasında gerçekleştirilen analiz çalışmalarında kullanılan yöntemler, algoritmalar, kütüphaneler, test sonuçlarına yönelik veriler, indeksler aşağıdaki tabloda gösterilmiştir. Analiz çalışmaları sırasında aşağıdaki tabloda belirtilen R dilindeki kütüphaneleri de kullanılmıştır.

Tablo 3.1: Analizlerde kullanılan yöntem, algoritmalar ve kütüphaneler.

Çalışma İsmi	Kütüphane İsmi/ Test Sonuçları	Yöntem/ Model/ Algoritma/İndeks
Güvenilirlik, Hata Oranı	TM-21, LM80	Arrhenius-Weibul, TM21, ALTA
İzleme Oranı-Gelişmişlik Oranı Analizi	HDI	Anova, Levene, Games_Howell
Birliktelik Analiz	arules, arulesViz	Apriori
Müşteri Segmentasyonu	Amap, Cluster, clvalid, fpc, ClusterCrit, ggpubr, factoextra	K-means, Xie-Beni, Tau, Silhouette, Dunn, C, Calinski_Harabasz, Davies_Bouldin, PCA

3.2 BİRLİKTELİK KURALLARININ GÖRESELLEŞTİRİLMESİ

Veriler arasındaki ilişkilerin daha kolay görülebilmesi, kural dağılımlarının daha rahat anlaşılması için birliktelik kurallarına yönelik kullanılan farklı görselleştirme teknikleri mevcuttur. Birliktelik kurallarının görselleştirmesi için farklı araştırmacılar tarafından kullanılan teknikler arasında serpilme grafiği, iki kriterli serpilme grafiği (two-key plot), mozaik grafiği, çift katlı grafik (double decker), paralel koordinatlar grafiği, matris tabanlı grafik ve grup matris tabanlı grafik gibi araçlar sayılabilirler (Hahsler ve Karpienko, 2017).

Birliktelik Kuralları çalışmalarında sıklıkla kullanılan açık kaynak kodlu R kütüphaneleri arasında arules ve arulesViz sayılabilir. arules kütüphanesi apriori algoritmasına yönelik çalışmalarda kullanılırken, arulesViz analiz çalışmalarının görselleştirilmesinde kullanılmaktadır (Hahsler, 2005 ; Hahsler ve Chelluboina, 2011).

3.2.1 Serpme Grafiği

Serpme grafiği, birliktelik kuralı görselleştirmelerinde sıklıkla kullanılan basit bir tekniktir. Birliktelik kuralına ait destek, güven ve kaldıraç değerleri x, y eksenleri ile grafiğin sağında bulunan renk skalasında gösterilir. Renk skalası ve x, y eksenlerinde destek, güven ve kaldıraç değerlerinden hangisinin gösterileceği değişkendir, programlama sırasında isteğe göre

düzenlenebilir. ArulesViz içinde kullanılan serpilme grafiğinin varsayılan değerine göre kaldıraç değeri renk skalasında belirtilirken, destek ve güven değerleri x ve y eksenlerinde belirtilirler (Hahsler ve Chelluboina, 2011).

3.2.2 İki Kriterli Grafik (Two Key Plot)

Unwin ve diğ. (2001) tarafından serpilme grafiğinden türetilen görselleştirme tekniğidir. Bu grafikte x ekseni destek, y ekseni güven değerini belirtirken, renk skalası ise order değerini belirtmektedir. Order kural içinde bulunan öğe sayısını ifade etmektedir. Renk skalasında bulunan renk koyulaştıkça kuraldaki öğe sayısı artmaktadır. Serpilme grafiğinin etkileşimli modunda, grafikteki istenilen bölgenin uzaklaştırılıp, yakınlaştırılabilmesi, noktaların belirttiği kuralları görüntüleyebilme, filtreleyebilme ve görüntüyü sonlandırma seçenekleri mevcuttur (Unwin ve diğ., 2001).

3.2.3 Çift Katlı Grafik (Double Decker Graph)

Çift katlı görselleştirme tekniği mozaik görselleştirme tekniğinin bir çeşididir. Mozaik grafik tekniğinde koşul tablosu kullanılır. Grafik üzerinde yer alan dikdörtgen biçimindeki her bir desenin boyutu koşul tablosundaki değeri ile ilişkilidir. Mozaik grafik yönteminde yinelemeli yatay ve dikey bölmeler varken, Çift katlı grafik yönteminde sadece yatay bölmeler kullanılır. Çift katlı grafik tekniği tek bir birliktelik kuralının görselleştirilmesine imkân verir (Hahsler ve Chelluboina, 2011).

3.2.4 Paralel Koordinatlar Grafiği

Paralel koordinatlar grafiği, çok boyutlu verilerin x ve y eksenlerinde gösterimine yönelik tasarlanmış görselleştirme tekniğidir. Her veri noktası, her boyut için değerleri birbirine bağlayan çizgilerle gösterilir. Çizgiler öncül ve sonuç değerlerini birbirine bağlar. Çizgilerin ucundaki oklar sonuç değerini işaret eder.

Okların genişliği destek değerini, okların renk yoğunlukları ise güven değerini ifade etmektedir. Kuralın destek değeri arttıkça ilgili okun genişliği, kuralın güven değeri arttıkça ise okun renk yoğunluğu grafikte artmaktadır. Paralel koordinatlar grafiğinin olumsuz noktası, kural sayısının artmasına bağlı olarak çizgilerin birbirlerine karışması kaynaklı oluşacak görselleştirme analizi zorlaşmasıdır (Yang, 2003).

3.2.5 Matris Tabanlı Grafikler

Matris tabanlı grafikler, öncül ve sonuç öğelerinin x ve y eksenlerinde gösterimine dayalı görselleştirme teknikleridir. Belirlenen kuralın öncül ve sonuç değerlerinin grafik üzerindeki kesişim noktasında seçilen ilgilenilen değer görüntülenir. Eğer seçilen öncül ve sonuç öğelerinin kombinasyonu için bir kural tanımlı değilse grafik üzerindeki kesişim yeri boşluk olarak gösterilmektedir. Minimum güven ve destek eşik değerlerini sağlayamayan birliktelik kuralları da matris grafiğinde boş hücreler şeklinde belirtilmektedir. Grafiğin sağ kısmında renk skalası yer alır. Güven değerini gösteren renk skalasının rengi, güven değeri artıkça açık renkten koyu renge doğru dönüşür. Güven değeri 0-1 arasında değerler almaktadır (Hahsler ve Chelluboina, 2011).

2 boyutlu matris tabanlı grafikler bire bir ikili ilişkiyi göstermek için etkili bir tekniktir. Farklı disiplinlerdeki çok çeşitli verileri analiz etmek için kullanılmışlardır. Bununla birlikte birden çok öncül öğeye sahip birliktelik kuralının olduğu durumlarda çoktan bire ilişkinin görselleştirilmesi gerektiğinde 2 boyutlu yapı gücünü kaybetmektedir. 2 boyutlu matris yapısının diğer bir dezavantajı ise farklı meta veri değerlerini göstermek için birden fazla simge kullanıldığında matris desenleri üst üste oturduğundan kapama sorunu oluşmaktadır. Bu olumsuzlukların önüne geçebilmek için Wong ve diğ. (1999) tarafından 3 boyutlu matris tabanlı grafik yapısı önerilmiştir. 3 boyutlu matris grafik yapısında öncül ve sonuç değerlerinin kesişimleri 3 boyutlu barlar ile gösterilmektedir. Öncül öğeler ve sonuç öğeler x ve y eksenlerinde gösterilirken, z ekseninde kuralın güven değeri 3 boyutlu bar şeklinde gösterilmektedir. Kuralın güven değeri arttıkça 3 boyutlu barın yüksekliği de artmaktadır (Hahsler ve Chelluboina, 2011).

Kural sayısının arttığı hallerde tipik olarak öncül ve sonuç değerleri sayılarının da artmasından dolayı matris tabanlı grafik ile etkili bir şekilde görselleştirme, kural sayısına bağlı sınırlanmaktadır. Grafik üzerindeki renkli karelerin veya 3 boyutlu barların boyutları küçülmekte ve görülebilmeleri zorlaşmaktadır. Bu durumun önüne geçebilmek için gruplama yöntemi önerilmiştir. Gruplama için K-ortalamalar sınıflandırma yöntemi kullanılmıştır (Hahsler ve Chelluboina, 2011; Hahsler ve Karpienko, 2017).

3.2.6 Grafik Tabanlı Görselleştirme

Grafik tabanlı görselleştirme tekniğinde düğümler ve oklar kullanılmaktadır. Düğümler nesne ve nesne kümelerini belirtirken, oklar da kurallar içerisindeki öncül ve sonuçları birbirine bağlayarak ilişkiyi gösterirler. Okların kalınlığı, renkleri veya etiketler birliktelik kuralının gücü hakkında bilgi vermektedirler. Kurala ait destek değeri arttıkça grafikteki okun kalınlığı, kaldıraç, güven değeri arttıkça okun renk koyuluğu artmaktadır.

Grafik tabanlı görselleştirme tekniği küçük boyutlu kural kümelerinde başarılı olurken, kural kümesinin boyutu artıkça karmaşık hale gelmektedir. Kural kümesinin boyutunun büyüdüğü durumlar için gelişmiş etkileşimli özelliklere sahip, düğümlerin renklerinin değiştirilebildiği, görüntünün yakınlaştırılıp, uzaklaştırılabildiği, filtreleme veya gruplama yapılabilen Gephi gibi platformlar ya da gruplandırılmış matris tabanlı görselleştirme teknikleri önerilmiştir (Hahsler ve Chelluboina, 2011; Hahsler ve Karpienko, 2017).

3.2.7 Gruplandırılmış Matris Tabanlı Görselleştirme

Matris tabanlı görselleştirme tekniğinin büyük boyutlu kural kümeleri üzerinde etkili olamaması sebebi ile gruplandırılmış matris tabanlı görselleştirme tekniği geliştirilmiştir. Daha fazla sayıda kural üzerinde görselleştirme işlemi yapabilmek için kuralların öncül değerleri kümeleme yöntemi ile gruplanır veya kurallar ilgi durumlarına göre sıralanır. Kümeleme yöntemi kullanılarak büyük farklı öncül kümeler (sütunlar) küçük grup kümesi halinde düzenlenir. Gruplama sırasında küme içi kareler toplamı minimum olacak şekilde öncüller kümesi k adet gruba sahip kümeye ayrılır $s = \{s_1, s_2, s_3, s_4 \ldots . s_k\}$.

$$argmin_s \sum_{i=1}^{k} \sum_{m_j \in S_i} \|m_j - \mu_i\|^2 \tag{3.1}$$

$j = 1, \ldots, A$ olmak üzere m_j, aynı öncül değerli bütün kuralları gösteren sütun vektörüdür, μ_i ise S_i içindeki vektörlerin ortalama değerini ifade etmektedir. Kümeleme işlemi yapılmadan önce minimum destek ve güven eşik değerinin altında kalan kayıp değere sahip öncül/sonuç kombinasyonları doğal eleman değeri ile doldurulur. Sonuçta sütun olarak öncül grupları içeren daha küçük matris oluşur. Gruplandırılışmış matris tabanlı görselleştirme tekniğinde grup kuralları balonlar şeklinde gösterilir (Hahsler, 2017).

Gruplandırılmış matris tabanlı görselleştirme tekniğinde balonların boyutu destek değerini, renkleri ise kaldıraç değerini göstermektedir. Balon büyüklüklerinin ve renklerinin temsil ettiği değerler uygulamaya göre farklı şekilde de seçilebilmektedirler. Gruplandırılmış matris tabanlı görselleştirme tekniği ayrıca kullanıcılara yakınlaştırma, uzaklaştırma ve gruplama gibi özellikleri etkileşimli olarak da sunabilmektedir.

Büyük boyutlu kural kümeleri üzerinde etkili olarak çalışabilen açık kaynak kodlu Gephi grafik tabanlı görselleştirme platformu, veri analizi, ilişki analizi, sosyal ağ analizi ve biyolojik ağ analizlerinde kullanılabilmektedir. Etkileşimli şekilde yapı, şekiller ve renkler değiştirilerek gizli paternlerin ortaya çıkarılmasına imkân verilmektedir (Bastian ve diğ., 2009).

arulesViz kural ilişkilerini belirlemek için kolay bir yöntemdir. arulesViz'e alternatif olarak kullanılabilecek Java betik tabanlı HTML uygulamaları veya R markdown ile yaratılmış seçenekler de mevcuttur (Hahsler, 2017).

3.3 İNSANİ GELİŞMİŞLİK İNDEKSİ (HUMAN DEVELOPMENT INDEX, HDI) VE POST-HOC ANALİZİ

Çalışma sırasında Birleşmiş Milletler Gelişim Programı (UNDP, United Nations Development Programme) tarafından yayınlanan 2020 yılı İnsani gelişmişlik raporu verileri kullanılmıştır.

UNDP tarafından İnsani gelişmişlik indeksi (HDI) değerinin elde edilmesinde 3 boyut dikkate alınmıştır. Bu boyutlar; uzun ve sağlıklı yaşam süresi, bilgiye erişim değeri ile yaşam standardıdır. HDI indeksi bu üç parametrenin normalize edilen değerlerinin geometrik ortalaması alınarak hesaplanmıştır. Bilgiye erişim değeri, beklenen eğitimi süresi ve ortalama eğitim süresi verilerinden elde edilmiştir. Yaşam standardı değeri, kişi başına düşen gayri safi milli gelir verisinden sağlanmıştır (Human Development Report 2020).

İndeks değerlerini 0 ve 1 arasındaki değerlere dönüştürmek için hesaplamalarda sağlık, eğitim ve yaşam standardı alanlarında kabul edilen maksimum ve minimum değerleri aşağıdaki tablodaki gibidir.

Tablo 3.2: Gösterge maksimum ve minimum değerleri.

Boyut	Gösterge	Minimum	Maksimum
Sağlık	Yaşam beklentisi (yıl)	20	85
Eğitim	Beklenen eğitim süresi	0	18
	Ortalama eğitim süresi	0	15
Yaşam Standardı	Kişi başına düşen milli gelir (2017 PPP)	100	75000

Boyut indeksi değerleri aşağıdaki formüle göre normalize edilmişlerdir.

HDI değerleri ile izleme oranları arasındaki ilişkiyi incelemek için çalışma sırasında Anova analizi kullanılmıştır.

$$boyut\ endeksi = \frac{gerçek\ değer\ - minimum\ değer}{maksimum\ değer\ - minimum\ değer} \tag{3.2}$$

HDI indeksi aşağıdaki formülde gösterildiği gibi 3 parametrenin geometrik ortalaması alınarak hesaplanmıştır.

$$HDI = \left(I_{yaşam} \cdot I_{eğitim} \cdot I_{gelir}\right)^{1/3} \tag{3.3}$$

Tablo 3.3: Bölge bazlı yıllara göre insani gelişmişlik düzeyi puanları.

Yıllar	Arap Ülkeleri	Doğu Asya ve Pasifik	Avrupa ve Orta Asya	Latin Amerika ve Karayipler	Güney Asya	Sahra Altı Afrika
1990	0,556	0,517	0,662	0,632	0,437	0,404
2000	0,614	0,595	0,675	0,69	0,501	0,426
2010	0,676	0,688	0,739	0,736	0,58	0,501
2014	0,687	0,718	0,772	0,756	0,612	0,53
2015	0,691	0,724	0,775	0,759	0,62	0,535
2017	0,699	0,735	0,785	0,762	0,635	0,542
2018	0,702	0,740	0,787	0,764	0,637	0,544
2019	0,705	0,747	0,791	0,766	0,641	0,547

UNDP tarafından bildirilen bölgelere göre 1990, 2000, 2014, 2015, 2016, 2017, 2018 ve 2019 yıllarını kapsayan İnsani gelişmişlik düzeyi puanları yukarıdaki tabloda verildiği şekildedir.

Anova analizi ile grupların ortalamaları arasında istatistiksel olarak anlamlı bir fark olup olmadığı anlaşılabilir. Anova testi sonucuna göre yokluk (Ho) hipotezi reddedilip, Ha hipotezi kabul edildiği durumda grupların ortalamaları arasında anlamlı bir fark olduğu söylenir. Bu durumda gruplardan en az birinin ortalaması diğerlerinden farklıdır. Fakat hangi grup ya da grupların ortalamalarının farklı olduğu Anova testi ile belirlenemez. Gruplar arasında farklı grup veya grupları tespit etmek için post-hoc analizleri kullanılır. Gruplar arası varyans homojenliği varsayımının sağlanması (gruplar arası varyansın eşit olup olmaması), örneklem büyüklüğü ve karşılaştırılacak grup sayısı, analiz için seçilecek post-hoc testinin belirlenmesinde önemli kriterlerdir. Yanlış post-hoc testi seçimi hipotez testleri için tip1 ve tip2 hatasına yol açabilir.

Gruplar arası varyansların eşitliği halinde seçilebilecek post-hoc testleri çoklu karşılaştırma testleri ve çoklu aralık testleri olmak üzere iki gruba ayrılırlar. Çoklu karşılaştırma tipindeki testler arasında LSD, Tukey HSD, Gabriel, Scheffe, Bonferroni, Sidak, Hochberg's GT2 sayılabilir. Çoklu aralık tipindeki testleri arasında ise SNK (Student-Newman-Keuls) Duncan, Dunnett, Waller Duncan, Tukey's B, yer almaktadır (Patel ve diğ., 2015; Karakaş, 2017).

Gruplar arası varyansların farklı olması durumunda seçilebilecek post-hoc testleri, çoklu aralık testleri tipindedir. Games-Howell, Tamhane's T2, Dunnet's C, Dunnet's T3 testleri varyansların eşit olmaması halinde kullanılan post-hoc testleridir.

Anova testi normal olmayan verilere karşı dirençli olsa da normalliğin sağlanamadığı durumlarda sonuçlar yanlı olabilir. Normal olmayan veri setinin olduğu durumlarda veriyi normal hale getirebilmek için örneklem büyüklüğü arttırılabilir, veri dönüştürülerek, dönüştürülmüş veri üzerinden Anova analizi yapılabilir veya parametrik olmayan Kruskal Wallis testi kullanılabilir.

Kruskal Wallis testi sonrasında gruplar arasında anlamlı bir fark olduğu tespit edilirse, çiftli olmayan (unpaired) veriler için Mann-Whitney U testi, çiftli veriler için ise Wilcoxon Sign Rank test ile gruplar arasında kıyaslama yapılabilir. Grup sayısının 3 veya daha fazla olduğu durumda çiftli veriler için Friedman testi yapılır (Khan ve Goel 2021; Măruşteri ve Bacârea, 2010).

Anova varsayımları; varyansların homojenliği, verilerin bağımsız olması, grup içi verilerin normal olması, bağımlı değişkenin eşit aralıklı ölçeğe sahip olmasıdır. Normallik testi için Kolmogorov–Smirnov testi ve Shapiro–Wilk testi kullanılabilir (Guglielmetti ve diğ., 2021).

Grup büyüklükleri eşit olduğunda Anova normallik ve grup varyans homojenliği ihlallerine karşı dirençlidir. Grupların popülasyon varyanslarının ve eleman sayılarının eşit olmaması post-hoc testlerini etkilemektedir.

Grup sayısının ve örneklem büyüklüğünün artması sonucu kullanılan post-hoc analiz yöntemlerinden bazılarında p değerinin küçülmesi yönünde hassasiyet oluşabilmektedir, bu durumda hatalı şekilde Ho hipotezinin reddedilerek, Ha hipotezinin kabul edilmesine yol açabilmektedir. Örneğin grup sayısının 3'ün üzerinde olduğu durumlarda Tukey testi önerilmemektedir.

Numune sayısı parametrik ve parametrik olmayan testlerin seçiminde bir etkendir. Örnek sayısının 15'ten küçük olduğu durumda parametrik testlerin yapılması sağlıklı değildir, parametrik olmayan testlerin yapılması gerekir. Dağılım Gaussian değilse, parametrik testlerde p değeri yanıltıcı olabilir. Eğer dağılım Gaussian ise non parametrik test ile elde edilen p değeri, parametrik test ile elde edilen p değerinden büyük olabilir. Bu durumda non parametrik testin gücü zayıftır (Măruşteri ve Bacârea, 2010).

Varyans analizi ve çoklu karşılaştırma testleri için kullanılabilecek paket program arasında IBM SPSS, Minitab, Jamovi, R, NCSS, GraphPad, SAS sayılabilir.

Tukey HSD (Tukey Honestly Significant Difference) testi, varyansların homejen olduğu durumda yaygın olarak kullanılan bir çoklu karşılaştırma testidir. Gruplardaki örneklem sayılarının eşit olduğu varsayımını dikkate alan ve kontrol grubu bulundurmayan bir post-hoc testidir. Sütten- t istatistiğini kullanır (Karakaş, 2017).

Tukey-Kramer testi, örneklem eşitliği varsayımı gerektirmeyen karşılaştırma testidir. Tukey HSD'den farklı olarak standart hata değeri Tukey-Kramer testinde hesaplanmaktadır (Karakaş, 2017).

Bonferroni yöntemi, eşit örneklem sayısı varsayımını gerektirmeyen çoklu karşılaştırma testlerindendir. Bonferroni düzeltmesi olarak da adlandırılmaktadır. 3 veya daha fazla

değişkenin olduğu durumda çiftler arasında yapılan karşılaştırma testlerinin sayısı arttıkça Tip1 olarak da adlandırılan yokluk hipotezinin (Ho) yanlışlıkla reddedilme oranı artar. Bonferroni düzeltmesi sırasında anlamlılık düzeyi, test sayısına (n) bölünerek yeni bir anlamlılık düzeyi (α/n) hesaplanır. İkili karşılaştırmalar sırasında hesaplanan yeni anlamlılık düzeyine göre testler yapılır. Bonferroni düzeltmesi yapılarak karşılaştırmalar sırasında oluşacak Tip1 hata oranı azaltılır (Karakaş, 2017).

Dunnett testi, varyansların homejen olduğu durumda kullanılan bir çoklu aralık testidir. Kontrol grubu ortalaması ile deneysel grupların ortalaması belirli bir anlamlılık düzeyinde karşılaştırır. Dunett testi için gruplarda yer alan örnek sayısının eşit olması varsayımı söz konusudur. Student-t istatistiğini kullanır (Karakaş, 2017; Gündoğdu, 2014).

Fisher LSD (Least Significant Difference) Varyansların eşit olması durumunda, grup içindeki eşit örneklem varsayımını gerektirmeyen çoklu karşılaştırma testidir. Grup sayısının 2'den fazla olduğu durumda yapılan karşılaştırma testlerinde grup sayısı artıkça Tip 1 hata oranı artmaktadır. Tip 1 hata türüne karşı korumasızdır, diğer Post-hoc testleri ile kıyaslandığında daha zayıftır (Gündoğdu, 2014).

Scheffe testi, varyansların eşitliği durumunda kullanılan çoklu karşılaştırma testlerindendir. Gruplardaki örneklem sayılarının eşitliği varsayımını dikkate almaz . Gruplar arasında sadece ikili karşılaştırmaları değil mümkün olan bütün kombinasyonları Tip1 hatayı da kontrol altında tutarak karşılaştırır. İkili karşılaştırmalarda Tukey HSD kadar güçlü olmamasına karşın karmaşık karşılaştırmalarda esnek ve çok güçlüdür (Scheffe, 1953).

Games_Howell testi, varyansların eşit olmaması durumunda tercih edilen çoklu aralık testlerindendir. Liberal çoklu karşılaştırma testi olarak da isimlendirilmektedir. Tamhane's T2'den farklı olarak Student-t ve genişletilmiş t modülü tabanlarının her ikisinde de çalışabilmektedir.

3.4 İZLENME SÜRESİNE GÖRE LED GÜVENİLİRLİK ANALİZİ

Çalışma kapsamında televizyon ekranlarının arka aydınlatmasında kullanılan LED'lerin parlaklık seviyelerinin azalmasına yönelik ömür analizi faaliyetleri yürütülmüştür. Hızlandırılmış ömür testleri sonucunda elde edilen LED'ler ile ilgili verilerin Weibull dağılımı ile analiz edilmesi amaçlanmıştır.

LED (Light Emitting Diode) ışık kaynakları günümüzde aydınlatma elemanları olarak yoğun şekilde kullanılmaya başlanmışlardır. Doğrudan iç ve dış aydınlatma LED sistemlerinde kullanılmalarının yanı sıra beyaz eşya ve kahverengi eşya sektörlerinde arka ışık kaynağı olarak da sıklıkla kullanılabilmektedirler. Konvansiyonel ışık kaynaklarına göre daha verimli ve uzun ömürlüdürler. Özellikle kahverengi eşya sektöründe televizyonların arka ışık kaynağı olarak, CCFL (Cold-Cathode Fluorescent Lamp) teknolojisine göre LED ışık kaynakları tercih edilmektedir. Televizyonlarda arka ışık kaynağı olarak LED lambaların kullanımı ile birlikte güç tüketim değerleri daha aşağı çekilmiş, görüntünün homojenliği artırılmıştır.

Eklem noktasındaki sıcaklık olan jonksiyon sıcaklığı, ortam sıcaklığı yanında geçen akım değeri ve ısıl empedans değerine bağlıdır. Jonksiyon sıcaklığının artması kullanılan LED'lerin performansları üzerinde olumsuz etkilere yol açmaktadır. Jonksiyon sıcaklık değerindeki değişime bağlı olarak LED'lerin ömür değerleri etkilenir, ışık akısı ve renk gamutu gibi paremetrelerinde değişimler meydana gelir. Jonksiyon sıcaklığı artıkça ömür değeri azalmaktadır.

Yaşam süresi-stres modellemesi sırasında sıcaklık, başınç, gerilim, nem gibi stres faktörleri uygulanarak ürünün normal durumlarda yüzyüze kaldığı stres seviyesinin üzerindeki koşullarda testler gerçekleştirilir. Buradaki amaç gerçek çalışma koşullarında yapılan testlerde ürünün bozulma süresi uzun süreceğinden, stres faktörü uygulanarak test ve analiz süresini kısaltmaktır. Stres uygulandığı koşullarda elde edilen değerlerden yararlanılarak ektrapolasyon yöntemi vasıtasıyla gerçek çalışma koşulları altındaki ömür tahmini gerçekleştirilir. Arrhenius yaşam modeli çalışmalarında hızlandırma etkeni olarak sıcaklık kullanılmaktadır. Sıcaklık arttıkça yaşam süresi kısalmaktadır. Çalışma sırasında Arrhenius modelinin kullanılması düşünülmüştür. LED lambalarda lüminans değeri üstel olarak azalmaktadır.

Kimyasal reaksiyon denkleminde E_a'nın reaksiyonun aktivasyon enerjisini (eV), k'nın Boltzmann sabitini (8,6171 x 10^{-5} eV/Kelvin), T'nin termodinamik Kelvin sıcaklık derecesini (°C + 273), A'nın ürün veya malzeme karakteristiğini belirttiği durumda, Arrhenius kimyasal reaksiyon oranı aşağıdaki şekilde ifade edilir (Escobar ve Meeker 2006).

$$Reaksiyon\ Oranı = Aexp\left(\frac{-E_a}{kT}\right) \tag{3.4}$$

2 farklı termal stres seviyesinde ömür karakteristiğine göre aktivasyon enerjisi aşağıdaki denklemdeki şekilde gösterilir.

$$E_a = \frac{kln(\frac{\tau}{\tau'})}{\frac{1}{T}-\frac{1}{T'}} \tag{3.5}$$

τ değerinin ürünün *T* sıcaklığındaki yaşam ömrünü, τ' değerinin ürünün *T'* sıcaklığındaki yaşam ömrünü ifade ettiği durumda hızlandırma faktörüne yönelik bağlantı aşağıdaki şekilde gösterilir. İki farklı sıcaklık değerindeki yaşam ömürlerinin oranı hızlandırma faktörü HF_A'a eşittir.

$$HF_A = \frac{\tau}{\tau'} = \exp\left[\left(\frac{E_a}{k}\right)\left(\frac{1}{T} - \frac{1}{T'}\right)\right] \tag{3.6}$$

25°C ortam sıcaklığındaki ürünün yaşam ömrü τ_0 aşağıdaki şekilde elde edilir. Aşağıdaki eşitlikte T_0, 25°C'deki jonksiyon sıcaklığı iken *T* hızlandırılmış koşullardaki ortam sıcaklığını ifade etmektedir.

$$\tau_0 = \tau' \exp\left[\left(\frac{E_a}{k}\right)\left(\frac{1}{T_0} - \frac{1}{T}\right)\right] \tag{3.7}$$

Uzun ömürleri, düşük güç tüketimleri ve boyut avantajlarından dolayı 2000'li yıllarda LED teknolojisi aydınlatma endüstrisinde birçok uygulamada yaygın olarak kullanılmaya başlanmıştır. LED aydınlatma sistemlerinin gelişmesi ile birlikte LED sistemlerin ömürlerinin hesaplanması ve test edilebilmelerine yönelik ihtiyaçlar ortaya çıkmış ve bu gereksinimler doğrultusunda standartlar oluşturulmaya başlanmıştır.

Katı Hal Aydınlatma Sistemleri ve Teknolojileri Birliği (ASSIST-The Alliance for Solid State Illumination System and Technologies) LED ışık kaynaklarının ömür ölçümlerine yönelik çalışmalar yapmış, standartlar önermiştir. Aydınlatma Mühendisliği topluluğu (Illuminating Engineering Society- IES) ASSIST tarafından sunulan önerileri göz önünde bulundurarak LED parlaklık seviyesi ölçümüne yönelik IES LM-79, IES LM-80 ve LED parlaklığı ömür tahminine yönelik IES TM-21 standartlarını oluşturulmuştur. Bu standartlara göre yapılan çalışmalarda doğru ömür öngörüsü için ; LED test ölçüm zamanının oda sıcaklığında en az 6000 saat olması önerilmektedir.6000 saatlik test zamanına alternatif olacak şekilde farklı metotlar içeren hızlandırılmış ömür testleri de bulunabilmektedir. Yapılan çalışmalarda sıcaklık, nem, akım değerlerinin artırılması ile veya adım stres ömür testleriyle test süresinin kısaltılması amaçlanmıştır.

LED sistemleri birden fazla bileşenden oluştuğu için sistemi oluşturan bileşenlerden bir tanesinin bozulması ve fonksiyonel özelliğini yitirmesi tüm sistemin arızalanmasına yol

açacaktır. Litaratürde LED sistemlerindeki arızalara katastrofik ve parametrik olarak iki gruba ayrılmaktadır. Lümen seviyesinin azalması ve renk değişimi parametrik hata tipleridir. (Lighting Research Center, 2021; Fan ve diğ., 2012). Florasan ve akkor ışık kaynaklarından farklı olarak LED ışık kaynaklarında görülen temel hatalar katastrofik arızalar değildir. LED yaşlanmasının en belirgin göstergesi yayılan ışık akısı miktarının başlangıçtaki seviyesine göre azalmasıdır (Hegedüs ve diğ., 2020).

Katastrofik arıza durumunda, elektriksel bir hata oluştuğundan sistem ışık veremez hale gelmektedir. LED'lerin üzerine dizildiği baskı devrede oluşan çatlaklar sonucu devre bağlantısının kesilmesi veya padlerin lehimlerinde oluşacak kısa devre, açık devre hataları, LED'leri besleyen güç kartı ünitelerindeki komponelerin özellikle güç mosfetleri ve elektrolitik kapasitörlerdeki bozulmalar bu tür hatalara yol açabilmektedir. Paremetrik hata durumu ise zamana bağlı olarak LED akı miktarının belirlenen limit değerinin altına düşmesi veya renklerin istenilen değerlerden sapması ile oluşur (Lighting Research Center, 2021). LM-80 standardının 2008, 2015 ve 2020 yıllarında farklı verisyonları yayınlaşmıştır.

IES tarafından oluşturulan LM-80-08 standardı LED ışık kaynağının Lümen seviyesini ölçmeye yönelik onaylı bir metottur. LM-80-08, Led paketlerinin, dizilerinin ve modullerinin lümen sviyelerinin ölçümüne yönelik gerekli test düzeneğinin yapısını,ölçüm koşullarını ve ölçüm prosedürünü tanımlamaktadır. LED tabanlı aydınlatma sistemlerinin ömür tahmininde kullanılan temel onaylı kaynaklar IES LM-80-08 yöntemi ve IES TM-21-11 teknik genelgesidir (LG8 Hegedüs ve diğ., 2020).Işık akısı ölçümüne yönelik yöntem olarak IES LM-80-08 ve ömür tahmini metodu olarak da IES TM-21-11 kullanılmaktadır.

ASSIST tarafından yayınlanan öneri dokümanında ışık akılarının korunma yüzdesine göre LED ömrü tanımı yapılmıştır. L80, L70 ve L50 değerleri sırasıyla ışık kaynağının akısının başlangıç anındaki seviyesine göre %80'e, %70'e ve %50'ye düştüğü süreyi saat cinsinden ifade eder. Aydınlatma amaçlı kullanımlar için eşik değeri, ilk ışık akısının %70'i seviyesi olarak tanımlanırken, renksel algının önemli olduğu durumlarda ise ilk ışık akısının %80 seviyesi eşik değeri olarak belirtilmiştir. Kritik olmayan dekorasyon amaçlı kullanımlarda ise ışık akısı miktarının yarıya düştüğü %50 seviyesi kabul edilebilir eşik seviyesi olarak ifade edilmektedir.

IES LM-80-08 yöntemine göre 3 farklı LED kılıf sıcaklığında ölçüm yapılması gerekmektedir. IES LM-80-08 yöntemine göre 55 °C ' ve 85 °C derecelerdeki ölçümler gerekli tutulmuştur.

Üçüncü ölçümün kaç derecede yapılacağı kararı üreticinin tercihine bırakılmıştır. Optik ölçümler 25 °C ortam sıcaklığında en fazla her 1000 saatlik aralıklarla ölçüm yapılacak şekilde tariflenmiştir. IES TM-21-11'de tariflenen ışık akısı veri toplama yöntemine göre toplam test süresi en az 6000 saat olmalı, mümkünse toplam test süresinin 10000 saate kadar uzatılması daha kesin öngörü yapabilmek için tercih edilmelidir. LM-80-08 yönteminin yeni versiyonu olan LM-80-15'de ise kılıf ölçümünün 3 farklı sıcaklık yerine 2 farklı sıcaklık değerinde ölçülmesi yeterli görülmüş, minimum ölçüm süresinin 6000 saat olması gerekliliği iptal edilmiştir (Hegedüs ve diğ., 2020).

IES tarafından oluşturulan TM-21 metodu, IES tarafından onaylı LED ışık kaynaklarının lümen bakımına yönelik LM-80 standardına göre elde edilen test verilerini kullanarak İnterpolasyon yöntemi ile L70 değeri için ışık akısının düşeceği zamanı tahmin etmeyi sağlamaktadır. LED üreticeleri LM-80 verileri paylaşmaktadırlar. Üreticiler tarafından paylaşılan LM-80 veri kümeleri çoğunlukla üç farklı sıcaklık ve 3 farklı akım değerindeki akı değerlerini içermektedir.

IES TM-21-11 metodu tarafından, LM-80-08 ile elde edilen optik test sonuçları kullanılarak üstel eğri uydurma yöntemine dayalı İnterpolasyon tekniği uygulanır. TM-21-11'e göre ilk 1000 saat öncesi veriler öngörü hesaplamalarında kullanılmamalıdır.LM-80 veri kümesine göre elde edilen zamanın 6 katı süre hesaplamalarda kullanılarak öngörü yapılabilir. [LG5 Driel ve diğ., 2015]. TM21 yöntemi, laboratuvarda yapılan ölçüm zamanının 6 katından fazlasını ömür öngörü hesabında istatistiksel hata artacağı için kullanmaz.

LM-80'e göre toplanan veriler 20 test numunesi içermelidir. Test numune sayısının 20 olması durumunda sürenin 6 katı çarpım faktörü kullanılabilir. Örneklem büyüklüğü 10 ile 19 arasında olduğu durumda sürenin 5.5 katı çarpım faktörü kullanılabilir. Bu yöntemin kullanılmabilmesi için örneklem büyüklüğününü 10'un altına düşmememesi gerekir. Belirli zaman aralıklarında elde edilen tüm akı verileri 0 ile 1 arasında olacak şekilde normalize edilir. Normalize edilen verilerin ortalaması alınır. 1000 saatin altındaki veriler kullanılmaz. 6000 ile 10000 saat arasındaki verilerden son 5000 saat içindeki veriler kullanılır.

Saha verileri tutulan ekranın arka aydınlatması için ışık kaynağı olarak LED barlar kullanılmaktadır. Arka aydınlatmada kullanılan LED'ler ile aynı çalışma aralığına sahip 55 °C ve 85 °C derecelerde LM-80 test sonuçları verisi olan LED'lerin test sonuçları analiz çalışmalarında kullanılmıştır.

4. BULGULAR

4.1 MÜŞTERİ SEGMENTASYONU ANALİZİ

Çalışma sırasında 1500 farklı gözleme ait farklı veri tiplerindeki 32 niteliğin veri kümesi üzerinde çalışma yapılmıştır. Kullanıcılar tarafında kullanılan kaynaklar üzerinde işlem yapılacağından 32 niteliğin içerisindeki 13 nitelik üzerinde işlem yapılmış, kaynak verisini içermeyen diğer niteliklerin bulunduğu sütunlar veri kümesinden çıkarılmıştır.

```
> str(dataMonitorKP);
'data.frame':   1500 obs. of  32 variables:
 $ X                              : int  1 2 3 4 5 6 7 8 9 10 ...
 $ arrival_time                   : chr  "2019-01-04 15:08:45" "2019-01-23 11:00:52" "2019-01-20 22:43:22"
 $ serial                         : num  8.08e+12 6.16e+13 8.01e+13 8.01e+13 8.01e+13 ...
 $ Logger.SW_Version              : chr  "KP" "KP" "KP" "KP" ...
 $ Logger.PanelCode               : chr  "4d5a4d3030303030303000" "6161616161616161616100" "525241303030303
 $ Logger.ChasisSerial            : chr  "4d5a4d435a5a46140010ffffffff" "30342e3032392e303000ffffffff" "524
 $ Logger.StartEndTimeDiff        : int  12795384 53188262 24040890 27579239 1548614807 20568626 1547229648
 $ Logger.TVLifeTime              : int  4006856 1100449 1215108 5463096 2234465 1455282 8797436 3703016 56
 $ SourceSettings.Country         : chr  "DEU" "TUR" "DEU" "DEU" ...
 $ Duration.Air                   : int  284 257 722 143 5881 1409 154 831 46 125 ...
 $ Duration.AirAnalog             : int  0 429 0 0 0 0 0 0 0 54 ...
 $ Duration.Cable                 : int  0 444903 1163225 5294564 0 143365 0 249461 142 0 ...
 $ Duration.CableAnalog           : int  0 1080 522 0 0 0 0 230 0 0 ...
 $ Duration.Component             : int  0 22 0 0 0 0 0 0 0 0 ...
 $ Duration.HDMI1                 : int  236585 452591 1564 7 354045 726983 8608908 752232 0 3628 ...
 $ Duration.HDMI2                 : int  67 12817 1549 0 1838377 55461 574 213392 0 971 ...
 $ Duration.HDMI3                 : int  391 933 4713 15571 1601 0 244 1201 0 1062382 ...
 $ Duration.HDMI4                 : int  60 28 0 0 500 0 1973 190 299 34769 ...
 $ Duration.Usb                   : int  15829 152924 46263 157117 34697 30668 189507 32053 2263055 118617
 $ Duration.SVideo                : int  0 2 0 0 0 0 0 0 0 0 ...
 $ Duration.Satellite             : int  3780310 36785 6982 0 0 497944 0 2541210 3383475 6418685 ...
 $ Duration.Scart                 : int  0 0 0 0 0 0 0 0 0 0 ...
 $ SourceSettings.CableOperator   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ SourceSettings.SatelliteOperator : int  0 0 0 0 0 0 0 8 0 8 ...
 $ SoundSettings.ARCSetting       : int  0 1 0 1 1 1 1 0 0 0 ...
 $ SoundSettings.AudioOutputSetting : int  2 2 2 2 2 0 1 2 2 2 ...
 $ SoundSettings.BluetoothDeviceName: chr  "----" "----" "----" "----" ...
 $ SoundSettings.BluetoothSetting : int  0 1 0 0 1 0 0 1 0 0 ...
 $ SoundSettings.ScartOutputSetting : int  1 0 1 1 1 1 1 1 1 1 ...
 $ SoundSettings.TVSpeakerSetting : int  1 1 1 1 1 0 1 1 1 1 ...
 $ SourceSettings.CAModuleName1   : chr  "Not Inserted" "Not Inserted" "HD+ Modul" "Conax CSP CIPLUS CAM" .
 $ SourceSettings.CAModuleName2   : chr  "Not Inserted" "Not Inserted" "Not Inserted" "Not Inserted" ...
```

Şekil 4.1: Gözlemlere ait nitelikler.

Farklı nitelikler arasında dağılımsal farklılıkları gidermek için min-max normalizasyon işlemi yapılmıştır. İlgisiz nitelikler çıkartıldıktan ve normalizasyon işlemi uygulandıktan sonra elde edilen veri kümesinin özet görünümü aşağıda yer verilmiştir.

```
> summary(dataMonitorN);
  Duration.Air       Duration.AirAnalog  Duration.Cable     Duration.CableAnalog Duration.Component
 Min.   :0.0000000   Min.   :0.0000000   Min.   :0.000000   Min.   :0.000000     Min.   :0.000000
 1st Qu.:0.0000000   1st Qu.:0.0000000   1st Qu.:0.000000   1st Qu.:0.000000     1st Qu.:0.000000
 Median :0.0000699   Median :0.0000000   Median :0.000000   Median :0.000000     Median :0.000000
 Mean   :0.0041793   Mean   :0.0016134   Mean   :0.018902   Mean   :0.001221     Mean   :0.002538
 3rd Qu.:0.0002591   3rd Qu.:0.0000013   3rd Qu.:0.000007   3rd Qu.:0.000000     3rd Qu.:0.000000
 Max.   :1.0000000   Max.   :1.0000000   Max.   :1.000000   Max.   :1.000000     Max.   :1.000000
 Duration.HDMI1      Duration.HDMI2      Duration.HDMI3      Duration.HDMI4       Duration.Usb
 Min.   :0.0000000   Min.   :0.0000000   Min.   :0.0000000   Min.   :0.0000000   Min.   :0.000000
 1st Qu.:0.0000009   1st Qu.:0.0000000   1st Qu.:0.0000000   1st Qu.:0.0000000   1st Qu.:0.003254
 Median :0.0007044   Median :0.0000211   Median :0.0000003   Median :0.0000000   Median :0.011161
 Mean   :0.0518881   Mean   :0.0268926   Mean   :0.0262830   Mean   :0.0187808   Mean   :0.033636
 3rd Qu.:0.0304023   3rd Qu.:0.0093225   3rd Qu.:0.0001608   3rd Qu.:0.0000307   3rd Qu.:0.030730
 Max.   :1.0000000   Max.   :1.0000000   Max.   :1.0000000   Max.   :1.0000000   Max.   :1.000000
 Duration.SVideo     Duration.Satellite  Duration.Scart
 Min.   :0.000000    Min.   :0.0000000   Min.   :0.000000
 1st Qu.:0.000000    1st Qu.:0.0000005   1st Qu.:0.000000
 Median :0.000000    Median :0.0082920   Median :0.000000
 Mean   :0.000775    Mean   :0.1112493   Mean   :0.003829
 3rd Qu.:0.000000    3rd Qu.:0.1741796   3rd Qu.:0.000000
 Max.   :1.000000    Max.   :1.0000000   Max.   :1.000000
```

Şekil 4.2: Dönüştürülmüş veri özet görünümü.

```
> str(dataMonitorN);
'data.frame':   1500 obs. of  13 variables:
 $ Duration.Air        : num  1.35e-04 1.23e-04 3.44e-04 6.82e-05 2.81e-03 ...
 $ Duration.AirAnalog  : num  0.00 4.69e-05 0.00 0.00 0.00 ...
 $ Duration.Cable      : num  0 0.04 0.105 0.477 0 ...
 $ Duration.CableAnalog: num  0 0.000242 0.000117 0 0 ...
 $ Duration.Component  : num  0 0.00141 0 0 0 ...
 $ Duration.HDMI1      : num  1.18e-02 2.26e-02 7.79e-05 3.49e-07 1.76e-02 ...
 $ Duration.HDMI2      : num  3.01e-06 5.75e-04 6.95e-05 0.00 8.25e-02 ...
 $ Duration.HDMI3      : num  2.10e-05 5.00e-05 2.53e-04 8.34e-04 8.58e-05 ...
 $ Duration.HDMI4      : num  3.63e-06 1.70e-06 0.00 0.00 3.03e-05 ...
 $ Duration.Usb        : num  0.00411 0.03969 0.01201 0.04078 0.00901 ...
 $ Duration.SVideo     : num  0.00 4.09e-06 0.00 0.00 0.00 ...
 $ Duration.Satellite  : num  0.139028 0.001353 0.000257 0 0 ...
 $ Duration.Scart      : num  0 0 0 0 0 0 0 0 0 0 ...
```

Şekil 4.3: Dönüştürülmüş veri kaynak kullanım süreleri.

R dili kullanılarak kümeleme çalışması için K-ortalamalar metodu ile çalışma yapılmıştır. Çalışma kapsamında cluster, clusterCrit, amap paketleri kullanılmıştır. Elde edilmek istenen küme sayısı sırasıyla 3'ten başlayıp 15'e kadar sırasıyla arttırılarak analiz çalışmaları yapılmıştır. K-ortalamalar metodu ile yapılan analiz çalışmaları sırasında küme yerleştirme tekrar sayısı 1000 seçilerek işlemler gerçekleştirilmiştir.

K-ortalamalar yöntemi kullanılarak istenen küme sayısı 5 seçildiğinde elde edilen küme boyutları aşağıdaki tabloda belirtildiği şekilde oluşmuştur. 1500 gözleme ait 13 nitelik için 5 farklı kümede sırasıyla 313, 915, 176, 57, 39 gözlem yer almaktadır.

Tablo 4.1: Küme sayısı 5 seçildiğinde elde edilen küme boyutları.

R Studio Komutu	> k-model<-kmeans (dataMonitorN1, 5, iter.max= 1000, nstart=1);
R Studio Komutu	> k-model;
K-ortalamalar yöntemi ile elde edilen küme boyutları	313, 915, 176, 57, 39

Tablo 4.2: K-ortalamalar analizi küme merkezi sonuçları.

Küme No	Kaynak	Means	Kaynak	Means
1	Duration.Air	2,932675E-04	Duraiton.AirAnalog	3,234209E-03
2	Duration.Air	6,353806E-03	Duraiton.AirAnalog	1,519608E-03
3	Duration.Air	4,302977E-05	Duraiton.AirAnalog	6,945620E-05
4	Duration.Air	5,178903E-03	Duraiton.AirAnalog	8,201784E-06
5	Duration.Air	1,555197E-03	Duraiton.AirAnalog	1,176514E-04
1	Duration.Cable	7,499975E-04	Duration.CableAnalog	6,100483E-06
2	Duration.Cable	9,261238E-03	Duration.CableAnalog	1,579696E-03
3	Duration.Cable	2,347574E-04	Duration.CableAnalog	3,569344E-05
4	Duration.Cable	1,797050E-03	Duration.CableAnalog	2,648745E-04
5	Duration.Cable	5,000307E-01	Duration.CableAnalog	9,318783E-03
1	Duration.Component	1,192578E-03	Duration.HDMI1	9,425681E-03
2	Duration.Component	1,078602E-03	Duration.HDMI1	7,794660E-02
3	Duration.Component	7,445706E-03	Duration.HDMI1	7,968957E-03
4	Duration.Component	9,960305E-03	Duration.HDMI1	1,845319E-02
5	Duration.Component	1,458367E-02	Duration.HDMI1	2,837053E-02
1	Duration.HDMI2	5,902892E-03	Duration.HDMI3	6,680037E-03
2	Duration.HDMI2	3,836868E-02	Duration.HDMI3	1,126507E-02
3	Duration.HDMI2	7,110508E-03	Duration.HDMI3	1,001245E-03
4	Duration.HDMI2	2,631091E-02	Duration.HDMI3	4,660896E-01
5	Duration.HDMI2	1,622356E-02	Duration.HDMI3	7,251426E-03
1	Duration.HDMI4	5,103236E-02	Duration.USB	5,313055E-02
2	Duration.HDMI4	2,540958E-02	Duration.USB	1,832701E-01
3	Duration.HDMI4	1,417482E-03	Duration.USB	7,754076E-02
4	Duration.HDMI4	4,836812E-02	Duration.USB	4,189112E-02
5	Duration.HDMI4	8,146449E-03	Duration.USB	2,617113E-02
1	Duration.Svideo	8,328886E-05	Duration.Satellite	2,062870E-01
2	Duration.Svideo	1,310866E-04	Duration.Satellite	1,378987E-02
3	Duration.Svideo	6,655196E-05	Duration.Satellite	5,017417E-01
4	Duration.Svideo	3,711798E-05	Duration.Satellite	2,191985E-02
5	Duration.Svideo	2,571000E-02	Duration.Satellite	3,395472E-03
1	Duration.Scart	2,488821E-03		
2	Duration.Scart	4,843592E-03		
3	Duration.Scart	1,440701E-04		
4	Duration.Scart	8,774054E-03		
5	Duration.Scart	1,736372E-04		

K-ortalamalar metodu ile yapılan çalışma sonucunda hangi gözlemin hangi kümede yer aldığını gösteren kümeleme vektörü de elde edilmiştir. İlk 154 gözleme ait kümeleme vektörü görünümü aşağıda belirtilmiştir.

Tablo 4.3: K-ortalamalar analizi kümeleme vektörü görünümü.

Clustering Vector																					
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**
1	2	2	5	2	2	2	2	1	1	2	2	4	2	2	2	2	2	2	2	1	2
23	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**	**37**	**38**	**39**	**40**	**41**	**42**	**43**	**44**
1	1	2	1	2	2	2	1	1	4	5	2	1	2	3	2	2	2	2	2	2	1
45	**46**	**47**	**48**	**49**	**50**	**51**	**52**	**53**	**54**	**55**	**56**	**57**	**58**	**59**	**60**	**61**	**62**	**63**	**64**	**65**	**66**
2	2	2	2	5	1	2	1	1	5	2	2	2	2	2	2	1	2	5	2	2	2
67	**68**	**69**	**70**	**71**	**72**	**73**	**74**	**75**	**76**	**77**	**78**	**79**	**80**	**81**	**82**	**83**	**84**	**85**	**86**	**87**	**88**
2	1	2	3	1	1	2	2	1	2	2	1	2	2	1	2	1	2	2	4	2	2
89	**90**	**91**	**92**	**93**	**94**	**95**	**96**	**97**	**98**	**99**	**100**	**101**	**102**	**103**	**104**	**105**	**106**	**107**	**108**	**109**	**110**
1	1	1	2	2	1	4	2	2	2	2	2	2	2	2	3	1	2	2	2	2	1
111	**112**	**113**	**114**	**115**	**116**	**117**	**118**	**119**	**120**	**121**	**122**	**123**	**124**	**125**	**126**	**127**	**128**	**129**	**130**	**131**	**132**
5	2	2	2	2	5	2	1	1	2	2	2	2	2	2	2	2	2	2	2	2	1
133	**134**	**135**	**136**	**137**	**138**	**139**	**140**	**141**	**142**	**143**	**144**	**145**	**146**	**147**	**148**	**149**	**150**	**151**	**152**	**153**	**154**
3	2	2	2	1	2	2	5	5	1	4	2	4	2	2	1	2	5	4	2	2	2

Küme içi kareler toplamı ve kümeler arası uzaklığın toplam uzaklığa oranı da K-ortalamalar çalışması sonucunda elde edilmiştir. Küme sayısı 5 seçildiğindeki uzaklık değerleri aşağıdaki şekildedir.

Tablo 4.4: Kümeleme sonucu uzaklık değerleri oranı.

Program Çıktısı	**Küme_1**	**Küme_2**	**Küme_3**	**Küme_4**	**Küme_5**
Within cluster sum of squares by cluster	7,4763670	50,734106	6,237378	4,103158	3,314766
Between_SS/ total_SS	46.7%				

Amap paketi içerisinde yer alan euclidean, maximum, manhattan, canberra, binary, pearson, abspearson, abscorrelation, correlation, spearman ve kendall uzaklık yöntemleri içerisinden sırası ile euclidean, manhattan, pearson yöntemleri seçilerek, kümelemeye ait durumlar

gözlemlenmiştir. Küme boyutlarının, küme koordinatlarının, küme vektörlerinin, küme içi uzaklık değerlerinin seçilen uzaklık yöntemine göre değiştiği gözlemlenmiştir. Çalışmanın sonraki aşamalarında k_means fonksiyonunun varsayılan uzaklık değeri ile analizler yürütülmeye devam etmiştir.

```
> k_model<-Kmeans(dataMonitorN1,5,iter.max=1000,nstart=1,method="euclidean"); # amap paketi
> k_model;
K-means clustering with 5 clusters of sizes 296, 940, 165, 60, 39

Cluster means:
  Duration.Air Duration.AirAnalog Duration.Cable Duration.CableAnalog Duration.Component Duration.HDMI1
1 2.640594e-04       3.416491e-03   0.0006955215         7.583798e-06        0.001503668    0.009894895
2 6.197758e-03       1.480262e-03   0.0090453560         1.538011e-03        0.001081018    0.075569217
3 4.568451e-05       7.408448e-05   0.0002503872         3.417522e-05        0.007329692    0.008062982
4 4.945070e-03       8.134347e-06   0.0017118112         2.516307e-04        0.009462290    0.023856637
5 1.555197e-03       1.176514e-04   0.5000306864         9.318783e-03        0.014583672    0.028370532
  Duration.HDMI2 Duration.HDMI3 Duration.HDMI4 Duration.Usb Duration.SVideo Duration.Satellite Duration.Scart
1    0.005909541    0.006499475    0.003631698   0.05547343    7.160841e-05        0.222092148   0.0026354157
2    0.037495273    0.010381050    0.024891869   0.01855876    1.333718e-04        0.016420567   0.0047156812
3    0.007303303    0.001067892    0.001511928   0.07941005    6.764373e-05        0.511012126   0.0001419428
4    0.025104900    0.454724610    0.052178789   0.04110027    3.526208e-05        0.020832136   0.0083353512
5    0.016223556    0.007251426    0.008146449   0.02617113    2.571000e-02        0.003395472   0.0001736372

Clustering vector:
   1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16   17   18   19   20   21   22
   1    2    2    5    2    2    2    2    1    1    2    2    4    2    2    2    2    2    2    2    1    2
  23   24   25   26   27   28   29   30   31   32   33   34   35   36   37   38   39   40   41   42   43   44
   1    1    2    1    2    2    2    1    2    4    5    2    1    2    3    2    2    2    2    2    2    1
  45   46   47   48   49   50   51   52   53   54   55   56   57   58   59   60   61   62   63   64   65   66
   2    2    2    2    5    1    2    1    1    5    2    2    2    2    2    2    1    2    5    2    2    2
  67   68   69   70   71   72   73   74   75   76   77   78   79   80   81   82   83   84   85   86   87   88
   2    2    2    3    1    2    2    2    1    2    2    1    2    2    1    2    1    2    2    4    2    2
  89   90   91   92   93   94   95   96   97   98   99  100  101  102  103  104  105  106  107  108  109  110
```

Şekil 4.4: Euclidean uzaklık yöntemine göre kümeleme analizi sonuçları.

```
> k_model<-Kmeans(dataMonitorN1,5,iter.max=1000,nstart=1,method="manhattan"); # amap paketi
> k_model;
K-means clustering with 5 clusters of sizes 117, 910, 380, 53, 40

Cluster means:
  Duration.Air Duration.AirAnalog Duration.Cable Duration.CableAnalog Duration.Component Duration.HDMI1
1 0.0029283447       7.497973e-06   0.0004209776         8.311408e-06       0.0009140482    0.410420626
2 0.0060392564       1.528613e-03   0.0092394901         1.587667e-03       0.0011146128    0.026550534
3 0.0001982301       2.692282e-03   0.0001496917         2.069530e-05       0.0040773698    0.009294835
4 0.0055573900       8.633022e-06   0.0019289743         2.848650e-04       0.0107120261    0.019122758
5 0.0015184993       1.147102e-04   0.4934354867         9.085814e-03       0.0142190806    0.027661269
  Duration.HDMI2 Duration.HDMI3 Duration.HDMI4 Duration.Usb Duration.SVideo Duration.Satellite Duration.Scart
1    0.034525748    0.007254025    0.012419724   0.03054278    4.460328e-04        0.016109446   0.0035838281
2    0.035366523    0.012763370    0.025137433   0.02072876    8.113817e-05        0.027380392   0.0044105141
3    0.006004678    0.002910384    0.002022096   0.06525001    8.343477e-05        0.365014443   0.0021141270
4    0.022654600    0.482497686    0.052018243   0.03978590    3.991934e-05        0.023303543   0.0094362467
5    0.015832958    0.007070140    0.007942788   0.02785959    2.506725e-02        0.003310585   0.0001692963

Clustering vector:
   1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16   17   18   19   20   21   22
   2    2    2    5    2    2    1    2    2    3    2    2    2    2    1    1    2    2    1    2    3    2
  23   24   25   26   27   28   29   30   31   32   33   34   35   36   37   38   39   40   41   42   43   44
   2    3    2    2    2    2    2    3    2    4    5    1    3    1    3    2    2    2    2    2    2    3
  45   46   47   48   49   50   51   52   53   54   55   56   57   58   59   60   61   62   63   64   65   66
   2    1    2    2    5    3    1    2    3    5    1    2    2    2    2    2    3    2    5    2    2    2
  67   68   69   70   71   72   73   74   75   76   77   78   79   80   81   82   83   84   85   86   87   88
   2    2    2    3    2    2    1    2    2    2    1    3    2    2    3    1    3    1    2    4    2    1
  89   90   91   92   93   94   95   96   97   98   99  100  101  102  103  104  105  106  107  108  109  110
```

Şekil 4.5: Manhattan uzaklık yöntemine göre kümeleme analizi sonuçları.

```
> k_model<-Kmeans(dataMonitorN1,5,iter.max=1000,nstart=1,method="pearson"); # amap paketi
> k_model;
K-means clustering with 5 clusters of sizes 430, 617, 138, 190, 125

Cluster means:
  Duration.Air Duration.AirAnalog Duration.Cable Duration.CableAnalog Duration.Component Duration.HDMI1
1 0.0039596100       2.628210e-03   0.0016441403         3.305155e-03       0.0006691806    0.157667655
2 0.0005590076       1.669830e-03   0.0003209198         1.503333e-05       0.0027512252    0.007620788
3 0.0190386692       5.695208e-05   0.0032639486         8.247056e-06       0.0030147938    0.008234998
4 0.0073738303       7.211757e-05   0.0021772648         1.242992e-04       0.0043982817    0.011936274
5 0.0015443570       1.904410e-03   0.2126770821         3.015175e-03       0.0045613900    0.015430029
  Duration.HDMI2 Duration.HDMI3 Duration.HDMI4 Duration.Usb Duration.SVideo Duration.Satellite Duration.Scart
1    0.070429135    0.004155212    0.007843400   0.01759196    1.308560e-04       0.008432425   0.0013106056
2    0.006546959    0.002820189    0.002746348   0.03275573    5.524584e-05       0.248129969   0.0013063313
3    0.011423055    0.006807174    0.001568438   0.12593150    5.243763e-06       0.051767949   0.0279544324
4    0.017176487    0.180299411    0.116348948   0.01618122    3.618950e-04       0.013462717   0.0026794029
5    0.009399384    0.005611518    0.006250741   0.01781531    8.021554e-03       0.003599478   0.0000541748

Clustering vector:
   1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16   17   18   19   20   21   22
   2    5    5    5    1    1    1    2    3    2    4    1    3    3    1    1    2    1    1    4    2    2
  23   24   25   26   27   28   29   30   31   32   33   34   35   36   37   38   39   40   41   42   43   44
   2    2    3    2    2    1    5    2    2    4    5    1    2    1    2    1    5    5    5    3    1    2
  45   46   47   48   49   50   51   52   53   54   55   56   57   58   59   60   61   62   63   64   65   66
   3    1    2    5    5    2    1    2    2    5    1    1    1    1    5    1    2    4    5    2    3    2
  67   68   69   70   71   72   73   74   75   76   77   78   79   80   81   82   83   84   85   86   87   88
   4    3    5    2    2    2    1    1    3    1    1    2    2    1    2    1    2    1    1    4    2    1
  89   90   91   92   93   94   95   96   97   98   99  100  101  102  103  104  105  106  107  108  109  110
```

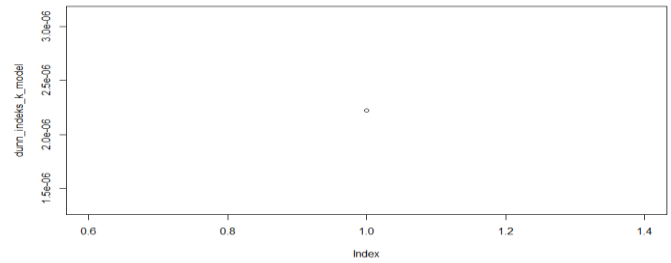
Şekil 4.6: Pearson uzaklık yöntemine göre kümeleme analizi sonuçları.

Kümeleme modeli performansının ölçülebilmesi için içsel indeksler kullanılmıştır. Kümeleme kalitesi çalışmaları da R dili ile yapılmış, bu amaçla cluster, clvalid, fpc, ClusterCrit paketleri kullanılmıştır. Kümeleme kalitesi ölçmek için çalışma sırasında Xie_Beni, Tau, Silhouette, Dunn, C, Calinski_Harabasz ve Davies_Bouldin indekslerinden yararlanılmıştır.

Küme sayısı 5 olarak seçildiğinde Dunn indeksi değeri ve grafiksel gösterimi aşağıdaki şekilde verilmiştir.

Tablo 4.5: Küme sayısı 5 olarak alındığında Dunn indeksi değeri.

R Studio Komutu	> dunn_indeks_k_model;
Dunn indeksi değeri	2,223454E-06



Şekil 4.7: Küme sayısı 5 için Dunn indeksi değeri.

Ele alınan küme sayısının 5 olduğu durumda, ilk 25 değerin Silhouette genişliği aşağıdaki tabloda gösterilmiştir.

Tablo 4.6: Küme sayısı 5 olarak alındığında Silhouette genişliği.

Rstudio Komutu	>distance <- dist(dataMonitorN,method="euclidean"); >sil_indeks_k_model <- silhouette(k_model$cluster,distance); >sil_indeks_k_model;		
No	**Küme**	**Komşu**	**Silhoutte Genişliği**
1	2	3	3,028490E-01
2	5	3	-2,227245E-02
3	5	3	2,392644E-01
4	5	3	3,754259E-01
5	1	3	7,175768E-03
6	1	3	9,649989E-03
7	1	3	3,088627E-01
8	2	3	1,402892E-01
9	3	2	1,169339E-01
10	2	3	4,303659E-01
11	4	3	-1,546115E-01
12	1	3	8,998282E-02
13	3	4	1,687157E-02
14	3	1	-1,857213E-02
15	1	3	3,106669E-01
16	1	3	3,463274E-01
17	2	3	5,122905E-02
18	1	3	1,473419E-01
19	1	3	3,453506E-01
20	4	2	-4,608140E-02
21	2	3	4,507590E-01
22	2	3	-1,385446E-01
23	2	3	2,417201E-01
24	2	3	3,808075E-01
25	3	1	2,275679E-02

Küme sayısı 5 olarak seçildiğinde Davies_Bouldin indeksinin kümelemeye yönelik oluşturduğu sonuç değerleri aşağıdaki şekildedir.

```
>  Davies_Bouldin_Indeks_k_model<-index.DB(dataMonitorN,k_model$cluster,p=2);
>  Davies_Bouldin_Indeks_k_model
$DB
[1] 2.524427

$r
[1] 2.742496 2.308753 2.742496 2.514577 2.313816

$R
         [,1]     [,2]     [,3]     [,4]     [,5]
[1,]      Inf 1.590532 2.742496 2.155962 1.973084
[2,] 1.590532      Inf 2.308753 1.625857 1.457599
[3,] 2.742496 2.308753      Inf 2.514577 2.313816
[4,] 2.155962 1.625857 2.514577      Inf 1.926101
[5,] 1.973084 1.457599 2.313816 1.926101      Inf

$d
          1         2         3         4         5
1 0.0000000 0.2904083 0.2010884 0.2587382 0.2619247
2 0.2904083 0.0000000 0.2198731 0.3161275 0.3244704
3 0.2010884 0.2198731 0.0000000 0.2400239 0.2431172
4 0.2587382 0.3161275 0.2400239 0.0000000 0.2953499
5 0.2619247 0.3244704 0.2431172 0.2953499 0.0000000

$S
[1] 0.2528777 0.2090260 0.2986065 0.3049520 0.2639219

$centers
             [,1]         [,2]         [,3]         [,4]         [,5]         [,6]         [,7]         [,8]
[1,] 0.0039596100 2.628210e-03 0.0016441403 3.305155e-03 0.0006691806 0.157667655 0.070429135 0.004155212
[2,] 0.0005590076 1.669830e-03 0.0003209198 1.503333e-05 0.0027512252 0.007620788 0.006546959 0.002820189
[3,] 0.0190386692 5.695208e-05 0.0032639486 8.247056e-06 0.0030147938 0.008234998 0.011423055 0.006807174
[4,] 0.0073738303 7.211757e-05 0.0021772648 1.242992e-04 0.0043982817 0.011936274 0.017176487 0.180299411
[5,] 0.0015443570 1.904410e-03 0.2126770821 3.015175e-03 0.0045613900 0.015430029 0.009399384 0.005611518
             [,9]      [,10]        [,11]       [,12]        [,13]
[1,] 0.007843400 0.01759196 1.308560e-04 0.008432425 0.0013106056
[2,] 0.002746348 0.03275573 5.524584e-05 0.248129969 0.0013063313
[3,] 0.001568438 0.12593150 5.243763e-06 0.051767949 0.0279544324
[4,] 0.116348948 0.01618122 3.618950e-04 0.013462717 0.0026794029
[5,] 0.006250741 0.01781531 8.021554e-03 0.003599478 0.0000541748
```

Şekil 4.8: Küme sayısı 5 için Davies_Bouldin indeksi değeri.

Küme sayısının sırasıyla 3'ten başlayarak 15'e kadar arttırıldığı durumdaki küme içi mesafenin toplam mesafeye oranı ve oluşan küme boyutlarının değerleri aşağıdaki tabloda gösterilmiştir.

Tablo 4.5: Farklı küme sayıları için küme dağılımları.

Küme Sayısı	Betweens/Total SS %	Küme Dağılımı
3	33,2	322,1139,39
4	41,7	322,1079,39,60
5	46,7	313,915,176,57,39
6	56,5	289,984,2,54,39,111
7	57,4	289,957,23,54,18,111,48
8	62,3	289,919,2,45,18,109,48,49
9	67,1	281,875,23,44,18,57,47,49,106
10	68,4	280,44,23,50,18,57,47,27,106,848
11	73,4	53,265,22,44,18,123,47,49,36,689,154
12	74,5	263,44,22,49,18,12,47,27,36,681,141,49
13	75,8	228,44,22,48,18,122,46,27,36,643,145,5,68
14	77,0	225,44,22,47,18,121,46,27,14,616,45,71,68,36
15	77,6	181,44,22,47,18,119,46,27,14,518,164,71,59,36,134

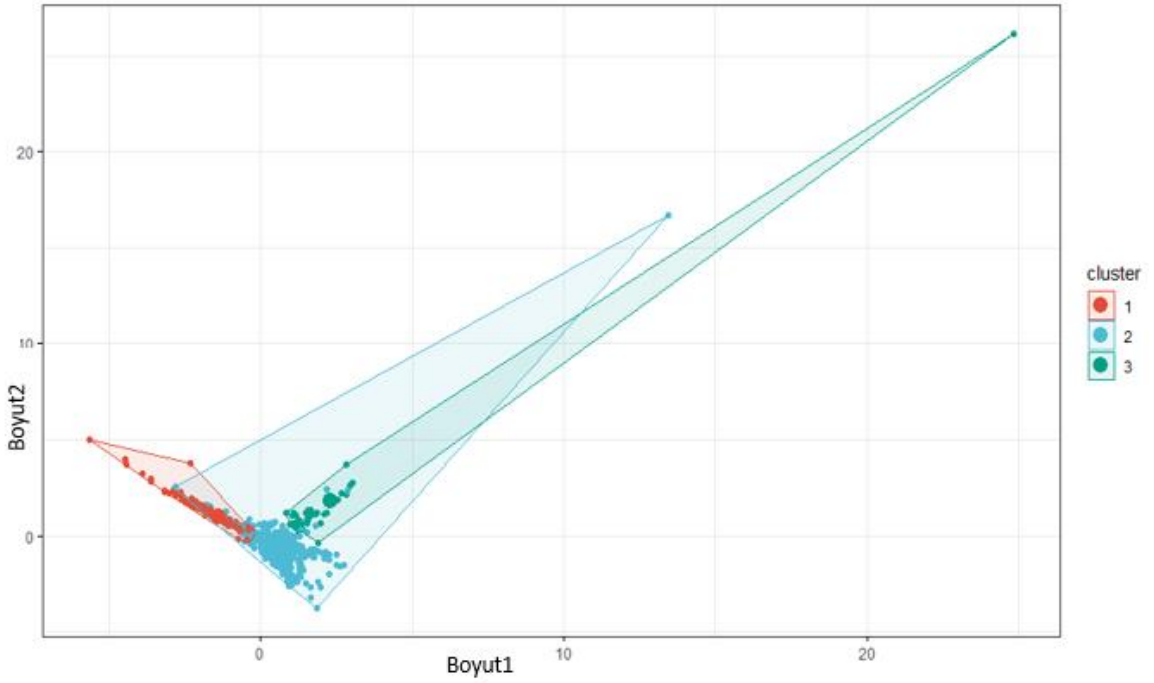
Küme sayısının sırasıyla 3'ten başlayarak 15'e kadar arttırıldığı durumdaki kümeleme performansının değerlendirilmesine yönelik içsel indekslerin aldığı değerler aşağıdaki tabloda gösterilmiştir.

Tablo 4.6: Farklı küme sayıları için iç indeks sonuçları.

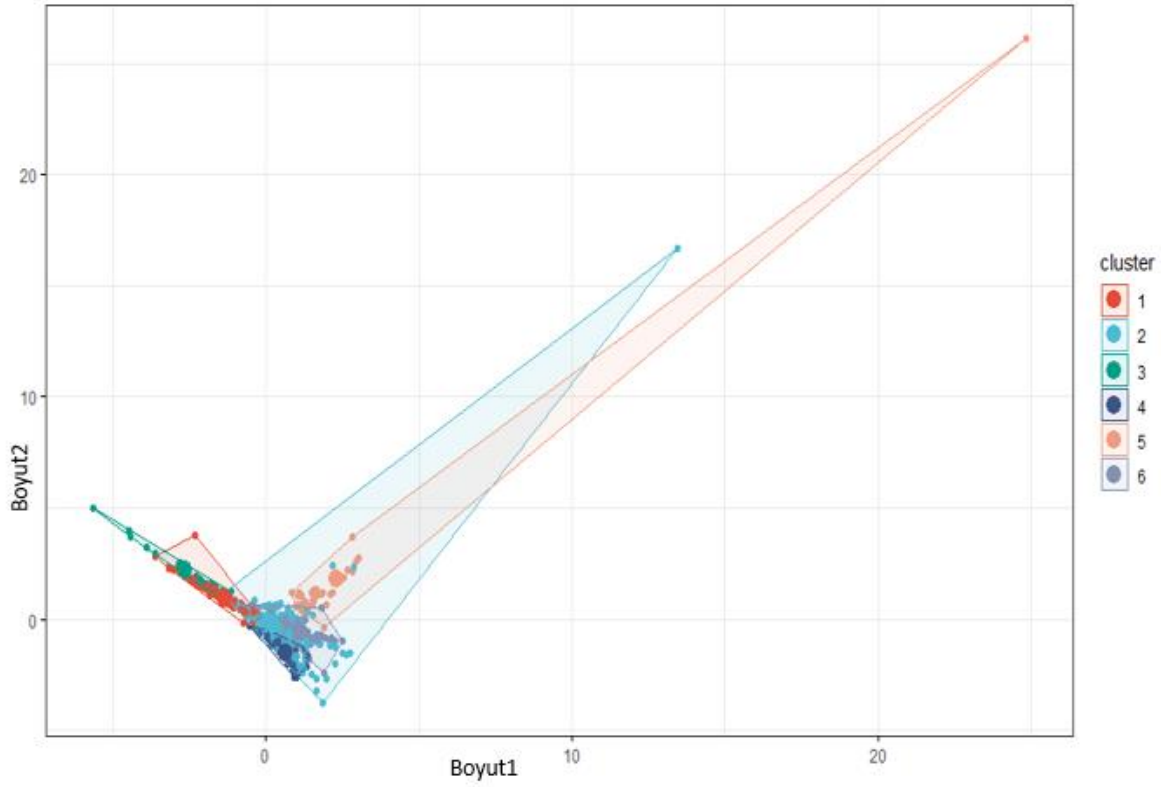
Küme Sayısı	Xie_Beni	Tau	Silhouette	Dunn	C	Calinski_ Harabasz	Davies_ Bouldin
3	449,6237	0,3724709	0,4343416	0,007978768	0,26786440	372,3968	0,9006570
4	392,5213	0,4449775	0,4098272	0,007978768	0,20529620	356,7351	0,8758131
5	558,0062	0,3722764	0,3497631	0,006531640	0,23112250	327,2638	1,0398590
6	815,6079	0,5435104	0,4031697	0,004784005	0,12257440	387,3367	0,7820150
7	797,2749	0,5303648	0,3490183	0,004784005	0,12747860	335,7036	0,8900937
8	797,2749	0,5303648	0,3490183	0,004784005	0,12747860	335,7036	0,8900937
9	1.162,765	0,5991570	0,3549432	0,003481109	0,06477902	380,5885	0,8104737
10	1.162,765	0,5991570	0,3549432	0,003481109	0,06477902	380,5885	0,8104737
11	314,0123	0,5385042	0,3454909	0,006164697	0,06135253	410,2786	0,7842734
12	467,4917	0,5427158	0,3373918	0,004941298	0,05563998	395,8320	0,8118719
13	510,8765	0,5209905	0,3252889	0,004603406	0,05581438	389,0406	0,8367109
14	487,3593	0,5107502	0,3311321	0,004603406	0,05474588	381,7060	0,8318354
15	2.353,392	0,4448661	0,3082235	0,002067569	0,06616055	366,4374	0,8956475

Optimum küme sayısını bulmak için, indeksin optimum küme sayısını verdiği değer Silhouette indeksinin maksimum olduğu değerde, Dunn indeksinin maksimum olduğu değerde, Davies_Bouldin indeksinin de minimum olduğu değerde gerçekleşmektedir. En iyi kümeleme performansı Silhouette ve Dunn indeksine göre küme sayısının 3 olduğu durumda gerçekleşirken, Davies_Bouldin indeksine göre ise küme sayısının 6 olduğu durumda elde edilmiştir.

K-ortalamalar yöntemi ile uygulanan kümeleme işleminin sonuçlarını görselleştirmek için R dili kullanılmış, ggpubr ve factoextra paketlerinden yararlanılmıştır. PCA (Principal component analysis) kullanılarak veriler daha küçük boyutlara indirilerek görselleştirme sağlanmıştır. PCA dönüşümünden sonra küme sayısının 3 ve 6 seçildiği durumda kümelemelerin iki boyutlu vektörlerinin grafik görselleri aşağıdaki şekillerde gösterilmiştir.



Şekil 4:9: Küme sayısı 3 alındığında küme görünümü.



Şekil 4.10: Küme sayısı 6 alındığında küme görünümü.

Tablo 4.7: Küme sayısı 3 için küme ağırlıkları.

	Küme1	Küme2	Küme3
Küme Sayısı	322	1139	39
Air	1,60E-04	5,41E-03	1,56E-03
AnalogAir	3,16E-03	1,23E-03	1,18E-04
Cable	1,46E-04	7,73E-03	5,00E-01
CableAnalog	2,31E-05	1,28E-03	9,32E-03
Component	4,67E-03	1,52E-03	1,46E-02
HDMI1	9,81E-03	6,46E-02	2,84E-02
HDMI2	6,42E-03	3,30E-02	1,62E-02
HDMI3	2,80E-03	3,36E-02	7,25E-03
HDMI4	2,19E-03	2,38E-02	8,15E-03
USB	6,94E-02	2,38E-02	2,62E-02
Svideo	6,33E-05	1,22E-04	2,57E-02
Satellite	3,97E-01	3,41E-02	3,40E-03
Scart	2,39E-03	4,36E-03	1,74E-04

R dili kullanılarak, K-ortalamalar yöntemi ile kümeleme çalışması gerçekleştirilmiştir. Silhouette indeksine göre en iyi küme performansının elde edildiği 3 küme sayısı için kümelerin merkezlerinin dağılımı yukarıdaki tabloda verilmiştir.

Müşteri kullanım değerlerine göre yapılan tablodaki segmentasyon çalışmasının sonuçlarına göre Küme1'de yer alan müşteriler en fazla sayısal uydu yayını izlemektedirler. Diğer kümelerle de karşılaştırıldığında en fazla sayısal uydu yayını izlemeyi tercih eden müşteri grubu bu küme içinde yer almaktadır. Uydu yayını haricinde video, ses ve resim oynatma imkânı veren USB kaynağını yaygın olarak kullandıkları görülmektedir. HDMI kaynağı üzerinden harici oynatıcı kullanım değeri diğer gruplara göre azdır. Ortalama kullanım süresine sahip müşteriler bu grup içerisindedir. Küme2 içinde yer alan müşteriler harici oynatıcı üzerinden video ve ses kullanımını sıklıkla tercih etmişlerdir. Sayısal uydu yayını izleme alışkanlığı ve USB üzerinde medya içeriklerinin oynatımı bu grup içerisinde yaygın bir kullanım türleridir. En az kullanım hacmine sahip müşteriler bu grup içerisinde yer almıştır. Küme3 içerisinde yer alan müşteriler sayısal kablolu yayın izlemeyi tercih etmişlerdir. Harici oynatıcı kullanımı ve USB kullanımı tercih ettikleri kullanım alışkanlıkları arasındadır. En fazla kullanım süresine sahip müşteri grubu bu küme içerisinde yer almıştır. Küme1'de yer alan müşterileri uydu yayını izleyicileri, küme2'de yer alan müşterileri harici oynatıcı izleyicileri, küme3'teki izleyicileri de daha çok sayısal kablo yayını izleyicileri olarak tanımlanabilirler. Her 3 küme içerisinde yer

alan kişilerde de analog yayın veya analog kaynaktan içerik izlemek öncelikli tercihleri arasında yer almamaktadır.

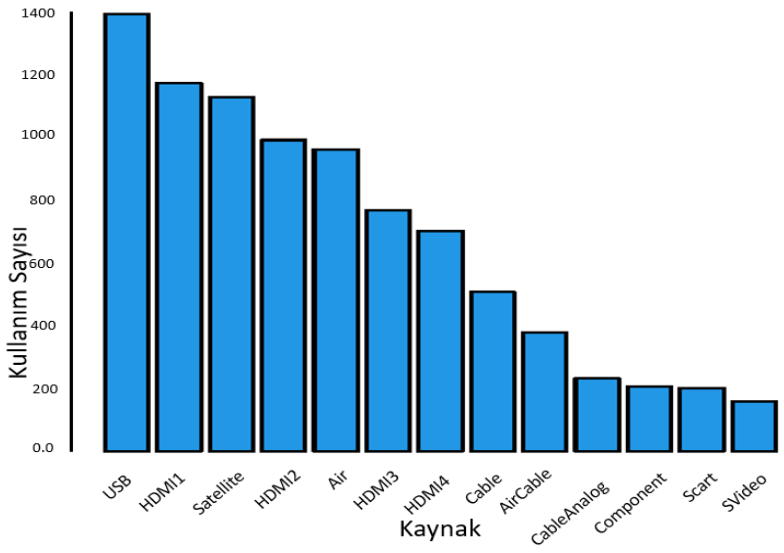
4.2 BİRLİKTELİK KURALI UYGULAMASI

1500 adet kullanıcının ürün üzerinde bulunan 13 farklı kaynağı kullanım alışkanlıkları analiz edilmiştir. Analiz çalışmaları Rstudio programı ile gerçekleştirilmiştir. Arules ve ArulesViz kütüphane dosyaları kullanılmıştır.

Tablo 4.8: Kaynak kullanım sayıları.

Kullanılan Kaynak Sayısı	1	2	3	4	5	6	7	8	9	10	11	12	13
Kullanıcı Sayısı	19	121	143	181	163	214	256	184	131	56	20	8	4

Yapılan incelemede bütün kaynakları da kullanan kullanıcıların sayısının 4 kişi olduğu görülmektedir. Sadece 1 kaynağı kullanmış olan kullanıcıların sayısının 19 olması sebebiyle kullanıcıların çoğunlukla 1'den fazla kaynak kullanım alışkanlığına sahip kişilerden oluştuğu anlaşılmaktadır. Yukarıdaki tablodaki sonuçlara bakıldığında 10 ve üzerindeki kaynak kullanım sayılarının, tek kaynak kullanım sayıları gibi düşük müşteri tercihine sahip olduğu görülmektedir.

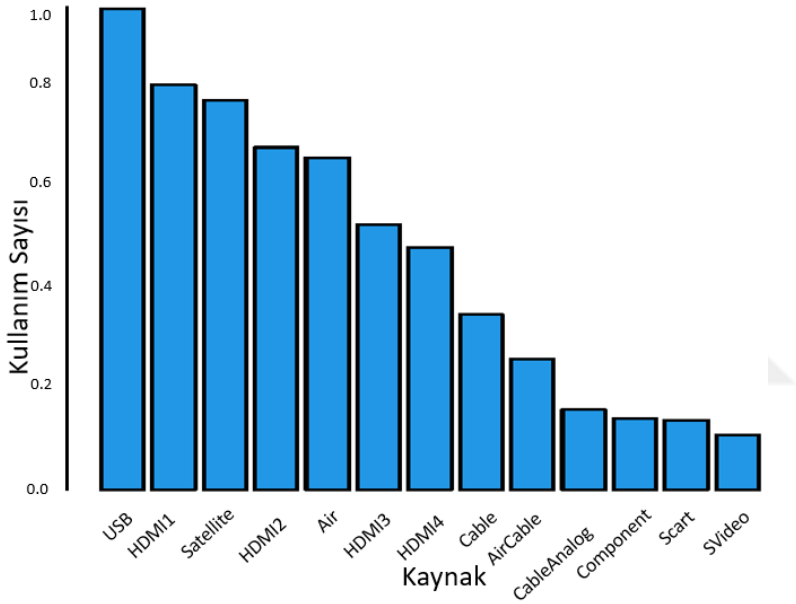


Şekil 4.11: Kaynak kullanım sayıları grafiği.

Kullanıcılar tarafından kaynakların sıklık kullanım değerleri, dağılım grafiklerinde gösterilmiştir. Kullanıcılar tarafından en sık kullanılan ilk 5 kaynak USB, HDMI1, Satellite, HDMI2 ve Air olmuştur. Kullanıcılar tarafında daha az kullanılan kaynaklar ise Analog kablo, Component, Scart ve Svideo olmuştur.

Tablo 4.9: En sık kullanılan kaynakların sayıları.

Kaynak Adı	USB	HDMI1	Satellite	HDMI2	Air
Kullanım Sayısı	1407	1184	1139	1001	971

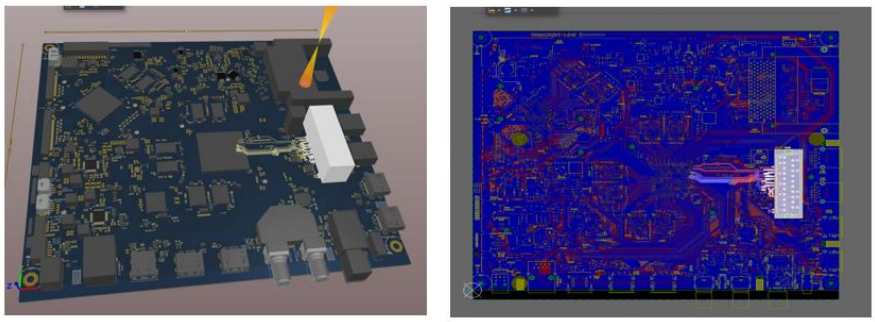


Şekil 4.12: Normalize edilen kaynak kullanım sayıları dağılımı grafiği.

En az kullanılan 3 kaynak component, Scart ve S video olmuştur. Bu kaynaklar ürün üzerinde tekil sayıda bulundukları halde kullanıcılar tarafından tercih edilmemişlerdir. Düşük oranda tercih edilen bu kaynaklar yerine dijital görüntü ve ses aktarımı sağlayan HDMI kaynağının yaygın olarak kullanıldığı görülmektedir.

Kullanım oranları en düşük olan S video, Scart, Component ve CableAnalog kaynaklarının ürün konfigürasyonundan çıkarılması ya da opsiyonel hale getirilmesi durumunda donanım ve yazılım tasarım çalışmaları ile test onay işlemlerinde iş gücü kullanımı açısından avantaj sağlanacaktır. Çalışma kapsamında değerlendirilen ürünün elektronik ana kartı üzerinde yer alan bu komponentlerin devreden iptal edilmesi durumunda, donanımsal olarak 59 adet daha az

komponent kullanılacak, ana kartın olduğu baskı devre kartının boyutu yaklaşık %4,1 küçülebilecektir.



Şekil 4.13: Elektronik baskı devre kartının 3 boyutlu ve 2 boyutlu görünümü.

Destek için belirlenen minimum değerin 0.30, güven için belirlenen minimum değerin 0.50 olduğu koşula göre birliktelik kuralı analizi yapılmıştır. 283 adet birliktelik kuralı oluşmuştur. Oluşan ikili, üçlü, dörtlü ve beşli kural sayıları aşağıdaki şekildedir.

Tablo 4.10: Birliktelik kuralı sayıları.

Kural Tipi	2'li Kural	3'lü Kural	4'lü Kural	5'li Kural
Kural Sayısı	41	97	100	45

Aşağıdaki kural tablosundaki ilk kurala baktığımızda Cable kaynağını kullanan kullanıcıların içinde Air kayağını da kullananların olma olasılığı %92'dir. Cable ve Air kaynaklarını birlikte kullanan kullanıcıların bütün kullanıcılar içindeki oranının ise %32 olduğu görülmektedir. Kaldıraç değerinin ise 1.43 olduğu görülmektedir. Kural tablosundaki ikinci kuralı incelediğimizde Cable kaynağını kullanan kullanıcıların içinde USB kaynağını da kullananların olma olasılığı %96'dır. Cable ve USB kaynaklarını birlikte kullanan kullanıcıların bütün kullanıcılar içindeki oranı ise %33 olarak görülmektedir. İkinci kural için kaldıraç değeri 1.02'dir. Kural tablosundaki üçüncü kuralı incelediğimizde HDMI4 kaynağını kullanan kullanıcıların içinde HDMI3 kaynağını da kullanan kullanıcıların olma olasılığı %88'dir. HDMI4 ve HDMI3 kaynaklarını birlikte kullananların bütün kullanıcılar içindeki oranı ise %42'dir. Üçüncü kural için kaldıraç değerinin 1.71 olduğu görülmektedir. Üçüncü kuralda yer alan her iki kaynak benzer işleri yapan özdeş kaynaklardır. Her iki kaynağı birlikte kullanan kişilerin oranının %42 olması HDMI kaynağı ses ve video bilgilerini ekrana yansıtmak için kullanıcılar arasında yaygın olarak kullanıldığını göstermektedir.

Tablo 4.11: Birliktelik kuralı değerlerinden ilk beşi.

Komut	> inspect(BasketRules[1:5]						
Kural	**Lhs**	**Rhs**	**Destek**	**Güven**	**Kapsam**	**Kaldıraç**	**Sayı**
1	{Cable}	=>{Air}	0,32	0,92	0,34	1,43	473
2	{Cable}	=>{USB}	0,33	0,96	0,34	1,02	491
3	{HDMI4}	=>{HDMI3}	0,42	0,88	0,47	1,71	624
4	{HDMI3}	=>{HDMI4}	0,42	0,81	0,52	1,71	624
5	{HDMI4}	=>{Air}	0,37	0,77	0,47	1,19	548

Destek değeri en yüksek on kural aşağıdaki şekildedir. HDMI1 ve USB konektörlerini birlikte kullanan kullanıcıların bütün kullanıcılar içindeki oranı %76. Bu durum video, ses ve resim gibi sayısal medya içeriklerinin sıklıkla kullanıldığını ve harici oynatıcı veya flash bellek gibi harici hafıza birimleri üzerinden yaygın olarak ekrana yansıtıldığını göstermektedir.

Tablo 4.12: Destek değeri en yüksek ilk on kural.

Kural	**Lhs**	**Rhs**	**Destek**	**Güven**	**Kapsam**	**Kaldıraç**	**Sayı**
1	{HDMI1}	=>{USB}	0,76	0,96	0,79	1,0	1135
2	{USB}	=>{HDMI1}	0,76	0,81	0,94	1,0	1135
3	{Satellite}	=>{USB}	0,72	0,95	0,76	1,0	1083
4	{USB}	=>{Satellite}	0,72	0,77	0,94	1,0	1083
5	{HDMI2}	=>{USB}	0,64	0,96	0,67	1,0	965
6	{USB}	=>{HDMI2}	0,64	0,69	0,94	1,0	965
7	{HDMI2}	=>{HDMI1}	0,64	0,95	0,67	1,2	955
8	{HDMI1}	=>{HDMI2}	0,64	0,81	0,79	1,2	955
9	{HDMI1,HDMI2}	=>{USB}	0,62	0,97	0,64	1,0	925
10	{HDMI2,USB}	=>{HDMI1}	0,62	0,96	0,64	1,2	925

Ürün üzerinde birbirine özdeş 4 farklı HDMI kaynak girişi bulunmaktadır. Kullanıcılar tarafından HDMI kaynaklarından sadece bir tanesi kullanılabildiği gibi ikisini, üçünü ve dördünü de kullanan kullanıcılarında olduğu görülmektedir.

Tablo 4.13: HDMI kaynak girişleri için birliktelik kuralı tablosu.

Kural	Destek	Güven	Kaldıraç	Sayı
{HDMI2} & {HDMI1}	0,64	0,95	1,21	955
{HDMI1} & {HDMI2}	0,64	0,81	1,21	955
{HDMI3} & {HDMI1}	0,49	0,95	1,20	732
{HDMI1} & {HDMI3}	0,49	0,62	1,20	732
{HDMI3} & {HDMI2}	0,47	0,91	1,36	703
{HDMI2} & {HDMI3}	0,47	0,70	1,36	703
{HDMI2,HDMI3} & {HDMI1}	0,46	0,98	1,25	692
{HDMI1,HDMI3} & {HDMI2}	0,46	0,95	1,42	692
{HDMI1,HDMI2} & {HDMI3}	0,46	0,72	1,40	692
{HDMI4} & {HDMI1}	0,45	0,95	1,20	673
{HDMI1} & {HDMI4}	0,45	0,57	1,20	673
{HDMI4} & {HDMI2}	0,43	0,91	1,36	645
{HDMI2} & {HDMI4}	0,43	0,64	1,36	645
{HDMI2,HDMI4} & {HDMI1}	0,42	0,98	1,25	635
{HDMI1,HDMI4} & {HDMI2}	0,42	0,94	1,41	635
{HDMI1,HDMI2} & {HDMI4}	0,42	0,66	1,41	635
{HDMI4} & {HDMI3}	0,42	0,88	1,71	624
{HDMI3} & {HDMI4}	0,42	0,81	1,71	624
{HDMI3,HDMI4} & {HDMI1}	0,41	0,98	1,24	609
{HDMI1,HDMI4} & {HDMI3}	0,41	0,90	1,75	609
{HDMI1,HDMI3} & {HDMI4}	0,41	0,83	1,76	609
{HDMI3,HDMI4} & {HDMI2}	0,40	0,95	1,42	593
{HDMI2,HDMI4} & {HDMI3}	0,40	0,92	1,78	593
{HDMI2,HDMI3} & {HDMI4}	0,40	0,84	1,78	593
{HDMI2,HDMI3,HDMI4} & {HDMI1}	0,39	0,99	1,26	589
{HDMI1,HDMI3,HDMI4} & {HDMI2}	0,39	0,97	1,45	589
{HDMI1,HDMI2,HDMI4} & {HDMI3}	0,39	0,93	1,80	589
{HDMI1,HDMI2,HDMI3} & {HDMI4}	0,39	0,85	1,80	589

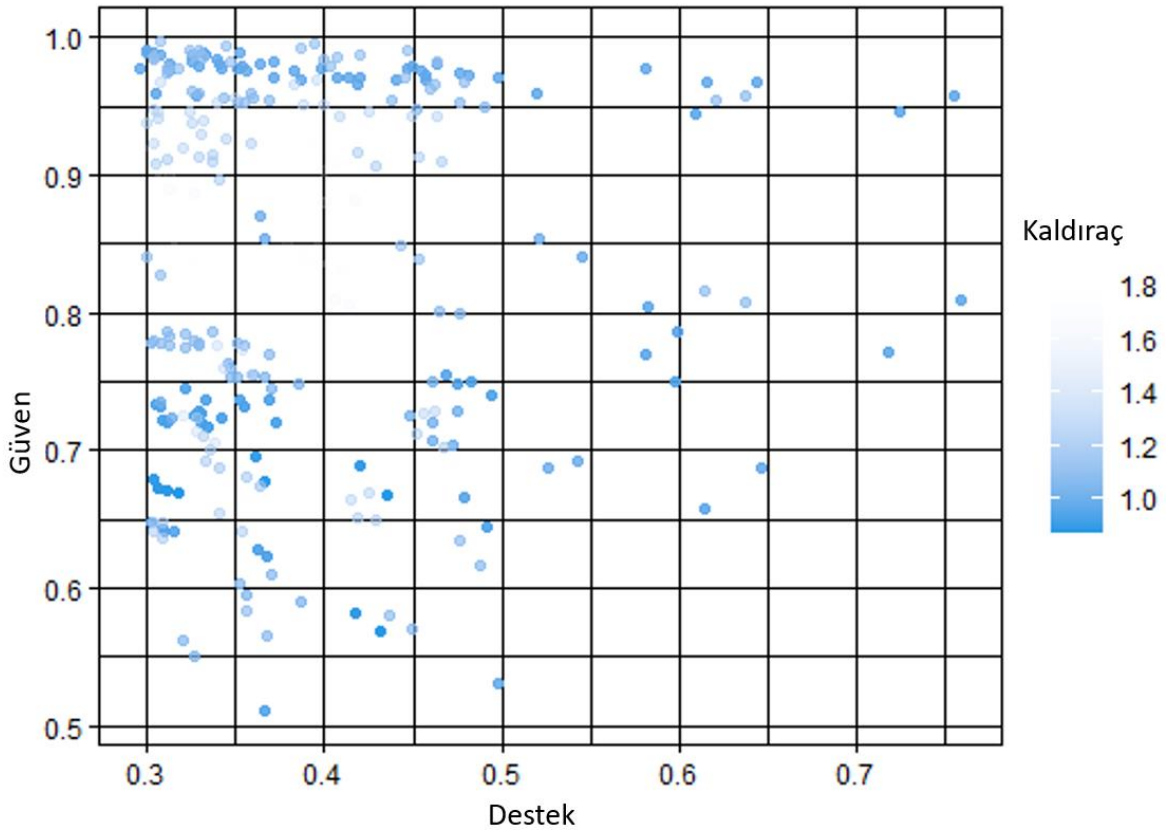
4 HDMI kaynağını da kullanan kullanıcıların toplam kullanıcılar içindeki oranı %39'dur. 3 HDMI kaynağı kullanan kullanıcıların toplam kullanıcılar içindeki en yüksek oranı ise %41'dir, bu değer HDMI1, HDMI3 ve HDMI4 kaynaklarının birlikte kullanılmış olduğu kullanım durumunda gerçekleşmektedir. 2 HDMI kaynağını kullanan kullanıcıların toplam kullanıcılar içindeki en yüksek oranı ise %64 dür, bu değer HDMI1 ve HDMI2 kaynaklarının birlikte kullanıldığı durumda gerçekleşmektedir. Tek HDMI kaynağı kullanan kullanıcıların ise en fazla HDMI1 kaynağını kullanmayı tercih ettikleri görülmektedir. Destek değerlerinin en büyük olduğu farklı kural tipindeki birliktelikler aşağıdaki tabloda gösterilmiştir. Destek değeri en büyük dörtlü birliktelik HDMI1, HDMI2, Satellite ve USB kullanım durumunda oluşurken,

destek değeri en büyük beşli birliktelik HDMI1, HDMI2, HDMI2, HDMI4 ve USB kullanımı durumunda oluşmuştur.

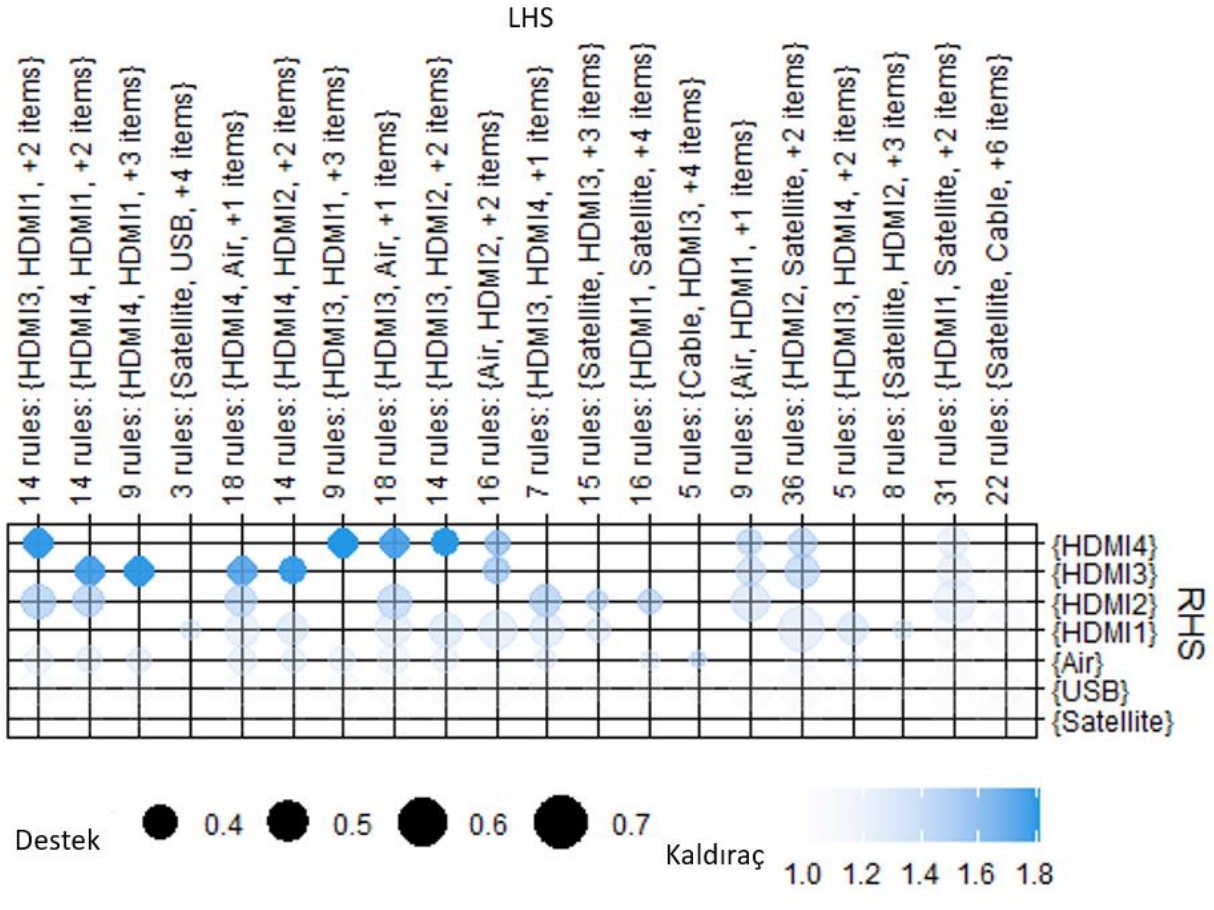
Tablo 4.14: Destek değeri en büyük farklı birliktelik durumları.

Kural Türü	Kural	Destek	Güven	Kaldıraç	Sayı
İkili Birliktelik	{HDMI1} & {USB}	0,757	0,959	1,022	1135
Üçlü Birliktelik	{HDMI1, HDMI2} & {USB}	0,617	0,969	1,033	925
Dörtlü Birliktelik	{HDMI1, HDMI2, Satellite} & {USB}	0,465	0,979	1,044	697
Beşli Birliktelik	{HDMI1, HDMI2, HDMI3, HDMI4} & {USB}	0,383	0,975	1,039	574

Destek değerlerine göre güven değerinin serpme grafiği aşağıdaki şekilde gösterilmiştir. Kaldıraç değeri, renk skalası ile gösterilmiştir. Serpme grafiğinden de anlaşıldığı gibi kuralların çoğunlukla destek değerinin 0.5'ten küçük, güven değerinin 0.55'ten büyük olduğu bölgede yoğunlaştığı görülmektedir.



Şekil 4.14: Birliktelik kuralına göre serpme grafiği.

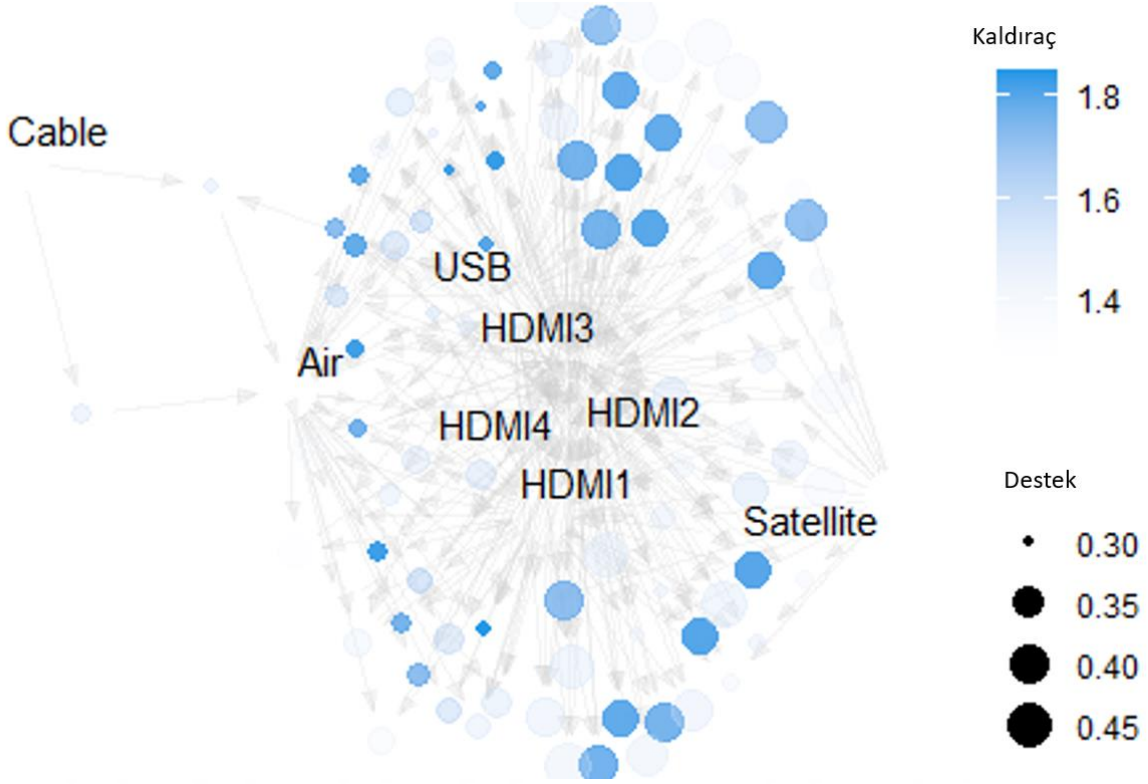


Şekil 4.15: Birliktelik kurallarına göre grup grafiği.

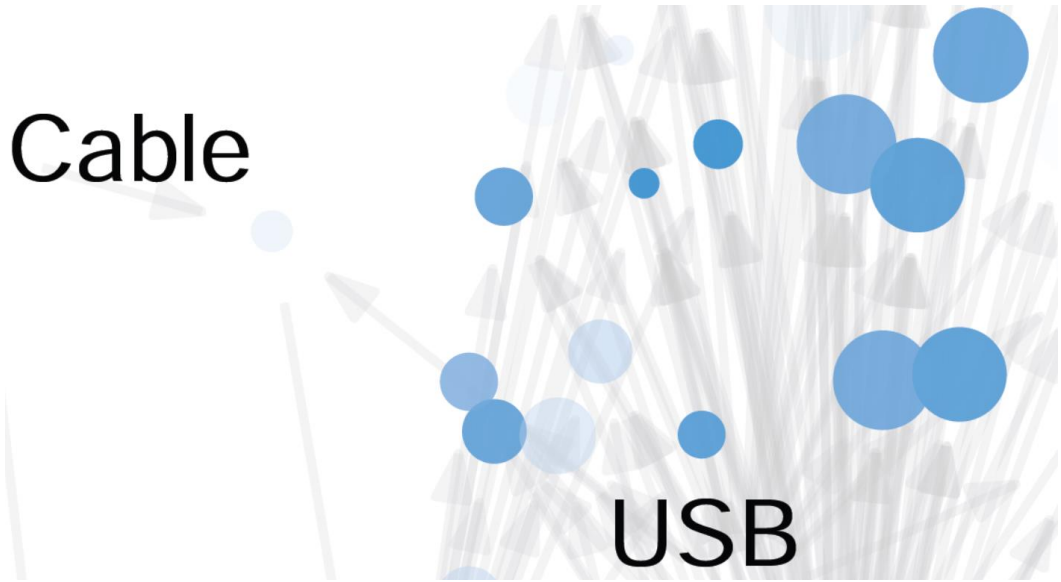
Yukarıdaki grup grafiğinde gruplandırılmış öncül değerler sütunlar üzerinde, sonuç değerleri satırlar üzerinde yer almakta, destek değeri artıkça öncül değerler ile sonuç değerlerinin kesişiminde yer alan birlikteliği simgeleyen baloncuğun boyutu büyümektedir. Grafikte kaldıraç değerinin artması renk koyulaşması ile gösterilmiştir. Minimum güven ve destek değerinin üzerindeki kurallar balonlar şeklinde gösterilmiştir.

Aşağıdaki grafik çizimi gösteriminde kaldıraç değeri artıkça renk koyuluğu destek değeri artıkça da düğüm balonunun boyutsal büyüklüğü artmaktadır. Kural kümesi çok sayıda birlikteliği içerdiğinden öncül ve sonuçlar arasındaki bağlantı yönünü gösteren oklar belirgin olarak görülememektedir. Yakınlaştırılmış grafik çiziminde oklar ve düğümler daha belirgin

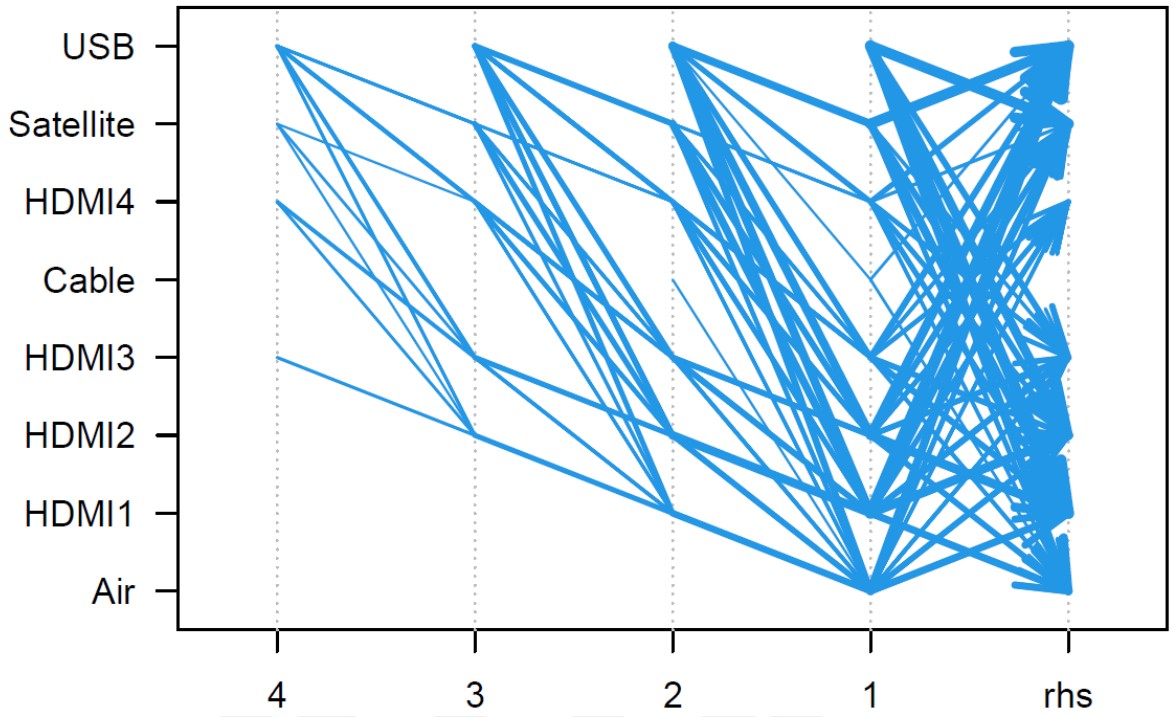
halde gösterilmiştir. Grafik çiziminden de görülebildiği gibi USB ve HDMI kaynaklarını içeren farklı birliktelik kuralları mevcuttur.



Şekil 4:16: Kaynakların kaldıraç ve destek değerlerine göre grafik çizimi.



Şekil 4.17: Yakınlaştırılmış grafik gösterimi.



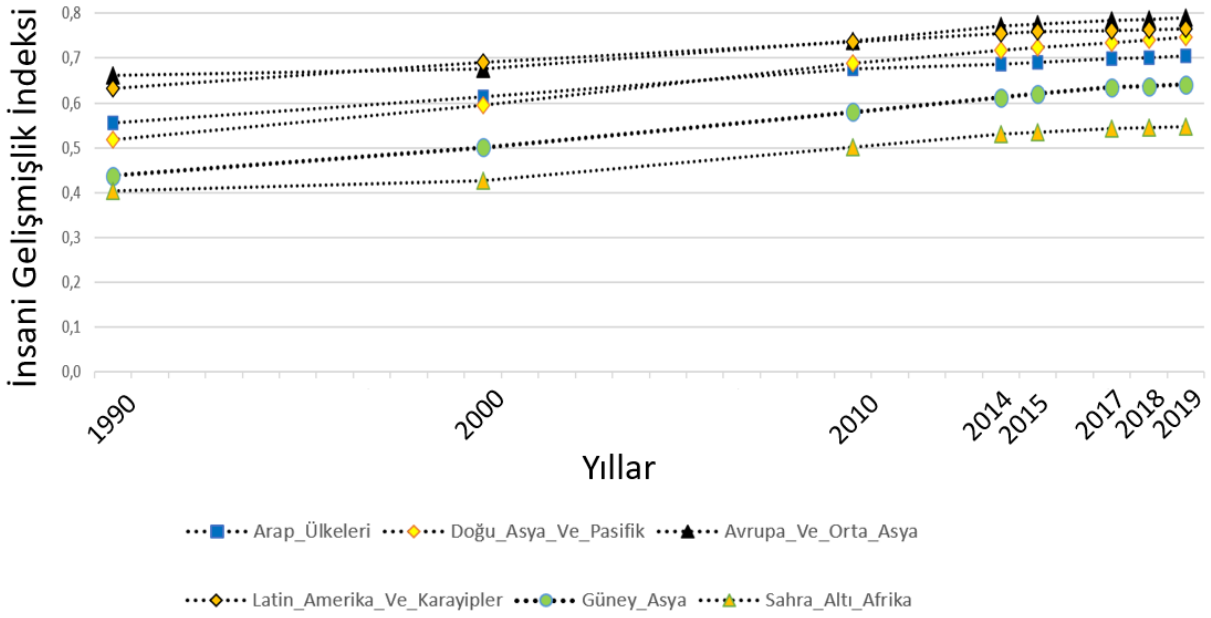
Şekil 4.18: Paralel koordinatlar grafiği.

Paralel koordinatlar grafiğinde destek değeri ve güven değeri sırasıyla ok genişliği ve ok renk yoğunluğu ile gösterilmiştir. Ok yönü sonuç değerini göstermektedir. Çizgiler, öncül ve sonuç değerlerini birbirine bağlamaktadır. USB ve HDMI kaynaklarını içeren birliktelik kuralı sayısı fazla olduğundan grafikte bu değerlerin olduğu kısımlarda çizgisel yoğunluk artmaktadır.

4.3 HDI ANALİZİ SONUÇLARI

Tablo 3.3'te verilen HDI puanları kullanılarak elde edilen bölgesel HDI puanlarının yıllara göre değişimini gösteren grafik aşağıdaki şekildedir. 2010 yılı sonrasında insani gelişmişlik düzeyi bakımından Avrupa ve Orta Asya bölgesinin en yüksek puanlara, Sahra Altı Afrika bölgesinin en düşük puanlara sahip olduğu görülmektedir. Analiz hesaplamalarında ele alınan yıllar içerisindeki verilere göre yıllar geçtikçe bütün bölgelerde insani gelişmişlik puanının arttığı görülmektedir. 2000 yılı verilerine göre en yüksek HDI puanına sahip olan Latin Amerika ve Karayip bölgesi, 2019 yılı gelişmişlik düzeyi puanına göre ikinci sıraya gerilemiştir.1990 yılında HDI puanlarına göre üçüncü ve dördüncü sırada yer alan Arap Ülkeleri bölgesi ile Doğu

Asya ve Pasifik bölgesi 2019 HDI puanlarına göre sıralamada yer değiştirmişlerdir. Doğu Asya ve Pasifik bölgesi 2019 yılı verilerine göre üçüncü sıraya yükselmiştir.



Şekil 4.19: Bölgesel insani gelişmişlik puanının zamana göre değişimi.

Avrupa kıtasında yer alan ülkelerin ortalamaları arasında anlamlı bir fark olup olmadığının analizi için Anova analizi kullanılmıştır. Hangi ülkelerin ortalamasının birbirinden farklı olduğunu anlamak için ise Games_Howell post-hoc testleri kullanılmıştır.

UNDP tarafından paylaşılan 31 adet Avrupa bölgesindeki ülkenin 1990, 2000, 2010, 2014, 2015, 2017, 2018, 2019 yıllarına ait HDI puanları verilmiştir. Parametrik analiz için 8 yıl ait veri yeterli olmayacağı için UNDP tarafından verilen ülke bazlı HDI trendi grafiğinden 1995 ve 2021 yılları arasındaki HDI değerleri elde edilmiştir. Ülke bazlı HDI puanlarının, Anova aracı ile analizi yapılmıştır. Yokluk hipotezi (Ho), $\mu 1=\mu 2$ ve alternatif hipotez (Ha), $\mu 1\neq\mu 2$ olmak üzere Anova analizi sonucu aşağıdaki şekilde gösterilmiştir. Analiz sonucunda %95 güven düzeyinde elde edilen p değeri 0.05'ten küçük olduğu için Ho hipotezi reddedilip, Ha hipotezi kabul edilir, en az bir ülkenin ortalama değeri diğerlerinden farklıdır denilebilir. Gruplar arasında ($p < 0.05$) anlamlı bir fark vardır. Gruplar arası değişkenlik, toplam değişkenliğin %66,45'ini oluşturmaktadır. Grup içi değişkenliğin nedeni yıllar içinde ülkelerin HDI puanlarının değişiminden kaynaklanmaktadır.

Gruplar arası değişim, grup içi değişimden fazla olduğu için (F değeri 1'den büyük) en az bir bölgenin HDI puan ortalaması diğer bölgelerden farklıdır denilebilir.

Factor Information

Factor	Levels	Values
Ülke	31	Almanya; Avusturya; Belçika; Birleşik Krallık; Bulgaristan; Çekya; Danimarka; Estonya; Finlandiya; Fransa; Hırvatistan; Hollanda; irlanda; ispanya; isveç; isviçre; italya; Letonya; Litvanya; Lüxemburg; Macaristan; Norveç; Polonya; Portekiz; Romanya; Rusya; Sırbistan; Slovakya; Slovenya; Türkiye; Yunanistan

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Ülke	30	2,473	0,082430	53,22	0,000
Error	806	1,248	0,001549		
Total	836	3,721			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0,0393569	66,45%	65,20%	63,82%

Şekil 4.20: F değeri.

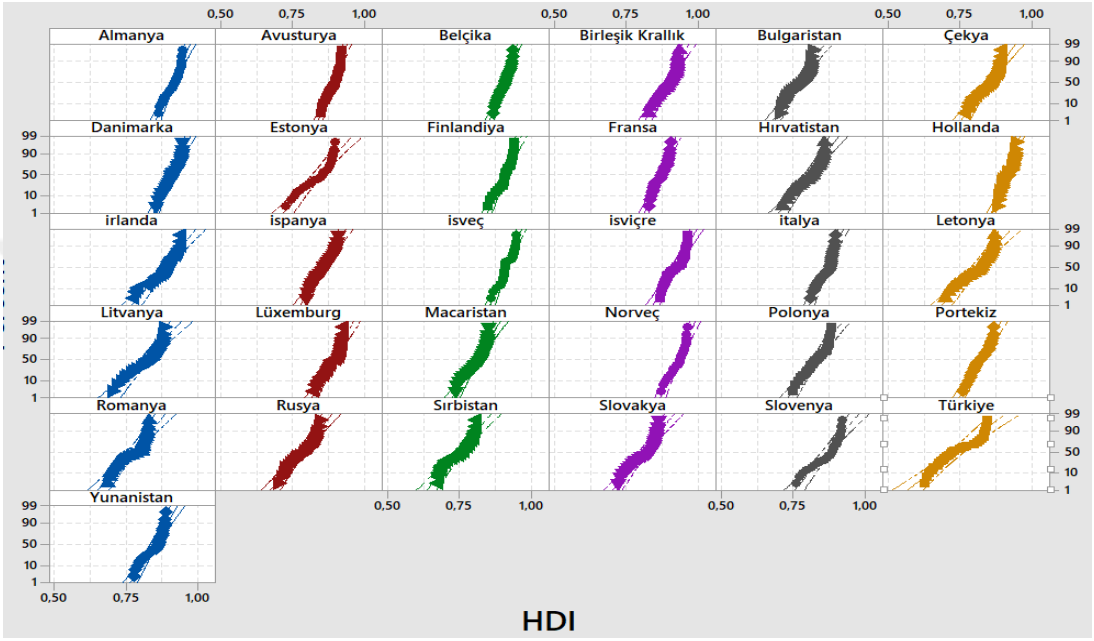
Anova analizi sonucunda 27 yıllın verilerine göre 31 ülke içinde HDI puanı ortalaması en yüksek olan ülkenin Norveç, en düşük olan ülkenin ise Türkiye olduğu görülmektedir.

Tablo 4.15: Avrupa ülkeleri HDI değerleri.

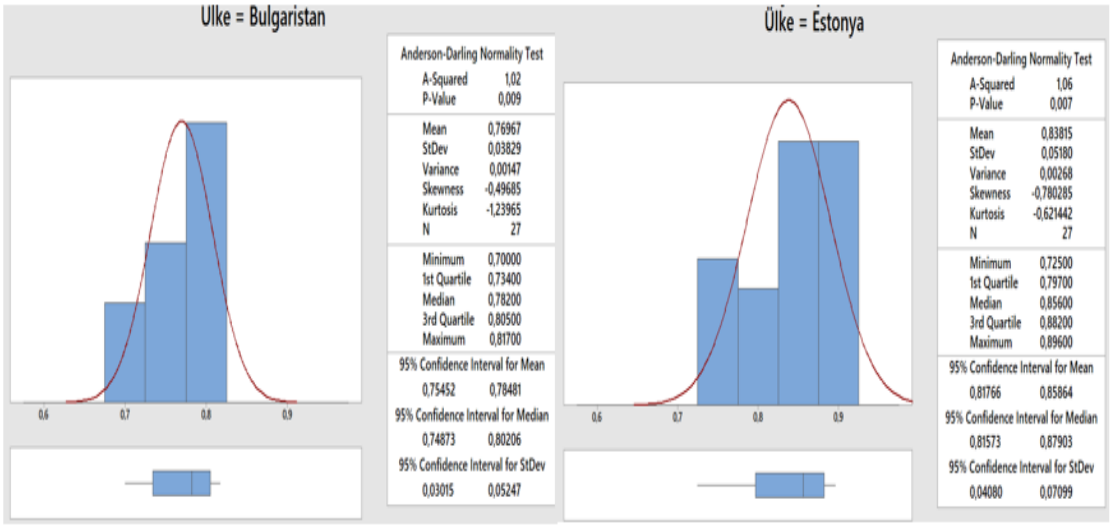
Ülke	N	Mean	StDev	95% CI
Almanya	27	0,91607	0,02592	(0,90121; 0,93094)
Avusturya	27	0,89215	0,02144	(0,87728; 0,90702)
Belçika	27	0,90604	0,02071	(0,89117; 0,92090)
Birleşik Krallık	27	0,89415	0,03261	(0,87928; 0,90902)
Bulgaristan	27	0,76967	0,03829	(0,75480; 0,78453)
Çekya	27	0,85063	0,04061	(0,83576; 0,86550)
Danimarka	27	0,91226	0,02737	(0,89739; 0,92713)
Estonya	27	0,83815	0,05180	(0,82328; 0,85302)
Finlandiya	27	0,90770	0,02572	(0,89284; 0,92257)
Fransa	27	0,87015	0,02411	(0,85528; 0,88502)
Hırvatistan	27	0,80493	0,04518	(0,79006; 0,81979)
Hollanda	27	0,91204	0,02131	(0,89717; 0,92690)
irlanda	27	0,88511	0,04799	(0,87024; 0,89998)
ispanya	27	0,85793	0,03401	(0,84306; 0,87279)
isveç	27	0,91315	0,02327	(0,89828; 0,92802)
isviçre	27	0,92126	0,03308	(0,90639; 0,93613)
italya	27	0,86667	0,02559	(0,85180; 0,88153)
Letonya	27	0,8076	0,0521	(0,7927; 0,8224)
Litvanya	27	0,8190	0,0543	(0,8042; 0,8339)
Lüxemburg	27	0,89159	0,03211	(0,87673; 0,90646)
Macaristan	27	0,80978	0,03647	(0,79491; 0,82465)
Norveç	27	0,93100	0,02682	(0,91613; 0,94587)
Polonya	27	0,82826	0,04085	(0,81339; 0,84313)
Portekiz	27	0,82093	0,03173	(0,80606; 0,83579)
Romanya	27	0,77219	0,05037	(0,75732; 0,78705)
Rusya	27	0,78048	0,04553	(0,76561; 0,79535)
Sırbistan	27	0,74856	0,04809	(0,73369; 0,76342)
Slovakya	27	0,81096	0,04564	(0,79610; 0,82583)
Slovenya	27	0,86748	0,04817	(0,85261; 0,88235)
Türkiye	27	0,7373	0,0732	(0,7225; 0,7522)
Yunanistan	27	0,84911	0,03544	(0,83424; 0,86398)

Ülke verilerinin ayrı olarak normal dağılıp, dağılmadığı kontrol edilmiştir. Normalite test sonucuna göre p değeri 0.05'ten küçük ülkeler olduğu için verileri normal dağılmayan ülkeler

mevcuttur. Bu nedenle varyansların eşitlik durumunun kontrolü için Levene testi kullanılmıştır. Yokluk hipotezi (Ho) σ1=σ2 ve alternatif hipotez (Ha) σ1≠σ2 olmak üzere Levene testi sonucuna göre p değeri 0.05 den küçük olduğu için Ha hipotezi kabul edilir. Varyans homojenliği sağlanmamıştır. Varyansların eşit olmaması durumuna göre Post-Hoc analizi yapılmıştır.



Şekil 4.21: Ülke bazlı normalite dağılımları.



Şekil 4.22: Normal dağılıma sahip olmayan ülke örnekleri.

Varyanslar eşit olmadığında yapılabilecek testlerden bir tanesi de Games_Howell testidir. Games_Howel yöntemine göre analiz edilen ülkeler gruplandırılmıştır.

Games_Howel yöntemine göre yapılan analiz sonucuna göre 31 Avrupa ülkesi HDI puanlarına göre 11 farklı gruba ayrılmıştır. Aşağıdaki tabloda ortalama değerleri aynı harf ile gösterilmeyen ülkelerin ortalama değerleri arasında anlamlı bir fark vardır. K harfi ile belirtilen grupta yer alan ülkelerin HDI puanlarının ortalaması diğer harflerle belirtilen gruplarda yer alan ülkelerin puanlarının ortalamasından daha düşüktür.

Nordig ülkeleri ve Avrupa Birliğini oluşturan birliğinin önemli kurucu ülkelerinin A ve B grupları içerisinde yer aldığı görülmektedir. Avrupa Birliği dışında kalan ülkeler arasında yer alan Rusya, Sırbistan, Türkiye ile çoğunluğu Avrupa Birliğine sonradan katılan üye ülkeler içinde bulunan Bulgaristan ve Romanya K grubunda yer almıştır.

Tablo 4.16: Games_Howell testi sonucu.

Grouping Information Using the Games-Howell Method and 95% Confidence

C1	N	Mean	Grouping									
Norveç	27	0,93100	A									
isviçre	27	0,92126	A	B								
Almanya	27	0,91607	A	B								
isveç	27	0,91315	A	B								
Danimarka	27	0,91226	A	B								
Hollanda	27	0,91204	A	B								
Finlandiya	27	0,90770	A	B	C							
Belçika	27	0,90604	A	B	C							
Birleşik Krallık	27	0,89415		B	C	D						
Avusturya	27	0,89215		B	C	D						
Lüxemburg	27	0,89159		B	C	D	E					
irlanda	27	0,88511		B	C	D	E	F				
Fransa	27	0,87015				D	E	F				
Slovenya	27	0,86748			C	D	E	F	G			
italya	27	0,86667				D	E	F	G			
ispanya	27	0,85793					E	F	G			
Çekya	27	0,85063						F	G	H		
Yunanistan	27	0,84911						F	G	H		
Estonya	27	0,83815						F	G	H		
Polonya	27	0,82826							G	H		
Portekiz	27	0,82093								H	I	
Litvanya	27	0,8190							G	H	I	J
Slovakya	27	0,81096								H	I	J
Macaristan	27	0,80978								H	I	J
Letonya	27	0,8076								H	I	J
Hırvatistan	27	0,80493								H	I	J
Rusya	27	0,78048									I	J	K
Romanya	27	0,77219										J	K
Bulgaristan	27	0,76967										J	K
Sırbistan	27	0,74856											K
Türkiye	27	0,7373											K

Games_Howell testinde farklı ülkelerin ikili kıyaslama değerleri ele alınmıştır. Kıyaslama amacıyla seçilen ülkelerden birinin Almanya olduğu durumdaki Games_Howell testi sonucunda da ülkeler arasında gelişmişlik düzeyi arasında fark olduğu görülebilmektedir. Almanya ile gelişmişlik düzeyi puanı ortalaması açısından fark olan ülkeler %95 güven düzeyinde $p< 0.05$ koşulunu sağlamaktadır. Games_Howell yöntemi ile kıyaslama sonuçlarına göre Bulgaristan, Çekya, Estonya, Fransa, İspanya, İtalya, Hırvatistan, Letonya, Litvanya, Macaristan, Polonya, Portekiz, Romanya, Rusya, Sırbistan, Slovakya, Slovenya, Türkiye ve Yunanistan'ın HDI puanları ortalaması ile Almanya'nın HDI puanları ortalaması arasında fark vardır denilebilir.

Games-Howell Simultaneous Tests for Differences of Means

Difference of Levels	Difference of Means	SE of Difference	95% CI	T-Value	Adjusted P-Value
Avusturya - Almanya	-0,02393	0,00647	(-0,04979; 0,00194)	-3,70	0,106
Belçika - Almanya	-0,01004	0,00639	(-0,03555; 0,01547)	-1,57	0,999
Birleşik Kra - Almanya	-0,02193	0,00802	(-0,05396; 0,01011)	-2,73	0,617
Bulgaristan - Almanya	-0,14641	0,00890	(-0,18215; -0,11067)	-16,45	0,000
Çekya - Almanya	-0,06544	0,00927	(-0,10275; -0,02814)	-7,06	0,000
Danimarka - Almanya	-0,00381	0,00725	(-0,03275; 0,02512)	-0,53	1,000
Estonya - Almanya	-0,0779	0,0111	(-0,1233; -0,0326)	-6,99	0,000
Finlandiya - Almanya	-0,00837	0,00703	(-0,03635; 0,01961)	-1,19	1,000
Fransa - Almanya	-0,04593	0,00681	(-0,07310; -0,01875)	-6,74	0,000
Hırvatistan - Almanya	-0,1111	0,0100	(-0,1516; -0,0707)	-11,09	0,000
Hollanda - Almanya	-0,00404	0,00646	(-0,02984; 0,02176)	-0,63	1,000
irlanda - Almanya	-0,0310	0,0105	(-0,0735; 0,0116)	-2,95	0,470
ispanya - Almanya	-0,05815	0,00823	(-0,09109; -0,02521)	-7,07	0,000
isveç - Almanya	-0,00293	0,00670	(-0,02966; 0,02381)	-0,44	1,000
isviçre - Almanya	0,00519	0,00809	(-0,02713; 0,03750)	0,64	1,000
italya - Almanya	-0,04941	0,00701	(-0,07731; -0,02150)	-7,05	0,000
Letonya - Almanya	-0,1085	0,0112	(-0,1541; -0,0630)	-9,69	0,000
Litvanya - Almanya	-0,0970	0,0116	(-0,1442; -0,0499)	-8,38	0,000
Lüxemburg - Almanya	-0,02448	0,00794	(-0,05621; 0,00725)	-3,08	0,377
Macaristan - Almanya	-0,10630	0,00861	(-0,14082; -0,07177)	-12,34	0,000
Norveç - Almanya	0,01493	0,00718	(-0,01365; 0,04350)	2,08	0,955
Polonya - Almanya	-0,08781	0,00931	(-0,12528; -0,05035)	-9,43	0,000
Portekiz - Almanya	-0,09515	0,00789	(-0,12665; -0,06365)	-12,07	0,000
Romanya - Almanya	-0,1439	0,0109	(-0,1881; -0,0996)	-13,20	0,000
Rusya - Almanya	-0,1356	0,0101	(-0,1764; -0,0948)	-13,45	0,000
Sırbistan - Almanya	-0,1675	0,0105	(-0,2101; -0,1249)	-15,93	0,000
Slovakya - Almanya	-0,1051	0,0101	(-0,1460; -0,0643)	-10,41	0,000
Slovenya - Almanya	-0,0486	0,0105	(-0,0912; -0,0059)	-4,62	0,011
Türkiye - Almanya	-0,1787	0,0149	(-0,2402; -0,1172)	-11,96	0,000
Yunanistan - Almanya	-0,06696	0,00845	(-0,10078; -0,03315)	-7,92	0,000

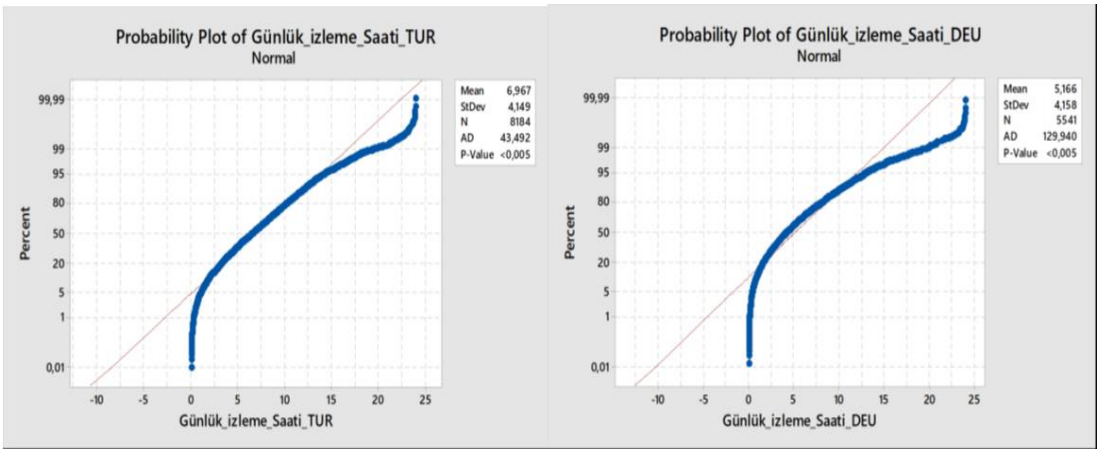
Şekil 4.23: Games_Howell testi kıyaslama sonucu.

UNDP tarafından yayınlanan 2020 yılı verilerine göre Avrupa ülkelerinin HDI puanları aşağıdaki tabloda belirtildiği şekildedir.

Tablo 4.17: Avrupa ülkeleri 2020 yılı HDI verileri.

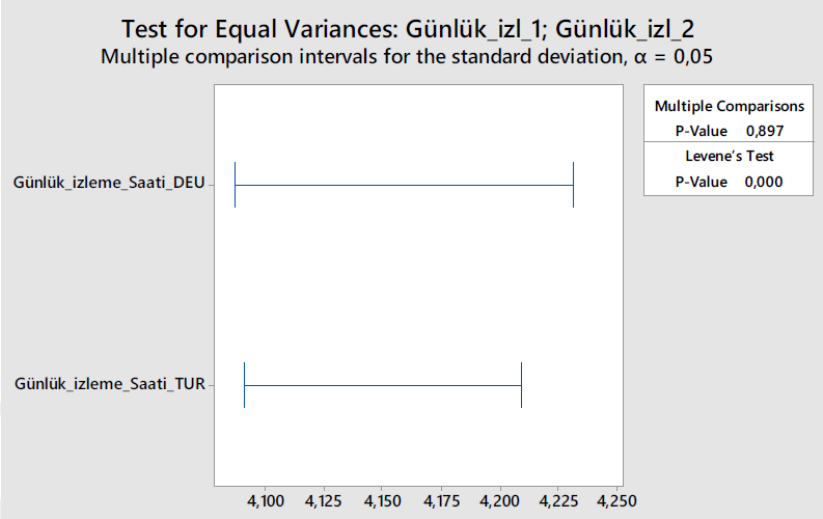
Ülke	HDI Puanı	Ülke	HDI Puanı
AUT	0,922	ITA	0,892
BEL	0,931	LTU	0,882
BGR	0,816	LUX	0,916
CHE	0,955	LVA	0,866
CZE	0,9	NLD	0,944
DEU	0,947	NOR	0,957
DNK	0,94	POL	0,88
ESP	0,904	PRT	0,864
EST	0,892	ROU	0,828
FIN	0,938	RUS	0,824
FRA	0,901	SCG	0,806
GBR	0,932	SVK	0,86
GRC	0,888	SVN	0,917
HRV	0,851	SWE	0,945
HUN	0,854	TUR	0,82
IRL	0,955		

Ülkeler arasında kullanıcıların televizyon izleme süreleri arasında fark olup olmadığı incelenmek amacıyla kullanıcı sayılarının daha fazla olduğu Türkiye ve Almanya izlenme süreleri açısından kıyaslanmıştır. İki ülkenin normal dağılım grafikleri kontrol edilmiştir. %95 güven düzeyinde elde edilen p değeri 0.05'ten küçük olduğu için Ho hipotezi reddedilip, Ha hipotezi kabul edilir. İki ülkenin de dağılımları normal değildir.



Şekil 4.24: İzlenme süresi olasılık dağılımı grafikleri.

İki ülkenin varyans değerleri kontrol edilmiştir. Levene testi sonucunda varyanslarının eşit olmadığı görülmüştür.



Şekil 4.25: Varyans değerlerinin kıyaslanması.

İki ülkenin ortalamalarını kıyaslamak için Two Sample T testi yapılmıştır. Yokluk hipotezi (Ho) μ1=μ2 ve alternatif hipotez (Ha) μ1≠μ2 olmak üzere yapılan test sonucuna göre iki ülke izlenme oranları arasında fark yoktur denilemez.

Method

μ_1: mean of Günlük_izleme_Saati_DEU
μ_2: mean of Günlük_izleme_Saati_TUR
Difference: $\mu_1 - \mu_2$

Equal variances are not assumed for this analysis.

Estimation for Difference

Difference	95% CI for Difference
-1,8013	(-1,9429; -1,6596)

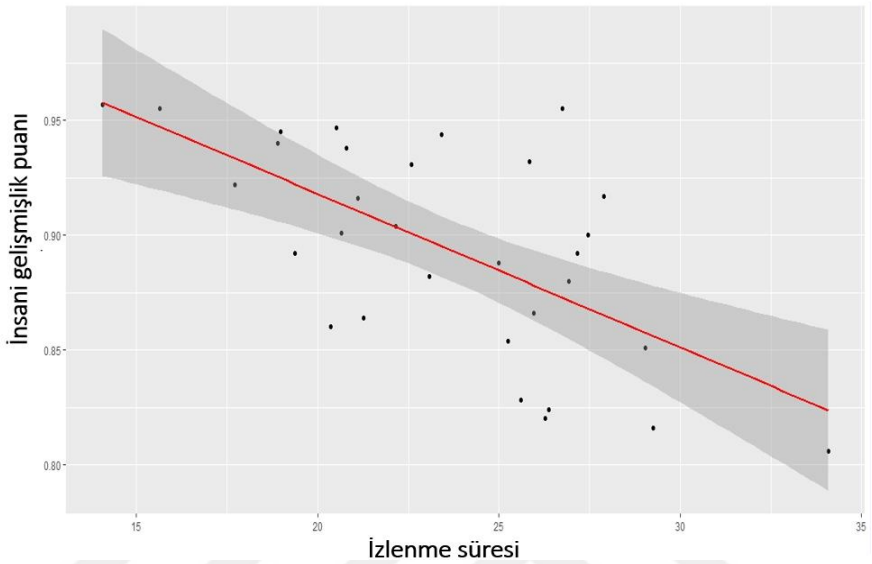
Test

Null hypothesis $H_0: \mu_1 - \mu_2 = 0$
Alternative hypothesis $H_1: \mu_1 - \mu_2 \neq 0$

T-Value	DF	P-Value
-24,92	11874	0,000

Şekil 4.26: Two sample t testi analizi.

Ülke gelişmişlik düzeyi ile televizyon izleme oranları arasındaki ilişki incelenmiştir. Ülke gelişmişlik düzeyi için 2020 yılına ait HDI puanları kullanılmıştır. Yapılan korelasyon analizine göre ülke gelişmişlik oranları ile televizyon izleme süreleri arasında zayıf negatif bir ilişki vardır. Gelişmişlik düzeyi arttıkça televizyon izleme süreleri azalmaktadır.



Şekil 4.27: İnsani gelişmişlik puanı ile izlenme süresi serpilme diyagramı.

Analiz sonucunda elde edilen korelasyon katsayı değeri -0,64'tür.

```
data22 <- read.xlsx("D:/Santez/Doktora/Tez/TV_Analizi/Deneme/Denemelerhaziran20
21/Life_HDI_Analysis.xlsx",sheetName ="2019");
  str(data22);

  library(ggplot2);

  compLifeTimeHDI <- ggplot(data=data22, aes(x=TvViewRatio, y= HumanDevelopment
Index)) + geom_point();
  compLifeTimeHDI;

  korL<- data22 %>%
  summarize(N = n(), r = cor(TvViewRatio,HumanDevelopmentIndex));

  korL;
```

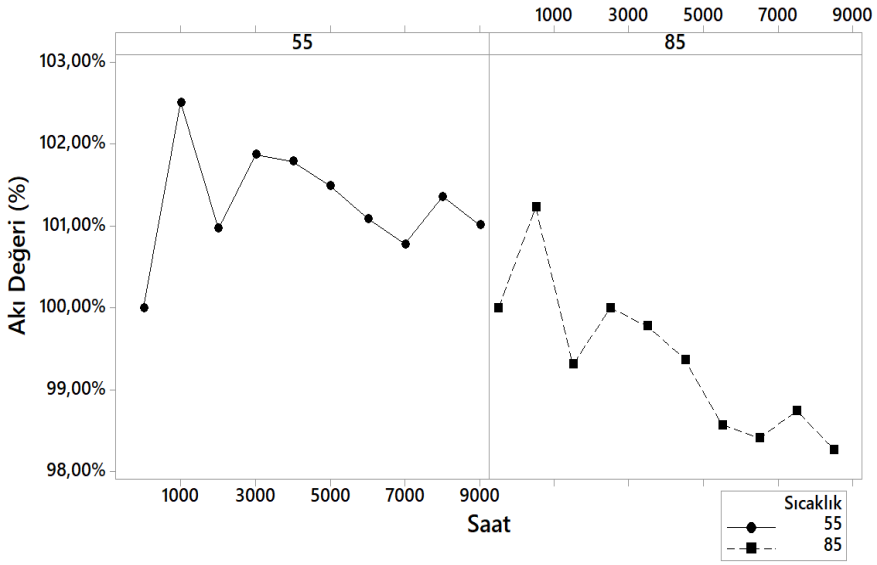
```
TV izleme oranları ile  Ülke bazlı gelişmişlik indeksi

'data.frame':  31 obs. of  9 variables:
 $ Country             : Factor w/ 31 levels "AUT","BEL","BGR",..: 1 2 3 4 5 6
7 8 9 10 ...
 $ Number              : num  7704 156 139 808 151 ...
 $ TimeDiff            : num  5914 5829 5410 7695 4781 ...
 $ LifeTime            : num  1048 1317 1583 1204 1313 ...
 $ TimeDiff.Total      : num  45561456 909324 751990 6217560 721931 ...
 $ LifeTime.Total      : num  8073792 205452 220037 972832 198263 ...
 $ Ratio..             : num  17.7 22.6 29.3 15.6 27.5 ...
 $ HumanDevelopmentIndex: num  0.922 0.931 0.816 0.955 0.9 0.947 0.94 0.904 0.8
92 0.938 ...
 $ TvViewRatio         : num  17.7 22.6 29.3 15.6 27.5 ...
>   library(ggplot2);
>   compLifeTimeHDI <- ggplot(data=data22, aes(x=TvViewRatio, y= HumanDevelopme
ntIndex)) + geom_point();
>   compLifeTimeHDI;
>   korL<- data22 %>%
+   summarize(N = n(), r = cor(TvViewRatio,HumanDevelopmentIndex));
>   korL;
   N          r
1 31 -0.6376266
```

Şekil 4.28: HDI ile izlenme süresi arasındaki korelasyon ilişkisi.

4.4 LED BAR GÜVENİLİRLİK ANALİZİ

Çalışma kapsamında LM-80 test sonuçları kullanılmıştır. LM-80 test sonuçları 1000'er saat arayla 9000 saate kadar olan akı değişimi verilerini içermektedir. LM-80 testleri LED ışık kaynağı numunelerinin 1A lik sabit akım kaynağı ile sürüldüğü durumda yapılmıştır.



Şekil 4.29: LM-80 testi zamana bağlı akı değerleri değişimi.

Farklı akım değerlerinin test sonuçları mevcut değildir. 28 adet numune 55 °C'de, 28 adet numune 85°C'de test edilmişlerdir. Test edilen 28'er numunenin 55 °C'deki ve 85 °C'deki akılarının yüzdesel değer ortalamalarının zamana bağlı değişimi yukarıdaki grafikte verilmiştir.

Yukarıdaki şekilden de görüldüğü gibi kılıf sıcaklığının artması akı değerini olumsuz yönde etkilemekte, malzemenin yararlı ömrünü kısaltmaktadır. LM-80 testleri sırasında optik, elektriksel ve mekaniksel bir arıza ile karşılaşılmadığı belirtildiği için bozulma sayıları her ikisi sıcaklıkta da 0 olarak alınmıştır.

Tablo 4.18: LM-80 test parametreleri.

Parametre	55°C LM-80 Testi	85°C LM-80 Testi
Sıcaklık(°C)	55	85
Sıcaklık(K)	328	358
LM-80 test akımı (mA)	1000	1000
Numune sayısı	28	28
Bozulma sayısı	0	0

IES tarafından yayınlanan TM-21 metodu kullanılarak Arrhenius sıcaklık yaşam modeli parametleri hesaplanmıştır.

Tablo 4.19: LM-80 test koşullarında analiz özeti.

Sıcaklık(°C)	Akım (mA)	Numune		Test süresi (saat)	Öngörü veri aralığı		α	B	L70
		Sayısı	Bozulma sayısı		Başlangıç saati	Bitiş saati			
55	1000	28	0	9000	4000	9000	2.00e-6	1.0284	>54.000
85	1000	28	0	9000	4000	9000	2.77e-6	1.0065	>54.000

550 mA akım seviyesi ile sürülen LED barın Tj jonksiyon sıcaklık değeri 82,5 °C derece olarak hesaplanmıştır. Elde edilen Arrhenius parametreleri kullanılarak L70 kriterine göre 82,5 °C derece için interpolasyon yapılarak akı düşümü zamanı değerlendirilmiştir.

Tablo 4.20: İnterpolasyon sıcaklık akım değerleri.

Maksimum değeri (mA)	1500
Sıcaklık (interpolasyon)	82,5 °C
Nominal Vf (V)	3,3
in-situ akımı (mA)	550

Yapılan hesaplamaya göre sadece sıcaklık interpolasyonuna göre L70 değeri için öngörülen süre 54000 saatten fazladır. Ortalama televizyon izleme süresi 8 saat olması durumda yaklaşık 18,5 yılda lümen seviyesi %87 seviyelerine düşecektir. 2019 yılı Ocak ayı ilk gün verisine göre ortalama izleme süresi 6,2 saattir, bu durumda yaklaşık 23,9 yılda lümen seviyesi %87 seviyelerine düşecektir. Akım değerinin 550mA düşmesi durumunda ise bu süre daha da uzayacaktır.

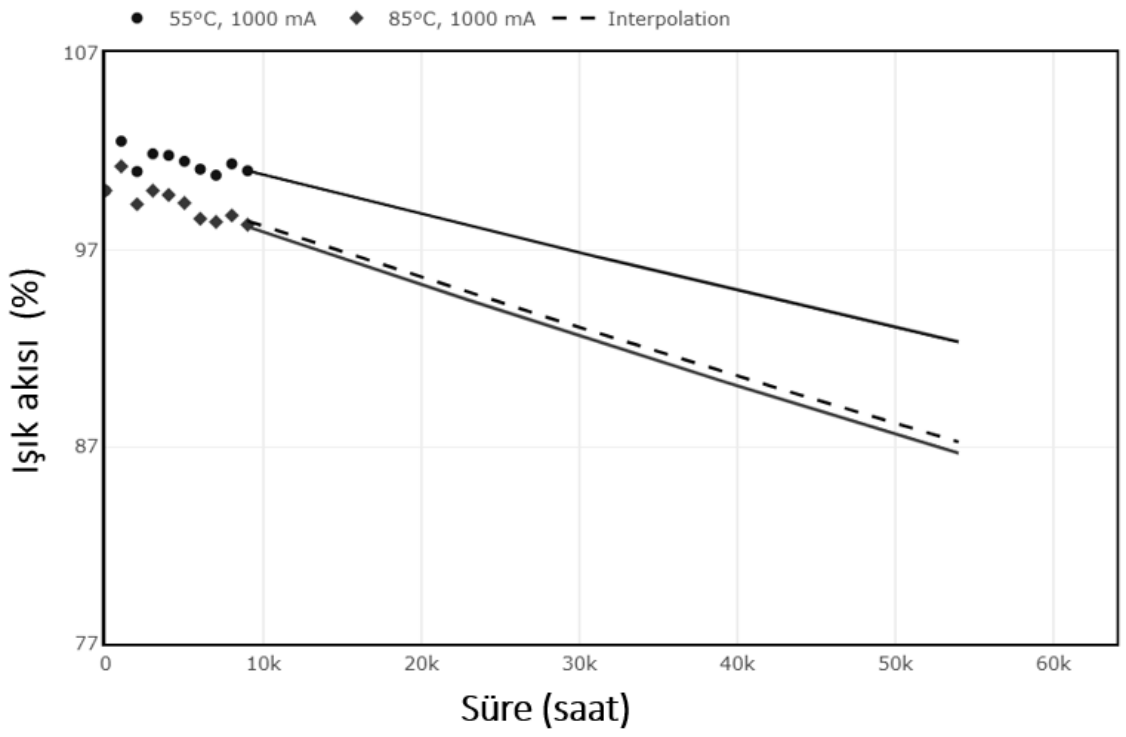
Tablo 4.21: İnterpolasyon L70 öngörüsü.

Sıcaklık interpolasyon 82,5 °C 1000 mA	
A	2.688e-6
B	1.087
L70 (saat)	>54.000

LED ışık kaynağı için 55 °C ve 85 °C'de elde edilen verilerden 4000 saatten 9000 saate kadar olanları kullanılarak, 82,5 °C'deki kılıf sıcaklığına göre yapılan sıcaklık interpolasyon sonucuna yönelik ömür tahmini grafiği şekil 4.27'de gösterilmiştir.

Arka aydınlatmasına yönelik kullanılan LED'lerin elektriksel güvenilirliğine yönelik de çalışma yapılmıştır. 550 mA akım seviyesi ile sürülen LED barın Tj jonksiyon sıcaklığı için normal koşullar için hesaplanan 82,5 °C değeri çalışmada kullanılmıştır.

LED bar üzerinde kullanılan LED'lerin bozulmasına yönelik hızlandırılmış yaşam testi analizi yapılmıştır. Bu amaçla LED ler üzerinden 700 mA, 750 mA ve 800 mA geçirilmesi durumunda elde edilen 3 farklı jonksiyon sıcaklığı ve açma kapama çevrim değerleri kullanılmıştır. Akım değerlerine bağlı olarak elde edilen jonksiyon sıcaklık seviyeleri artmaktadır. Analiz sırasında kullanılan jonksiyon sıcaklıkları değerleri 432 K, 446 K ve 461 K 'dir.



Şekil 4.30: 82,5 °C'de ömür tahmini.

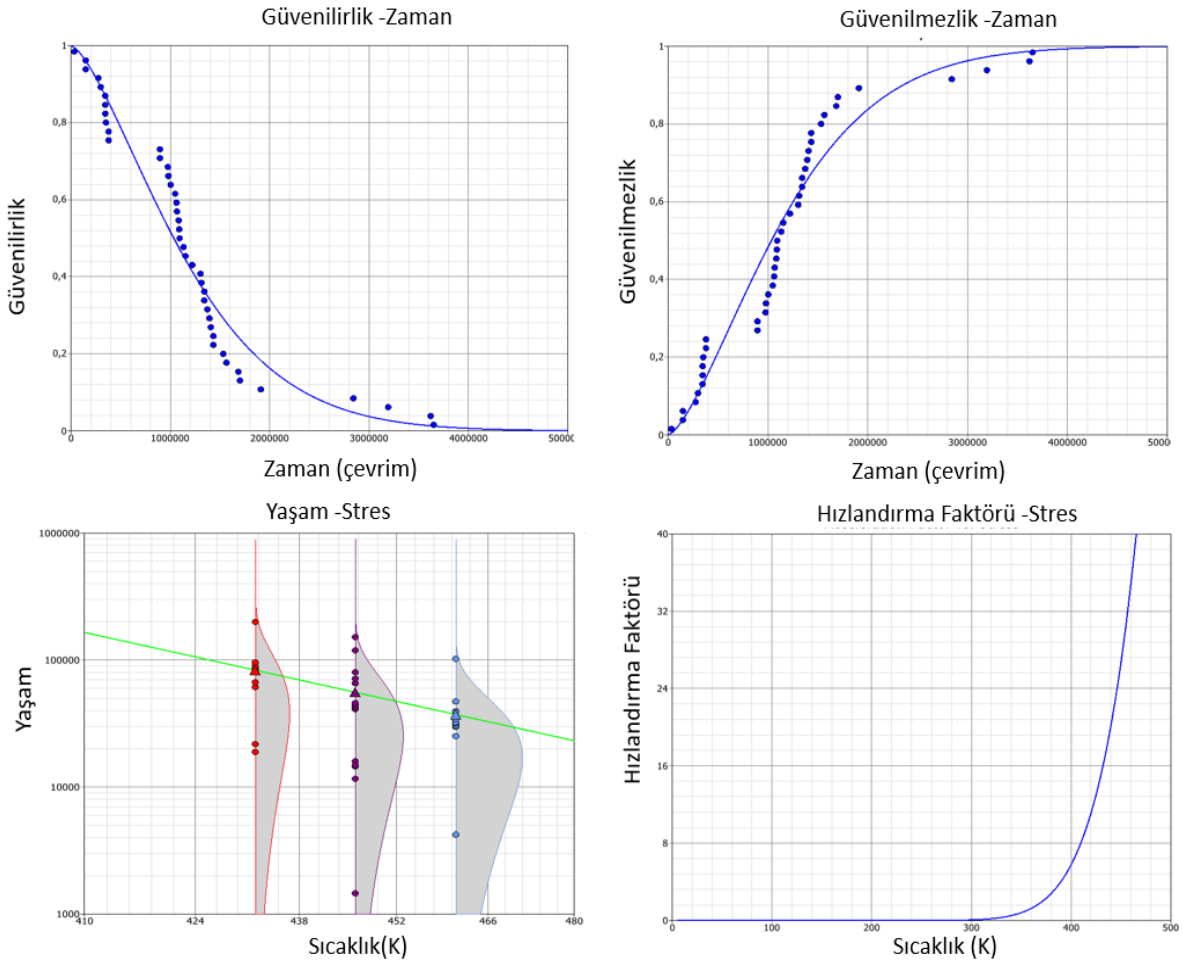
Analizler için Reliasoft programının ALTA modulü kullanılmıştır. ALTA modülü kullanılarak yapılan analiz sonucunda verilerin Weibull dağılımına uygun olduğu görülmüştür. Yapılan analiz sonucunda jonksiyon sıcaklığının artışına yönelik stres faktörü değeri arttıkça ürün ömrünün azaldığı görülmüştür.

Yılda 2263 saat çalışma değerlerine göre saatte 1 çevrim olması durumda arka aydınlatmada kullanılan LED'ler için elde edilen güvenilirlik değerleri aşağıdaki tabloda belirtildiği şekildedir.

Tablo 4.22: Güvenilirlik değerleri.

	2263 saat 1.Yıl	4526 saat 2.Yıl	6789 saat 3.Yıl	9052 saat 4.Yıl	11315 saat 5.Yıl
Güvenilirlik	0,9999	0,9997	0,9995	0,9993	0,9990

ALTA modulü kullanılarak yapılan analiz sonucuna göre elde edilen güvenilirlik-zaman, güvenilmezlik-zaman, yaşam- stres ve hızlandırma faktörü-stres grafikleri şekil 4.28'de verilmiştir. Şekilden de görüldüğü gibi zaman bağlı olarak güvenilirlik değeri azalırken, güvenilmezlik değeri artmaktadır. Stres faktörü olarak sıcaklığın kullanıldığı durumda, sıcaklık artmasına bağlı olarak hızlandırma faktörü değeri artmaktadır.



Şekil 4.31: Hızlandırılmış yaşam testi analizi sonuçları.

Büyükkör ve Özler (2016) tarafından gerçekleştirilen hızlandırılmış yaşam testi analizine yönelik stres faktörü olarak sıcaklığın seçildiği çalışmada veriler Arrhenius-Weibull modeli kullanılarak ele alınmış ve parametreleri tahmin edilmiştir. Testler 40°C ve 50°C'de yapılmıştır. Büyükkör ve Özler (2016) tarafından elde edilen çıktılar kullanılarak, sahadaki ortalama televizyon izleme süresi olan 6,2 saat olması koşulundaki güvenilirlik ve hata oranı değerleri hesaplanmıştır. Ortalama günlük 6,2 saatlik izleme süresi durumunda 1 yıl sonraki güvenilirlik ve hata oranı değerleri aşağıdaki şekilde olacaktır.

$$R(2263) = exp\left[-\left(\frac{x}{\eta}\right)^{\beta}\right] = exp\left[-\left(\frac{2263}{109170,85}\right)^{1,25}\right] = 0,9921 \tag{4.1}$$

1 yıl sonrasındaki (2263 saat) güvenilmezlik değeri *F(x);*

$$F(2263) = 1 - R(2263) = 0,0078 \tag{4.2}$$

1 yıl sonraki hata oranı h(x) yılda 2263 saat kullanım koşulana göre aşağıdaki şeklide hesaplanmaktadır.

$$h(2263) = \frac{f(x)}{R(x)} = \frac{\beta}{\eta}\left(\frac{x}{\eta}\right)^{\beta-1} = \frac{1,25}{109170,85}\left(\frac{2263}{109170,85}\right)^{0,25} \tag{4.3}$$

$$h(2263) = \frac{f(x)}{R(x)} = \frac{\beta}{\eta}\left(\frac{x}{\eta}\right)^{\beta-1} = \frac{1,25}{109170,85}\left(\frac{2263}{109170,85}\right)^{0,25} = 0,00000434 \tag{4.4}$$

1 yıl sonraki kümülatif arıza değeri aşağıdaki şekildedir.

$$H(2263) = \left(\frac{x}{\eta}\right)^{\beta} = \left(\frac{2263}{109170,85}\right)^{1,25} = 0,0079 \tag{4.5}$$

Günde ortalama 6,2 saat televizyon izlenme değerlerine göre hesaplanan 5 yıl içindeki güvenilirlik değeri, hata oranı ve kümülatif hata değerleri aşağıdaki tabloda verilmiştir. Yıllar içerisinde arıza oranı ve kümülatif arıza değerleri artmaktadır. İzleme süresi kriterinin 6,2 saat olmasının güvenilirlik ve arıza oranı değerlerini etkilediği görülmektedir, bu süre ilerleyen çalışmalarda kriter olarak kullanılabilir.

Tablo 4.23: Yıllara göre güvenilirlik, hata oranı ve kümülatif hata değerleri.

	2263 saat 1.Yıl	**4526 saat 2.Yıl**	**6789 saat 3.Yıl**	**9052 saat 4.Yıl**	**11315 saat 5.Yıl**
Güvenilirlik	0,9922	0,9815	0,9694	0,9565	0,9429
Güvenilmezlik	0,0078	0,0185	0,0306	0,0435	0,0571
Arıza Oranı	4,34458E-06	5,16661E-06	5,71779E-06	6,14416E-06	6,49666E-06
Kümülatif Arıza	0,0079	0,0187	0,0311	0,0445	0,0588

5. TARTIŞMA VE SONUÇ

Tez çalışması kapsamında veri madenciliği süreç adımları izlenmiş, veri madenciliği yöntemlerinden yararlanılarak internet erişimli televizyon kullanıcılarına ait veriler arasında ilişkili kurulması, verilerin kümelenmesi ve öngörüde bulunulmasına yönelik çalışmalar gerçekleştirilmiştir. Veri madenciliği yöntemleri kullanılmadan önce eksik ve gürültü verilerin temizlenmesine yönelik önişleme çalışmaları yapılarak veri kalitesi arttırılmıştır. Dönüştürme, normalizasyon ve boyut indirgeme faaliyetleri gerçekleştirilmiştir. 17 adet nitelik kullanılarak veri madenciliği analiz faaliyetleri gerçekleştirilmiştir.

Müşteri segmentasyonuna yönelik yapılan çalışmada hiyerarşik olmayan kümeleme yöntemlerinden olan k-ortalamalar yöntemi kullanılmıştır. Küme sayısının sırasıyla 3 adetten başlayarak 15 adede kadar arttırıldığı kümeleme işlemleri gerçekleştirilmiştir. Kümeleme performansını ölçmek ve optimum küme sayısını belirlemek amacıyla iç indekslerden yararlanılmıştır. Bu amaçla Xie_Beni, Tau, Silhouette, Dunn, C, Calinski_Harabasz ve Davies_Bouldin iç indeksleri kullanılmıştır. Silhouette indeksine göre optimum küme performansı, kullanıcıların üç kümeye ayrıldığı durumda elde edilmiştir. Küme sayısının üç olduğu durumda müşteri grupları bazında hangi kaynakların sıklıkla tercih edildiği ele alınmıştır. Kullanıcıların üç gruba ayrılması durumunda, en yüksek kullanıcı sayısına sahip gruptaki kişi sayısının toplam kişi sayısına oranının %75,9 olduğu görülmüştür. Bu küme içinde yer alan kullanıcıların kaynak olarak çoğunlukla harici oynatıcıları kullanmayı ve uydu yayınlarını izlemeyi tercih ettikleri saptanmıştır. Diğer küme grupları da incelendiğinde kullanıcıların sayısal yayınları alabilecekleri kaynakları sıklıkla kullandıkları görülmüştür. Kullanıcılar tarafından analog kaynakların kullanımının tercih edilmediği saptanmıştır. İleride ürün konfigürasyonlarına göre analog kaynakların ürünlerde kullanımlarının kaldırılması değerlendirilebilir. Giriş, orta ve gelişmiş segment ürün gruplarına ait yol haritaları oluşturulurken benzeri müşteri segmentasyonu çalışmasından yararlanılabilir.

Kaynakların birlikte kullanımına yönelik ilişkiyi incelemek için sıralı algoritmalar arasında yer alan apriori algoritmasından yararlanılmıştır. Birliktelik kuralı pazarlama faaliyetlerinde, market çalışmalarında sıklıkla müşteri alışkanlıklarının belirlenmesinde ve rafların yerleşiminde kullanılmaktadır. Birlikte sıkça talep edilen ürünler yakın raflarda olacak şekilde yerleşim düzeni oluşturulmaktadır. Çalışma kapsamında ise kaynakların kullanım durumları

dikkate alınarak Apriori algoritması vasıtasıyla birliktelik kuralı analizi yapılmıştır. Analiz çalışması sırasında destek için belirlenen minimum değerin 0.30, güven için belirlenen minimum değerin 0.50 olduğu kriter koşuluna göre analiz çalışması yürütülmüştür. İkili, üçlü, dörtlü ve beşli grupları içeren 283 adet birliktelik kuralı oluşturulmuştur. İlgili kurallara ait destek, güven , kapsam ve kaldıraç değerleri hesaplanmıştır. Destek değerinin en büyük olduğu farklı sayılardaki birliktelik durumları ortaya çıkarılmıştır. Hangi kaynakların birlikte kullandığı saptanmıştır. Kullanıcıların sadece tek bir kaynak yerine birden fazla kaynağı kullanmayı tercih ettiği belirlenmiştir. Kullanıcıların 10 ve üzerindeki sayıda kaynağı kullanmayı tercih etmedikleri görülmüştür. Birbirine özdeş olan 4 HDMI kaynağının da birlikte kullanım oranının %39 olduğu saptanmıştır.

Benzer şekilde birliktelik kuralı ileride gerçekleştirilecek ürün donanım ve arayüz yazılım tasarım çalışmalarında da kullanılabilir.

- Birlikte sık olarak kullanılan özelliklere ait donanımların baskı devre kartına veya cihaz yüzeyine yerleşimi birliktelik kuralı dikkate alınarak yapılabilir. Örneğin sıklıkla birlikte kullanılan kaynaklar birbirine çok yakın olduklarında harici takılan kabloların ve farklı donanım birimlerinin kaynaklara takıp çıkarılması zorluklara yol açabilir. Bu nedenle tasarım safhasında kaynaklar arasındaki mesafe belirlenirken birlikte kullanım durumu dikkate alınabilir,
- Destek değeri düşük olan özelliklere ait donanımlar ana kart üzerinde değil de bir modül üzerinde olacak şekilde tasarım yapılabilir. Bu modül bütün ürün segmentlerinde kullanılmayıp sadece opsiyon olarak belirlenen ürün grubunda müşterilere sunulabilir. Bu durum ürün segmentasyonuna destek olacak ve kullanılmadığı ürünlerde maliyet avantajı elde edilmesini sağlayacaktır,
- Birliktelik kuralı kullanıcı arayüzü tasarımında da kullanılabilir. Sık kullanılan menülerin tasarımı destek ve kaldıraç değerleri dikkate alınarak yapılabilir, kullanım kolaylığı sağlanabilir. Kullanıcıların ardışıl olarak tercih ettiği menüler saptanarak, kullanıcıya bir sonraki adımda kullanım olasılığı yüksek olan menüler öneri olarak sunulabilir,

- Birliktelik kuralına göre yapılan analiz sonucunda yeni bir kumanda veya klavye tasarımı yapılması değerlendirilebilir. Birliktelik kuralında hangi tuşların kullanıldığı yanında ne sıklıkla kullanıldığının da dikkate alındığı bir yaklaşım geliştirilmesi gerekebilir.

Analiz çalışmaları sonucunda dikkate alınan zaman aralığında televizyonun günlük izleme süresinin ortalama 6,2 saat olduğu tespit edilmiştir. Elde edilen ortalama izlenme süresinden yararlanılarak, ürün güvenilirliğine yönelik yapılan yaşam süresi-stres modellemesi sırasında sıcaklık etken olarak alındığından Arrhenius yöntemi kullanılmıştır. İnterpolasyon yöntemi vasıtası ile 82,5 °C' deki jonksiyon sıcaklığındaki LED ışık akısına yönelik öngörü çalışması gerçekleştirilmiştir. Hesaplamalar sonucunda LED ışık akısının %70 seviyesine düşeceği zamanın 54.000 saatten fazla olacağı öngörülmüştür. Yaşam-stres analizinde sıcaklığın hızlandırıcı stres etkisi gözlemlenmiştir. Güvenirlik analizleri ve malzeme ömür hesaplamalarında tespit edilen ortalama izleme süresi kriter olarak alınabilir. Gerek tasarım çalışması sırasında gerekse de sonrasında gerçekleştirilen alternatif malzeme seçimi sırasında bu değer göz önünde bulundurularak malzeme seçimi yapılabilir. Kullanım süresindeki değişim tasarım ve test kriterlerini etkilediği gibi malzeme temin ve saha sonrası hizmet süreçlerini de etkileyecektir.

Çalışma kapsamında Avrupa kıtasında yer alan ülkelerdeki ortalama televizyon izleme süreleri ile gelişmişlik düzeyi arasındaki ilişki incelenmiştir. Birleşmiş Milletler Gelişim Programı tarafından yayınlanan HDI verilerinden yararlanılarak yapılan korelasyon analizi sonucunda gelişmişlik düzeyi arttıkça izleme oranlarının düştüğü görülmüştür. Az gelişmiş ülkelerdeki ortalama televizyon izleme oranının yüksek olmasından dolayı, bu dezavantaj televizyonun bir eğitim aracı arayüzüne dönüştürülmesi fikri ile avantajlı bir hale evirilebilir. Televizyonun izlenme oranının yüksek olduğu ülkelerde sağlık, bilgilendirme ve anons platformu olarak kullanılması da değerlendirilebilir.

KAYNAKLAR

Acuna, E. and Rodriguez, C., 2004, The treatment of missing values and its effect on classifier accuracy, *Classification, clustering, and data mining applications*, 639-647.

Agrawal, R., Imieliński, T. and Swami, A., 1993, Mining association rules between sets of items in large databases, *ACM sigmod record*, 22 (2), 207-216.

Aggarwal, C.C., Procopiuc, C., Wolf, J.L., Yu, P.S. and Park, J.S., 1999, Fast algorithms for projected clustering, *In proceedings of the ACM sigmod conference*, Philadelphia, PA., 61-72,

Agrawal, R. and Srikant, R., 1994, Fast algorithms for mining association rules, *Proceeding of the 20th int'l conference on very large databases*, Santiago, Chile.

Agrawal, R., Gehrke, J., Gunopulos, D. and Raghavan, P., 1998, Automatic subspace clustering of high dimensional data for data mining applications, *In proceedings of the ACM sigmod conference*, Seattle, WA., 94-105.

Aggarwal, C.C. and Yu, P.S. 2000, Finding generalized projected clusters in high dimensional spaces, *Sigmod record*, 29 (2), 70-92.

Aggarwal, C., 2015, *Data mining : the textbook*, Springer, New York.

Aggarwal, C.C., 2017, *Outlier analysis*, Springer, New York, USA.

Akpınar, H., 2017, *Veri madenciliği veri analizi*, 2nd ed., Papatya Yayıncılık Eğitimi, İstanbul.

Altunkaynak, B., 2019, *Veri madenciliği yöntemleri ve R uygulamaları*, 2nd ed., Seçkin Yayıncılık, Ankara.

ASSIST, 2005, ASSIST recommends: LED life for general lighting, Troy, NY: lighting research center,1, 1-7.

Bhatt, V., Dhakar, M. and Chaurasia, B., 2016, Filtered clustering based on local outlier factor in data mining, *International journal of database theory and application*, 9 (5), 275-282.

Allison, P.D., 2009, *Missing Data*, Quantitative methods in psychology, In: Millsao, R. and Maydeu-Olivares, A. (Ed), London, Sage Publication, 72-89.

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J.M. and Perona, I., 2013, An extensive comparative study of cluster validity indices, *Pattern recognition.*, 46, 243-256.

Ault, S. V., 2018, On speech recognition algorithms, *International journal of machine learning and computing*, 8 (6), 518–523.

Azevedo, A. and Santos, M., 2008, KDD, SEMMA and CRISP-DM: A parallel overview, *IADIS european conference data mining*, 182-185.

Baher, S. and Lobo L.M.R., J., 2012, A comparative study of association rule algorithms for course recommender system in e-learning, *International journal of computer applications*, 39, 48-52.

Bastian, M., Heymann, S. and Jacomy, M., 2009, Gephi: An open source software for exploring and manipulating networks, 361-362.

Borgelt, C., 2005, An implementation of the FP-growth algorithm, *Proceedings of the 1st international workshop on open source data mining frequent pattern mining implementations - OSDM '05*.

Büyükkör, Y. and Özler C., 2016, Hızlandırılmış yaşam testi verilerinin analizi üzerine bir araştırma, *The journal of academic social sciences*, 34, 383-383.

Bradley, P., Fayyad, U., and Reina, C., 1998, Scaling clustering algorithms to large databases, *In proceedings of the 4th ACM SIGKDD*, New York, NY, 9-15.

Caliński, T. and Harabasz, J., 1974, A dendrite method for cluster analysis, *Communications in statistics-theory and methods*, 3 (1), 1-27.

Cateni, S., Colla, V. and Vannucci, M., 2008, Outlier detection methods for industrial applications, *Advances in robotics, automation and control*.

Cebeci, Z., Yıldız, F. and Kayaalp, G.T., 2015, K-ortalamalar kümelemesinde optimum k değeri seçilmesi, *2. ulusal yönetim bilişim sistemleri kongresi*, 8-10 Ekim 2015, Erzurum, Orka Ofset Matbaacılık, ISBN:978-975-442-738-7, 231-242.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Wirth, R., 2000, CRISP-DM 1.0. Step-by-step data mining guide, *CRISP-DM consortium*.

Charrad, M., Ghazzali, N., Boiteau, V. and Niknafs, A., 2014, NbClust: An R package for determining the relevant number of clusters in a data set, *Journal of statistical software*, 61, 1–36.

Chen, Y., Kim, J. and Mahmassani, H., 2019, Operational scenario definition in traffic simulation-based decision support systems: pattern recognition using a clustering algorithm, *Journal of transportation engineering*, part a: systems, 145 (4).

Dalli, A., 2003, Adaptation of the f-measure to cluster based lexion quality evalution, *Proceedings of the european association for computational linguistics*, 51-56, 10.3115/1641396.1641404.

Davies, D. L. and Bouldin, D.W., 1979, A cluster separation measure, *IEEE transactions on pattern analysis and machine intelligence*, 2, 224-227.

Davies_Bouldin evaluation, https://www.mathworks.com/help/stats/clustering.evaluation.daviesbouldinevaluation.html;jsessionid=2087d22a58110457246968150d32, [Ziyaret Tarihi : 10.03.2022]

Day, W. and Edelsbrunner, H., 1984, Efficient algorithms for agglomerative hierarchical clustering methods, *Journal of classification,* 1 (7), 7-24.

Dhar, V., 1998, Data mining in finance: using counterfactuals to generate knowledge from organizational information systems, *Information systems,* 23 (7), 423-437.

Dilts, D., Khamalah, J. and Plotkin, A., 1995, Using cluster analysis for medical resource decision making, *Medical decision making*, 15 (4), 333-346.

Dunn, J., 1974, Well-separated clusters and optimal fuzzy partitions, *Journal of cybernetics,* 4 (1), 95-104.

Erpolat, S., 2022, Otomobil yetkili servislerinde birliktelik kurallarının belirlenmesinde apriori ve FP-growth algoritmalarının karşılaştırılması, *Anadolu üniversitesi sosyal bilimler dergisi,* 12 (1), 151-166.

Escobar, L.A. and Meeker, W.Q., 2006, A Review of accelerated test models, *Statistical science,* 21, 552-577.

Ester, M., Kriegel, H-P., Sander, J. and Xu, X., 1996, A density-based algorithm for discovering clusters in large spatial databases with noise, *In proceedings of the 2nd ACM SIGKDD,* Portland, Oregon, 226-231.

Ester, M., Krigel, H-P., and Xu, X., 1995, A database interface for clustering in large spatial databases, *In proceedings of the 1st ACM SIGKDD,* Montreal, Canada, 94-99.

Fan, J., Yung, K.C. and Pecht, M.G., 2012, Lifetime estimation of high-power white LED using degradation-data-driven method, *IEEE transactions on device and materials reliability,* 12, 470-477.

Fayyad, U., Piatetsky-Shapiro, G. and Padhraic Smyth, P., 1996, From data mining to knowledge discovery in databases, *AI magazine,* 17 (3), 37.

Flach, P.A. and Lachiche, N., 2001, Confirmation-guided discovery of first-order rules with Tertius, *Machine learning 42,* 61–95 . https://doi.org/10.1023/A:1007656703224

Fox, G., Negrete-Yankelevich, S. and Sosa, V., 2015, *Ecological statistics : contemporary theory and application,* 1st ed., Oxford University Press, 4, ISBN: 9780199672547.

Ganti, V., Gehrke, J. and Ramakrishnan, R., 1999, CACTUS-clustering categorical data using summaries, *In proceedings of the 5th ACM SIGKDD*, San Diego, CA,73-83.

Ghafoor, N. and Ahmad, M., 2021, Prioritizing effectiveness of algorithms of association rule mining, *Journal of computational learning strategies & practies,* 1 (1).

Gibson, D., Kleinberg, J. and Raghavan, P., 1998, Clustering categorical data: an approach based on dynamic systems, *In proceedings of the 24th international conference on very large databases,* New York, NY, 311-323.

Goethals, B., Nijssen, S. and Zaki, M., 2005, Open source data mining, *ACM SIGKDD explorations newsletter*, 7 (2), 143-144.

Goil, S., Nagesh, H., and Choudhary, A., 1999, MAFIA: Efficient and scalablesubspace clustering for very large data sets, *Technical report CPDC-TR-9906-010*, Northwestern university.

Guglielmetti, L.C., Faber-Castell, F., Fink, L. and Vuille-Dit-Bille, R.N., 2021, Statistics decrypted—a comprehensive review and smartphone-assisted five-step approach for good statistical practice, *Langenbeck's archives of surgery*, 407, 529 - 540.

Guha, S., Rastogi, R., and Shim, K. 1999. ROCK: A robust clustering algorithm for categorical attributes, *In proceedings of the 15th ICDE*, Sydney, Australia, 512-521.

Gupta, R.D. and Kundu, D., 2001, Exponentiated exponential family: an alternative to gamma and weibull distributions, *Biometrical journal*, 43, 117-130.

Gupta, R.D. and Kundu, D., 2007, Generalized exponential distribution: existing results and some recent developments, *Journal of statistical planning and inference*, 137, 3537-3547.

Gündoğdu, S., 2014, The Usage of common multiple comparison tests (Post-Hoc) in fisheries sciences, *Journal of fisheriessciences.com*, 8 (4), 310-316.

Han, J., Kamber, M. and Pei, J., 2012, *Data mining concepts and techniques*, 3rd ed., Elsevier Inc.

Halkidi, M., Batistakis, Y. and Vazirgiannis, M., 2002, Clustering validity checking methods: part II, *ACM SIGMOD record*, 31 (3), 19-27.

Hahsler, M. and Chelluboina S., 2011, Visualizing association rules in hierarchical groups, *In 42nd symposium on the interface: statistical, machine learning, and visualization algorithms (interface 2011)*, The Interface Foundation of North America.

Hahsler, M. and Karpienko, R., 2017, Visualizing Association Rules in hierarchical groups, *J bus econ*, 87, 317–335 . https://doi.org/10.1007/s11573-016-0822-8.

Hahsler, M., 2017, arulesViz: interactive visualization of association rules with R, *The R journal*, 9 (2), 163-175.

Hahsler, M., Grün, B. and Hornik, K., 2005, arules – A Computational environment for mining association rules and frequent item sets, *Journal of statistical software*, 14 (15), 1–25.

Hahsler, M. and Chelluboina, S., 2011, Visualizing association rules : introduction to the R-extension package arulesViz.

Hastie, T., Tibshirani, R. and Friedman, J.H., 2009, The elements of statistical learning: data mining, inference, and prediction, 2nd ed., Springer.

Hegedüs, J., Hantos, G. and Poppe, A., 2020, Lifetime modelling issues of power light emitting diodes, *Energies*.

Hinneburg, A. and Keim, D., 1999, Optimal grid-clustering: towards breaking the curse of dimensionality in high-dimensional clustering, *In proceedings of the 25th conference on VLDB*, Edinburgh, Scotland, 506-517.

Houtsma, M. and Swami, A., 1995, Set-oriented mining for association rules in relational databases, *Proceedings of the 11th IEEE International conference on data engineering*, Taipei, Taiwan, 25-34,

Huang, Z., 1998, Extensions to the k-means algorithm for clustering large data sets withcategorical values, *Data mining and knowledge discovery*, 2 (3), 283-304. https://fultonfindings.com/ [Ziyaret tarihi: 13/04/2021]

Human development report 2020, United nations development programme, http://hdr.undp.org/sites/default/files/hdr2020.pdf, [Ziyaret tarihi: 16.01.2022].

Human development report technical notes, 2020, http://hdr. undp.org/sites/default/files /hdr2020_technical_notes.pdf, [Ziyaret tarihi: 16.01.2022].

IESNA, 2011, IES TM-21-11: Projecting long term lumen maintenance of LED light sources, *Illuminating engineering society*, New York.

IESNA, 2008, IES LM-80-08 : approved method: measuring lumen maintenance of LED light sources, *Illuminating engineering society*, New York.

IESNA, 2015, Approved method: measuring luminous flux and color maintenance of LED packages, arrays and modules, IES LM-80-15, *Illuminating engineering society*, New York.

IESNA, 2020, Approved method: measuring luminous flux and color maintenance of LED packages, arrays, and modules. IES LM-80-20, *Illuminating engineering society*, New York.

Ireson, W., Coombs, C. and Moss, R., 1996, *Handbook of reliability, engineering and management*, 2nd ed., McGraw Hill, New York.

José-García, A. and Gómez-Flores, W., 2021, A survey of cluster validity indices for automatic data clustering using differential evolution. In *proceedings of the genetic and evolutionary computation conference* , 314-322.

Junsawang, P., Promwongsa, M. and Srisodaphol, W., 2021, Robust outliers detection method for skewed distribution, *Thailand statistician*, 19 (3), 450-471.

Karakaş, S., 2017, *Prof. Dr. Sirel Karakaş psikoloji sözlüğü:bilgisayar programı ve veritabanı*, https://www.psikolojisozlugu.com/post-hoc-test-post-hoc-test, [Ziyaret tarihi: 17.07.2022].

Karypis, G. and Kumar, V., 1999, A fast and highly quality multilevel scheme for partitioning irregular graphs, *SIAM journal on scientific computing*, 20, 1.

Karypis, G., Han, E.H. and Kumar, V., 1999, CHAMELEON: A hierarchical clustering algorithm using dynamic modeling, *Computer*, 32, 68-75.

Kaufman, L. and Rousseeuw, P., 1990, *Finding groups in data: an introduction to cluster analysis*, John Wiley and Sons, New York.

Kaur, G. and Chauhan, R.K., 2014, Comparative study of association rule mining algorithms, *International journal for scientific research and development*, 2, 412-416.

Kim, M. and Ramakrishna, R., 2005, New indices for cluster validity assessment, *Pattern recognition letters*, 26 (15), 2353-2363.

Keller, A.Z., Kamath, A. and Perera, U.D., 1982, Reliability analysis of CNC machine tools, *Reliability engineering*, 3, 449-473.

Khan, A. and Goel, A., 2021, Basics of statistical comparisons, *Indian pediatrics*, 58 (10), 987-990.

KP, M., Singh, R. and Kumar, S., 2012, Apriori-Hybrid algorithm as a tool for colon cancer microarray data classification, *International journal of engineering research and development*, 4 (7), 53-57.

Kumar, R. and Sharma, A., 2017, Data mining in education: a review, *International journal of mechanical engineering and information technology*, 5 (1), 1843-1845.

Kumbhare, T. and Chobe, S., 2014, An overview of association rule mining algorithms, *International journal of computer science and information technologies*, 5 (1), 927-930.

Kuş, C. and Akdoğan, Y., 2010, BURR XII dağılımının parametrelerinin ilerleyen tür ilk bozulma sansürlemeye dayalı güven aralıkları ve güven bölgeleri, *VII. istatistik günleri sempozyumu*, 3, 16-24.

Lighting Research Center, 2021, *Literature summary of lifetime testing of light emitting diodes and LED products*, https://www.iea-4e.org/wp-content/uploads/publications/2022/05/SSL-Annex-Lifetime-Literature-Review-Report-by-the-LRC_final.pdf, [Ziyaret Tarihi : 17.07.2022].

Lloyd, S.P., 1982, Least squares quantization in PCM, *IEEE transactions information theory*, 28 (2), 128–137.

Lorr, M.,1983, *Cluster analysis for social sciences*, San Francisco:Jossey-Bass.

Lu, Y., Phillips, C. and Langston, M., 2019, A robustness metric for biological data clustering algorithms, *BMC bioinformatics*, 20, 15.

Mahdavi, A. and Kundu D., 2017, A new method for generating distributions with an application to exponential distribution, *Communication in statistics- theory and methods*, 46 (13), 6543-6557.

MacQueen, J., 1967, Some methods for classification and analysis of multivariate observations, *Proc. of the 5th berkeley symposium math. stat. prob.*, Berkeley, CA, 281–297.

McCue, C., 2015, *Data mining and predictive analysis: intelligence gathering and crime analysis*, 2nd ed., Elsevier.

McLachlan, G., 1992, Cluster analysis and related techniques in medical research, *Statistical methods in medical research*, 1 (1), 27-48.

Mann, N. R., 1972, Design of over-stress life-test experiments when failure times have a two parameter weibull distribution, *Technometrics*, 14 (2), 437-451.

Marshall, A.W. and Olkin, I., 1997, A new method for adding a parameter to a family of distributions with application to the exponential and weibull families, *Biometrika*, 92, 505-505.

Măruşteri, M. and Bacârea, V.C., 2010, Comparing groups for statistical differences: how to choose the right statistical test, *Biochemia medica*, 20, 15-32.

Meeker, W. Q. and Nelson, W. B., 1978, Theory for optimum censored accelerated life tests for the weibull and extreme value distributions, Technometrics, 20 (2), 171-177

Menčík, J., 2016, Software for reliability analysis, *Concise reliability for engineers*.

Mudholkar, G. S. and Srivastava, D. K., 1993, Exponentiated weibull family for analyzing bathtub failure-rate data, *IEEE transactions on reliability*, 42 (2), 299–302.

Nelson, W. and Kielpinski, T., 1976, Theory for optimum censored accelerated life tests for normal and lognormal life distributions, *Technometrics*, 18 (1), 105.

Nelson, W., 1982, *Applied life data analysis*, John Wiley & Sons, New York.

Newman, D., 2014, Missing data, *Organizational research methods*, 17 (4), 372-411.

Ng, R. and Han, J., 1994, Efficient and effective clustering methods for spatial data mining, *In proceedings of the 20th conference on VLDB*, Santiago, Chile, 144-155.

Osborne., J. W., 2013, *Best practices in data cleaning*, Sage Publication. Inc., California.

Özkan, Y., 2020, *Veri madenciliği yöntemleri*, 4th ed., Papatya Yayıncılık Eğitim, İstanbul.

Pala, O. and Aksaraylı, M., 2020, Kümeleme için değiştirilmiş dunn indeksi ile bir parçacık sürü optimizasyon yaklaşımı, Journal of Yaşar university, 15 (58), 236-245.

Pang, G., Cao, L. and Aggarwal, C., 2021, Deep learning for anomaly detection: challenges, methods, and opportunities, *Proceedings of the 14th ACM international conference on web search and data mining.*

Pascual, F. and Montepiedra, G., 2005, Lognormal and weibull accelerated life test plans under distribution misspecification, *IEEE transactions on reliability,* 54 (1), 43-52.

Patel, S., Naik, V. and Patel, P.B., 2015, An analysis of application of multiple comparison tests (post-hoc) in ANOVA in recently published medical research literature, *National journal of community medicine*, 6, 117-120.

Peng, C.Y.J., Harwell, M., Liou, S.M. and Ehman L.H., 2007, *Advances in missing data methods and implication for educational research*, Real data analysis, In: Sawilowski, S.(ed.), USA:IAP-Information Age Publishing, 31-77.

Xppower.com, *Reliabity in Electronics*, https://hu.farnell.com/wcsstore/ExtendedSitesCatalogAssetStore/cms/asset/pdf/common/xp-power/reliability-in-electronics.pdf, [Ziyaret tarihi: 09.08.2022].

ReliaSoft Corporation, *Life data analysis reference book*, 2014, http://reliawiki.com/index.php/Life_Data_Analysis_Reference_Bookrom ReliaWiki.org,[Ziyaret tarihi: 07.08.2022.

Rendón, E, Abundez, I., Arizmendi, A., and Quiroz, E.M., 2011, Internal versus external cluster validation indexes, *International journal of computers and communications,* 5, 27-34.

Rijsbergen, C.V., 1979, *Information retrieval*, 2nd ed., Butterworth, London.

Rousseeuw, P., 1987, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, *Journal of computational and applied mathematics*, 20, 53-65.

Rubin, DB, 1976, Inference and missing data, *Biometrika*, 63, 581–592.

Salgado, C., Azevedo, C., Proença, H. and Vieira, S., 2016, Missing data, *Secondary analysis of electronic health records*, 143-162.

Satyavathi, N., Rama, D. and Nagaraju, D., 2019, Present state of the art of association rule mining algorithms, *International journal of engineering and advanced technology*, 9 (1), 6398-6405.

Schafer, J. L., 1999, Multiple imputation: a primer, *Statistical methods on medical resarch*, 8 (1), 3-15.

Scheffe, H.,1953, A method of judging all contrasts in the analysis of variance, *Biometrika*, 40, 87-104.

Schröer, C., Kruse, F. and Gómez, J., 2021, A systematic literature review on applying CRISP-DM process model, *Procedia computer science*, 181, 526-534.

Sheikh, B., Suke, A. and Tighare, K., 2021, Study of association rule mining techniques, *International journal of advanced innovative technology in engineering*, 6 (3).

Sheikholeslami, G., Chatterjee, S., and Zhang, A., 1998, WaveCluster: A multi resolution clustering approach for very large spatial databases, *In proceedings of the 24th conference on VLDB*, New York, NY, 428-439.

Shi, C., Wei, B., Wei, S., Wang, W., Liu, H. and Liu, J., 2021, A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm, *EURASIP journal on wireless communications and networking*, 2021 (1).

Silahtaroğlu, G., 2022, *Veri madenciliği kavram ve algoritmaları*, 4th ed., Papatya Yayıncılık, İstanbul.

Steinley, D., 2006, K-means clustering: A half-century synthesis, *British journal of mathematical and statistical psychology*, 59 (1), 1-34.

Suzuki, R. and Shimodaira, H., 2006, Pvclust: an R package for assessing the uncertainty in hierarchical clustering, *Bioinformatics*, 22 (12), 1540-1542.

Symynck, J. and Bal, F.D., 2011, Monte Carlo pivotal confidence bounds for weibull analysis with implementations in R, *TEHNOMUS-new technologies and products in machine manufacturing technologies*.

Şeker, S., 2018, CRISP-DM: Endüstriler arası standart işleme – veri madenciliği için (cross industry standard processing – data mining), YBS ansiklopedi, 5 (2).

Tabachnick, B. G. and Fidell, L. S., 1996, *Using multivariate statistics*, 3rd ed., Harper Collins, New York.

Tan, P., Steinbach, M. and Kumar, V., 2019, *Introduction to data mining*, 2nd ed., Pearson.

Tang, J. and Ngan, H., 2016, Traffic outlier detection by density-based bounded local outlier factors, IT in industry, 4 (1).

Taylor, F. W., 1911, *The principles of scientific management*, Harper & Brothers.

Tekin, S. and Özel, G., 2022, Alfa güç BURR-XII dağılımı ve bir uygulama, Journal of quantitative sciences, 2 (1).

Unwin, A., Hofmann, H. and Bernt, K., 2001, The two key plot for multiple association rules control, *PKDD LNAI*, 2168, 472-483.

Upadhyay, S.K. and Sharma, R., 2018, A bayes analysis and comparison of weibull and lognormal based accelerated test models with actual lifetimes unknown, *RT&A*, 4 (51).

Van Driel, W.D., Schuld, M. and Jacobs, B., Commissaris, F., Van der Eyden, J. and Hamon, B., 2015, Lumen maintenance predictions for LED packages using LM80 data, 2015, *16th international conference on thermal, mechanical and multi-physics simulation and experiments in microelectronics and microsystems*, 1-5.

Vinay, S., Nisarg, S., Samay P., Nirali D., Nimisha C., Umang P., Ankit P., Vishvash C. and Ashish P., 2013, A study of various projected data based pattern mining algorithm, *Research journal of material sciences*,1 (2), 1-5.

Wang, F., Franco-Penya, H., Kelleher, J., Pugh, J. and Ross, R., 2017, An analysis of the application of simplified silhouette to the evaluation of k-means clustering validity, *Machine learning and data mining in pattern recognition*, 291-305.

Wang, W., Yang, J. and Muntz, R., 1997, STING: a statistical information grid approach to spatialdata mining. *In proceedings of the 23rd conference on VLDB,* Athens, Greece,186-195.

Waters, M., 1995, *Globalization*, 1st ed., Routledge. London and New York.

Wong, P.C., Whitney, P. and Thomas, J.J., 1999, Visualizing association rules for text mining, *Proceedings 1999 IEEE symposium on information visualization (InfoVis'99*), 120-123.

Xie, X. and Beni, G., 1991, A validity measure for fuzzy clustering, *IEEE transactions on pattern analysis and machine intelligence*, 13 (8), 841-847.

Xu, A. and Tang, Y., 2015, A Bayesian method for planning accelerated life testing, *IEEE transactions on reliability*, 64 (4), 1383-1392.

Yang, L., 2003, Visualizing frequent itemsets, association rules, and sequential patterns in parallel coordinates, *In computational science and its applications – ICCSA, lecture notes in computer science*, 21–30.

Yoo, I., Alafaireet, P., Marinov, M., Pena-Hernandez, K., Gopidi, R., Chang, J. and Hua, L., 2011, Data mining in healthcare and biomedicine: a survey of the literature, *Journal of medical systems*, 36 (4), 2431-2448.

Zaki, M. J., Meira Jr, W., 2014, *Data mining and analysis: fundamental concepts and algorithms*, Cambridge University Press.

Zhang, T., Ramakrishnan, R., and Livny, M., 1997, BIRCH: A new data clustering algorithm and its applications, *Journal of data mining and knowledge discovery*, 1 (2), 141-182.

Zimmer, W.J., Keats, J.B. and Wang, F.K., 1998, The Burr XII distribution in reliability analysis, *Journal of quality technology*, 30 (4), 386-394.

ÖZGEÇMİŞ

Kişisel Bilgiler	
Adı Soyadı	Tevfik ÖRKÜN
Doğum Yeri	
Doğum Tarihi	
Uyruğu	☑ T.C. ❒ Diğer:
E-Posta Adresi	
Web Adresi	

Eğitim Bilgileri	
Lisans	
Üniversite	Hacettepe Üniversitesi
Fakülte	Mühendislik Fakültesi
Bölümü	Elektrik ve Elektronik Mühendisliği Bölümü
Mezuniyet Yılı	2001

Yüksek Lisans	
Üniversite	Hacettepe Üniversitesi
Enstitü Adı	Mühendislik Fakültesi
Anabilim Dalı	Elektrik ve Elektronik Mühendisliği Bölümü
Programı	Elektronik Programı

Doktora	
Üniversite	İstanbul Üniversitesi
Enstitü Adı	Fen Bilimleri Enstitüsü
Anabilim Dalı	Enformatik Anabilim Dalı
Programı	Enformatik Programı

Makale ve Bildiriler