T.C.
İSTANBUL BEYKENT ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

ENDÜSTRİ MÜHENDİSLİĞİ ANABİLİM DALI
ENDÜSTRİ MÜHENDİSLİĞİ BİLİM DALI

AKARYAKIT SEKTÖRÜNDE İŞLEM ANORMALLİKLERİNİN VERİ MADENCİLİĞİ YÖNTEMLERİ İLE SINIFLANDIRILMASI

Yüksek Lisans Tezi

Tezi Hazırlayan
Sabahattin Mert BERKMEN

İstanbul, 2024

T.C.
İSTANBUL BEYKENT ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

ENDÜSTRİ MÜHENDİSLİĞİ ANABİLİM DALI
ENDÜSTRİ MÜHENDİSLİĞİ BİLİM DALI

AKARYAKIT SEKTÖRÜNDE İŞLEM ANORMALLİKLERİNİN VERİ MADENCİLİĞİ YÖNTEMLERİ İLE SINIFLANDIRILMASI

Yüksek Lisans Tezi

Tezi Hazırlayan
Sabahattin Mert BERKMEN

Öğrenci No
2120006010

ORCID NO
0000-0001-7273-954X

Danışman
Dr. Öğr. Üyesi Sabahattin Kerem AYTULUN

İstanbul, 2024

YEMİN METNİ

Yüksek Lisans Tezi olarak sunduğum **"Akaryakıt Sektöründe İşlem Anormalliklerinin Veri Madenciliği ile Sınıflandırılması**" başlıklı bu çalışmanın, bilimsel ahlak ve geleneklere uygun şekilde tarafımdan yazıldığını, bu tezdeki bütün bilgileri akademik ve etik kurallar içinde elde ettiğimi, yararlandığım eserlerin tamamının kaynaklarda gösterildiğini ve çalışmanın içerisinde kullanıldıkları her yerde bunlara atıf yapıldığını, patent ve telif haklarını ihlal edici bir davranışımın olmadığını belirtir ve bunu onurumla doğrularım. 24.05.2024

Sabahattin Mert BERKMEN

T.C.
İSTANBUL BEYKENT ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ MÜDÜRLÜĞÜ
TEZLİ YÜKSEK LİSANS SINAV TUTANAĞI

24./05/2024

Enstitümüz ***Endüstri Mühendisliği*** Anabilim Dalı ***Endüstri Mühendisliği*** Programı Yüksek Lisans öğrencilerinden ***2120006010*** numaralı ***Sabahattin Mert BERKMEN'in*** "İstanbul Beykent Üniversitesi Eğitim-Öğretim Yönetmeliği"nin ilgili maddesine göre hazırlayarak Enstitümüze teslim ettiği ***"Akaryakıt Sektöründe İşlem Anormalliklerinin Veri Madenciliği Yöntemleriyle Sınıflandırılması"*** konulu tezini, Yönetim Kurulumuzun 14/05/2024 tarih ve 2024/22 sayılı toplantısında seçilen ve Taksim yerleşkesinde toplanan biz jüri üyeleri huzurunda, İstanbul Beykent Üniversitesi Lisansüstü Eğitim ve Öğretim Yönetmeliğinin 29. maddesinin 3. fıkrası gereğince 45 dakika süre ile aday tarafından savunulmuş ve sonuçta adayın tezi hakkında ***"OYBİRLİĞİ"*** ile ***"KABUL"*** kararı verilmiştir.

İşbu tutanak, 2 nüsha olarak hazırlanmış ve Enstitü Müdürlüğü'ne sunulmak üzere tarafımızdan düzenlenmiştir.

DANIŞMAN
Dr. Öğr. Üyesi Sa*** Ke*** AY***
(İstanbul Arel Üniversitesi)

ÜYE
Dr. Öğr. Üyesi Fu*** Dİ***
(İstanbul Beykent Üniversitesi)

ÜYE
Prof. Dr. Nu*** ÇO***
(İstanbul Arel Üniversitesi)

Adı ve Soyadı : Sabahattin Mert BERKMEN

Danışmanı : Dr. Öğr. Üyesi Sabahattin Kerem AYTULUN

Derecesi ve Tarihi : Yüksek Lisans (Tezli), 2024

Alanı : Endüstri Mühendisliği

Anahtar Kelimeler : Veri Madenciliği, Sınıflandırma, Akaryakıt Sektörü

ÖZ

AKARYAKIT SEKTÖRÜNDE İŞLEM ANORMALLİKLERİNİN VERİ MADENCİLİĞİ İLE SINIFLANDIRILMASI

Akaryakıt sektöründe veri madenciliği uygulamaları gün geçtikçe gelişmekte ve uygulamalar yaygınlaşmaktadır. Akaryakıt sektöründe kullanılan yöntemler ve çeşitli analizler ile beraber akaryakıt hırsızlığı, sızıntı, dolum anında miktar aşımları, aşırı dolum sonrası taşma gibi önemli konular takip edilerek aksiyonlar alınmaktadır. Bu çalışmada belirlenen 4 önemli kategori doğrultusunda analizleri yapılmış ve kategorize edilmiş olan verinin farklı veri madenciliği sınıflandırma yöntemleri ile sınıflandırarak analizlerin doğruluğunu ve başarımlarını ölçmek amacıyla Türkiye'de faaliyet gösteren bir petrol şirketinin 2022-2023 tarihleri arasındaki verileri kullanılarak farklı veri madenciliği sınıflandırma yöntemleri uygulanmıştır.

Uygulama öncesinde veri analize uygun hale getirilmiştir. Analizde etkisi olmayan değişkenler çıkarılmıştır. Açılış, kapanış, dolum, fark, satış, SEL değeri, azalma miktarı ve kategori analizde belirleyici etkene sahip olduğundan analizde bu veriler kullanılmıştır.

Uygulama aşmasında RAPIDMINER programından yararlanılmıştır. Veri madenciliği sınıflandırma yöntemlerinden k en yakın komşu algoritması, Rastgele Orman Algoritması, Gradient Boosted Algoritması, ADABOOST Algoritması ve Karar Ağacı (J48) Algoritması ile sınıflandırmalar yapılarak çeşitli başarı ölçütleri ile modellerin başarıları ölçülmüştür. En başarılı sınıflandırma yönteminin Karar Ağacı (J48) Algoritması olduğu görülmüştür.

Name and Surname : Sabahattin Mert BERKMEN

Supervisor : Assist. Prof. Dr. Sabahattin Kerem AYTULUN

Degree and Date : Master's (Thesis), 2024

Major : Industrial Engineering

Keywords : Data Mining, Classification, Fuel İndustry

ABSTRACT

CLASSIFICATION OF PROCESS ABNORMALITIES IN THE FUEL INDUSTRY BY DATA MINING

Data mining applications in the fuel industry are developing day by day and the applications are becoming widespread. With the methods used in the fuel industry and various analyses, important issues such as fuel theft, leakage, quantity excesses at the time of filling, and overflow after overfilling are monitored and actions are taken. In this study, different data mining classification methods were applied using the data of an oil company operating in Turkey between 2022-2023 in order to measure the accuracy and performance of the analyzes by classifying the analyzed and categorized data with different data mining classification methods in line with the 4 important categories determined in this study.

Before the application, the data was made suitable for analysis. Variables that had no effect in the analysis were removed. Since opening, closing, filling, difference, sales, SEL value, decrease amount and category are the determining factors in the analysis, these data were used in the analysis.

RAPIDMINER program was used during the implementation phase. Classifications were made with the k nearest neighbor algorithm, Random Forest Algorithm, Gradient Boosted Algorithm, ADABOOST Algorithm and Decision Tree (J48) Algorithm, which are among the data mining classification methods, and the success of the models was measured with various success criteria. It has been observed that the most successful classification method is the Decision Tree (J48) Algorithm

İÇİNDEKİLER

Sayfa No

TABLOLAR LİSTESİ

Sayfa No.

ŞEKİLLER LİSTESİ

Sayfa No.

KISALTMALAR

CART : Classification and Regression Trees (Sınıflandırma ve Regresyon Ağaçları)

CHAID : Chi-Square Automatic Interactıon Detector (Otomatik Ki-Kare Etkileşim Belirleme Analizi)

CRISP-DM : Cross Industry Standard Process for Data Mining (Çapraz Endüstri Veri Madenciliği Standart Süreci)

EPDK: Enerji Piyasası Düzenleme Kurumu

FN : False Negative (Yanlış Negatif)

FP: False Positive (Yanlış Pozitif)

KNN : K Nearest Neighbors (K-En Yakın Komşu Algoritması)

ROC : Receiver Operating Characteristics (Olasılık Eğrisi)

RO : Random Forest Algorithm (Rastgele Orman Algoritması)

SEL : Sızıntı Eşik Litresi

TP : True Positive (Doğru Pozitif)

TN : True Nagative (Doğru Negatif)

QUEST : Quick Unbiased Efficient Statistical Trees (Hızlı, Yansız, Etkili İstatistiksel Ağaçlar)

XGBOOST : Gradient Boosting Trees (Aşırı Aşamalı Yükseltme Ağaçları)

WSM: Wet Stock Management (Yakıt Stok Yönetimi)

SÖZLÜK

Akaryakıt Sektörü: Küresel piyasalarda petrolün aranması, bulunması, işlenmesi, depolanması, taşınması ve stok kontrollerinin takiplerinin yapılması süreçlerini kapsayan terimdir.

İşlem Anormallikleri: İşlem anormallikleri genel olarak bir sektörde beklenmeyen tahmin edilemeyen ve normal olmayan davranışları ifade etmek amacıyla kullanılan bir terimdir. İşlem anormallikleri sektörün işleyişinde aksamalara, o sektöre olan güvenilirliğin azalmasına yol açabilmektedir. Akaryakıt sektöründe işlem anormallikleri fiyat dalgalanmaları, yolsuzluk ve kaçakçılık, tedarikçi sorunları ve mevcut politikalar olarak sıralanabilir.

Manipülasyon: Akaryakıt sektöründe manipülasyon, istasyonun bağlı olduğu otomasyon sisteminin mekanik veya elektronik aksamlarına müdahale edilerek verilerin hatalı oluşmasına veya oluşmamasına etki eden durumlar olarak ifade edilir.

Sıcaklık Değişimi: Akaryakıt sektöründe sıcaklık değişimi, tank sıcaklığının dolum ile veya mevsim şartlarına bağlı olarak değişimini ifade etmek amacıyla kullanılan terimdir.

Tank Kalibrasyonu: Kalibrasyon, belirlenmiş koşullar altında doğruluğu bilinen bir ölçüm standardını veya sistemini kullanarak diğer test ve ölçüm aletinin doğruluğunun ölçülmesi, sapmalarının belirlenmesi ve doküman haline getirilmesi için kullanılan ölçümler dizisidir. Akaryakıt sektöründe kalibrasyon ise, elirli bir depolama sisteminin doğru hacim değerinin, belirli bir ölçüm değerine göre belirlenmesi bilimidir. Tankın verileri sıralanır ve oranlarla karşılaştırma yapılır, yüklü olan cetvel ile satış yapan tankın arasındaki farkı ifade etmek amacıyla kullanılmaktadır.

Veri Madenciliği: Veri madenciliği süreci, problemin tanımlanması adımı ile başlayarak değerlendirme adımı ile son bulmaktadır. Süreç sonunda elde edilen örüntüler ve kurallar ile geleceğe yönelik planların ve kararların doğru bir şekilde alınması sağlanmaktadır.

GİRİŞ

Günümüz dünyasında, işletmeler gelişen teknoloji ile birlikte elde ettikleri ham verileri veri tabanlarında tutmaktadırlar. Bu veriler gün geçtikçe artmaktadır. Veriler, işletmelerin daha kaliteli hizmet verebilmesi, işletmelerin geleceğe taşınmalarında önemli rol oynamaktadır. Ham veriler kullanıma fazla elverişli değildir. Ham verilerden anlam taşıyan bilgilerin çıkarılması adına işlenmesi gerekmektedir. Ham verilerin işlenmesinde günümüzde oldukça sık kullanılmaya başlayan veri madenciliği yöntemlerinden faydalanılmaktadır.

Veri madenciliği süreci, problemin tanımlanması adımı ile başlayarak değerlendirme adımı ile son bulmaktadır. Süreç sonunda elde edilen örüntüler ve kurallar ile geleceğe yönelik planların ve kararların doğru bir şekilde alınması sağlanmaktadır.

Veri madenciliği yöntemleri 2000 yılından sonra oldukça sık kullanılmaya başlanmıştır. Günümüzde de yapılan çalışmaların sayısının arttığı görülmüştür. Bu sektörlerden biri de ülkemizde ve dünyada en çok kullanılan ve enerji ihtiyacının bir bölümünü teşkil eden akaryakıt sektörüdür. Akaryakıt sektöründe veriler otomasyon sistemleri üzerinden veri tabanlarına aktarılmaktadır. Akaryakıt sektöründe veriler işlenerek farklı anlam ve örüntüler çıkarılarak kategori bazında sınıflandırılmaktadır.

Çalışmanın ilk bölümünde, veri madenciliği tanımıyla başlanarak, veri madenciliğinde kullanılan temel kavramlar ile birlikte uygulama alanları, literatür çalışmalarından, veri madenciliği süreçleri adımları anlatılmıştır. Ayrıca bu bölümde veri madenciliğinin tanımlayıcı yöntemlerinden olan birliktelik kuralları, kümeleme analizi kavramlarından ve tahmin edici yöntemlerinden olan zaman serileri ve genetik algoritma kavramlarından bahsedilmiştir.

İkinci bölümde ise ilk olarak veri madenciliğinde sınıflandırma süreci hazırlık ve işlem süreci olarak iki başlık altında incelenmiştir. Ardından sınıflandırma metotları ve algoritmaları, istatistiğe dayalı algoritmalar, mesafeye dayalı algoritmalar ve hiyerarşik sınıflayıcı algoritmalar olarak alt başlık altında kategorize edilerek ayrıntılı bir şekilde anlatılmıştır.

Üçüncü bölümde analizde kullanılacak olan RAPIDMINER programı üzerinde durulmuştur. Dördüncü ve son bölüm ise uygulama bölümüdür. Uygulama Türkiye'de akaryakıt sektörüne veri ve analiz desteği sağlayan otomasyon firmasının veri tabanında tutulan ve kayıt altına alınan veriler kullanılarak yapılmıştır Günlük olarak istasyonda meydana gelen sorunlar otomasyon sistemi sayesinde alarm üretmekte ve bu kayıtların veri tabanına yansımasını sağlamaktadır. Veri tabanına yansıyan kayıtlar günlük olarak analiz edilmekte ve sorunun kaynağına göre yaklaşık olarak 50 kategori altında sınıflandırılıp sorunun çözümüne yönelik aksiyon alınmaktadır. Bu kapsamda 2022-2023 yıllarına ait veri seti kullanılarak daha önce analiz edilmiş ve kategori bazında sınıflandırılmış olan veri setlerinin çeşitli veri madenciliği yöntemleri kullanılarak sınıflandırılmış ve yöntemlerin sınıflandırma başarısı ve yapılan analizlerin doğruluğu değerlendirilmiştir.

Sonuç ve öneriler kısmında ise çalışmamızda diğer çalışmalar ile karşılaştırılmış eksik yönler değerlendirilerek farklı olarak nelerin üzerinde durulması gerektiği konusunda öneride bulunulmuştur.

1.VERİ MADENCİLİĞİNE GENEL BAKIŞ

Günümüzde, dünyada ortaya çıkan ve ham veri olarak depolanan bilgiler mevcuttur. Gelişen teknoloji ile birlikte bu veriler işlenerek anlamlı sonuçlar ve örüntüler elde edilerek kullanılabilir ve yorumlanabilir haline dönüştürülmüştür. İşleme sonucu elde edilen bilgiler işletmelerin problemlerin çözümlenmesini, daha kaliteli hizmet verebilmesini ve işletmelerin geleceğe taşınmasında önemli rol oynamaktadır. Bu durum işletmeleri veri tabanlarında bulunan ham bilgiyi işleyebilmek adına veri madenciliğine yönlendirmektedir.

Veri madenciliği yöntemleri, bilgisayarlı veri, toplama teknolojilerinin gelişmeye başladığı 2000 yıllardan itibaren kullanılmaya başlanmış ve veriler kayıt altına alınmaya başlanmıştır. Bu dönem işletmelerin ve bireysel verilerin elektronik ortamda kayıt altına alındığı büyük veri (big data) dönemi olarak adlandırılmaktadır (Tuzcu, 2018, s. 3).

Bu dönemde büyük verinin kullanılmaya başlanması işletmeler için devrim niteliğindedir. İşletmeler geleneksel veri işleme yöntemlerini kullanırken büyük veri kavramının ortaya çıkması ile birlikte işletmelerin iş yapma biçimleri ve yönetim şekillerinde birtakım değişiklikler meydana gelmiştir. Büyük verinin bileşenleri aşağıdaki gibidir (Karaboğa, 2020, s. 11).

- Hacim (volüme): Veri miktarını ifade eder.
- Hız (velocity): Verilerin üretim hızını ve toplanma hızını ifade eder. Gerçek zamanlı, anlık ve süreklidir.
- Çeşitlilik (variety): Büyük veri içerisinde var olan yapılandırılmamış, yarı yapılandırılmış ve yapılandırılmış olan farklı veri gruplarını temsil eder.
- Değer (value): Büyük ham veri içerisinden verilerin işlenerek ve işletmeler için değer yaratacak bilginin elde edilmesi sürecini ifade eder.
- Gerçeklik (Veracity): Verilerin güvenilirliğini ve tutarlılığını ifade eder.
- Değişkenlik (veriability): Yapılandırılmamış veri setindeki değişkenliği ifade eder.

- Görselleştirme (visualization): Veri madenciliği işlemi sonucu ortaya konulan modellerin çeşitli modeller ile yorumlanabilmesini ifade eder.

1.1. Veri Madenciliği Tanımı

Veri madenciliği yöntemleri 2000 yılları başından itibaren yaygınlaşmaya başlamış ve gün geçtikçe kullanıldığı sektörler ve alanlar artmıştır. Bu nedenle veri madenciliğinin tanımı kullanıldığı sektörlere göre değişiklik göstermiştir. En yaygın kullanılan tanım şöyledir; Veri madenciliği işletmelerin veri tabanlarında bulunan gerekli ve gereksiz bilgilerin işlenerek kullanılabilir hale getirilmesi ve bu bilgilerin işletme kararlarında faydalı şekilde kullanılmasıdır (Silahtaroğlu, 2016, s. 9).

Yukarıdaki tanıma ek olarak yapılan farklı tanımlamalar aşağıda verilmiştir:

Veri madenciliği, veri tabanlarında bulunan büyük ve anlamsız verilerin anlamlı bir yapı haline getirebilmek amacıyla gerekli formülleri kullanarak analiz etmeye ve uygulamaya yönelik çalışmalar bütünüdür (Ceran, 2006, s. 25).

Veri madenciliği, gerekli matematiksel modellerin ve gelişen teknolojiler ile birlikte ortaya çıkan örüntü tanıma teknolojileri kullanılarak veri tabanlarında yer alan ve oldukça büyük miktardaki verilerin içerisinden yeni anlamlar ve örüntüler elde edilme sürecidir (Uyumaz, 2017, s. 1-2).

Veri madenciliği, işletmelerin veri tabanlarında bulunan verilere odaklanarak çeşitli yöntemlerle depolanan verilerin analiz edilerek gelecek için işletmelere yarar sağlama sürecini ifade eder (Şekeroğlu, 2010, s. 10).

Veri madenciliği, oldukça büyük verilerin bilgisayar programları kullanılarak, gelecek dönemler için tahmin yapılmasını sağlayacak verilerin aranması ve ortaya çıkan verilerin kullanılmasıdır (Alan, 2016, s. 14).

Yukarıdaki tanımlamalara bakıldığında kısacası veri madenciliği verilerin çeşitli işlemler sonucunda işlenerek veriden bilgi çıkarma işlemidir.

1.2. Veri Madenciliği Tarihsel Gelişimi

Veri madenciliği geçmişten günümüze incelendiğinde 1960 yılından itibaren gelişme göstermiş olmasına rağmen günümüzde kullanılan veri madenciliği yöntemleri 2000 yılından sonra sıklıkla kullanılmaya başlanmıştır.

Veri madenciliği, zaman içerisindeki gelişimi tablo 1'de gösterilmiştir.

Tablo 1. Veri Madenciliği Tarihsel Gelişimi

Tarih	Teknolojiler
1950	• Bilgisayarın keşfi
1960	• Veri Toplanması • Bilgisayarlar • Diskler
1980	• İlişkisel Veri Tabanları • Veri Erişimi
1990	• Veri Ambarı • Büyük Boyutlu Veriler • Karar Destek Sistemleri
Günümüz	• Büyük Veri Tabanı • İleri Algoritmalar • Gelişmiş İşletim Sistemleri

Kaynak: Uyumaz, Ö. (2017). *Bankacılık sektöründe pazarlama stratejilerinin belirlenmesinde sınıflandırma ve veri madenciliği* [Yüksek lisans tezi]. Beykent Üniversitesi.

1950 yılında ilk bilgisayarlar sayımlar için kullanılmaya başlanmıştır.1960 gelindiğinde, verilerin toplanması ile basit öğrenmeli bilgisayarlar geliştirilmiştir. 1980 yıllarda ilişkisel veri tabanı kavramının geliştirilmesiyle birlikte veri madenciliği yöntemleri yaygınlaşmış ve birçok farklı sektörde kullanılmaya başlanmıştır. Bu dönemde şirketler var olan veriler ile ilk veri tabanlarını bu dönemde oluşturmuştur.

1990 yılına gelindiğinde büyük boyutlu veriler işletmelerin veri tabanlarında birikmeye başlamış ve işletmeler, veri tabanlarına kayıtlı olan verileri kullanarak bu verilerden faydalı veriler elde etmenin yollarını aramaya başlamışlardır. Bu dönemden sonra yapılan çalışmalar sonucunda 1992 yılında veri madenciliği kavramı ve yazılımı geliştirilmiştir. 2000 yılından sonra ise hemen hemen her alanda kullanılmaya başlanmıştır (Savaş vd., 2012, s.4-5).

1.3. Veri Madenciliği Uygulama Alanları

Veri madenciliği uygulama alanlarına bakıldığında, verinin kalitesi, bilgisayarların işletim sistemleri ve veri madenciliğinde kullanılan teknolojilerin gelişmesi ile birlikte veri madenciliği uygulama alanları genişlemiş ve yaygınlaşmıştır. Veri madenciliği, devlet uygulamaları, bankacılık ve finans uygulamaları, mühendislik uygulamaları, perakende, üretim ve imalat, pazarlama, ulaştırma, eğitim vb. amaçlarla kullanılmaktadır.

Pazarlama;

- Satış tahmini
- Pazar sepeti analizi
- Birliktelik kurallarının belirlenmesi
- İşletmelerin var olan müşterilerinin tutulması ve yeni müşterilerin işletmeye kazandırılması
- Müşterilerin durumlarının değerlendirilmesi

Bankacılık

- Dolandırıcılık tespiti
- Gelecek dönemdeki kredi kartı taleplerinin tahmini
- Bankada yer alan farklı tablolar arasında korelasyon tespiti

Biyoloji-Genetik

- Genetik hastaların tespiti
- Gen analizleri ve haritalarının çıkarılması
- Kanser hastalığının tespiti

1.4. Literatürde Veri Madenciliği

Veri madenciliği 1990 yıllında ortaya çıkmış bir kavramdır. Ancak 2000 yıllından itibaren çeşitli sektörlerde kullanıldığı görülmüştür. Kullanılan çeşitli sektörlerle ilgili çalışmalar kronolojik olarak aşağıda açıklanmıştır.

Boland ve arkadaşları yapmış oldukları çalışmada ABD ordusu Mühendisler Birliği'ndeki su kullanımını veri madenciliği yöntemleri ile tahminlenerek açıklamışlar ve bu konu hakkında tavsiyelerde bulunmuşlardır (Boland vd., 1981, s. 140).

Koonce ve Tsai, genetik algoritma ve veri madenciliği kavramlarını birlikte kullanmışlardır. Yapmış oldukları çalışmada bir imalathanede çizelgeleme işlemi için geliştirilen bir genetik algoritma tarafından üretilen verilerin analizi ve veri kalıplarının keşfi için veri madenciliği yöntemlerinden yararlanmışlardır (Koonce ve Tsai, 2000, s. 373).

Sforna, enerji alanında faaliyet gösteren işletmenin, müşteri faturalandırma işlemlerini desteklemek amacıyla oluşturulan büyük ve karmaşık veri tabanından veri madenciliği yöntemlerini kullanarak gizli bilgileri elde etmek ve keşfedilmemiş kuralları ortaya çıkarmaya çalışmıştır (Sforna, 2000, s. 207).

Hui ve Jha yapmış oldukları çalışmada, geleneksel üretim tekniklerini kullanan bir işletmede müşteri hizmetleri veri tabanında yer alan büyük boyutlu verilerden anlamlı kurallar çıkarılabilmesi için veri madenciliği tekniklerinin nasıl uygulanacağını araştırmaktadır (Hui ve Jha, 2000, s. 1).

Ulaş yapmış olduğu sepet analizi çalışmasında, perakende alanında faaliyet gösteren bir işletmede veri tabanında yer alan verileri kullanarak satış analizi gerçekleştirmiş ve Apriori algoritmasını detaylıca anlatmıştır. Satış analizi sonucunda en yüksek sıklığın sebze ve ekmek arasında olduğunu gözlemlemiştir (Ulaş, 2001, s. 21).

Güvenç yapmış olduğu çalışmada, yüksek öğretim kurumlarında yer alan büyük verilerin ilişkilendirme kurallarını, karar ağaçlarını ve istatistiksel veri madenciliğine dayalı teknikleri kullanarak öğrencilerin ders içerisindeki başarılarını ve hangi dersleri seçmeleri ile ilgili veri tabanı geliştirerek var olan verilerin öğrenci

ve yüksek öğretim kurumu açısından yararlı bilgiye dönüştürülmesini amaçlamıştır (Güvenç, 2001, s. 6).

Chung Hu ve arkadaşları, Apriori algoritmasına dayalı bulanık sınıflandırma kurallarını ortaya çıkarmak amacıyla veri madenciliği yöntemlerini ve sınıf sayısını belirlemek amacıyla yardımcı olarak genetik algoritmaları kullanmışlardır (Chen vd., 2003, s. 453).

Özekes yapmış olduğu çalışmada, veri madenciliği içerisinde yer alan işlevlerine göre farklılık gösteren birliktelik kuralları, sınıflama, regresyon ve kümeleme kavramlarını detaylıca açıklayarak uygulama alanlarına değinilmiştir (Özekes, 2003, s. 66).

Kiremitçi yapmış olduğu yüksek lisans tezi çalışmasında, sıklıkla kullanılan ve bilinen veri analiz yöntemlerinin analiz kısmında yetersiz kaldığı ve iyi bir performans göstermediği gerekçesiyle veri madenciliği yöntemlerinin kullanılması gerektiğini önermiştir. Yapmış olduğu uygulamada büyük veri tabanına sahip bir işletmede veri seti üzerine kümeleme ve karar ağacı yöntemlerini uygulayarak veri setinden çeşitli kümeler ve örüntüler oluşturarak, işletme için karar destek sistemi oluşturmuştur (Kiremitçi, 2005, s. 1-2).

Chen ve arkadaşları yapmış oldukları çalışmada, tedarik zinciri yöntemlerinden bahsetmişlerdir. Uygulamada ise paralel koridor düzenine sahip dağıtım merkezindeki sipariş toplama problemi için veri madenciliği yöntemlerini kullanarak dağıtıma çıkacak ürün gruplarını kümelemişlerdir (Chen vd., 2005, s. 453).

Lu ve arkadaşları yapmış oldukları çalışmada, veri madenciliği yöntemlerini kullanarak elektrik piyasası fiyat artış tahmini yapmışlardır. Çalışmada daha önce yazılmış elektrik piyasası fiyat artışı araştırmalarından farklı olarak fiyat artışlarının yanı sıra normal fiyatı da tahmin edebilen bir model önermektedir (Lu vd., 2005, s. 19).

Ouali ve arkadaşları önerdikleri çalışmada, şizofreni hastalığının tanısında kullanılacak ve tespit sağlayabilecek bir veri madenciliği modeli önermişlerdir. Modelde şizofreni teşhisi için seçilen adayların genotiplerini içeren bir veri seti kullanılarak bayes ağları sınıflandırıcısı destek sistemi ile hastaları orta başlangıç ve ileri düzey olarak 3 grupta sınıflandırmıştır (Ouali vd., 2006, s. 1278).

Kuşaksızoğlu, yapmış olduğu tez çalışmasında, bir telekomünikasyon şirketinde normal kullanıcılar ile sahtekarlık yapan kullanıcıları ayıran bir model geliştirmek için abonelerin konuşma detaylarını, demografik özellikleri incelenmiştir (Kuşaksızoğlu, 2006, s. 1-5).

Damle ve arkadaşları, ABD yer alan Mississippi nehri üzerinde yer alan ölçüm istasyonundan aldığı verilerle zaman serisi ve veri madenciliği tekniklerini kullanarak taşkın tahminlerini yaparken aynı zamanda nehrin gelecekteki deşarj değerlerini tahmin etmek için model önermişlerdir (Damle ve Yalçın, 2007, s. 305).

Tunçelli yapmış olduğu tez çalışmasında, büyük ve tanınır bir işletmede istatistik ve veri madenciliği yöntemlerini kullanarak talep tahmini yapmayı amaçlamıştır. Talep tahmini modelinde fiyat indirimi, reklam vs. göz önünde bulundurulmuştur. Talep tahmini, öncelikle üstel düzeltme ve lineer regresyon gibi istatistik modelleri ile daha sonra regresyon ve regresyon ağacı gibi veri madenciliği yöntemleri ile çözülmüş, en son olarak istatistik ve veri madenciliği yöntemleri kıyaslanmıştır (Ayşe, 2007, s. 1).

Aydoğan ve arkadaşları, veri madenciliğinde büyük ve eksik veri setleri üzerinden bilgi keşfi yapabilmek için kullanılan kaba küme teorisi üzerinde durmuşlar ve temel kavramlarından bahsetmişlerdir. Ayrıca veri madenciliğinde sınıflandırma yöntemleri ile ilgili yapılan çalışmaları incelemişler ve sonuçlarını sunmuşlardır (Aydoğan ve Gencer, 2007, s. 17).

Rupnik ve arkadaşları 2007 yılında yapmış olduğu çalışmada veri madenciliği bilgi keşif süreçlerinden detaylı olarak bahsetmiştir. Aynı zamanda veri madenciliği yöntemlerinin karar destek sistemleriyle entegrasyonu konusunda model önermiştir (Rupnik vd., 2007, s. 89).

Olafsson ve arkadaşları veri madenciliği ve yöneylem araştırması konularının birbirleriyle olan ilişkilerini açıklamışlardır. Bu kapsamda makalenin ilk bölümünde yapılan örnek çalışmaları detaylı olarak incelemişler ve diğer çalışmalara referans sağlamayı amaçlamışlardır. İkinci bölümde ise yönetim alanında yapılmış çalışmalar detaylı olarak incelenmiştir (Olafsson vd., 2008, s. 1429).

Dönmez yapmış olduğu tez çalışmasında, dayanaklı tüketim sektöründe faaliyet gösteren bir işletmede var olan birbirlerine benzer bayilerin veri madenciliği yöntemleriyle kümelenmiştir. Kümeleme sonucunda ortaya bayilerin pazar stratejileri belirlenerek cirolarının arttırılması amaçlanmıştır. Dönmez çalışmasında kümeleme analizi olarak Kareli Öklid Uzaklığı ve Ward Yöntemlerini kullanmışlardır. Uygulama sonucu bayiler beş farklı küme grubuna ayrılmışlardır (Dönmez, 2008, s. 1).

Gün yapmış olduğu tez çalışmasında ürün çeşitliliği çok olan perakende sektöründe en iyi satış performansına sahip ürünü veya ürünleri belirlemek hem de perakende sektöründe ortaya çıkabilecek tüm stratejileri göz önüne alarak bu stratejilerin doğru uygulanabilmesi için veri madenciliği yöntemlerinden ilişkilendirme algoritmalarını kullanmış ve ilgili modeller geliştirmiştir (Gün, 2008, s. 1).

Ngai ve arkadaşları yapmış oldukları çalışmada veri madenciliği yöntemlerinin müşteri ilişkileri yönetimi (CRM) ile ilişkisini araştırmıştır. Bu kapsamda 2000-2006 yılları arasında 24 dergi baz alınarak yapılan çalışmada ortaya çıkarılan çalışmalar incelenmiş ve literatür oluşturarak sınıflandırma yapılmıştır (Ngai vd., 2009, s. 2592).

Mastrogiannis ve arkadaşları veri madenciliği ve yöneylem araştırması kavramlarını kullanarak veri madenciliğinde sınıflandırma algoritmalarının doğruluğunu arttırmak için model önermişlerdir. Önerilen modelde veri madenciliğinde verilerin sınıflandırılması sonucunda ortaya çıkan en iyi kuralları seçer ve ortaya çıkan sınıfları sınıflandırılması gereken nesnelere atar. Modelin doğruluğu yaygın olarak kullanılan beş veri tabanında denenerek test edilmiştir (Mastrogiannis vd., 2009, s. 2829).

Şahin yapmış olduğu tez çalışmasında, günümüzde çözülmesi zor olan ve np-zor problem grupları içerisinde yer alan araç rotalama problemlerini çözmek için veri madenciliği yöntemlerinden faydalanmıştır. Yapmış olduğu çalışmada bir otomobil firması tarafından üretilen araçların ülkemizdeki çeşitli bayilere ulaştırmada maliyeti en aza indirecek bir model önermiştir. Araçların teslimatında tırlar kullanılmaktadır. Bir tırı dolduran nakliyatlarda tırın rotası bellidir ancak tırın dolu olmadığı nakliyatlarda tır birden fazla lokasyona uğrayarak teslimatlarını tamamlar. İlk olarak

hangi rotaya uğrayacağı tarihsel verilerden yararlanılarak veri madenciliği yöntemleri ile belirlemiştir (Şahin, 2010, s. 1-4).

Köksal ve arkadaşları imalat sektöründeki kalite süreçlerinde veri madenciliği yöntemlerinin kullanımı hakkında yaptıkları çalışmada veri madenciliği yöntemlerini tanımlanmış ardından 1997-2007 yılları arasını kapsayan literatür çalışması yapmışlardır. Uygulama bölümünde ise bir döküm fabrikasında döküm hatalarına neden olan değişkenlerin belirlenmesinde veri madenciliği yöntemlerinden olan karar ağacı algoritması yöntemi kullanılmıştır (Köksal vd., 2010, s. 10).

Zengin ve arkadaşları yapmış oldukları çalışmada eğitim sektöründe bulunan ve kurumların veri tabanlarında bulunan yüksek miktardaki verilerin veri madenciliği yöntemleriyle analiz edilerek kurum için anlamlı bir yapıya dönüştürülmesi hedeflenmiştir. Çalışmada Bilgisayar Öz Yeterlilik ölçeği kullanılmış ve çalışma grubuna uygulanmıştır. Çalışmada kümeleme varyans testleri ve karar ağacı algoritmasından yararlanılmıştır (Zengin vd., 2011, s. 4028).

Fu yapmış olduğu çalışmada veri madenciliği yöntemleri ile zaman serisi kavramları üzerinde durmuştur. Çalışmada son on yılda yıl da yapılan çalışmaların karmaşıklığı nedeniyle çalışmaları inceleyerek indeksleme, benzerlik ölçümü ve madencilik kavramları altında kategorize etmiştir (Fu, 2011, s. 164).

Irmak ve arkadaşları, gelecekte hastanelerin hasta yoğunluklarının belirlenmesinde veri madenciliği yöntemlerinden faydalanmışlardır. Hastane veri tabanından elde edilen veriler temizlenerek analize uygun hale getirilmiştir. Ardından çeşitli tahmin modelleri kullanılarak modeller arası karşılaştırılmalar yapılmıştır. Uygulama sonuçlarına göre Üstel düzgünleştirme yöntemlerinden Winter Modeli ve ARIMA yöntemlerinden ARIMA (3,1,0)(1,0,0) modelleri iyi sonuç üretmişlerdir. İki yöntem karşılaştırıldığında ise Winter Modelinin diğer modellere göre daha iyi çalıştığı ve gerçekleşen değerlere en yakın değerler ürettiği görülmüştür (Irmak vd., 2012, s. 101).

Çataloluk yapmış olduğu tez çalışmasında tıp alanında medikal verilerin veri tabanlarında oldukça büyük yer kapladığı ve bu verileri veri madenciliği yöntemleri ile anlamlı kural çıkarılarak hekimlere ve hastalara fayda sağlamayı amaçlamıştır. Bu kapsamda k-en yakın komşu algoritmasından ve k-means algoritmaları detaylı olarak incelenmiştir. Ayrıca hastalıkların daha önceden teşhisini sağlayacak ve hastalıklar arasındaki ilişkiyi tespit etmek amacıyla karar destek sistemi tasarlamıştır (Çataloluk, 2012, s. 1).

Dudas ve arkadaşları yapmış oldukları çalışmada, işletmelerin herhangi bir konuda stratejik karar verme aşamasında veri madenciliği yöntemlerinden ve optimizasyon yöntemlerinin ortaklaşa kullanımından daha iyi sonuçlar elde edebileceğini ortaya koymuşlardır (Dudas vd., 2013, s. 824).

Şengür yapmış olduğu tez çalışmasında öğrencilerin akademik başarılarını veri madenciliği yöntemleri ile tahminlemeye çalışmıştır. Veriler Fırat Üniversitesi Bilgisayar Öğretim Teknolojileri bölümünün veri tabanından alınmıştır. Tahminlemeler yapay sinir ağları ve karar ağaçları yöntemleri ile yapılmış ve daha sonra karşılaştırılmıştır. Yapay sinir ağlarının karar ağaçları yöntemine oranla daha başarılı olduğu görülmüştür (Şengür, 2013, s. 3).

Martin ve arkadaşları yaptıkları çalışmada İspanya yollarında güvenliğin iyileştirilmesi amacı ile veri madenciliği tekniklerini kullanarak kaza sayısı ile tehlikeli yollar arasındaki ilişki saptanarak kazaların en aza indirilmesi amaçlanmıştır (Martin vd., 2014, s. 607).

Dominic ve arkadaşları yapmış oldukları çalışmada veri madenciliği yöntemlerini kullanarak inşaat maliyetlerinin tahminlenmesi üzerine model geliştirmişlerdir. Toplam 1600 adet inşaat projesi tahmin edilmiş ve modelin başarılı olduğu görülmüştür (D ve Dagbui, s. 682).

Son yıllarda veri madenciliği yöntemlerinin gelişmesiyle birlikte tüm sektörlerde kullanılmaya başlanmıştır. Bunlardan biri de perakende sektörüdür. Perakende sektöründe son yıllarda veri tabanlarında oldukça büyük veriler birikmiştir. Oluşan verilerin kullanılarak işletmeye katma değer sağlaması gerekmektedir.

Bu kapsamda Sağlam yapmış olduğu tez çalışmasında bir perakende işletmesinin veri tabanında bulunan verileri kullanarak işletmenin mağazalarını gruplandırmaya çalışmıştır. Sınıflandırma yöntemi olarak K-means algoritması, k-medoids algoritması ve Random Forest algoritmaları kullanılmıştır (Sağlam Hadık, 2015, s. 1).

Jothi ve arkadaşları, veri madenciliği yöntemleri ile sağlık alanında yapılmış olan bilimsel çalışmalar ve algoritmaları inceleyip sonuçlarını özetler şeklinde sunmuşlardır (Jothi vd., 2015, s. 306).

Deniz yapmış olduğu tez çalışmasında, veri madenciliği yöntemlerinden K-means ve Two-Step algoritmalarını ve Euro 6 emisyon sertifikası verilerini kullanarak yakıt tüketimi ve emisyon seviyelerini etkileyen parametreleri sınıflandırmıştır (Deniz, 2016, s. 29).

Dalman yapmış olduğu tez çalışmasında, akaryakıt sektöründe en önemli problemlerinden biri sızıntı tespitidir. Sızıntı tespiti tanklarda meydana gelen kalibrasyon kaynaklı farklar ile karıştırılmaktadır. Çalışmada bu kapsamda veri madenciliği yöntemlerini kullanarak sızıntı tespitini kolaylaştırmak ve iyileştirmek amaçlanmıştır (Dalman, 2017, s. 1).

Gürbüz ve arkadaşları yapmış oldukları çalışmada, çeşitli sınıflandırma yöntemlerini kullanarak meydana gelebilecek tramvay arızalarının en aza indirilmesi için veriler üzerinden kurallar oluşturmuş ve sınıflandırma tekniklerini karşılaştırmışlardır (Gürbüz ve Turna, 2018, s. 568).

Dos Santos ve arkadaşları yapmış oldukları çalışmada 2009-2018 yılları arasında çeşitli bilimsel veri tabanlarından incelenen bilimsel çalışmalar veritabanlarına, kullanılan tekniklere, kullanılan program dillerine göre sınıflandırılmış ve sonuçlar ayrıntılı olarak verilmiştir (Dos santos vd., 2019, s. 1).

Schuh ve arkadaşları yapmış oldukları çalışmada üretim sektöründe süreç planlamada veri madenciliği yöntemlerinin kullanılmasının öneminden bahsetmişlerdir (Schuh vd., 2020, s. 48).

Hindistan tarıma büyük ölçüde bağımlı bir ülkedir. Önceden tarımı etkileyecek faktörlerin belirlenmesi bu yüzden çok önemlidir. Bu kapsamda Kamath ve arkadaşları çeşitli madencilik yöntemlerini kullanarak ürünün verimini tahmin etmeyi amaçlamışlardır (Kamath ve Patil, 2021, s. 402).

Mitroshin ve arkadaşları yapmış olduğu çalışmada karayolu taşımacılığında veri madenciliği yöntemlerini kullanarak çeşitli altyapı firmaları ve devlet kurumları için destek modelleri önermişlerdir (Mitroshin vd., 2022, s. 462).

Almuttar ve arkadaşları yapmış oldukları çalışmada, COVID-19 değerlendirmek amacıyla k-means yaklaşımı geliştirmişlerdir (Almuttar ve Alkhafaji, 2023, s. 1).

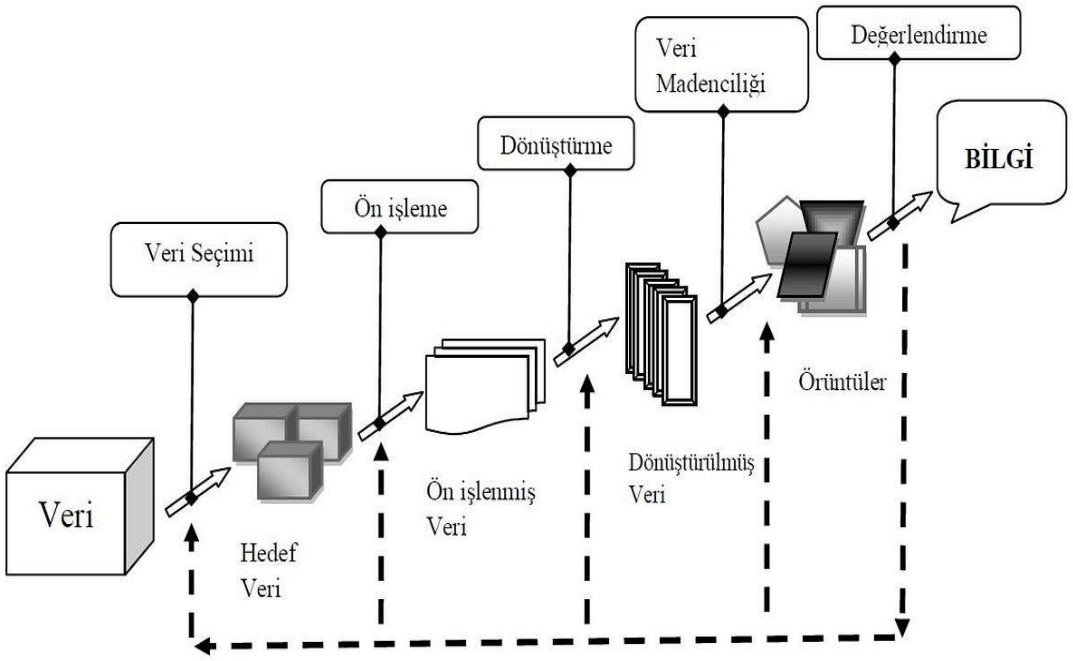
1.5. Veri Madenciliği Süreci

Veri madenciliği süreci, kendi içerisinde çeşitli adımlardan oluşmakta ve her adımda yapılabilecek bir hata diğer adıma yansımakta ve sonucu olumsuz etkileyebilmektedir. Bu nedenle veri madenciliği süreci dikkatli ve titizlikle yürütülmesi gereken bir süreçtir.

Veri madenciliği süreci, ilk öncelikle başlangıçta farklı alanlar için uygulamacılar tarafından farklı şekillerde yürütülmekteydi. Ancak karmaşıklığın önüne geçmek, sürecin standartlaşması ve meydana gelen farklılıkları en aza indirmek için 1996 yıllarının sonlarına doğru standart bir süreç olan ve en yaygın olarak kullanılmaya başlanan CRISP-DM (Cross Industry Standard Process for Data Mining) veri madenciliği süreç modeli geliştirilmiştir.

Veri madenciliği süreç modeli ile ilgili ilk çalışmalar 1996 yıllarının sonlarına doğru SPSS, NCR ve Daimler-Benz tarafından başlatılmıştır. Yapılan çalışmalar ve düzenlemeler sonucunda süreç ile ilgili ortaya çıkan ilk sürümü Brüksel'de düzenlenen 4.CRISP-DM workshop'ında sunulmuştur. CRISP-DM için 2006-2008 yılları arasında ikinci sürüm için çalışmalara başlanmış olsa da başarısız olunmuştur (Akpınar, 2017, s. 9).

CRISP-DM veri madenciliği süreçlerinde tüm işletmeler için atılması gereken adımları en ayrıntısına kadar açıklayan bir süreçtir. CRISP-DM süreci verilerimizin en iyi şekilde analiz etmemizi sağlarken aynı zamanda hata oranını minimum seviyeye indirerek zaman kaybı yaşamamıza engel olmaktadır.



Şekil 1. CRISP-DM Süreci

Kaynak: Tuzcu, S. (2018). *Ders yönetim sistemi tabanlı veri madenciliği ve öğrenme analitiği* [Yüksek lisans Tezi]. Eskişehir Osmangazi Üniversitesi.

1996 yılında tüm sektörlerin kullanabilmesi adına geliştirilmiş olan CRISP-DM veri madenciliği süreci aşağıdaki aşamalardan oluşmaktadır. İlgili sıralama kesin olamamakla birlikte aşamalar arasında yer değiştirmeler yaşanabilmektedir.

- Problemin tanımlanması
- Verilerin toplanması ve anlaşılması
- Verilerin hazır hale getirilmesi
- Modelleme
- Değerlendirme
- Uygulama

1.5.1. Problemin Tanımlanması

Problemin tanımlanması, veri madenciliği sürecinin en önemli adımıdır. İşletmelerin ve organizasyonların veri madenciliği sürecine başlamadan önce var olan problemlerini, fayda kriterlerini amaçlarını açık bir şekilde tanımlamalı ve açık bir dille ifade etmelidir. Yapılacak olan doğru tanımlamalar veri madenciliği sürecinin aksamamasına, model başarı düzeyinin etkilenmemesine ve katlanılacak olan maliyetlerin minimum düzeyde olmasına katkı sağlayacaktır (Ünsal, 2023, s. 21).

CRISP-DM sürecinde temel olarak problemin tanımlanması aşamasında aşağıdaki faaliyetlerin gerçekleşmesi beklenir.

- Problem için iş hedeflerinin belirlenmesi
- Problemin mevcut durumun değerlendirilmesi
- Kullanılan yöntemin amacının belirlenmesi
- Uygulanacak yöntem doğrultusunda işlem adımlarının oluşturulması

1.5.2. Verilerin Toplanması ve Anlaşılması

Verilerin toplanması ve anlaşılması adımından önce problemin en iyi şekilde tanımlanmış olması gerekmektedir. Problemin en iyi şekilde tanımlandığına emin olunduktan sonra çeşitli yöntemler ile verilerin toplanılması sağlanır. Toplanan verinin anlaşılabilmesi adımında verilerin kalitesinin ve öz niteliklerinin belirlenmesi veri ile gerçekleştirilebilecek hipotezlerin geliştirilmesinin yanında örüntülerin ve alt veri kümlerinin belirlenmesi gerekmektedir.

1.5.3. Verilerin Hazırlanması

Verilerin hazırlanması adımı, ham verinin modelde kullanılabilmesi amacıyla işlenerek modele uygun hale getirilmesi sürecidir. Başka bir ifadeyle verilerin uygun formatta dönüştürülmesi işlemidir. Süreç içerisindeki en çok hatanın yapıldığı ve sık sık verinin tekrardan düzenlenmesine ve zaman kaybına neden olan kısımdır. Verilerin hazırlanma aşaması; veri toplama, değer biçme ve veri temizleme, veri dönüştürme ve veri indirgeme aşamalarından oluşmaktadır.

Veri Toplama

Problemde kurulacak olan modeli en iyi ifade edecek olan veri setlerinin toplandığı ve veri kaynakların belirlendiği adımdır. Verinin toplanmasında kullanılacak olan kaynaklar kuruluşun kendi veri kaynakları olabileceği gibi kurum dışı veri kaynaklarından da faydalanılabilir. Veri seçiminde dikkat edilmesi gereken bir diğer konu ise seçilecek olan veri setinin büyüklüğüdür. Seçilecek olan veri setinin küçük olması yapılacak olan çalışmanın eksik kalmasını ve amacına ulaşmasını engelleyebileceği gibi veri setinin oldukça geniş seçilmesi çalışmanın süresini uzatabilmektedir. Bu yüzden veri seçimi adımında oldukça titiz davranılmalıdır.

Değer Biçme ve Veri Temizleme

Veri toplama sonucunda farklı kaynaklardan elde edilen veri dizileri, farklı kodlama yapısına ve farklı ölçü sistemlerine sahip olduklarından birbirleri arasında veri uyumsuzluğu oluşacaktır. Oluşan veri uyumsuzluğu düzeyi oldukça fazla ise elde edilecek sonuç o kadar başarısız olmakta ve işlem süresi uzamaktadır. Oluşan uyumsuzluğu en aza indirmek içi çeşitli basit yöntemler kullanılmaktadır. Seçilen veriler içerisinde uyumsuzluk olabileceği gibi veri setleri içerisinde eksik verilerde olabilmektedir. Bu gibi durumlarda verilerin incelenerek temizlenmesi gerekmektedir.

Veri Dönüştürme

Toplanan ve temizlenen veri setinin modelde kullanılacak olan algoritmalara göre kendi içinde düzenlenmesidir. Ayrıca veri üzerinde yapılan düzenlemeler sonucunda problemde kullanılan algoritmanın başarısı etkilenecektir. Örneğin yapılan bir çalışmada yapay sinir ağları algoritması kullanılarak kategorik değişkenler evet/hayır olarak kategorize edilmesi, aynı çalışmanın karar ağaçlı algoritması ile yapılan çalışmasında ise iyi, orta ve kötü olarak gruplanmış olması modelin başarısını arttıracaktır (Akpınar, 2017, s. 19).

Veri İndirgeme

Problemlerde kaynak olarak kullanılan veri tabanları oldukça büyük gerekli/gereksiz veri yığınları ile doludur. Gerekli/gereksiz saklanan veri yığınları

yapılacak olan veri analizi çalışmalarını olumsuz etkileyecektir. Veri analizine başlamadan önce çeşitli veri indirgeme yöntemleri kullanılarak gereksiz veriler analizden çıkarılmalıdır. Bu yöntemler boyut sayısının azaltılması makine öğrenmesi ve faktör analizi teknikleridir.

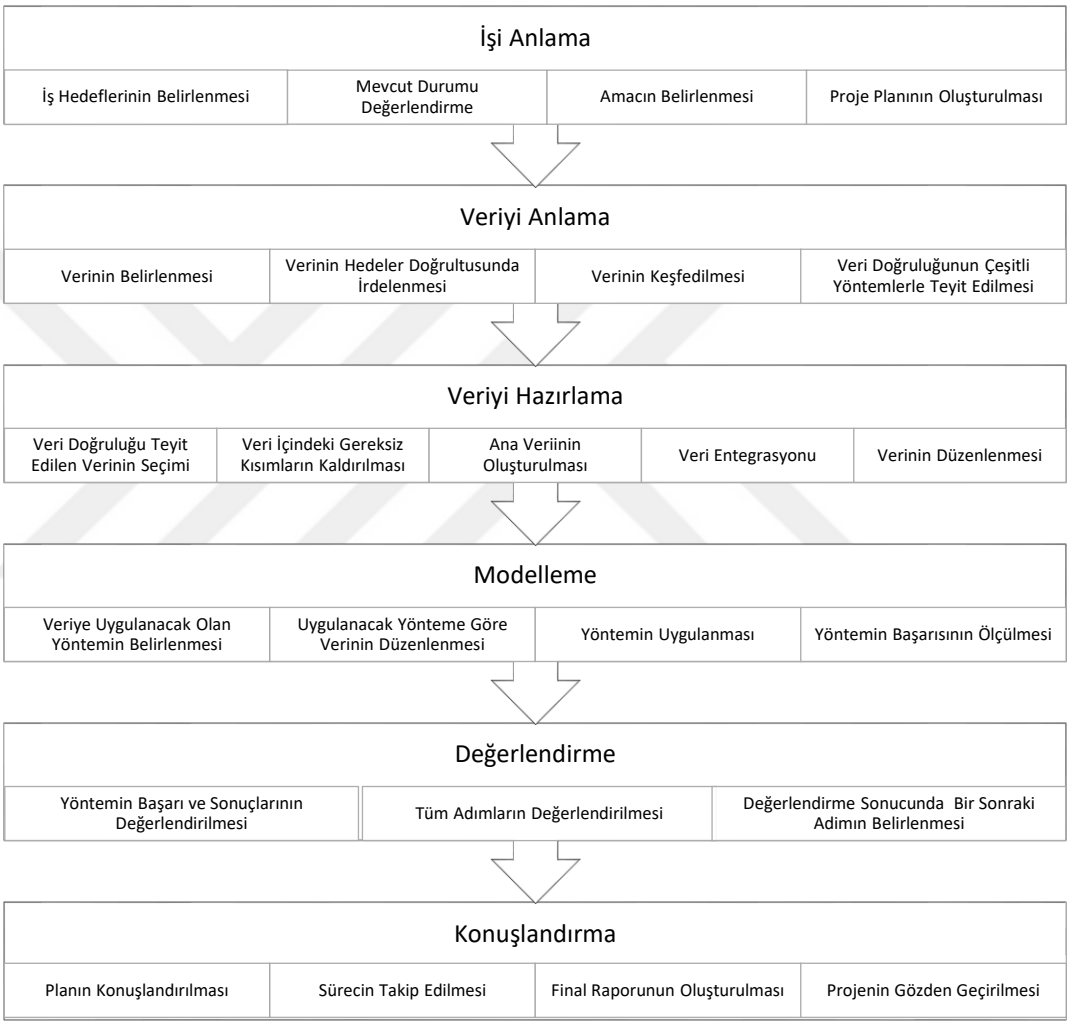
1.5.4. Modelleme

Modelleme adımına bakacak olursak ilk olarak oluşturulacak olan modelde kullanılacak olan algoritmalar belirlenmelidir. Ardından kullanılacak olan algoritmanın çeşitli deneysel testleri yapılarak sonuçlar değerlendirilmektedir. Ardından sonuçlara göre modelin kurulması ve sonuçların değerlendirilmesi olarak modelleme adımları sıralanabilmektedir. Modelin kurulmasında dikkat edilmesi gereken önemli noktaları aşağıdaki şekilde sıralamak mümkündür.

- Problemin amacına uygun veri setinin ve uygun yöntemin seçilmesi. Seçilen yöntemin ve veri analizi yapan uzmanın bilgi birikimi modelin başarı düzeyini etkileyecektir.
- Analizde kullanılacak olan özniteliklerin seçilmesi. Seçilen öznitelikler (bağımlı, bağımsız değişken vb..) modelin başarısını belirleyecek önemli noktalardan biridir. Bu yüzden öznitelikler probleme uygun olarak belirlenmelidir. Problemin amacına uygun olmayan özniteliklerin belirlenmesi hem zaman kaybına hem de modelin başarı düzeyine olumsuz etki yaratacaktır.
- Problemin amacına uygun olarak toplanan verinin, seçilen uygun yöntem ile uyumlu çalışması için verinin dönüştürülmesi.
- Seçilen verinin içerisinde modelin başarısını etkileyecek olan sıra dışı değerlerin (outlier) kontrol edildikten sonra veri içerisinden çıkarılması veya değerlerin farklı yöntemlerle (kayıp değer analizi) yeniden tanımlanması.
- Veri seçimi yapıldıktan sonra seçilen uygun yöntemler ile modelin test edilmesi gerekmektedir. Test sonucunda başarılı ve güvenilir sonucu veren modelin seçilmesi gerekir.

1.5.5. Değerlendirme

Modelin kurulduktan sonra uygulamaya geçmeden önceki gerekli kontrollerin yapıldığı son adımdır. Bu yüzden kurulan modelin son kez titizlikle kontrol edildiği ve kurulan modelin işletme/organizasyon amaçlarına ulaşmasına yardımcı olacağına emin olunmalıdır. Kısacası veri madenciliği çıktılarının kullanılabileceğine dair kararın verildiği adımdır.



Şekil 2.Veri Madenciliği Süreci

Kaynak: Uyumaz, Ö. (2017). *Bankacılık sektöründe pazarlama stratejilerinin belirlenmesinde sınıflandırma ve veri madenciliği* [Yüksek lisans tezi]. Beykent Üniversitesi.

Kısacası veri madenciliği süreci, uzun bir süreç olup dikkatli ve titizlikle yürütülmesi gereken bir süreçtir. Yukarıda Şekil.2 de veri madenciliği süreçlerinin ayrıntılı olarak verilmiştir.

1.6. Veri Madenciliği Yöntemleri ve Algoritmaları

Veri madenciliği modelleri, kullanıldıkları alanlara göre farklı değişik modellere ayrılsalar da genel olarak tahminleyici ve tanımlayıcı olarak iki yaklaşıma ayrılmaktadırlar. Tahminleyici modellerde, çeşitli yöntemler kullanılarak analiz edilmiş olan veri grupları değerlendirilerek daha başarılı modellerin ortaya konması ve ortaya koyulan modeller doğrultusunda sonuçları bilinmeyen veri grupları için sonuçların doğru olarak tahminlenmesi amaçlanmaktadır. Tahminleyici modeller, sınıflama ve regresyon modelleri olarak ikiye ayrılmaktadırlar (Dünder, 2021, s. 11).

Tanımlayıcı modeller ise daha çok sonuçların değerlendirirken karar vermeye yardımcı olmak amacıyla verilerdeki örüntülerin belirlenmesi işlemi olarak tanımlanabilmektedir (Şekeroğlu, 2010, s. 35).

Ayrıntılı olarak veri madenciliği modelleri aşağıdaki Tablo.2'de gösterilmiştir.

Tablo 2. Veri Madenciliği Modelleri ve Teknikleri

Veri Madenciliği Modelleri	
Tanımlayıcı Modeller	**Tahminleyici Modeller**
Birliktelik Kuralları	Karar Ağaçları
Kümeleme	Bayes
Ardaşıklık Keşfi	Yapay Sinir Ağları
Özetleme	Zaman Serisi Analizi
Tanımsal İstatistik	Karar Destek Makineleri
İstisnai Analiz	

Kaynak: Şekeroğlu, S. (2010). *Hizmet sektöründe veri madenciliği uygulaması* [Yüksek lisans tezi]. İstanbul Teknik Üniversitesi.

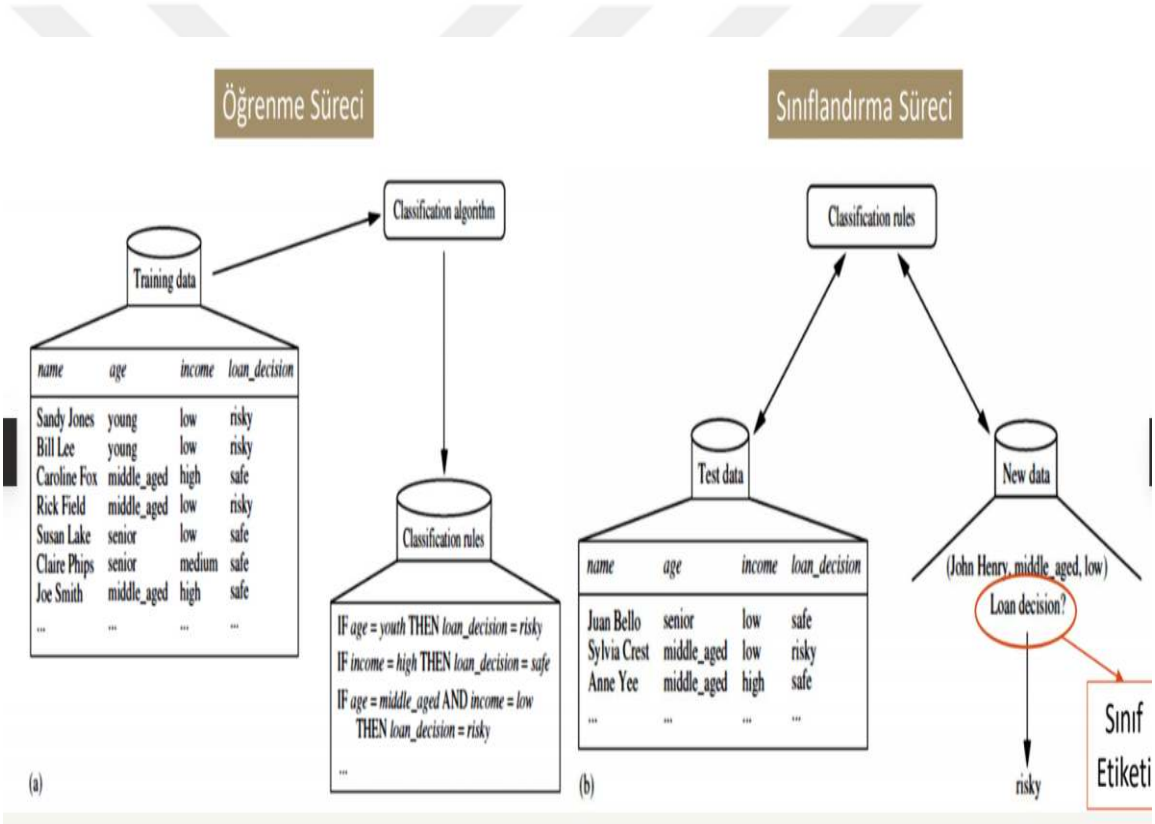
Veri madenciliği modelleri fonksiyonlarına göre 3'e ayrılır;

- Sınıflama Modelleri
- Kümeleme Modelleri
- Birliktelik Kuralları ve Modelleri

1.6.1. Sınıflandırma Modelleri

Veri madenciliğinde sınıflandırma önemli veri analiz konularından bir tanesidir. Sınıflandırma bir bankanın müşterisine kredi verme işleminde müşterinin güvenli biri olup olmadığının tespitinde, kanser tedavisinde uygulanacak olan tedavinin tespit edilmesi gibi önemli veri sınıflarınım oluşturulmasında kullanılmaktadır. Sınıflandırma yapılacak olan sınıflar kesikli değerler ile ifade edilir.

Sınıflandırma sürecine bakıldığında temel olarak iki adımlı bir süreç olduğu görülmektedir. Birinci adımda var olan sınıflar kullanılarak algoritmalar inşa edilirken ikinci adımda algoritma ve veriler kullanılarak veri sınıflandırması işlemi tamamlanır.



Şekil 3. Sınıflandırma Süreci

Kaynak: Han, J., Kamber, M. ve Pei, J. (2012). Data mıning concepts and Techniques. Morgan Kaufmann.

Sınıflandırma sürecinde hazırlık aşamasında sonuçların doğru ve sorunsuz olarak sınıflandırılabilmesi için daha öncesinde CRISP-DM sürecinde bahsedilen adımların tek tek olarak uygulanması gerekmektedir.

1.6.2. Kümeleme Modelleri

Kümeleme algoritmalarında veri setlerinde bulunan veriler arasından birbirlerine en yakın özelliğe sahip olan elemanların bulunması amaçlanmaktadır (Ünsal, 2023, s. 29).

Kümeleme modelleri, veri setleri içerisinde bulunan elemanların tüm diğer elemanlar üzerindeki değerlerini hesaplayarak ortaya çıkacak olan grup sayılarına odaklanırlar. Veri setlerinde bulunan elemanların benzerlik oranlarını saptamak amacıyla uzaklık ölçüleri ve korelasyon ölçütleri kullanılmaktadır (Kalaycı, 2018, s. 349).

Veriler arasındaki uzaklıklar Öklid Manhattan Minkovski gibi oldukça sık kullanılan uzaklık hesaplama yöntemleri kullanılarak belirlenmektedir. Bunun dışında kullanılan veri benzerliğini ölçen yöntemler de mevcuttur. Bu yöntemler DICE, JACCARD, COSINE, OVERLAP olarak sıralanabilir.

Kümeleme yöntemleri DNA analizi, iş zekası, biyoloji, görüntü ve patern belirleme vb. bir çok farklı alanda kullanılmaktadır.

Kümeleme yöntemleri hiyerarşik ve bölümlemeli (partitioning) yöntemler olarak sınıflandırılmaktadır.

Tablo 3. Kümeleme Modelleri Sınıflandırma

Kümeleme Modelleri	
Hiyerarşik Yöntemler	**Bölümlemeli (Partitioning) Yöntemler**
Toplaşım Kümeleme Algoritmaları	Yer Değiştirme Algoritmaları
Bölünür Kümeleme Algoritmaları	Olasılıksal Algoritmalar
	k-means Yöntemler

Kaynak: "yazar tarafından oluşturulmuş/derlenmiştir"

Hiyerarşik Yöntemler

Hiyerarşik yöntemlerde kümeleme yapılırken küme ağacı oluşturulur.

Toplaşım Kümeleme Algoritmaları

Toplaşım algoritmalarında uzman kişi tarafından önceden küme sayısı belirlenir. Başlangıçta veri tabanında var olan nesnelerin her birini bir küme olarak görür. Bu kümeleri birleştirerek önceden uzman tarafından belirlenen küme sayısı kadar küme oluşturulur.

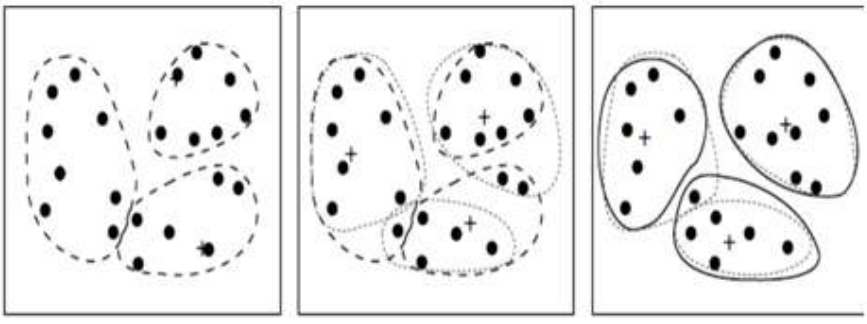
Bölünür Kümeleme Algoritmaları

Bölünür kümeleme algoritmalarına bakıldığında ise toplaşım kümeleme algoritmalarından farklı olarak başlangıçta veri tabanında yer alan tüm elemanları bir kümeye toplayarak birbirine benzemeyen elemanları kümeden dışarı atarak uzman tarafından belirlenmiş küme sayısı kadar kümeye bölünür.

Bölümlemeli (Partitioning) Yöntemler

k-means Algoritması

k-means algoritmasında amaç küme içerisinde benzerliği artırırken diğer kümeler ile olan benzerliği azaltmaktır. k-means yönteminde daha önceden belirlenmiş n sayıda kümeyi temsilen bir küme merkezi bulunur. Bu merkez kümenin tam ortasında bulunur. Bu merkezin bulunmasında ortalama veya medoid yöntemlerden faydalanılır. Kümeler arasındaki benzerlik oranının belirlenmesinde Öklid uzaklığı kullanılır.



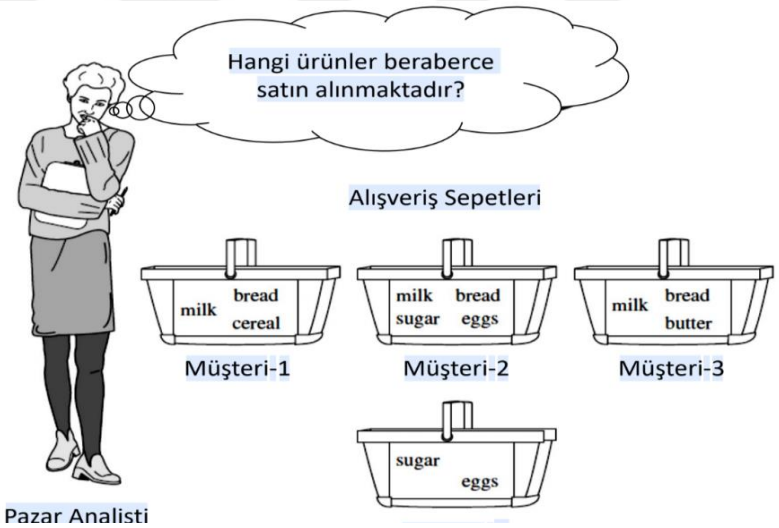
Şekil 4.k-means Algoritması

Kaynak: Silahtaroğlu, G. (2016). *Veri Madenciliği Kavram ve Algoritmaları.* Papatya Yayıncılık.

1.6.3. Birliktelik Kuralları ve Modelleri

Birliktelik kuralları, veri madenciliğinin en dikkat çeken konularından birisidir. Birliktelik kuralları bir veri tabanı içerisinde yer alan bir nesnenin veya bilginin diğer kayıt altına alınmış nesne veya bilgi ile bağlantısını açıklamaya çalışır. Günümüzde birçok alanda kullanılsa da barkod sistemlerinin gelişmesi ile birlikte en çok pazar sepeti analizi uygulaması ile karşımıza çıkmaktadır.

Pazar sepeti analizi ilk defa 1993 yılında Agrawal tarafından ele alınmıştır. Pazar sepeti analizi, müşteri satın alma alışkanlıklarının daha önceden veri tabanında var olan bilgiler doğrultusunda ortaya çıkarılması amaçlanmaktadır (Silahtaroğlu, 2016, s. 39). Analiz sonucu ortaya çıkan bulgular doğrultusunda alışveriş merkezlerinin alanlarının belirlenmesi, ürün yerleşimlerinin belirlenmesi ve satılacak ürünlerin belirlenmesinde karar destek sistemi olarak kullanılmaktadır.



Şekil 5.Market Sepet Analizi

Kaynak: Han, J., Kamber, M. ve Pei, J. (2012). *Data mınıng concepts and Techniques*. Morgan Kaufmann.

Örneğin, bilgisayar alan bir müşterinin bilgisayar güvenliği sağlamak amacıyla virüs programı alması arasında kuvvetli bir ilişki bulunmuş ise bilgisayar satış reyonuna virüs ile ilgili stantı koyularak iki ürününde satışı arttırılmış olur.

AIS,SETM,APRİORİ,APRİORİTİD,FP-GROWTH algoritmaları birliktelik kurallarının ortaya çıkarılması için modellenmiş algoritmalardır. Günümüzde oldukça sık kullanılan algoritma APRİORİ algoritmasıdır.

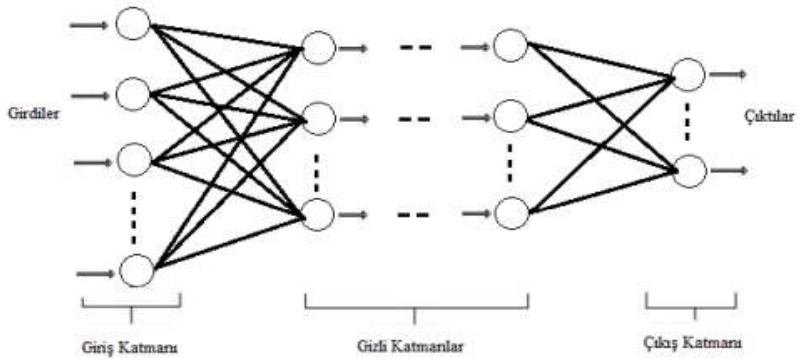
APRİORİ Algoritması

Apriori algoritması birliktelik kurallarının belirlenmesinde en çok kullanılan ve tercih edilen bir yöntemdir. Aprriori algoritması geniş kümeler içerisindeki birliktelik kurallarının ortaya çıkarılması amacıyla çok sayıda tarama uygulayan bir yöntemdir. Birinci taramada sık tekrarlanan kümeleri tespit eder. İkinci taramada ise aday gözlem kümelerinin oluşturulması için kullanılır. En son üçüncü tarama işleminde ise ikinci tarama sonucu elde edilen aday gözlem kümlerini kullanır. Bu işlem herhangi bir küme kalmayıncaya kadar tekrarlanır (Dünder, 2021, s. 25).

1.6.4. Yapay Sinir Ağları

Yapay sinir ağları, insan beyninin işleyişinin taklit edilmesi sonucu ortaya çıkmış bir yapıdır. İşleyiş olarak bakıldığında dışarıdan aldığı bilgileri işleyerek çıktıları dışarı aktarır. Dışarıdan aldığı bilgileri nasıl işlediği belli olmadığından yapay ve çıktı sonuçlarını nasıl oluşturulduğunu açıklama yeteneği olmadığından yapay sinir ağlarına olan güveni sarsmaktadır.

Yapay sinir ağlarının yapısına bakıldığında yapay sinir hücrelerinin birbirlerine bağlanması sonucu oluşmuş yapılardır. Yapısında giriş katmanı, gizli katmanlar ve çıkış katmanları bulunur.



Şekil 6.Yapay Sinir Ağları Yapısı

Kaynak: Elmas, Ç. (2016). *Yapay Zeka Uygulamaları* (3. b.). Seçkin Yayıncılık.

Giriş katmanında dışarıdan alınan bilgiler doğruca gizli katmana iletilir. Herhangi bir işlem yapılmaz.

Gizli katmanda ise giriş katmanından aldığı bilgileri işleyerek çıkış katmanına iletilir.

En son katman olan çıkış katmanında gizli katmandan gelen bilgi doğrultusunda bilgiyi işleyerek dış dünyaya aktarır.

Yapay Sinir Ağları Özellikleri

- Veri sayısı çoğaldıkça çözüm süresi ve doğruluk payı azalacaktır. Ancak yapay sinir ağları bu doğrultuda tasarlanmıştır.
- Sadece nümerik bilgiler ile çalışabilmektedirler.
- Veri içerisinde bozulmuş eksik vs. verileri kolayca işleyebilmektedirler.
- Hata paylarına sahiptirler.

Yapay Sinir Ağları Avantajları

- Veri sayısı çoğaldıkça analiz süreleri ve doğruluk payları diğer yöntemlere göre daha iyi sonuçlar vermektedir.
- Süreç tamamlandıktan sonra ilgili verileri hafızasında saklayabilme yeteneğine sahiptir.
- Diğer yöntemlerin oluşturamadığı ve çözemediği matematiksel modelleri oluşturmada ve çözmede oldukça başarılıdır.

Yapay Sinir Ağları Dezavantajları

- Gizli katmanda nasıl bir sürecin işlediği bilinememektedir.
- Yapay sinir ağlarının uzunluğunun ne kadar olması ve ne zaman bitirileceğine dair herhangi bir bilgi bulunmamaktadır.
- Ağ yapısının belirlenmesi ile ilgili herhangi bir kural yoktur.

1.6.5. Genetik Algoritmalar

Genetik algoritmalar, doğadaki canlıların geçirmiş oldukları biyolojik süreçleri örnek alarak geliştirilmiş bir yöntemdir. Genetik algoritmalar en çok matematiksel modelin kurulamadığı ve çözüme ulaşılamadığı durumlarda kesin yöntem olarak kullanılmaktadır. Elde edilen sonuçlardan daha iyi sonuçların elde edilmesi için genetik algoritmalar çaprazlama ve değiştirme operatörlerini kullanır. Grup üzerinde arama yaparak en iyi çözümü araması ile diğer yöntemlerden ayrılmaktadır (Elmas, 2016, s. 403).

Genetik algoritmalar, doğadaki canlılarda olduğu gibi doğal seçim ve genetik kurallar sürecine dayanmaktadır. Doğadaki canlıların en küçük yapı taşı gendir. Genler bir araya gelerek kromozomları meydana getirir. Genetik algoritmalarda da aynı süreç işlemektedir. Başlangıçta konu ile ilgili birimleri gen olarak alan genetik algoritmalar genler sayesinde kromozom üreterek en iyi sonuca ulaşmaya çalışır. Genetik algoritmanın işlem adımları aşağıda gösterilmiştir.

- Çözülmesi istenen problemin genetik gösterim ile ifade edilmesi
- Genetik gösterim sonrası başlangıç popülasyonun belirlenmesi
- Uygunluk değerinin ve kromozom uygunluk değerinin hesaplanması
- Oluşturulan her bir kromozom için seçilme olasılığının ifade edilmesi
- Çaprazlama operatörünün uygulanması
- Mutasyon operatörünün uygulanması
- İşleyişin devamı sonucu yeni popülasyonun oluşturulması

Genetik algoritmalar günümüzde oldukça sık olarak eğitim, ulaşım, lojistik, bilgi teknolojileri, üretim, planlama gibi birbirinden farklı sektörlerde kullanılmaktadır. Literatüre bakıldığında ise daha çok araç rotalama ile ilgili problemlerde çözüm önerisi sunmaktadır.

1.7. Veri Madenciliğinde Kullanılan Gelişmiş Teknikler

Günümüz dünyasında, işletmeler gelişen teknoloji ile birlikte elde ettikleri ham verileri veri tabanlarında tutmaktadırlar. Bu veriler gün geçtikçe artmaktadır. Veriler, işletmelerin daha kaliteli hizmet verebilmesi, işletmelerin geleceğe taşınmalarında ve diğer işletmeler ile rekabet edebilmelerinde önemli rol oynamaktadır. Bu yüzden işletmeler bu verileri en iyi şekilde işleyerek anlamlı ve kullanışlı bir hale getirmek zorundadırlar.

Gelişen teknoloji ve yazılım ile birlikte işletmeler yeni yöntemler geliştirerek yapılarında atıl olarak bulunan veriler arasındaki ilişkileri ortaya çıkarmışlar ve veri tabanlarında bu şekilde saklamışlardır. Bu durumu iyi değerlendiren işletmeler gelecekte diğer işletmeler ile rekabet avantajı elde ederken, ellerinde bulunan verileri işlemeyen şirketlerin diğer işletmeler ile rekabet etmesi zorlaşacaktır.

Veri madenciliğinde günümüzde oldukça kullanılan gelişmiş yöntemlere bakıldığında Web Madenciliği, Metin Madenciliği, Multimedya Madenciliği ve Büyük Veri kavramlarının oldukça sık şekilde kullanıldığı görülmektedir.

1.7.1. Web Madenciliği

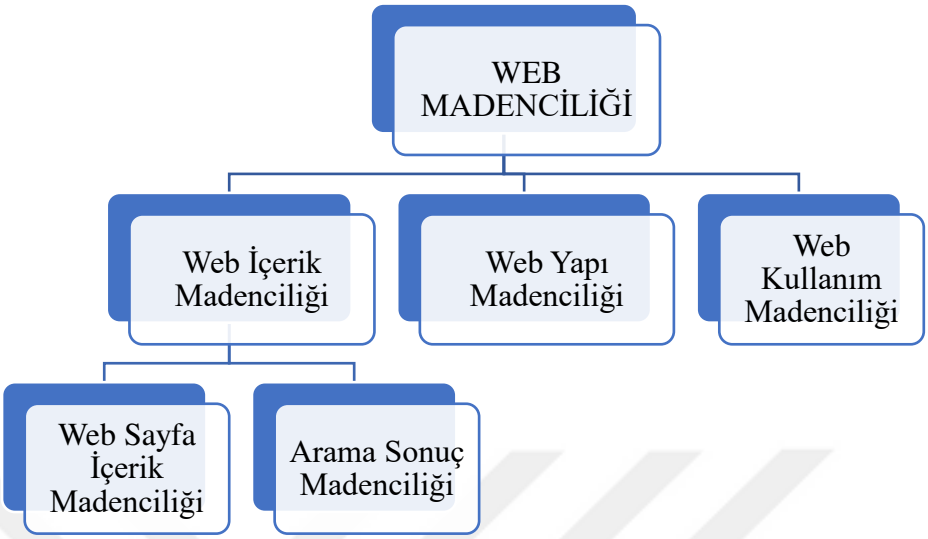
Web madenciliği bankacılık, e-ticaret, belge paylaşımı ve kurumsal işler olmak üzere farklı alanlarda kullanılmaktadır. Günümüzde web ortamındaki verilerin karmaşık ve analiz edilmesinin zor olduğunu düşünürsek web madenciliği bu konuda oldukça ayrı bir önem kazanmaktadır (Özseven ve Düğenci, 2011, s. 154).

Web madenciliği 1996 yılında Oren Etzioni tarafından ortaya atılmış bir yöntemdir. İlk ortaya çıktığında Web içerik Madenciliği ve Web Kullanım madenciliği olarak sınıflandırılsa da daha sonrasında Web Yapı Madenciliği olarak 3 bir sınıf eklenmiştir.

Web içerik madenciliği internet sitelerindeki içeriklerin en iyi şekilde analiz edilmesinde görev alır (Kantardziç, 2003, s. 358).

Web kullanım madenciliği ise sunucu üzerindeki kullanıcı verilerine erişir.

En son olarak Web yapı madenciliği ise internet siteleri arasındaki web yapı verisini inceleyerek anlamlı sonuç çıkarmaya çalışır (Özseven ve Düğenci, 2011, s. 154).



Şekil 7.Web Madenciliği ve Alt Bölümleri

Kaynak: Gezer, M., Erol, Ç. ve Gülseçen, S. (2007, 31 Ocak-2 Şubat). *Bir web sayfasının web madenciliği ile analizi* [Bildirimi sunumu]. IX. Akademik Bilişim Konferans Bildirileri, Kütahya

1.7.2. Metin Madenciliği

Metin madenciliği, metin içeriğini bir veri kaynağı olarak görüp işleyerek anlamlı bir hale getirmeye çalışan veri madenciliği çalışmasıdır (Şeker, 2015, s. 30).

Metin madenciliği çalışmaları günümüzde çok sıklıkla kullanılsa da genel itibarı ile veri yığınının çok olduğu, analizlerin yapılmadığı metinlerde metnin sınıflandırılmasının, kümelenmesinin, anlamlı bir hale getirilmesinin ve görselleştirilmesinde oldukça sık kullanılmaktadır (Artsın, 2020, s. 345).

Metin Madenciliği işlem adımlarına bakıldığında aşağıdaki gibi olduğu görülmektedir.

- Metin içerisinde yer alan büyük veriden anlam çıkarılmak istenen özelliğin belirlenmesi
- Belirlenen özelliğin çeşitli algoritmalar ve yöntemler ile işlenmesi
- İşlenen veri sonucunda anlamlı ve yapılandırılmış verinin ortaya çıkarılması

1.7.3. Multimedya Madenciliği

Multimedya madenciliği, görsel temelli yapılarda görsellerin uzaktan algılanabilmesi için verilerin sayısallaştırılmasına dayanan veri madenciliği yöntemlerinden biridir (Girija ve Srivatsa, 2006, s. 595). Günümüzde birçok alanda kullanılmakta olup web içerik madenciliği olarak da bilinmektedir.

Multimedya madenciliği web sayfa içerik madenciliği ve arama sonuç madenciliği olarak ikiye ayrılmaktadır. Web sayfa içerik madenciliği web sayfası üzerinde yer alan ses görüntü vs. gibi belgelerin içerisinden yararlı bilgiler elde etmeye çalışırken web arama sonuç madenciliği sonuçların getirilmesi adına verileri işlemektedir.

1.7.4. Büyük Veri

Günümüzde veri tabanlarında biriktirilmiş âtıl olarak bekleyen oldukça fazla veri mevcuttur. Bu veriler gün geçtikçe çeşitlilik ve hacim yönünden artış göstermekte olup gelişen teknoloji ile kullanılabilir hale getirilmesi ile birlikte büyük veri kavramı ortaya çıkmıştır (Doğan ve Arslantekin, 2016, s. 21). Günümüzde büyük veri kavramı önem kazanmasıyla birlikte çeşitli alanlarda kullanılmaya başlanmıştır. Aşağıda kullanım alanları ile ilgili detaylar paylaşılmıştır.

Eğlence sektöründe; ülkemizde son yıllarda eğlence sektöründe büyük gelişmeler meydana gelmiştir. Bu gelişmeler sonucunda eğlence sektörüne olan ilgi gün geçtikçe artmıştır. İlginin artması ile birlikte toplanan kişisel verilerde artmıştır. Büyük veri kavramı ile bu veriler işlenerek sektörün gelişmesine katkı sağlamaktadır (Aktan, 2018, s. 7).

Sağlık sektöründe; sağlık kuruluşlarına gelen bireylerin daha kolay hizmet alabilmeleri ve özel hizmetler sağlayabilmek adına veriler veri tabanlarında toplanarak işlenmektedir (Uyumaz, 2017, s. 22).

Taşımacılık sektöründe; GPS verilerinin veri tabanlarında biriktirilerek araç rotalaması için planların yapılması, maliyetlerin azaltılması vb konularda büyük veri kavramı kullanılmaktadır.

2. VERİ MADENCİLİĞİNDE SINIFLANDIRMA TEKNİKLERİ

Bu bölümde tez içerisinde kullanılacak olan ilgili teknikler ve literatür taraması sonucu literatürde kullanılan teknikler tablo4 de gösterilen şekilde 2 ana grup halinde detaylı olarak anlatılacaktır.

Tablo 4.Veri Madenciliğinde Sınıflandırma Teknikleri

İstatistiğe Dayalı Algoritmalar	Mesafeye Dayalı Algoritmalar
Naive Bayes Sınıflandırıcı	Karar Ağaçları
CHAID Algoritması	En Yakın Komşu (kNN) Algoritması
Regresyon	

Kaynak: "yazar tarafından oluşturulmuş/derlenmiştir"

2.1. İstatistiğe Dayalı Algoritmalar

Veri madenciliğinde sınıflandırma çeşitli içeriklere sahip değişkenlerin çeşitli yöntemler ile sınıflara ayrılarak sınıfların tahmin edilmesi olarak tanımlanabilir. Bayesyen yaklaşım, regresyon, zaman serileri analizi vb. istatistiksel yöntemler kullanarak sınıflandırma işlemleri gerçekleştirilir (Gemici, 2012, s. 28).

2.1.1. Bayesyen Sınıflandırma

Bayesyen sınıflandırma tekniği, istatistiğe dayalı sınıflandırma yöntemlerinden biridir. Daha önceden sınıflandırılmış olan verileri kullanarak gelecek olan yeni veri setlerinin hangi sınıfa gireceğini olasılıksal olarak hesaplayan bir yöntemdir. Başka bir tanıma bakacak olursak bir kayıt demetinin hangi sınıfa ait olup olmadığını belirleyen istatistik bir yöntem olarak açıklanmıştır.

Bayesyen sınıflandırma bayes teoremine dayanmaktadır.Yapılan sınıflandırma performansı , kolay uygulanabilirliğinden ve hızlı sonuçlar verdiğinden dolayı oldukça sık kullanılmaya başlanmıştır. Yöntemin avantajlı yanları olduğu gibi dezavantajlı yanları da bulunmaktadır.

Bayesyen sınıflandırma yönteminde kullanılan değişkenler kategorik olmak zorundadır. Sürekli değişkenler kullanılmadan önce kategorik veriye dönüştürülmesi gerekmektedir. Bu yöntemde sınıfların şartlı bağımsızlığı varsayımı geçerlidir. Bu varsayıma göre belirli bir sınıf içinde alan değer ve özelliklerin diğer alanların değerlerinden bağımsız olduğunu varsaymaktadır. Bu varsayım sayesinde yöntemin adımları kısalmaktadır (Uyumaz, 2017, s. 12).

Eşitlik 2.1'de C_{i} sınıf mantığı ile kurulmuş bir hipotez ise P(C/X) şartlı içeriğinde X biliniyorken C'nin gerçekleşme olasılığını verir. Eşitlik 2.1 de gösterilen ifade içerisinde yer alan P(X) ifadesi yerine eşitlik 2.2 de verilen ifade yerleştirilir ve gerekli işlemler uygulanırsa bayes teoremi eşitlik 2.3' deki gibi ifade edilmiş olur.

$$\mathrm{P}(C/\mathrm{X}) = \frac{P(X/C)P(C)}{P(X)} \tag{2.1}$$

$$\sum_{J=1}^{m} P(X_i / C_J)P(C_J) \tag{2.2}$$

$$P(C_1 / X_i) = \frac{P\left(X_i / C_1\right)P\left(C_1\right)}{P\left(X_i\right)} \tag{2.3}$$

P(C): C terimi için önsel olasılık

P(C/X): X için C' nin koşullu olasılığını verir.

P(X/C): C için X' nin koşullu olasılığını verir.

P(X): X terimi için önsel olasılık olarak ifade edilmektedir.

. Bayesyen algoritması daha iyi anlaşılabilmesi adına Tablo 5 ' de yer alan veriler kullanılarak algoritmanın tüm adımları ayrıntılı olarak anlatılacaktır.

Tablo 5. Naive Bayes Algoritması için Örnek Veri Seti

Yaş	Gelir	Öğrenci	Kredi Notu	Sınıf: Bilgisayar Alır mı?
Genç	Yüksek	Hayır	Orta	Hayır
Genç	Yüksek	Hayır	Çok iyi	Hayır
Orta Yaş	Yüksek	Hayır	Orta	Evet
Yaşlı	Orta	Hayır	Orta	Evet
Yaşlı	Düşük	Evet	Orta	Evet
Yaşlı	Düşük	Evet	Çok iyi	Hayır
Orta yaş	Düşük	Evet	Çok iyi	Evet
Genç	Orta	Hayır	Orta	Hayır
Genç	Düşük	Evet	Orta	Evet
Yaşlı	Orta	Evet	Orta	Evet
Genç	Orta	Evet	Çok iyi	Evet
Orta yaşlı	Orta	Hayır	Çok iyi	Evet
Orta yaşlı	Yüksek	Evet	Orta	Evet
Yaşlı	Orta	Hayır	Çok iyi	Hayır

Kaynak: Silahtaroğlu, G. (2016). *Veri Madenciliği Kavram ve Algoritmaları.* Papatya Yayıncılık.

.

Yukarıda verilen örnek veri seti kullanılarak yaş, gelir,öğrenci olup olmadığı ve kredi notu değişkenleri göz önünde bulundurularak naive bayes yöntemi ile X=(genç, orta gelir, öğrenci, orta) olarak ifade edilen x veri demeti sınıflandırılacaktır.

Sınıflandırmak istediğimiz sınıf etiketi bilgisayar satın alıp almayacağı olmak üzere iki eğer almaktadır. Bu iki değere C_1 ve C_2 diyelim.

P (C_1) bilgisayar alma olasılığını gösteriyorken P(C_2) bilgisayar almama olasılığını göstermektedir. Tablomuz yardımıyla P(C_1) ve P(C_2) değerleri hesaplanmış ve aşağıdaki sonuçlar elde edilmiştir.

P ($C_1 = bilgisayar\ alır$) =9/14= **0,642**

P ($C_2 = bilgisayar\ alımaz$)= 5/14= **0,357**

P (C_1) ve P (C_2) değerleri için sınıflandırmak istenen veri demetindeki her özellik için şartlı olasılıkları hesaplanır.

P ($Genç/C_1 = bilgisayar\ alır$) = 2/9= **0,222**

P ($Orta\ gelir/C_1 = bilgisayar\ alır$)= 4/9= **0,444**

P (Öğ$renci/C_1 = bilgisayar\ alır$)= 6/9= **0,666**

P ($Orta/C_1 = bilgisayar\ alır$)= 6/9= **0,666**

P ($Genç/C_2 = bilgisayar\ almaz$) = 3/5= **0,600**

P ($Orta\ gelir/C_2 = bilgisayar\ almaz$)= 2/5= **0,400**

P (Öğ$renci/C_2 = bilgisayar\ almaz$)= 1/5= **0,200**

P ($Orta/C_{12} = bilgisayar\ almaz$)= 2/5= **0,400**

Bundan sonraki adımda P ($X/C_1 = bilgisayar\ alır$) ve P ($X/C_2 = bilgisayar\ almaz$) değerleri ayrı ayrı olarak yukarıda bulunan değerlerin çarpılması ile elde edilmiş olur.

P ($X/C_1 = bilgisayar\ alır$)= 0,222*0,444*0,666*0,666= **0,0437**

P ($X/C_2 = bilgisayar\ almaz$)= 0,600*0,400*0,200*0,400= **0,0192**

Yukarıda bulunan değerler doğrultusunda ve bayes teoremi yardımıyla P($C_1 = bilgisayar\ alır$/X) ve P($C_2 = bilgisayar\ almaz$/X) değerleri hesaplanır ve hangi değer yüksek ise veri demetimiz o sınıfa atanmış olur.

$$P(C_1 bilgisayar\ alır/X) = \frac{P(X/C_1 = bilgisayar\ alır)P(C_1 = bilgisayar\ alır)}{P(X)}$$

P($C_1 bilgisayar\ alır$/X)= 0,0437*0,642= **0,028**

. P($C_2 bilgisayar\ almaz$/X)= $\frac{P(X/C_2 bilgisayar\ almaz)P(C_2 = bilgisayar\ almaz)}{P(X)}$

. P($C_2 bilgisayar\ almaz$/X)= 0,0192*0,357= **0,0068**

Naive bayes algoritmasına göre ilgili veri satırımızda yer alan özelliklere göre satır bilgisayar sınıfında sınıflandırılmıştır.

2.1.2. CHAID Algoritması

CHAID algoritması 1990'lı yıllarda ortaya çıkmış ve ''An Exploratory Technique for Invetigating Large Quantities of Categorical Data'' ismi ile duyurulmuş bir algoritmadır (Kass, 1980, s. 120). CHAID algoritması karar ağacı algoritması olarak bilinse de algoritmada karar ağacı ile birlikte istatistikte bilinen Ki-kare test yöntemini de kullanmaktadır.

CHAID algoritması karar ağacı algoritması ile dallanan değişken ile bağımlı değişken arasındaki ilişki Ki-kare testi ile test edilir. Ki-kare testi sonucu iki değişkenin arasında bağımlı olduğu yönünde bir sonu elde edilmiş ise karar ağacı dallanmaya devam ederken tersi bir sonuçta ise ağacın durmasına neden olur (Çelik, 2009, s. 50). CHAID algoritmasında kullanılan veriler kategorik olmak zorundadır.

2.1.3. Regresyon

Regresyon analizi bir değişkenin diğer değişken/değişkenler olan ilişkisini ifade edebilmek amacıyla matematiksel modelleri kullanan bir karar destek sistemi olarak açıklanabilir. Regresyon analizinde sınıflandırma bölme ve tahmin olmak üzere iki yaklaşım etrafında kullanılır. Regresyon modelleri içerdiği değişken sayısına farklı modeller olarak incelenebilmektedir.

Bir model içerisinde yer alan bağımlı değişken bağımsız değişken tarafından etkileniyor ve açıklanabiliyorsa bu tür modellerde regresyon modeli olarak basit doğrusal regresyon modeli tercih edilmelidir. Birden fazla bağımsız değişken bir bağımlı değişkeni etkiliyor ve açıklayabiliyorsa çoklu regresyon modelleri kullanılmalıdır. Ayrıca regresyon modellerinde değişkenler arasındaki ilişki doğrusal ise doğrusal regresyon değil ise doğrusal olmayan değişken olarak adlandırılmaktadır (Silahtaroğlu, 2016, s. 63).

Basit Doğrusal Regresyon

Basit doğrusal regresyon modelleri bir bağımlı değişkenin bir bağımsız değişken ile açıklanabildiği modellerdir. Örneğin aylık araba giderleri ile aylık gelir arasındaki ilişki doğrusal regresyon modeli ile açıklanabilir. Basit doğrusal regresyon denklemi aşağıda verilmiştir.

$$y = \beta_0 + \beta_1 x + \varepsilon \qquad (2.4)$$

β_0= Doğrunun y- eksenini kestiği nokta

β_1= Eğim

ε: Hata Terimi

Doğrusal regresyon modelinde y bağımlı değişkeni β_0 sabit değeri x bağımsız değişkeni β_1 ise y bağımlı değişkenin x bağımsız değişkeni tarafından açıklanma düzeyini ifade eder. ε değeri ise hata terimi olarak adlandırılır.

Doğrusal regresyon modellerinde parametre tahminleri en küçük kareler tekniği ile yapılmaktadır. Doğrusal regresyon modelinde en küçük kareler yöntemlerini kullanabilmek için gerekli olan bazı koşullar mevcuttur. Bu koşullar ε hata değeri ile x bağımsız değişken ile ilgilidir. Koşulun sağlanabilmesi için. ε hata teriminin ortalaması 0 eşit olmalı ve normal dağılıma uymalıdır. Aynı şekilde x bağımsız değişkeninin hata terimi ile ilişkisi olmamalı ve varyansı pozitif bir sayı olmalıdır.

Lojistik Regresyon

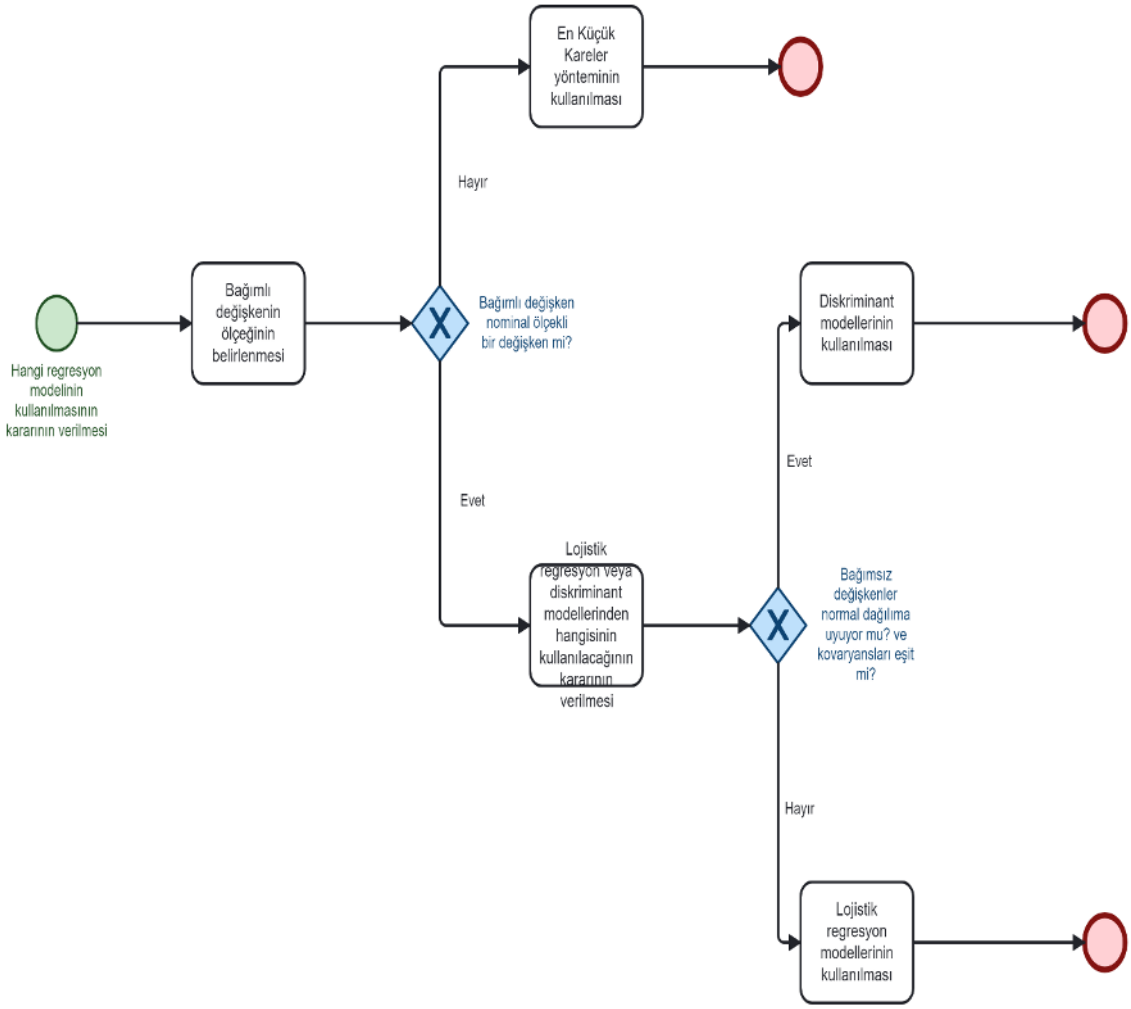
Lojistik regresyon analizinde, çok değişkenli bağımlı ve bağımsız değişkenler arasındaki ilişkiler incelenirken bağımlı değişken nominal ölçekli yapıya sahip olduğundan ve bağımlı değişken nominal yapıda normal dağılım koşulunu sağlamadığından dolayı basit doğrusal regresyonda olduğu gibi "En Küçük Kareler" yöntemi bağımsız ve bağımlı değişkenin arasındaki ilişkiyi açıklamada yetersiz kalmaktadır. Bu durumda çok değişkenli modellerde bağımlı değişken nominal bir yapıya sahip olduğunda En Küçük Kareler yöntemi dışında diskriminant ve lojistik regresyon modelleri daha başarılı sonuçlar vermektedir (Akpınar, 2017, s. 60).

Diskiriminant analizi regresyon modellerinde olduğu gibi bağımlı ve bağımsız çok değişkenli yapıların arasındaki ilişkiyi inceleyen bir modeldir. Ancak modelin başarılı bir sonuç vermesi için bağımsız değişkenlerin normal dağılıma uyması ve bağımsız değişkenlerinin kovaryanslarının her grup üzerinde eşit olması gerekmektedir. Bu durumda diskriminant analizinde nominal ölçekli değişkenler kullanıldığında yapılan analizler sağlıksız ve doğru olmayacaktır. Lojistik regresyon modellerinde ise bağımsız değişkenler için böyle bir koşul geçerli olmadığından daha çok lojistik regresyon modelleri kullanılmaktadır. Lojistik regresyon modelleri aşağıdaki gibi ifade edilmektedir.

$$L = In\left[\frac{P_i}{1 - P_i}\right] = b_0 + b_1 X_i + e_i \qquad (2.5)$$

Lojistik regresyon modellerinin sonuçları analitik olarak doğru ve verimli sonuçlar elde edilememektedir. Lojistik regresyon modellerinin sonuçları çeşitli istatistik programları sayesinde elde edilmektedir. Genellikle istatistik programları parametre tahminlerini iteratif bir yöntem olan maksimum olabilirlik (ML) yöntemi ile tahmin etmektedir.

Şekil.7 görüldüğü gibi hangi regresyon modellerinin nerelerde hangi koşullara göre kullanılacağı iş akış şeması şeklinde verilmiştir.



Şekil 8. Regresyon Modelleri İş Akış Şeması

Kaynak: "yazar tarafından oluşturulmuş/derlenmiştir"

2.2. Mesafeye Dayalı Algoritmalar

Veri Madenciliğinde sınıflandırma yöntemleri algoritmaları istatistiğe dayalı algoritmalar, mesafeye dayalı algoritmalar, karar ağaçları olmak üzere 3 ana grup içinde gruplandırılmaktadır. Mesafeye dayalı algoritmalar aşağıdaki gibi sınıflandırılmaktadır.

- *k* En Yakın Komşu Algoritması
- En Küçük Mesafe Sınıflandırıcısı

2.2.1. K En Yakın Komşu Algoritması

Mesafeye dayalı algoritmalar içerisinde en çok kullanılan algoritma olan K En Yakın Komşu algoritması yabancı kaynaklarda K-Nearst Neighbor veya KNN algoritması olarak ifade edilmektedir. Basit bir sınıflandırma algoritması olan KNN algoritması ile yeni bir veri grubu sınıflandırılmak istendiğinde daha önceden etrafında sınıflandırılmış olan k adet veri grupları içerisinde çeşitli uzaklık algoritmalarını kullanarak en yakın komşusuna olan uzaklığını hesaplar. Hesaplanan değerler doğrultusunda sınıflandırılacak olan veri/veri grupları mesafe olarak en yakın ve adet olarak fazla olan sınıfın içerisine dahil edilir (Şimşek, 2019, s. 45). Sınıflandırılacak olan veri seti ile daha önce sınıflandırılmış olan veri seti arasındaki uzaklık ölçümleri Öklid, Chebyschev, Minkowski, Manhattan uzaklıkları ile hesaplanmaktadır.

Algoritma yeni veri/veri grubunu k adet en yakın komşusuna göre sınıflandıracağından dolayı sınıflandırmanın doğru ve başarılı bir şekilde yapılması için k değerlerinin en iyi şekilde belirlenmesi gerekmektedir. Algoritmadaki k değeri sınıflandırılacak olan verinin, sınıflandırma sırasında kaç veriye bakarak sınıflandırılacağını gösteren sayıyı ifade etmektedir. Genellikle yapılan çalışmalar incelendiğinde seçilen sayıların 3,5 ve 7 gibi tek sayılar olduğu görülmüştür.

KNN algoritmaları coğrafi bilgi sistemleri gibi oldukça büyük sınıf sayısına ve karmaşık veri yapısına sahip olan çalışmalarda kullanılmakta ve başarılı bir performans göstermektedir. Kısacası KNN algoritmasının çalışma prensibi aşağıda verilen adımlar ile özetlenebilir (Silahtaroğlu, 2016, s. 65).

- Sınıflandırılacak olan nesnenin belirlenmesi ve uygun hale getirilmesi
- Çalışmaya uygun olan mesafe ölçüm yönteminin belirlenmesi (Literatürde en çok Öklid mesafe ölçüm tekniği kullanılmaktadır.)
- Probleme en uygun olan k değerinin yapılan çeşitli denemeler sonucunda belirlenmesi
- Belirlenen k adet nokta içerisinden sınıflandırılacak olan nesneyi en yakın ve adet olan fazla sınıfın belirlenmesi
- Belirlenen sınıfa nesnenin atanması

2.2.2. En Küçük Mesafe Sınıflandırıcısı

En Küçük Mesafe Sınıflandırıcısı algoritması, KNN algoritmasının çalışma mantığındaki gibi mesafeye dayalı olarak çalışan bir algoritmadır. Algoritma sınıflandırılacak olan nesnenin, daha önce sınıflandırılmış olan tüm nesnelerin merkezleri ile olan uzaklığı hesaplanarak nesnenin sınıflandırılması tekniğine dayanmaktadır. En küçük mesafe sınıflandırıcısı adımları aşağıda verilmiştir.

- Daha önce sınıflandırılmış olan nesnelerin merkezlerinin belirlenmesi
- Belirlenen merkezler ile sınıflandırılacak nesneler arasındaki uzaklıkların belirlenmesi
- Sınıflandırılacak olan nesnelerin en büyük değerli sınıfa atanması

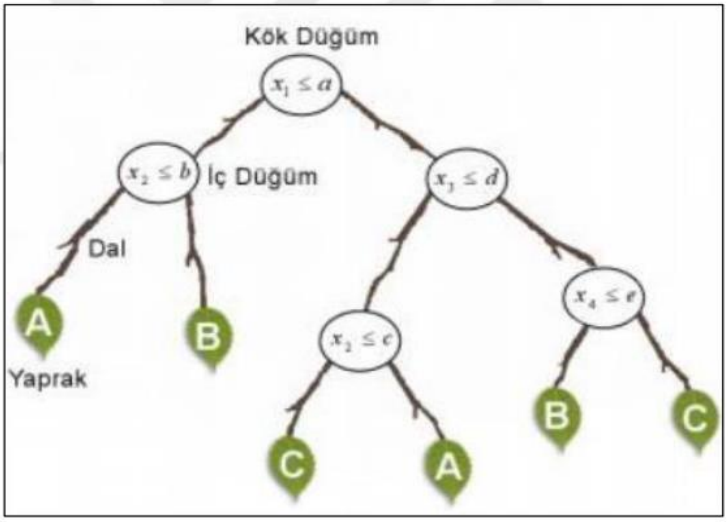
2.3. Karar Ağaçları ve Algoritmaları

Karar ağaçları ve algoritmaları sınıflandırma yöntemleri içerisinde en çok kullanılan yöntemler içerisinde yer almaktadır. Sınıflandırma yöntemleri içerisindeki bazı sınıflandırma yöntemleri ile kıyaslandığında başarı oranı daha yüksek ve yapılandırılması ve anlaşılabilmesi daha kolay bir yöntemdir (Agrawal vd., 1993, s. 920). Karar ağaçlarının bir diğer avantajı ise güvenli, olması ve işlenecek olan verilerin sisteme kolaylıkla adapte edilebilmesi olarak sıralanabilir. Karar ağacı algoritmalarının avantajları olduğu kadar dezavantajları da vardır. Karar ağacı algoritmalarında analiz ve aktarılan veri ne kadar fazla ise karar ağacı dallanma sayısı o kadar fazla olacak analiz süresi uzayacak ve takibi zorlaşacaktır. Bir diğer dezavantajı ise hesaplama maliyetlerinin yüksek olmasıdır (Demirel ve Yakut, 2019, s. 57).

Karar ağaçları yapısında bulundurduğu birden çok bağımlı ve bağımsız değişken ile regresyon modellerine benzemektedir. Karar ağaçları bağımlı değişkenin kategorik olduğu durumlarda lojistik regresyon ve diskriminant analizi yöntemlerine bir alternatif model olarak kullanılmaktadır. (Çelik, 2009, s. 29).

Karar ağaçları, tıpkı ağacın yapısında bulunan dal, yapraklar ve karar düğümlerinden oluşmaktadır. Karar ağaçlarından sınıflandırma adımları karar düğümü ile başlar. Karar düğümünde veriye çeşitli testler uygulanarak kararın verildiği adımdır. Karar düğümünden çıkan sonuçlar doğrultusunda dallar elde edilir.

Elde edilen her dal ile sınıflandırılması gereken nesnenin sınıflandırılması için gerekli olan tahmin sonuçlarının bulunduğu yapraklara gidilir. Sınıflandırmanın tamamlanabilmesi için karar düğümünden çıkan her dalın yapraklara ulaştırılması gerekmektedir. Fakat dalların sonrasında yapraklar oluşmuyorsa tekrar karar düğümlerine dönülür bu işlem tüm dalların yapraklara (sınıflara) atanana kadar devam eder (Şimşek, 2019, s. 39). Tipik bir karar ağacı Şekil. 9 de gösterilmiştir.



Şekil 9. **Karar Ağacı Yapısı**

Kaynak: Silahtaroğlu, G. (2016). *Veri Madenciliği Kavram ve Algoritmaları.* Papatya Yayıncılık.

Literatüre bakıldığında karar ağacı algoritmalarının çeşitlerine göre karar ağacı yapılarındaki düğüm noktaları ve dallanmaların nasıl yapılacağı değişiklik göstermektedir. Bu yüzden iki farklı karar ağacı algoritmasının sınıflandırma sonuçları aynı olarak çıkmaz (Şimşek, 2019, s. 40).

Karar ağaçları algoritmaları kullanılarak sınıflandırılacak olan bir verinin hangi özelliğinin kullanarak dallandırılacağının belirlenmesi için özellik seçim ölçütleri kullanılır. Başlıca yöntemler aşağıda sıralanmıştır.

- Bilgi Kazanımı (Information Gain)
- Kazanım Oranı (Gain Ratio)
- Gini İndeksi (Gını Index)

Bilgi Kazanımı (Information Gain); genellikle karar ağacı algoritmalarından olan ID3 algoritmasında kullanılan bir kavramdır. Yöntem karar ağaçları içerisinde bağımlı değişkenin kategorik olması durumunda dallanma için tercih edilir (Çelik, 2009, s. 24).Bilgi kazanımı kavramı entropi ilkesine dayanmaktadır. Entropi kavramı ise saflığın zıttı olarak nitelendirilir. Bilgi kazanımı yöntemini kullanarak dallandırılan karar ağaçlarında entropi azalmaktadır. Yöntemde başlangıçta X ile gösterilen bir düğüm oluşturulur. Oluşturulan her düğüm bir özelliği temsil etmektedir. Bu özellikler kullanarak dallanan verilerden en yüksek bilgi kazanımı elde edilmesi amaçlanmıştır. Yöntemdeki notasyonlar aşağıdaki gibi ifade edilmiştir.

- D: Eğitilmiş veri seti
- m: Sınıf etiketinin alabileceği değer
- C_i: Sınıf (i=1,2, 3…n)
- p_i: C_i sınıfının veri setindeki olasılığı
- m: Sınıf etiketinin alabileceği değer
- v: A özelliği kesikli ise v kadar çıktı çıkar

Herhangi bir veri setini sınıflandırmak istediğimizde veri setindeki belirsizliği (entropi) 2.6 da verilen eşitlik ile hesaplanmaktadır.

$$Info(D) = -\sum_{i=1}^{m} p_i log_2 (p_i) \tag{2.6}$$

Entropi kavramından sonra bilgi kazanımı değerinin bulunabilmesi için veri setinin A özelliği ile dallanması sonucu entropinin hesaplanması gerekmektedir. Eşitlik 2.7' de gösterilmiştir. Bu değer ne kadar küçük ise sınıflandırma o kadar başarılı olmaktadır.

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} * Info(D_j) \tag{2.7}$$

Bilgi kazanımı eşitlik 2.8 'de gösterildiği gibi ifade edilir. Bu değer ne kadar büyük ise o kadar iyidir.

$$\text{Gain (A)} = \text{Info(D)} - \text{Info}_\text{A}\text{(D)} \tag{2.8}$$

Kazanım Oranı (Gain Ratio); ilkesi bir özellik ne kadar fazla değer alıyorsa bu değerlere doğru bir yönelim gösterir ve karar ağaçları bu değerlere göre karar ağaçları oluşturma eğilimi gösterir. Kısacası kazanım oranında bir taraflılık söz konusudur. Bu taraflılıktan kurtulmak için yöntem içerisinde normalizasyon uygulanmaktadır.

Yöntemde ilk olarak ayrık bilgi eşitlik (2.9) verilen formül ile hesaplanmalıdır. Ayrık bilgi mutlaka sıfırdan farklı bir değer almalıdır.

$$Splitınfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} * \log_2\left(\frac{|D_j|}{|D|}\right) \tag{2.9}$$

Ayrık bilgi hesaplandıktan sonra kazanım oranı eşitlik (2.10)' da verilen formül ile hesaplanır. Test edilen özelliklerden hangisinin kazanım oranı yüksek ise o özellik seçilmelidir.

$$\text{GainRatio(A)} = \left(\text{Gain(A)}/Splitınfo_A(D)\right) \tag{2.10}$$

Gini İndeksi (Gını Index); veride oluşan kirliliğin hesaplanması için kullanılan bir yöntemdir. X veri kümesindeki herhangi bir özelliği sınıflandırmak için eşitlik (2.11) de verilen formül kullanılır. Formüldeki p değeri D veri kümesindeki bir kaydın X kümesine ait olma olasılığını ifade eder.

$$Gini(D) = 1 - \sum_{i=1}^{n} p_i^2 \tag{2.11}$$

Gini indeksi her bir özelliğin kendi ve boş kümesi hariç tüm alt kümelerini test eder. Bir özelliğin alabileceği alt küme sayısı $2^v - 2$ ile belirlenir. Belirlenen her bir alt küme için ayrı ayrı gini indeksi hesaplanır. A özelliğinin alt kümeleri olan D_1 ve D_2 için eşitlik (2.12) de gösterildiği gibi hesaplanır.

$$Gini_A D = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2) \tag{2.12}$$

Yöntemde kullanılacak olan özellik(değişken) seçimi için kullanılacak gini indeksi eşitlik (2.11) ve (2.12) farkı ile hesaplanır.

$$Gini(A) = Gini(D) - Gini_A D \tag{2.13}$$

Karar ağacı algoritmaları, 1950 yılların sonlarında ortaya çıkmış ve günümüze kadar birden çok çeşitli karar ağacı algoritmaları ortaya atılmış ve kullanılmıştır. Karar ağacı algoritmalarının bir kısmı sınıflandırma için kullanılırken bir kısmı tahminleme modelleri için bir kısmı da hem tahminleme hem de sınıflandırma modelleri için kullanılmaktadır. Literatüre bakıldığında birden çok algoritma kullanılsa da çalışmamızda Random Forest, Gradient Boosting Trees(XGBoost), CHAID, ID3, ADABOOST, CART, QUEST ve C4.5 algoritmalarıyla analizler yapıldığından ve literatürde en çok kullanıldıklarından dolayı bu yöntemlerden bahsedilecektir.

Random Forest, algoritması karar ağacı algoritması olarak kullanıldığında sınıflandırma sırasında birden fazla karar ağacı üreterek sınıflandırma başarısını arttırmaya çalışan bir algoritmadır. Bazı karar ağacı algoritmalarına göre iyi sonuçlar elde edebilmektedir. Random Forest algoritmasının avantajlarından biri ise karmaşık, düzensiz büyük veri grupları ve kategorik veri gruplarının sınıflandırılmasında oldukça iyi sonuçlar elde etmesidir (Aydın, 2018, s. 172).

Gradient Boosting Trees(XGBoost), algoritması random forest algoritması gibi karar ağaçları temelli bir algoritmadır. Random forest algoritmasından farklı olarak son sınıflandırma kararının tüm ağaçların toplamına bağlı olması iki algoritmayı birbirinden ayıran en önemli farktır. Random foresta olduğu gibi her koşulda başarılı sonuçlar verebilmektedir.

ADABOOST, algoritması ilk olarak 1996 yılında Freund ve Schapire tarafından önerilmiştir (Vezhnevets ve Vezhnevets, 2005, s. 1). Algoritmanın çalışma mantığı her karar ağacı turunda yeni bir model eğitme prensibine dayanmaktadır. Her karar ağacı turunda yanlış olarak sınıflandırılan veriler toplanarak yeni eğitim setinde tekrardan eğitilerek bir sonraki karar ağacı turu için geri beslenilmek amacıyla kullanılmaktadır. Buradaki fikir, sonraki modellerin önceki modellerin yaptığı hataları telafi edebilmesidir (Sammut ve Webb, 2017, s. 917).

ID3 Algoritması, bakıldığında çalışmalarda en çok kullanılan karar ağacı yöntemlerinden biridir. İlk defa 1983 yılında Ross Quinlan tarafından önerilmiş bir karar ağacı algoritmasıdır. Algoritma kategorik değişkenlerin sınıflandırılmasında kullanılmaktadır. Algoritma kullanılarak sınıflandırılacak olan verinin hangi özelliğinin kullanarak dallandırılacağının belirlenmesi için özellik seçim ölçütlerinden bilgi kazanımı (Information Gain) yararlanır. ID3 algoritması entropiye dayalı olarak çalışan bir algoritmadır. ID3 algoritmasında entropi eşitlik (2.6) gösterildiği şekilde hesaplanır. Entropi kavramı ve eşitlik (2.7) de verilen her özelliğin beklenen değerin yardımıyla bilgi kazanımı kavramı eşitlik (2.8) gibi hesaplanabilmektedir. Kazanç değeri her değişken için hesaplandıktan sonra en yüksek kazanç değerine sahip olan değişken karar düğümünü oluşturur. Karar düğümünü oluşturan değişkenin sınıfında yer alan diğer değişkenlerde bu düğümün dallarını oluşturur. Bu işlemler tüm dallandırma işlemleri sonucunda sınıflandırılacak olan değişkenlerin belirli sınıflara atanması sonucunda algoritma dallandırmayı durdurur.

CART Algoritması, Classification and Regression Trees kavramının baş harflerinin kısaltılması ile ortaya çıkan algoritma ilk defa 1984 yılında Brieman, Friedman, Olshen ve Stone tarafından ortaya atılmış bir yöntemdir. Literatüre bakıldığında karar ağacı algoritmaları içerisinde en çok kullanılan yöntemdir. Yöntem sonuçları algoritma fazla bölünme eğilimine sahip olduğundan dolayı karar ağacı üzerinden etkin olarak anlaşılmamaktadır. Algoritma ID3 algoritmasında olduğu gibi kategorik verilerle çalışırken aynı zamanda sürekli verilerle de iyi sonuçlar vermektedir. Algoritma kategorik verilerle çalışırken düğümün hangi noktadan bölüneceğini belirlerken gini, twoing ve ordered twoning kavramlarından yararlanır. Sürekli değişkenlerde ise bölen değişkenin bulunmasında en küçük kareler sapması yöntemi kullanılmaktadır. Algoritma da en iyi ayırma kriterlerini belirlemede ID3 algoritmasında olduğu gibi entropi kavramından yararlanılır (Silahtaroğlu, 2016, s. 58). Algoritma karar ağacının kök düğümünden başlayarak her düzeyi ikiye ayırarak ikili olarak büyür. Buradaki amaç bir sonraki düzeyde homojen kökler elde etmektir. Algoritmada diğer karar ağacı algoritmalarından farklı olarak hangi düğümün kök ya da düğüm olacağının kararı yanında seçilen düğümün bir sonraki düzeyde homojen kökler oluşturmak üzere düğümün hangi noktasından ikiye ayrılmasının kararını da vermektedir.

QUEST Algoritması, (Quick, Unbiased, Efficient Statistical Tree), ilk olarak 1997 yılında Loh ve Shih tarafından ortaya atılmış bir algoritmadır. Çok kategorili değişkenlerin analizinde diğer yöntemler kullanılmazken QUEST algoritması ile çok kategorili değişkenler başarılı ve kolay bir şekilde sınıflandırılmaktadır. Algoritma CART algoritmasında olduğu gibi ikili bölünmeler sağlar ve son ağaç doğrudan veya budama kurallarından biriyle seçilme eğilimi gösteren karar ağacı algoritmasıdır (Loh ve Shih, 1997, s. 815).

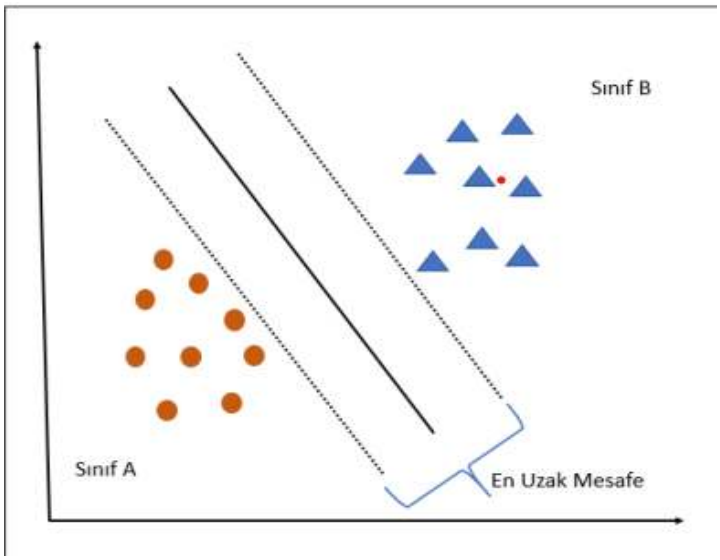
C4.5 Algoritması, Ross Quinlan tarafından 1993 yılında geliştirilmiş karar ağacı algoritmalarından biridir. Algoritmanın çalışma yapısı ID3 algoritmasına benzediğinden dolayı ID3 algoritmasının gelişmiş versiyonu olarak kabul edilen bir algoritmadır.

2.4. Karar Destek Makineleri

Karar Destek Makineleri, veri gruplarını sınıflandırmak ve sınıfları birbirinden ayırmak için kullanılan diğer yöntemler gibi geliştirilmiş bir algoritmadır. Karar destek sistemlerimin en önemli avantajları verileri doğrusal veya doğrusal olmayan biçimde ayırmadan her iki veri türünü de sınıflandırma özelliğine sahiptir. Ayrıca karar destek makineleri karmaşık ve orta ve küçük ölçekli veri grupları ile iyi sonuçlar vermektedir. Bu durum yöntemin dezavantajları arasında yer almaktadır. Destek sistemleri algoritmasının işlem adımları aşağıda sırlanmıştır.

- İlk olarak algoritma ayırılacak olan sınıfların aralarındaki en uzak mesafeyi çeşitli uzaklık algoritmaları ile tespit etmeye çalışır.
- En uzak mesafe tespit edildikten sonra bu noktanın tam ortasından geçecek şekilde ve düzlemi ikiye ayıracak şekilde bir çizgi çekilir.
- Çekilen bu düzlem sayesinde sınıflar birbirlerinden ayrılmış olur.

Şekil 10'da örnek bir destek vektör algoritmasının yapısı gösterilmiştir. Şekilde belirlenen en uzak mesafenin ortasından geçecek şekilde bir doğru çizilir. Sınıflandırmak istenen veri veya veri grubu bu çizginin ne tarafında kalıyorsa o sınıfa atanmış olur.



Şekil 10. Karar Destek Makineleri

Kaynak: Çelik, M. (2009). *Veri Madenciliğinde Kullanılan Sınıflandırma Yöntemleri ve Bir Uygulama* [Yüksek lisans tezi]. İstanbul Üniversitesi.

2.5. Sınıflandırma Yöntemlerinin Başarısının Ölçülmesi

Bir sınıflandırma problemi sınıflandırıldıktan sonra yapılan sınıflandırmanın diğer yöntemlere göre üstünlüğünü ve başarısını belirlemek için çeşitli yöntemler ve metrikler kullanılmaktadır. Bu yöntemlerden en çok bilinenleri sırasıyla kesinlik oranı, doğruluk-hata oranı, duyarlılık oranı ve F-ölçüt oranı yöntemleridir. Yöntemlerin kullanılabilmesi için bazı metriklerin belirlenmesi gerekmektedir. Bu metriklerin belirlenmesinde en çok gerçek sınıf ve tahmin edilen sınıfların karşılaştırıldığı Tablo-6'da gösterilen karmaşıklık matrisi kullanılmaktadır.

Tablo 6. Karmaşıklık Matrisi

		Tahmin Edilen Sınıf	
Gerçek Sınıf		**Pozitif**	**Negatif**
	Pozitif	TruePositives(TP)	FalsePositives(FP)
	Negatif	FalseNegatives(FN)	TrueNegatives(TN)

Kaynak: "yazar tarafından oluşturulmuş/derlenmiştir"

- **TruePositives(TP):** Sınıflandırma modelinin tahmini ile modelin eğitimi için kullanılan test verilerinin sınıf değerleri aynıdır. Sınıflandırıcı tarafından doğru sınıflandırılmıştır. TP pozitif olarak tahmin edilen demet sayısını ifade eder.
- **TrueNegatives(TN):** Gerçek değer negatif iken model tarafından tahmin edilen sınıfta negatiftir. Model tarafından doğru sınıflandırılmıştır. TN negatif demet sayısını ifade eder.
- **FalseNegatives(FN):** Gerçek değer pozitif iken model tarafından tahmin edilen sınıfta negatiftir. Model tarafından yanlış sınıflandırılmıştır.
- **FalsePositives (FP):** Gerçek değer negatif iken model tarafından tahmin edilen sınıfta pozitiftir. Model tarafından yanlış sınıflandırılmıştır.

2.5.1. Kesinlik (Precision)

Kesinlik değeri literatürde tamlık olarak da bilinmektedir. Kesinlik değeri Sınıflandırıcı modelin sonuçlarının ne kadar kesin olduğunu ölçmektedir. Kesinlik değeri eşitlik (2.13) te görüldüğü üzere sınıflandırıcı tarafından doğru sınıflandırılan pozitif demet sayısının (TP), sınıflandırıcı tarafından sınıflandırılmış tüm pozitif tahminlere bölünmesi ile bulunur.

$$\text{Kesinlik (Precision)}: \frac{TP}{TP + FP} \tag{2.13}$$

2.5.2. Duyarlılık (Recall)

Duyarlılık (recall) değeri, eşitlik (2.14) gösterildiği gibi doğru olarak sınıflandırılan pozitif demet sayısının (TP) tüm pozitif örnek sayısına bölünmesi ile bulunur.

$$\text{Duyarlılık (Recall)}: \frac{TP}{TP + FN} \tag{2.14}$$

2.5.3. Doğruluk-Hata Oranı

Doğruluk oranı veri madenciliği model başarısı değerlendirme yöntemleri içerisinde en çok kullanılan başarı ölçme yöntemidir. Doğruluk oranı model tarafından doğru olarak sınıflandırılan örnek sayısının (TP+TN) tüm örnek sayısına

(TP+FP+FN+TN) bölünmesi ile bulunur. Hata oranı ise 1-doğruluk olarak literatürde açıklanmıştır.

$$\text{Doğruluk:} \frac{TP + TN}{TP + FP + FN + TN} \tag{2.15}$$

$$\text{Hata Oranı:} 1 - \frac{TP + TN}{TP + FP + FN + TN} \tag{2.16}$$

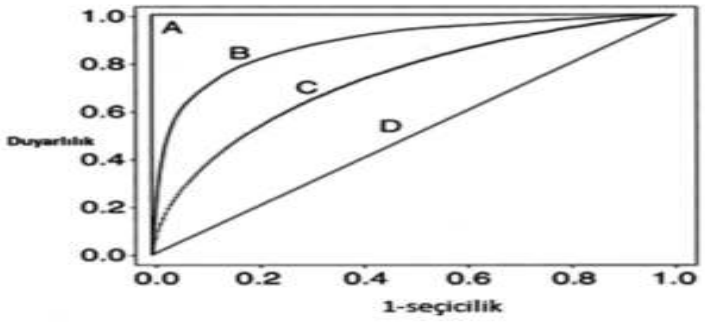
2.5.4. F-Ölçütü

Veri madenciliği yöntemlerinde bazı modeller için duyarlılık ve kesinlik değerleri ayrı ayrı yorumlanarak yeterli sonuca varılamamaktadır. Bunun için her iki ölçütü aynı anda değerlendirmek gerekmektedir. Bunun için F-ölçütü geliştirilmiştir. F-ölçütü duyarlılık ve kesinlik değerlerinin harmonik ortalaması olarak değerlendirilmektedir (Çoşkun ve Baykal, 2011, s. 55-61).

$$\text{F} - \text{Ölçütü:} \frac{2 * Duyarlılık * Kesinlik}{Duyarlılık + Kesinlik} \tag{2.17}$$

2.5.5. ROC Eğrisi

ROC Eğrisi ölçütü, sınıflandırma modellerinin başarısının karşılaştırılmasında bakılacak en önemli ölçütlerden biridir. Literatüre bakıldığında ROC ölçütünün daha çok düzensiz veri setleri tarafından oluşturulan sınıflandırma modellerinin başarı ölçütünü belirleme de kullanılmıştır. Şekil-11'de örnek bir ROC eğrisi grafiği gösterilmiştir.



Şekil 11. ROC Eğrisi Grafiği

Kaynak: Kılıç, S. (2013). Klinik karar vermede ROC analizi. *Journal of Mood Disorders, 3*(3), 135-140.

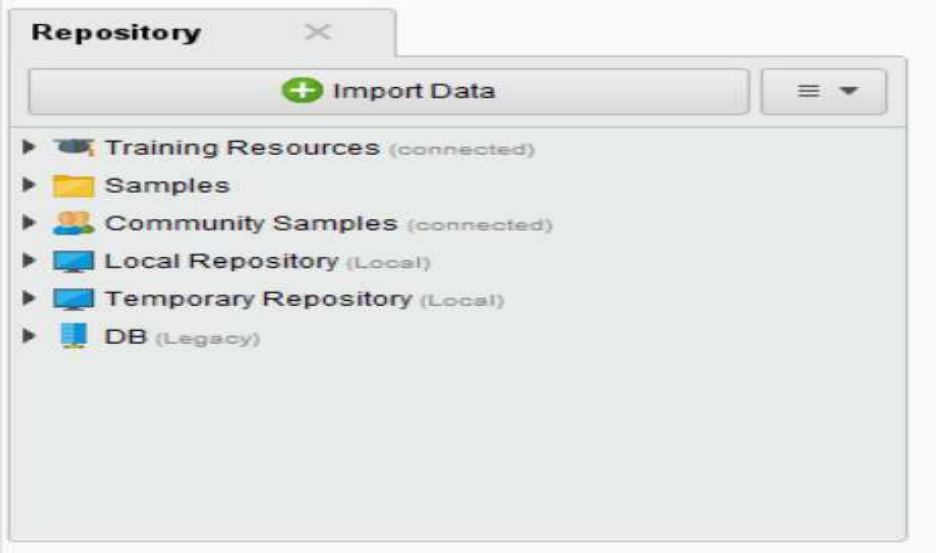
Grafikte görüldüğü üzere ROC eğrisi gerçek pozitif (TP) ile sahte pozitif (FP) değerleri yardımıyla bulunur. Grafikte yatay eksende sahte pozitif (kesinlik) oranı, dikey eksende ise gerçek pozitif(duyarlılık) oranı yer alır. Modelin başarısı grafikte gerçek pozitif oranın artması ve sahte pozitif oranın düşüklüğü artmaktadır. ROC eğrisi altında kalan alan ROC puanı olarak isimlendirilir. Bu alan 0-1 aralığında değer alır. Değerin 1 olması modelin sonucunun başarılı olduğu anlamına gelmektedir.

Şekil-11' de görüldüğü üzere A, B, C ve D modellerinin hangisi ROC eğrisine daha yakınsa bu modelin başarısı azalmaktadır. Grafikte görüldüğü üzere bu nedenle eğriye en uzak A modelinin daha doğru olduğu ortaya çıkmıştır.

3. RAPIDMINER

RAPIDMINER programı 2001 yılında ilk defa Dortmund Teknik Üniversitesi Ralf Klinkenberg, Ingo Mierswa ve Simon Fischer tarafından geliştirilmiş bir veri madenciliği aracıdır. RAPIDMINER uygulaması java dilinde yazılmış açık kodlu bir kaynak olup günümüzde birçok farklı alanda kullanılmaktadır. Bu sayede uygulama kullanımı kodlama bilgisi gerektirmez, veri aktarımı diğer veri madenciliği uygulamaları arasında kolay şekilde gerçekleştirilebilmektedir. Ayrıca aynı anda farklı analiz yöntemleri kullanılarak analizler gerçekleştirilebilmektedir (Çelik, vd., 2017, s. 1-2). RAPIDMINER programı ve diğer veri madenciliği programları arasında oldukça fazla benzerlik bulunsa da diğer uygulamalar ile arasında bazı önemli farklarda bulunmaktadır. Aşağıda RAPIDMINER diğer uygulamalarla (KNİME, R, WEKA) karşılaştırılmış olup temel farklar aşağıda paylaşılmıştır.

Sınıflandırma algoritmaları tarafından uygulamalar karşılaştırıldığında tüm sınıflandırma algoritmaları tüm uygulamalarda çalışırken KNN algoritması R'da RWeka paketinde bulunmaktadır. Kümeleme algoritmaları tarafından bakıldığında literatürde en çok kullanılan K-Means algoritması tüm uygulamalarda kullanılmaktadır. Birliktelik kuralları açısından bakıldığında APRİORİ algoritmaları tüm uygulamalarda bulunurken FP-Growth algoritması sadece KNİME ve RAPIDMINER uygulamalarında kullanılmaktadır. Karar ağacı algoritmalarında kullanılan algoritmanın hangi özelliklerinden dallanacağını belirleyen özellik seçim ölçütleri (bilgi kazananımı, kazanım oranı, gini indeksi) tüm uygulamalarda kullanılmaktadır.



Şekil 12. RAPIDMINER Repository Arayüzü

Kaynak: "yazar tarafından oluşturulmuş/derlenmiştir"

1.Repository kısmında diğer uygulamalar ile hazırlanmış tüm verilerin RAPIDMINER içerisine aktarılmasını sağlayan import data bölümü, hazır veri setleri ile hazır prosesler ve veri ve proseslerin kaydedilmesi için local ve Temporary repository bölümleri yer almaktadır.

Şekil 13. RAPIDMINER Operators Arayüzü

Kaynak: "yazar tarafından oluşturulmuş/derlenmiştir"

2.Operators kısmında sınıflandırma, kümeleme ve birliktelik kuralları panellerinin bulunduğu kısımdır. Operators kısmında veri önişleme işlemlerinin yapılması için gerekli olan ön işleme adımlarının yapıldığı kısımdır. Ayrıca farklı modeller için geçerli olan farklı yöntem ve modelleri içerisinde barındırır.

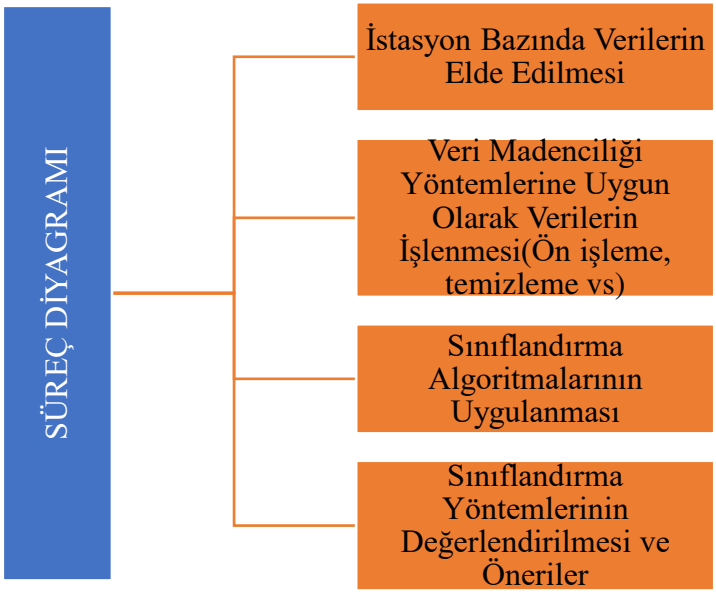
Şekil 14. RAPIDMINER Parameters Arayüzü

Kaynak: "yazar tarafından oluşturulmuş/derlenmiştir"

3.Parameters sekmesinde modelleme içerisinde kullanılan özelliklerin modele uygun hale getirilmesi için gerekli olan özelliklerin düzenlendiği bölümdür. Modele uygun olarak kuralların seçilmesi ve düzenlenmesi bu bölümden sağlanır.

4. UYGULAMA: AKARYAKIT SEKTÖRÜNDE İŞLEM ANORMALLİKLERİNİN VERİ MADENCİLİĞİ İLE SINIFLANDIRILMASI

Bu bölümde Türkiye'de petrol şirketlerine veri analizi desteği ve teknik destek hizmeti sağlayan ve otomasyon firması olan Turpak Elektromanyetik Yakıt İkmal Sis. Tic.A.Ş. veri tabanında bulunan ve Türkiye'de faaliyet gösteren bir petrol şirketinin istasyonlarına ait olan otomasyon verileri kullanılarak analizler yapılmıştır. Veriler veri madenciliği yöntemlerine uygun olacak şekilde revize edilmiş olup adımlar şekil.15'te gösterilmiştir. Günlük olarak ve tank bazında istasyon verileri otomasyon sistemi sayesinde veri tabanına aktarılmaktadır. Aktarılan veriler içerisinde istasyonda meydana gelen sorun doğrultusunda alarm üretmekte olan veriler incelenerek sorunun kaynağına göre çeşitli kategorilere göre sınıflandırılmaktadır. Çalışmamızda belirlenen 4 kategori doğrultusunda veri madenciliği sınıflandırma yöntemleri kullanılarak bundan sonra istasyonlarda ortaya çıkabilecek sorunların daha önce sınıflandırılan veriler yardımıyla kategorize edilmesi ve daha önceden yapılan analizlerin doğruluğunun ölçülmesi amaçlanmıştır. Seçilen kategoriler işletmede önemli analiz süreleri gerektiren ve önemli olan kategorilerden seçilmiştir. Tez çalışmamızda uygulama bölümünde analizler RAPIDMINER programı ile yapılmıştır.



Şekil 15. Uygulama Süreç Diyagramı

Kaynak: "yazar tarafından oluşturulmuş/derlenmiştir"

4.1. Şirket Hakkında Genel Bilgi

Turpak Elektromanyetik Yakıt İkmal Sis. Tic.A.Ş. 1999 yılında Türkiye'de akaryakıt sektöründe taşıt tanıma sistemleri ve otomasyon çözümleri sunan bir şirket olarak kurulsa da gün geçtikçe Türkiye'de faaliyet gösteren büyük akaryakıt ve LPG dağıtım şirketlerinin otomasyon, pompa, ödeme sistemleri ve taşıt tanıma sistemlerinin tek çözüm ortağı olarak faaliyetlerine devam etmektedir. Turpak Elektromanyetik Yakıt İkmal Sis. Tic.A.Ş. Enerji Piyasası Düzenleme Kurumu (EPDK) tarafından yetkilendirilmiş otomasyon şirketi lisansına sahip olup ayrıca maliye bakanlığı onaylı yazarkasa üreticisidir.

4.2. Sistem Üzerindeki İşleyiş Hakkında Genel Bilgi

Veri madenciliği uygulama adımlarına geçmeden önce sürecin nasıl işlediği ile ilgili bilgilendirme yapmak sağlıklı olacaktır. Süreç herhangi bir akaryakıt firmasına ait olan istasyonun kuruluş süreci ile başlar. Kuruluş sürecinde teknik destek ve istasyondan şirketin veri tabanına aktarım ve bundan sonraki destek süreci ile ilgili her faaliyet Turpak Elektromanyetik Yakıt İkmal Sis. Tic.A.Ş tarafından yürütülür. İstasyonların kurulum aşamasından sonraki ortaya çıkabilecek sorunlar çağrı merkezi ve saha destek ekipleri tarafından giderilmektedir. İstasyon kurulum aşamasından sonra istasyon ile ilgili günlük veriler (satışlar, envanterler, tank bilgileri, sıcaklık bilgileri, vs…) veri tabanına günlük olarak iletilir. İletilen veriler farklı sql tabloları üzerinden farklı departmanlara iletilir. Verinin iletildiği ve günlük olarak analizlerinin yapıldığı departmanlardan biri de WSM (Wet Stock Management) departmanıdır. Yakıt Stoğu Yönetimi (WSM), tüm veriler merkez ofise iletilecek şekilde, maksimum veri tutarlılığı ve minimum veri kaybı ile tüm geçmiş stokları takip etme imkânı sağlarken aynı zamanda geçmiş verileri istatistiksel yöntemlerle harmanlayarak gelecek verilerin tutarlı olarak aktarılmasına olanak sağlar. WSM (Wet Stock Management) sağladığı hizmetler petrol şirketlerine yer altı depolama tanklarında envanter kontrolünü doğru yapma olanağı sağlar. WSM departmanında analiz ve tespitler şirket içerisinde ve kendine özgü veri analiz yöntemlerini ve uygulamalarını kullanarak analizler yapar ve 50 temel kategori de verileri sınıflandırarak gerekli durumlarda aksiyonların alınmasını ve analiz sonuçlarının Enerji Piyasası Düzenleme Kurumu (EPDK) raporlanmasında görev alır.

4.3. Analiz Verilerinin Önişleme Uygulaması ve Test Ayarları

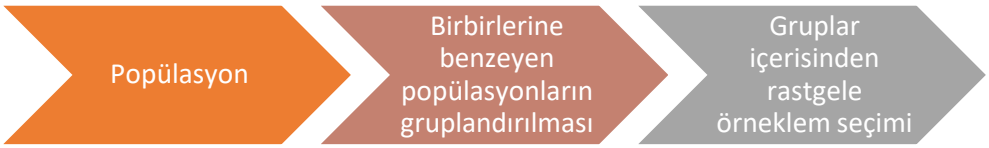
Veri tabanından alınan veriler ile uygulamaya başlamadan önce verilerin önişleme adımı gerçekleştirilmiştir. Veriler üzerinden analize herhangi bir önemi olmadığından Tarih, Tank No, İstasyon Adı, İstasyon Kodu, Agent sütunları çıkarılmıştır. Sınıflandırma uygulamasında kullanılacak olan kategoriler şirket için önemli olan ve tespit edilmesi zor olan kategoriler içerisinden 4 ana kategori olarak belirlenmiştir. Bu kategoriler Sıcaklık değişimi, Tank kalibrasyonu, Dolum-tesisat ve Arıza kategorileridir. Arıza kategorisi 4 farklı kategori olan ATG, probe, elektrik ve router arızaların birleşmesiyle oluşturulmuştur. Analizimizde her kategori ayrı olarak değerlendirilmiştir. Kategoriler ile ilgili genel bilgiler aşağıda tablo-7'de gösterilmiştir.

Tablo 7. Kategori Açıklama

KATEGORİ	AÇIKLAMA
Sıcaklık Değişimi	Tank içerisinde bulunan yakıtın mevsimsel nedenlerle veya tank içerisine dökülen yakıt ile etkileşime girmesi sonucunda sıcaklık farklılıkları sonucunda farklar oluşmakta olup bu kategori sıcaklık değişimi kategorisi olarak adlandırılmaktadır. Tankta bulunan yakıtın sıcaklığı arttıkça envanter artışı sıcaklık düştükçe envanter azalışı görülmektedir. Manipülasyon çıkarımlarında sıcaklık değişimi kategorisi önemli bir yer tutmaktadır.
Tank Kalibrasyonu	Tank kalibrasyonu, tank içerisinde bulunan yakıt hacminin belirli bir referans değerine göre değerlendirilmesi işlemi olarak adlandırılmaktadır.
Dolum-Tesisat	İstasyonlara tank bazlı olarak yapılan dolumlar sonrası oluşan alarm kategorisidir.
Arıza	Arıza kategorisi istasyonlarda meydana gelen arızaların tespiti sonrasında farklar yaratan alarm kategorisidir. Çalışmamızda arıza kategorisi ATG, probe, router ve elektrik arızaları olmak üzere 4 alt grup altında analizlere dahil edilmiştir.

Kaynak: "yazar tarafından oluşturulmuş/derlenmiştir"

Verilerin ön işlenmesi ve gerekli temizleme işlemlerinin yapılmasından sonra excel üzerinde bulunan veri grubu RAPIDMINER uygulamasına retrieve sekmesi sayesinde aktarılmıştır. Aktarılma işlemi sırasında açılış, kapanış, azalma miktarı, satış, fark, oran, sızıntı eşik değeri (SEL) gibi sayısal özellikler nümerik değerler olarak tanımlanırken dolum bilgisi ınteger ve kategori ise nominal değer olarak tanımlanmıştır. Kategori özelliğinde birden çok grup olduğundan dolayı RAPIDMINER uygulamasında yer alan Set Role ile nominal olan bu özellik polynominal değer olarak değiştirilmiştir. Bu işlemden sonra örnek veri grubunun istenen doğrultuda alt kümesini oluşturmak için Split Data kullanılmıştır. Split Data ayarlarında örneklem türü automatic olarak seçilirken bölümler parametreleri 0,3 ve 0,70 oranlarına sahip iki bölüm olacak şekilde yapılandırılmıştır. İstenen özelliklerde alt kümeler oluşturulduktan sonra analizimizde kullanılmayacak özellikler Filter Examples kullanılarak analizden çıkarılmıştır. Analizimize başlamadan önce RAPIDMINER sınıflandırma kütüphanesine eksik olan sınıflandırma algoritmaları eklenmiştir. Bazı sınıflandırma algoritmalarını kullanmadan önce algoritmanın yapısına göre öznitelik değişimleri gerçekleştirilmiştir. Uygun verinin test edilmesi için Cross Validation kullanılmıştır. Cross Validation özelliklerinde bulunan çaprazlama seçeneği 10 olarak seçilirken örneklem türü stratified sampling (katmanlı örneklem) olarak belirlenmiştir. Stratified Sampling (katmanlı örneklem) bir popülasyonun daha küçük alt gruplara bölünebilmesini sağlayarak incelenecek olan tüm popülasyonlar içerisinden popülasyonu en iyi şekilde temsil eden popülasyonun belirlenmesine yardımcı olur. Stratified Sampling (katmanlı örneklem) süreci şekil.16 gösterilmiştir.



Şekil 16. Stratified Sampling

Kaynak: "yazar tarafından oluşturulmuş/derlenmiştir"

4.4. Sınıflandırma Algoritmalarının Uygulanması

Uygun hale getirilen analiz verileri sonrasında verilere uygun olarak belirlenen veri madenciliği sınıflandırma yöntemleri verilere uygulanarak ayrı başlıklar altında değerlendirilerek sonuçlar elde edilmiş ve başarıları değerlendirilmiştir. Çalışmamızda kullanılan veri madenciliği sınıflandırma yöntemleri aşağıdaki gibidir.

- K-En Yakın Komşu Algoritması
- Gradient Boosted Trees Algoritması
- ADABOOST Algoritması
- Rastgele Orman Algoritması
- Karar Ağacı Algoritması (J48).

4.4.1. K-En Yakın Komşu Algoritması Uygulaması

RAPIDMINER programında sınıflandırma algoritmaları içerisinde yer alan K-En Yakın Komşu Algoritması veriler üzerine uygulanmıştır. Uygulamada 4 kategori ayrı ayrı değerlendirilmiş olup sonuçlar ayrı olarak hesaplanmıştır. Uygulama aşamasında algoritmanın parametreleri arasında yer alan k değeri 3 olarak belirlenmiştir. Parametrenin belirleme aşamasında model birden çok k değeri seçilerek çalıştırılmış ve en uygun değerin 3 olduğu görülmüştür. Literatüre bakıldığında ise yapılan çalışmalarda k değerinin 3-10 arasında bir değer olarak belirlendiği görülmüştür. Modelin sonuçları farklı başarı ölçütleri (TP, kesinlik, hassasiyet, F ölçütü, ROC değeri) hesaplanmış olup sonuçlar tablo.8 ‘de gösterilmiştir.

Tablo 8. KNN Algoritması Başarı Sonuçları

	Gerçek Pozitif (TP)	Kesinlik (Precision)	Hassasiyet (Recall)	F-Ölçütü	ROC Değeri
Tank Kalibrasyonu	0,6779	0,6779	0,6027	0,6375	0,8060
Dolum-Tesisat	0,8018	0,8018	0,8995	0,8473	0,8666
Sıcaklık Değişimi	0,6184	0,6184	0,4654	0,5291	0,8500
Arıza	0,963	0,9638	0,9989	0,9811	0,8780

Kaynak: “yazar tarafından oluşturulmuş/derlenmiştir”

```
positive_predictive_value: 67.79% +/- 5.08% (micro average: 67.58%) (positive class: 1)
ConfusionMatrix:
True:   0       1
0:      3259    503
1:      366     763
```

Şekil 17. KNN Algoritması Kalibrasyon Modeli Karmaşıklık Matrisi

Kaynak: "yazar tarafından oluşturulmuş/derlenmiştir"

Kalibrasyon modeli için KNN algoritması çalıştırılmış olup Şekil.17 de verilen karmaşıklık matrisi oluşturulmuştur. Toplamda kalibrasyon olarak nitelendirilen toplam 1266 alarm varken algoritma 763 tanesini doğru olarak sınıflandırmıştır. En yakın komşu kalibrasyon modelinin başarı oranı %67,79 olarak hesaplanmıştır.

```
positive_predictive_value: 80.18% +/- 1.88% (micro average: 80.15%) (positive class: 1)
ConfusionMatrix:
True:   0       1
0:      1782    258
1:      566     2285
```

Şekil 18. KNN Algoritması Dolum Modeli Karmaşıklık Matrisi

Kaynak: "yazar tarafından oluşturulmuş/derlenmiştir"

Dolum-tesisat modeli için KNN algoritması çalıştırılmış olup Şekil.18 de verilen karmaşıklık matrisi oluşturulmuştur. Toplamda dolum-tesisat kategorisi olarak nitelendirilen toplam 2543 alarm varken algoritma 2285 tanesini doğru olarak sınıflandırmıştır. En yakın komşu kalibrasyon modelinin başarı oranı %80,18 olarak hesaplanmıştır

```
positive_predictive_value: 61.84% +/- 4.95% (micro average: 61.79%) (positive class: 1)
ConfusionMatrix:
True:   0       1
0:      4166    301
1:      162     262
```

Şekil 19.KNN Algoritması Sıcaklık Değişimi Modeli Karmaşıklık Matrisi

Kaynak: "yazar tarafından oluşturulmuş/derlenmiştir"

Sıcaklık Değişimi modeli için KNN algoritması çalıştırılmış olup Şekil.19 de verilen karmaşıklık matrisi oluşturulmuştur. Toplamda sıcaklık değişimi kategorisi olarak nitelendirilen toplam 563 alarm varken algoritma 262 tanesini doğru olarak sınıflandırmıştır. En yakın komşu kalibrasyon modelinin başarı oranı %61,84 olarak hesaplanmıştır.

```
positive_predictive_value: 98.69% +/- 1.86% (micro average: 98.61%) (positive class: 1)
ConfusionMatrix:
True:   0       1
0:      4367    164
1:      5       355
```

Şekil 20. KNN Algoritması Arıza Modeli Karmaşıklık Matrisi

Kaynak: "yazar tarafından oluşturulmuş/derlenmiştir"

Arıza modeli için KNN algoritması çalıştırılmış olup Şekil.20 de verilen karmaşıklık matrisi oluşturulmuştur. Toplamda arıza kategorisi olarak nitelendirilen toplam 519 alarm varken algoritma 355 tanesini doğru olarak sınıflandırmıştır. En yakın komşu kalibrasyon modelinin başarı oranı %96,38 olarak hesaplanmıştır.

4.4.2. Rastgele Orman Algoritması Uygulaması

Rastgele Orman Algoritması, RAPIDMINER programı içerisinde tahminleme modelleri altında yer alan karar ağaçları algoritmaları içerisinde yer alan model Random Forest olarak bilinmektedir. Analize başlamadan önce Random Forest algoritmasının özelikleri tanımlanmalıdır. Rastgele orman algoritmasının özellikleri içerisinde yer alan oluşturulacak olan ağaç sayısı 100 olarak belirlenmiştir. Ayrıca sınıflandırılacak olan verinin hangi özelliğinden dallandırılacağına karar veren kriter özelliği gain ratio olarak seçilmiştir. Random forest algoritmasında 4 kategori ayrı ayrı değerlendirilmiş olup seçilen özellikler 4 kategori içinde aynı olarak belirlenmiştir. Modelin sonuçları farklı başarı ölçütleri (TP, kesinlik, hassasiyet, F ölçütü, ROC değeri) hesaplanmış olup sonuçlar tablo.9 'da gösterilmiştir.

Tablo 9. Random Forest Algoritması Başarı Sonuçları

	Gerçek Pozitif (TP)	Kesinlik (Precision)	Hassasiyet (Recall)	F-Ölçütü	ROC Değeri
Tank Kalibrasyonu	0,8034	0,8034	0,9076	0,8518	0,974
Dolum-Tesisat	0,9086	0,9086	0,9953	0,9499	0,997
Sıcaklık Değişimi	0,8228	0,8228	0,7974	0,8088	0,977
Arıza	0,9724	0,9724	0,9937	0,9846	0,974

Kaynak: "yazar tarafından oluşturulmuş/derlenmiştir"

```
positive_predictive_value: 80.34% +/- 3.46% (micro average: 80.18%) (positive class: 1)
ConfusionMatrix:
True:   0       1
0:      3341    117
1:      284     1149
```

Şekil 21.RO Algoritması Kalibrasyon Modeli Karmaşıklık Matrisi

Kaynak: "yazar tarafından oluşturulmuş/derlenmiştir"

Kalibrasyon modeli için Random Forest algoritması çalıştırılmış olup Şekil.21'de de verilen karmaşıklık matrisi oluşturulmuştur. Toplamda kalibrasyon olarak nitelendirilen toplam 1266 alarm varken algoritma 1149 tanesini doğru olarak sınıflandırmıştır. Random Forest algoritması kalibrasyon modelinin başarı oranı %80,18 olarak hesaplanmıştır.

```
positive_predictive_value: 90.86% +/- 1.09% (micro average: 90.85%) (positive class: 1)
ConfusionMatrix:
True:   0       1
0:      2093    12
1:      255     2531
```

Şekil 22. RO Algoritması Dolum Modeli Karmaşıklık Matrisi

Kaynak: "yazar tarafından oluşturulmuş/derlenmiştir"

Dolum-tesisat modeli için Random Forest algoritması çalıştırılmış olup Şekil.22'de de verilen karmaşıklık matrisi oluşturulmuştur. Toplamda dolum-tesisat olarak nitelendirilen toplam 2543 alarm varken algoritma 2531 tanesini doğru olarak

sınıflandırmıştır. Random Forest algoritması dolum-tesisat modelinin başarı oranı %90,86 olarak hesaplanmıştır.

```
positive_predictive_value: 82.28% +/- 4.39% (micro average: 82.08%) (positive class: 1)
ConfusionMatrix:
True:   0       1
0:      4230    114
1:      98      449
```

Şekil 23. RO Algoritması Sıcaklık Değişimi Modeli Karmaşıklık Matrisi

Kaynak: "yazar tarafından oluşturulmuş/derlenmiştir"

Sıcaklık değişimi modeli için Random Forest algoritması çalıştırılmış olup Şekil.23'te de verilen karmaşıklık matrisi oluşturulmuştur. Toplamda sıcaklık değişimi olarak nitelendirilen toplam 563 alarm varken algoritma 449 tanesini doğru olarak sınıflandırmıştır. Random Forest algoritması sıcaklık değişimi modelinin başarı oranı %82,28 olarak hesaplanmıştır.

```
positive_predictive_value: 97.11% +/- 3.24% (micro average: 97.05%) (positive class: 1)
ConfusionMatrix:
True:   0       1
0:      4360    124
1:      12      395
```

Şekil 24. RO Algoritması Arıza Modeli Karmaşıklık Matrisi

Kaynak: "yazar tarafından oluşturulmuş/derlenmiştir"

Arıza modeli için Random Forest algoritması çalıştırılmış olup Şekil.24'te de verilen karmaşıklık matrisi oluşturulmuştur. Toplamda arıza olarak nitelendirilen toplam 519 alarm varken algoritma 395 tanesini doğru olarak sınıflandırmıştır. Random Forest algoritması arıza modelinin başarı oranı %97,24 olarak hesaplanmıştır.

4.4.3. Gradient Boosted Trees Algoritması Uygulaması

Gradient Boosted Trees algoritması, RAPIDMINER programı içerisinde tahminleme modelleri altında yer alan karar ağaçları algoritmaları içerisinde yer almaktadır. Algoritma Random Forest algoritmasında olduğu gibi karar ağacı temeline dayanan bir algoritmadır. Algoritmanın her koşulda iyi sonuçlar verebilmesi yöntemin en avantajlı özelliklerinden biridir. Analize başlamadan önce Gradient Boosted Trees

algoritmasının özelikleri tanımlanmalıdır. Algoritmanın özellikleri içerisinde yer alan oluşturulacak olan ağaç sayısı 50 olarak belirlenmiştir. Özeliklerin belirlenmesinde literatürde yapılan çalışmalardan faydalanılmıştır. Gradient Boosted Tress algoritmasında 4 kategori ayrı ayrı değerlendirilmiş olup seçilen özellikler 4 kategori içinde aynı olarak belirlenmiştir. Modelin sonuçları farklı başarı ölçütleri (TP, kesinlik, hassasiyet, F ölçütü, ROC değeri) hesaplanmış olup sonuçlar tablo.9 'da gösterilmiştir.

Tablo 10. Gradient Boosted Trees Algoritması Başarı Sonuçları

	Gerçek Pozitif (TP)	Kesinlik (Precision)	Hassasiyet (Recall)	F-Ölçütü	ROC Değeri
Tank Kalibrasyonu	0,7534	0,7534	0,8957	0,8180	0,9560
Dolum-Tesisat	0,9104	0,9803	0,9953	0,9435	0,9820
Sıcaklık Değişimi	0,7521	0,7521	0,7654	0,7578	0,9580
Arıza	0,9350	0,9350	0,7957	0,8575	0,9420

Kaynak: "yazar tarafından oluşturulmuş/derlenmiştir"

```
positive_predictive_value: 75.34% +/- 3.80% (micro average: 75.25%) (positive class: 1)
ConfusionMatrix:
True:   0       1
0:      3252    132
1:      373     1134
```

Şekil 25. XGBOOST Algoritması Kalibrasyon Modeli Karmaşıklık Matrisi

Kaynak: "yazar tarafından oluşturulmuş/derlenmiştir"

Kalibrasyon modeli için Gradient Boosted Trees algoritması çalıştırılmış olup Şekil.25'te de verilen karmaşıklık matrisi oluşturulmuştur. Toplamda kalibrasyon olarak nitelendirilen toplam 1266 alarm varken algoritma 1134 tanesini doğru olarak sınıflandırmıştır. Gradient Boosted Trees algoritması kalibrasyon modelinin başarı oranı %75,34 olarak hesaplanmıştır.

```
positive_predictive_value: 91.04% +/- 3.47% (micro average: 90.89%) (positive class: 1)
ConfusionMatrix:
True:    0       1
0:       2098    50
1:       250     2493
```

Şekil 26. XGBOOST Algoritması Dolum Modeli Karmaşıklık Matrisi

Kaynak: "yazar tarafından oluşturulmuş/derlenmiştir"

Dolum-tesisat modeli için Gradient Boosted Trees algoritması çalıştırılmış olup Şekil.26'de de verilen karmaşıklık matrisi oluşturulmuştur. Toplamda dolum-tesisat olarak nitelendirilen toplam 2543 alarm varken algoritma 2493 tanesini doğru olarak sınıflandırmıştır. Gradient Boosted Trees algoritması dolum-tesisat modelinin başarı oranı %91,04 olarak hesaplanmıştır.

```
positive_predictive_value: 75.21% +/- 5.51% (micro average: 74.96%) (positive class: 1)
ConfusionMatrix:
True:    0       1
0:       4184    132
1:       144     431
```

Şekil 27. XGBOOST Algoritması Sıcaklık Değişimi Modeli Karmaşıklık Matrisi

Kaynak: "yazar tarafından oluşturulmuş/derlenmiştir"

Sıcaklık değişimi modeli için Gradient Boosted Trees algoritması çalıştırılmış olup Şekil.27'te de verilen karmaşıklık matrisi oluşturulmuştur. Toplamda sıcaklık değişimi olarak nitelendirilen toplam 563 alarm varken algoritma 431 tanesini doğru olarak sınıflandırmıştır. Gradient Boosted Trees algoritması sıcaklık değişimi modelinin başarı oranı %75,21 olarak hesaplanmıştır.

```
positive_predictive_value: 93.50% +/- 5.08% (micro average: 93.23%) (positive class: 1)
ConfusionMatrix:
True:    0       1
0:       4342    106
1:       30      413
```

Şekil 28. XGBOOST Algoritması Arıza Modeli Karmaşıklık Matrisi

Kaynak: "yazar tarafından oluşturulmuş/derlenmiştir"

Arıza modeli için Gradient Boosted Trees algoritması çalıştırılmış olup Şekil.28'te de verilen karmaşıklık matrisi oluşturulmuştur. Toplamda arıza olarak nitelendirilen toplam 519 alarm varken algoritma 413 tanesini doğru olarak sınıflandırmıştır. Random Forest algoritması arıza modelinin başarı oranı %93,50 olarak hesaplanmıştır.

4.4.4. ADABOOST Algoritması Uygulaması

ADABOOST algoritması, RAPIDMINER programı içerisinde tahminleme modelleri altında yer alan karar ağaçları algoritmaları içerisinde yer almaktadır. Algoritma en iyi sonuca ulaşmak için oluşturduğu her karar ağacı turunda yanlış olarak sınıflandırılan verileri toplayarak bir sonraki karar ağacı turunda geri beslemek için kullanıldığı bir modele dayanmaktadır. Algoritmanın her koşulda iyi sonuçlar verebilmesi yöntemin en avantajlı özelliklerinden biridir. Analize başlamadan önce ADABOOST algoritması için belirlenen değişken özelikleri tanımlanmalıdır. ADABOOST algoritmasını RAPIDMINER programı içerisinde kullanabilmemiz için ADABOOST algoritması içerisinde farklı bir sınıflandırma algoritması kullanılmalıdır. Analizimizde kullanılacak olan bu algoritma Gradient Boosted Trees algoritması olarak belirlenmiştir. Modelin sonuçları farklı başarı ölçütleri (TP, kesinlik, hassasiyet, F ölçütü, ROC değeri) hesaplanmış olup sonuçlar tablo.11 'da gösterilmiştir.

Tablo 11. ADABOOST Algoritması Başarı Sonuçları

	Gerçek Pozitif (TP)	Kesinlik (Precision)	Hassasiyet (Recall)	F-Ölçütü	ROC Değeri
Tank Kalibrasyonu	0,8260	0,8260	0,8831	0,8530	0,9720
Dolum-Tesisat	0,9680	0,9680	0,9862	0,9770	0,9960
Sıcaklık Değişimi	0,7791	0,7791	0,8080	0,7921	0,9670
Arıza	0,9222	0,9222	0,8187	0,8668	0,9650

Kaynak: "yazar tarafından oluşturulmuş/derlenmiştir"

```
positive_predictive_value: 82.60% +/- 3.44% (micro average: 82.51%) (positive class: 1)
ConfusionMatrix:
True:   0       1
0:      3388    148
1:      237     1118
```

Şekil 29. ADABOOST Kalibrasyon Modeli Karmaşıklık Matrisi

Kaynak: "yazar tarafından oluşturulmuş/derlenmiştir"

Kalibrasyon modeli için ADABOOST algoritması çalıştırılmış olup Şekil.29'te de verilen karmaşıklık matrisi oluşturulmuştur. Toplamda kalibrasyon olarak nitelendirilen toplam 1266 alarm varken algoritma 1118 tanesini doğru olarak sınıflandırmıştır. ADABOOST algoritması kalibrasyon modelinin başarı oranı %82,60 olarak hesaplanmıştır.

```
positive_predictive_value: 96.80% +/- 0.70% (micro average: 96.80%) (positive class: 1)
ConfusionMatrix:
True:   0       1
0:      2265    35
1:      83      2508
```

Şekil 30. ADABOOST Dolum-Tesisat Modeli Karmaşıklık Matrisi

Kaynak: "yazar tarafından oluşturulmuş/derlenmiştir"

Dolum-tesisat modeli için ADABOOST algoritması çalıştırılmış olup Şekil.30'de de verilen karmaşıklık matrisi oluşturulmuştur. Toplamda dolum-tesisat olarak nitelendirilen toplam 2543 alarm varken algoritma 2508 tanesini doğru olarak sınıflandırmıştır. ADABOOST algoritması dolum-tesisat modelinin başarı oranı %96,80 olarak hesaplanmıştır.

```
positive_predictive_value: 77.91% +/- 5.94% (micro average: 77.51%) (positive class: 1)
ConfusionMatrix:
True:   0       1
0:      4196    108
1:      132     455
```

Şekil 31. ADABOOST Sıcaklık Değişimi Modeli Karmaşıklık Matrisi

Kaynak: "yazar tarafından oluşturulmuş/derlenmiştir"

Sıcaklık değişimi modeli için ADABOOST algoritması çalıştırılmış olup Şekil.31'te de verilen karmaşıklık matrisi oluşturulmuştur. Toplamda sıcaklık değişimi olarak nitelendirilen toplam 563 alarm varken algoritma 455 tanesini doğru olarak sınıflandırmıştır. ADABOOST algoritması sıcaklık değişimi modelinin başarı oranı %77,91 olarak hesaplanmıştır.

```
positive_predictive_value: 92.22% +/- 3.22% (micro average: 92.19%) (positive class: 1)
ConfusionMatrix:
True:   0       1
0:      4336    94
1:      36      425
```

Şekil 32. ADABOOST Arıza Modeli Karmaşıklık Matrisi

Kaynak: "yazar tarafından oluşturulmuş/derlenmiştir"

Arıza modeli için ADABOOST algoritması çalıştırılmış olup Şekil.32'te de verilen karmaşıklık matrisi oluşturulmuştur. Toplamda arıza olarak nitelendirilen toplam 519 alarm varken algoritma 425 tanesini doğru olarak sınıflandırmıştır. ADABOOST algoritması arıza modelinin başarı oranı %92,22 olarak hesaplanmıştır.

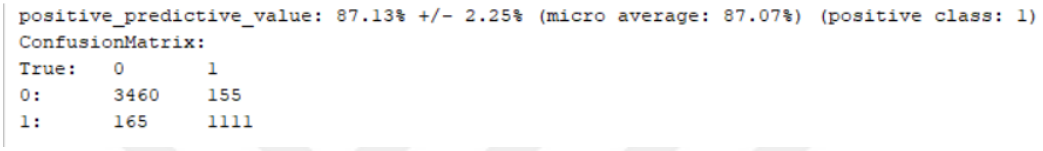
4.4.5.Karar Ağacı (J48) Algoritması Uygulaması

Uygulamada kullanılacak olan karar ağacı algoritması J48 algoritması olarak belirlenmiştir. J48 algoritması C 4.5 algoritması olarak da bilinmektedir. Algoritma analiz sırasında kayıp değer olmayan verilerden yararlanmaktadır. Algoritma karar ağacını oluştururken sınıflandırılacak olan verinin hangi özelliğinden dallanacağına karar veren özellik seçim ölçütlerinden kazanım oranı ölçütünü dikkate alarak dallandırma işlemini gerçekleştirmektedir. Analizimize başlamadan önce algoritmanın değişken özelliklerinin belirlenmesi gerekmektedir. Değişken özellikleri içerisinde yer alan güven değeri 0,25 olarak belirlenmiştir. Dört farklı model için karar ağaçları yapısı ayrı ayrı olarak gösterilmiştir. Modelin sonuçları farklı başarı ölçütleri (TP, kesinlik, hassasiyet, F ölçütü, ROC değeri) hesaplanmış olup sonuçlar tablo.12 'da gösterilmiştir.

Tablo 12. Karar Ağacı (J48) Algoritması Başarı Sonuçları

	Gerçek Pozitif (TP)	Kesinlik (Precision)	Hassasiyet (Recall)	F-Ölçütü	ROC Değeri
Tank Kalibrasyonu	0,8713	0,8713	0,8776	0,8740	0,9250
Dolum-Tesisat	0,9844	0,9844	0,9858	0,9851	0,9890
Sıcaklık Değişimi	0,8088	0,8088	0,8117	0,8090	0,9130
Arıza	0,9332	0,9332	0,8285	0,8768	0,9160

Kaynak: "yazar tarafından oluşturulmuş/derlenmiştir"



Şekil 33. Karar Ağacı (J48) Kalibrasyon Modeli Karmaşıklık Matrisi

Kaynak: "yazar tarafından oluşturulmuş/derlenmiştir"

Kalibrasyon modeli için karar ağacı (J48) algoritması çalıştırılmış olup Şekil.33'te de verilen karmaşıklık matrisi oluşturulmuştur. Toplamda kalibrasyon olarak nitelendirilen 1266 alarm varken algoritma 1111 tanesini doğru olarak sınıflandırmıştır. Karar ağacı (J48) algoritması kalibrasyon modelinin başarı oranı %87,13 olarak hesaplanmıştır. Karar ağacı (J48) Algoritmasında karar ağacı oluştururken sınıflandırılacak olan verinin hangi özellikten dallanacağına karar veren ölçüt bilgi kazanım oranıdır. Bilgi kazanım oranına göre şekil 34'te görüldüğü gibi kalibrasyon modeli için kök düğümün rate % (oran) olduğu görülmüştür.

Rate (%) <= 1.722134

| Rate (%) <= -1.272844

| | Difference <= -20.7211

| | | Close <= 224.75: 0 (54.0)

| | | Close > 224.75

| | | | Difference <= -52.3268

| | | | | Rate (%) <= -25.319752

| | | | | | Close <= 19451.4004

| | | | | | | Difference <= -1810.5673: 0 (9.0)

| | | | | | | Difference > -1810.5673

| | | | | | | | Refuel <= 0

| | | | | | | | | Tank No <= 5: 1 (22.0/7.0)

| | | | | | | | | Tank No > 5: 0 (2.0)

| | | | | | | | Refuel > 0

| | | | | | | | | Tank No <= 4: 0 (4.0)

| | | | | | | | | Tank No > 4: 1 (2.0)

| | | | | | Close > 19451.4004: 1 (13.0/1.0)

| | | | | Rate (%) > -25.319752: 1 (376.0/32.0)

| | | | Difference > -52.3268

| | | | | Close <= 4675.2963

| | | | | | Refuel <= 0: 1 (114.0/4.0)

| | | | | | Refuel > 0

| | | | | | | Open <= 683.9027: 0 (5.0)

| | | | | | | Open > 683.9027

| | | | | | | | Difference <= -27.5027: 1 (6.0)

| | | | | | | | Difference > -27.5027

| | | | | | | | | Reductıon <= 381.3198: 1 (2.0)

| | | | | | | | | Reductıon > 381.3198: 0 (5.0)

| | | | | Close > 4675.2963

| | | | | | Difference <= -40.0329

| | | | | | | Refuel <= 0

| | | | | | | | Oil Type = MOTORIN: 0 (52.0/24.0)

| | | | | | | | Oil Type = MotorinEkst

| | | | | | | | | Reductıon <= 2234.0061: 1 (7.0)

| | | | | | | | | Reductıon > 2234.0061: 0 (3.0)

| | | | | | | | Oil Type = KUR95

| | | | | | | | | Open <= 9633.1377

| | | | | | | | | | Open <= 8661.0488: 0 (2.0)

| | | | | | | | | | Open > 8661.0488: 1 (2.0)

| | | | | | | | | Open > 9633.1377: 0 (5.0)

| | | | | | | Refuel > 0

| | | | | | | | Oil Type = MOTORIN

| | | | | | | | | Difference <= -44.6737

| | | | | | | | | | Tank No <= 1: 0 (3.0)

| | | | | | | | | | Tank No > 1

| | | | | | | | | | | Tank No <= 3: 1 (6.0)

| | | | | | | | | | | Tank No > 3

| | | | | | | | | | | | Close <= 10505.6376: 1 (2.0)

| | | | | | | | | | | | Close > 10505.6376: 0 (3.0)

| | | | | | | | | Difference > -44.6737: 1 (10.0)

| | | | | | | | Oil Type = MotorinEkst

| | | | | | | | | Open <= 2938.2798: 1 (2.0)

| | | | | | | | | Open > 2938.2798: 0 (2.0)

| | | | | | | | Oil Type = KUR95

| | | | | | | | | Close <= 9696.9898: 1 (3.0/1.0)

| | | | | | | | | Close > 9696.9898: 0 (2.0)

| | | | | | Difference > -40.0329

| | | | | | | Rate (%) <= -1.962117

| | | | | | | Rate (%) <= -5.497698: 0 (85.0/12.0)

| | | | | | | Rate (%) > -5.497698

| | | | | | | | Refuel <= 0

| | | | | | | | | Tank No <= 1: 0 (30.0/6.0)

| | | | | | | | | Tank No > 1

| | | | | | | | | | Oil Type = MOTORIN

| | | | | | | | | | | Open <= 10709.3686: 1 (18.0/4.0)

| | | | | | | | | | | Open > 10709.3686

| | | | | | | | | | | | Reductıon <= 1194.0963

| | | | | | | | | | | | | Difference <= -29.6542

| | | | | | | | | | | | | | Difference <= -39.0603: 0 (2.0)

| | | | | | | | | | | | | | Difference > -39.0603

| | | | | | | | | | | | | | | Rate (%) <= -3.477129: 1 (4.0)

| | | | | | | | | | | | | | | Rate (%) > -3.477129: 0 (3.0/1.0)

| | | | | | | | | | | | | | Difference > -29.6542: 0 (6.0)

| | | | | | | | | | | | | Reductıon > 1194.0963: 0 (16.0)

| | | | | | | | | | | Oil Type = MotorinEkst: 1 (25.0/9.0)

| | | | | | | | | | Oil Type = KUR95

| | | | | | | | | | | Reductıon <= 940.0411: 0 (15.0/1.0)

| | | | | | | | | | | Reductıon > 940.0411

| | | | | | | | | | | | Tank No <= 4: 1 (10.0/1.0)

| | | | | | | | | | | | Tank No > 4

| | | | | | | | | | | | | Open <= 13286.9118: 0 (3.0)

| | | | | | | | | | | | | Open > 13286.9118: 1 (2.0)

| | | | | | | | Refuel > 0

| | | | | | | | | Oil Type = MOTORIN

| | | | | | | | | | Rate (%) <= -2.44396

| | | | | | | | | | | Reductıon <= 1161.1601

| | | | | | | | | | | | Close <= 8683.5928: 1 (4.0/1.0)

| | | | | | | | | | | | Close > 8683.5928: 0 (10.0)

| | | | | | | | | | | Reductıon > 1161.1601: 1 (3.0)

| | | | | | | | | | Rate (%) > -2.44396: 0 (10.0)

| | | | | | | | | Oil Type = MotorinEkst

| | | | | | | | | | Tank No <= 1: 1 (2.0)

| | | | | | | | | | Tank No > 1: 0 (5.0)

| | | | | | | | | Oil Type = KUR95

| | | | | | | | | | Tank No <= 2: 0 (6.0)

| | | | | | | | | | Tank No > 2

| | | | | | | | | | | Rate (%) <= -2.492317

| | | | | | | | | | | | Open <= 1654.072: 1 (3.0/1.0)

| | | | | | | | | | | | Open > 1654.072: 0 (8.0)

| | | | | | | | | | | Rate (%) > -2.492317: 1 (2.0)

| | | | | | | Rate (%) > -1.962117: 0 (36.0)

| | Difference > -20.7211: 0 (227.0/2.0)

| Rate (%) > -1.272844

| | Difference <= 47.7546: 0 (2484.0/5.0)

| | Difference > 47.7546

| | | Rate (%) <= 1.18634: 0 (117.0/2.0)

| | | Rate (%) > 1.18634

| | | | Rate (%) <= 1.351015

| | | | | Difference <= 75.5837: 0 (9.0)

| | | | | Difference > 75.5837

| | | | | | Tank No <= 4: 1 (9.0/1.0)

| | | | | | Tank No > 4: 0 (2.0)

| | | | Rate (%) > 1.351015: 1 (54.0/5.0)

Rate (%) > 1.722134

| Difference <= 19.5627: 0 (166.0/2.0)

| Difference > 19.5627

| | Rate (%) <= 38.275719

| | | Difference <= 30.9651

| | | | Refuel <= 0: 1 (67.0/6.0)

| | | | Refuel > 0

| | | | | Reductıon <= 439.4482: 1 (16.0/3.0)

| | | | | Reductıon > 439.4482: 0 (60.0/8.0)

| | | Difference > 30.9651

| | | | Difference <= 247.6391: 1 (456.0/26.0)

| | | | Difference > 247.6391

| | | | | Difference <= 1249.0319: 1 (67.0/14.0)

| | | | | Difference > 1249.0319

| | | | | | Open <= 30757.7825: 0 (13.0/2.0)

| | | | | | Open > 30757.7825: 1 (4.0)

| | Rate (%) > 38.275719: 0 (114.0/4.0)

Şekil 34. Karar Ağacı (J48) Algoritması Kalibrasyon Modeli Karar Çıktısı

Kaynak: "yazar tarafından oluşturulmuş/derlenmiştir"

```
positive_predictive_value: 98.44% +/- 0.91% (micro average: 98.43%) (positive class: 1)
ConfusionMatrix:
True:   0       1
0:      2308    36
1:      40      2507
```

Şekil 35. Karar Ağacı (J48) Dolum Modeli Karmaşıklık Matrisi

Kaynak: "yazar tarafından oluşturulmuş/derlenmiştir"

Dolum-tesisat modeli için karar ağacı (J48) algoritması çalıştırılmış olup Şekil.35'de de verilen karmaşıklık matrisi oluşturulmuştur. Toplamda dolum-tesisat olarak nitelendirilen toplam 2543 alarm varken algoritma 2507 tanesini doğru olarak sınıflandırmıştır. Karar ağacı (J48) algoritması dolum-tesisat modelinin başarı oranı %98, olarak hesaplanmıştır. Bilgi kazanım oranına göre şekil 36'te görüldüğü gibi dolum tesisat modeli için kök düğümün refuel (dolum) olduğu görülmüştür.

Refuel <= 0: 0 (1661.0)

Refuel > 0

| Difference <= -21.3402

| | Rate (%) <= -1.895693

| | | Difference <= -24.6402: 0 (233.0/1.0)

| | | Difference > -24.6402

| | | | Reductıon <= 887.2656: 0 (16.0/1.0)

| | | | Reductıon > 887.2656: 1 (5.0)

| | Rate (%) > -1.895693

| | | Rate (%) <= -0.066165

| | | | Rate (%) <= -1.224838

| | | | | Difference <= -42.2778

| | | | | | Difference <= -77.223: 0 (23.0)

| | | | | | Difference > -77.223

| | | | | | | Rate (%) <= -1.365766: 0 (12.0/2.0)

| | | | | | | Rate (%) > -1.365766: 1 (6.0)

| | | | | Difference > -42.2778: 1 (22.0)

| | | | Rate (%) > -1.224838: 1 (162.0)

| | | Rate (%) > -0.066165: 0 (43.0/1.0)

| Difference > -21.3402

| | Rate (%) <= 1.722134

| | | Reductıon <= -16.9028: 0 (35.0/1.0)

| | | Reductıon > -16.9028

| | | | Difference <= 123.8454

| | | | | Difference <= -16.7379

| | | | | | Sell <= 296.87: 0 (21.0)

| | | | | | Sell > 296.87: 1 (69.0/2.0)

| | | | | Difference > -16.7379

| | | | | | Rate (%) <= 1.285786: 1 (1916.0/4.0)

| | | | | | Rate (%) > 1.285786

| | | | | | | Difference <= 63.43: 1 (146.0/3.0)

| | | | | | | Difference > 63.43

| | | | | | | | Tank No <= 2

| | | | | | | | | Open <= 6778: 0 (2.0)

| | | | | | | | | Open > 6778: 1 (2.0)

| | | | | | | | Tank No > 2: 0 (5.0)

| | | | Difference > 123.8454

| | | | | Rate (%) <= 1.11502: 1 (4.0)

| | | | | Rate (%) > 1.11502: 0 (22.0)

| | Rate (%) > 1.722134

| | | Difference <= 31.0741

| | | | Difference <= 18.2107: 1 (142.0)

| | | | Difference > 18.2107

| | | | | Rate (%) <= 5.210174

| | | | | | Difference <= 26.188

| | | | | | | Open <= 8912.299: 1 (45.0)

| | | | | | | Open > 8912.299

| | | | | | | | Reductıon <= 712.7111: 0 (3.0)

| | | | | | | | Reductıon > 712.7111: 1 (8.0)

| | | | | | Difference > 26.188

| | | | | | | Rate (%) <= 2.275751: 1 (14.0/1.0)

| | | | | | | Rate (%) > 2.275751: 0 (9.0/1.0)

| | | | | Rate (%) > 5.210174: 0 (15.0/1.0)

| | | Difference > 31.0741: 0 (250.0/4.0)

Şekil 36. Karar Ağacı (J48) Dolum Modeli Karar Çıktısı

Kaynak: "yazar tarafından oluşturulmuş/derlenmiştir"

```
positive_predictive_value: 80.88% +/- 5.99% (micro average: 80.60%) (positive class: 1)
ConfusionMatrix:
True:   0       1
0:      4218    106
1:      110     457
```

Şekil 37. Karar Ağacı (J4) Sıcaklık Değişimi Modeli Karmaşıklık Matrisi

Kaynak: "yazar tarafından oluşturulmuş/derlenmiştir"

Sıcaklık değişimi modeli için karar ağacı (J48) algoritması çalıştırılmış olup Şekil.37'te de verilen karmaşıklık matrisi oluşturulmuştur. Toplamda sıcaklık değişimi olarak nitelendirilen toplam 563 alarm varken algoritma 457 tanesini doğru olarak sınıflandırmıştır. Karar ağacı (J48) algoritması sıcaklık değişimi modelinin başarı oranı %80,88 olarak hesaplanmıştır. Bilgi kazanım oranına göre şekil 38'te görüldüğü gibi dolum tesisat modeli için kök düğümün different (fark) olduğu görülmüştür

Difference <= -16.4223

| Reductıon <= 154.7327

| | Difference <= -52.4206

| | | Reductıon <= 83.5195: 1 (6.0)

| | | Reductıon > 83.5195: 0 (7.0/1.0)

| | Difference > -52.4206: 1 (229.0/3.0)

| Reductıon > 154.7327

| | Difference <= -61.4538: 0 (542.0/6.0)

| | Difference > -61.4538

| | | Rate (%) <= -1.661702

| | | | Close <= 3917: 0 (127.0/1.0)

| | | | Close > 3917

| | | | | Difference <= -19.1352

| | | | | | Sell <= 459.82

| | | | | | | Oil Type = MOTORIN

| | | | | | | | Tank No <= 4

| | | | | | | | | Refuel <= 0

| | | | | | | | | | Tank No <= 1

| | | | | | | | | | | Open <= 22815.9805: 1 (9.0)

| | | | | | | | | | | Open > 22815.9805

| | | | | | | | | | | | Difference <= -30.9804: 0 (4.0)

| | | | | | | | | | | | Difference > -30.9804: 1 (2.0)

| | | | | | | | | | Tank No > 1

| | | | | | | | | | | Reductıon <= 401.668: 1 (8.0)

| | | | | | | | | | | Reductıon > 401.668

| | | | | | | | | | | | Reductıon <= 434.2636: 0 (4.0/1.0)

| | | | | | | | | | | | Reductıon > 434.2636: 1 (4.0/1.0)

| | | | | | | | | Refuel > 0

| | | | | | | | | | Rate (%) <= -8.721367: 0 (2.0)

| | | | | | | | | | Rate (%) > -8.721367: 1 (7.0/1.0)

| | | | | | | | Tank No > 4

| | | | | | | | | Reductıon <= 331.22: 1 (3.0/1.0)

| | | | | | | | | Reductıon > 331.22: 0 (4.0)

| | | | | | | Oil Type = MotorinEkst: 1 (11.0/1.0)

| | | | | | | Oil Type = KUR95: 1 (41.0/6.0)

| | | | | | Sell > 459.82

| | | | | | | Close <= 6421.0111

| | | | | | | | Oil Type = MOTORIN

| | | | | | | | | Difference <= -39.0879: 0 (16.0)

| | | | | | | | | Difference > -39.0879

| | | | | | | | | | Difference <= -26.8272

| | | | | | | | | | | Refuel <= 0

| | | | | | | | | | | | Rate (%) <= -2.436956: 0 (10.0/2.0)

| | | | | | | | | | | | Rate (%) > -2.436956: 1 (2.0)

| | | | | | | | | | | Refuel > 0: 1 (2.0)

| | | | | | | | | | Difference > -26.8272: 0 (7.0)

| | | | | | | | | Oil Type = MotorinEkst: 0 (18.0/1.0)

| | | | | | | | | Oil Type = KUR95

| | | | | | | | | | Tank No <= 4: 0 (3.0/1.0)

| | | | | | | | | | Tank No > 4: 1 (6.0/1.0)

| | | | | | | | Close > 6421.0111

| | | | | | | | | Tank No <= 1

| | | | | | | | | | Refuel <= 0

| | | | | | | | | | | Difference <= -44.51: 0 (17.0/7.0)

| | | | | | | | | | | Difference > -44.51

| | | | | | | | | | | | Open <= 26706.2942: 1 (38.0/1.0)

| | | | | | | | | | | | Open > 26706.2942: 0 (3.0)

| | | | | | | | | | Refuel > 0

| | | | | | | | | | | Open <= 1099.3786: 1 (5.0)

| | | | | | | | | | | Open > 1099.3786

| | | | | | | | | | | | Open <= 10735.1143: 0 (12.0/3.0)

| | | | | | | | | | | | Open > 10735.1143: 1 (3.0)

| | | | | | | | | Tank No > 1

| | | | | | | | | | Oil Type = MOTORIN

| | | | | | | | | | | Close <= 8683.5928: 0 (25.0/4.0)

| | | | | | | | | | Close > 8683.5928

| | | | | | | | | | | Refuel <= 0: 1 (60.0/21.0)

| | | | | | | | | | | Refuel > 0

| | | | | | | | | | | | Tank No <= 5

| | | | | | | | | | | | | Tank No <= 2

| | | | | | | | | | | | | | Reductıon <= 1645.7303

| | | | | | | | | | | | | | | Difference <= -28.6762

| | | | | | | | | | | | | | | | Rate (%) <= -4.236611: 0 (2.0)

| | | | | | | | | | | | | | | | Rate (%) > -4.236611: 1 (6.0)

| | | | | | | | | | | | | | | Difference > -28.6762: 0 (3.0)

| | | | | | | | | | | | | | Reductıon > 1645.7303: 0 (5.0)

| | | | | | | | | | | | | Tank No > 2

| | | | | | | | | | | | | | Reductıon <= 1314.4257: 1 (6.0)

| | | | | | | | | | | | | | Reductıon > 1314.4257

| | | | | | | | | | | | | | | Open <= 9056.999: 0 (3.0)

| | | | | | | | | | | | | | | Open > 9056.999: 1 (3.0)

| | | | | | | | | | | | Tank No > 5: 0 (2.0)

| | | | | | | | | Oil Type = MotorinEkst

| | | | | | | | | | Refuel <= 0

| | | | | | | | | | | Reductıon <= 1575.5129

| | | | | | | | | | | | Open <= 16134.6453

| | | | | | | | | | | | | Difference <= -33.8304: 0 (4.0)

| | | | | | | | | | | | | Difference > -33.8304: 1 (6.0/1.0)

| | | | | | | | | | | | Open > 16134.6453: 0 (12.0)

| | | | | | | | | | | Reductıon > 1575.5129: 1 (4.0)

| | | | | | | | | | Refuel > 0

| | | | | | | | | | | Open <= 2876.1132: 0 (4.0/1.0)

| | | | | | | | | | | Open > 2876.1132: 1 (6.0)

| | | | | | | | | Oil Type = KUR95

| | | | | | | | | | Refuel <= 0

| | | | | | | | | | | Reductıon <= 940.0411: 1 (14.0/1.0)

| | | | | | | | | | | Reductıon > 940.0411

| | | | | | | | | | | | Reductıon <= 1043.4609: 0 (5.0)

| | | | | | | | | | | | Reductıon > 1043.4609: 1 (14.0/6.0)

| | | | | | | | | | Refuel > 0

| | | | | | | | | | | Close <= 14908.625

| | | | | | | | | | | | Open <= 4764.6053: 1 (4.0/1.0)

| | | | | | | | | | | | Open > 4764.6053: 0 (4.0)

| | | | | | | | | | | Close > 14908.625: 1 (3.0)

| | | | | Difference > -19.1352

| | | | | | Refuel <= 0: 1 (2.0)

| | | | | | Refuel > 0: 0 (22.0/1.0)

| | | Rate (%) > -1.661702: 0 (207.0/2.0)

Difference > -16.4223

| Reductıon <= -16.5313

| | Reductıon <= -89: 0 (55.0)

| | Reductıon > -89: 1 (56.0/4.0)

| Reductıon > -16.5313: 0 (3202.0/21.0)

Şekil 38. Karar Ağacı (J48) Sıcaklık Değişimi Modeli Karar Çıktısı

Kaynak: "yazar tarafından oluşturulmuş/derlenmiştir"

```
positive_predictive_value: 93.32% +/- 5.04% (micro average: 93.07%) (positive class: 1)
ConfusionMatrix:
True:   0       1
0:      4340    89
1:      32      430
```

Şekil 39. Karar Ağacı (J48) Arıza Modeli Karmaşıklık Matrisi

Kaynak: "yazar tarafından oluşturulmuş/derlenmiştir"

Arıza modeli için karar ağacı (J48) algoritması çalıştırılmış olup Şekil.39'te de verilen karmaşıklık matrisi oluşturulmuştur. Toplamda arıza olarak nitelendirilen toplam 519 alarm varken algoritma 430 tanesini doğru olarak sınıflandırmıştır. Karar ağacı (J48) algoritması arıza modelinin başarı oranı %93,32 olarak hesaplanmıştır. Bilgi kazanım oranına göre şekil 40'te görüldüğü gibi arıza modeli için kök düğümün open (açılış) olduğu görülmüştür.

Open <= 0: 1 (235.0/1.0)

Open > 0

| Close <= 32.2: 1 (75.0)

| Close > 32.2

| | Rate (%) <= 30.602737

| | | Difference <= -322.76

| | | | Sell <= 35.84: 1 (46.0/2.0)

| | | | Sell > 35.84

| | | | | Close <= 20890.6009

| | | | | | Oil Type = MOTORIN

| | | | | | | Sell <= 5093.12

| | | | | | | | Difference <= -456.4038: 1 (14.0)

| | | | | | | | Difference > -456.4038

| | | | | | | | | Close <= 9951.1309: 0 (4.0)

| | | | | | | | | Close > 9951.1309: 1 (5.0)

| | | | | | | Sell > 5093.12: 0 (16.0/3.0)

| | | | | | Oil Type = MotorinEkst: 0 (6.0/1.0)

| | | | | | Oil Type = KUR95: 1 (2.0/1.0)

| | | | | Close > 20890.6009: 0 (14.0)

| | | Difference > -322.76

| | | | Close <= 1238.2407

| | | | | Reductıon <= 255.6604

| | | | | | Refuel <= 0

| | | | | | | Sell <= 67.88: 1 (7.0)

| | | | | | | Sell > 67.88: 0 (3.0/1.0)

| | | | | | Refuel > 0: 0 (2.0)

| | | | | Reductıon > 255.6604: 0 (191.0/11.0)

| | | | Close > 1238.2407

| | | | | Difference <= -51.5059

| | | | | | Reductıon <= 614.9215

| | | | | | | Open <= 8249.0927

| | | | | | | | Close <= 3377.0591: 0 (4.0)

| | | | | | | | Close > 3377.0591: 1 (12.0/1.0)

| | | | | | | Open > 8249.0927: 0 (17.0/2.0)

| | | | | | Reductıon > 614.9215: 0 (356.0/9.0)

| | | | | Difference > -51.5059: 0 (3806.0/42.0)

| | Rate (%) > 30.602737

| | | Difference <= 23.1367: 0 (8.0)

| | | Difference > 23.1367

| | | | Reductıon <= 316.7425: 1 (46.0/1.0)

| | | | Reductıon > 316.7425

| | | | | Sell <= 928.85: 0 (4.0)

| | | | | Sell > 928.85

| | | | | | Tank No <= 2: 1 (7.0)

| | | | | | Tank No > 2

| | | | | | | Rate (%) <= 48.505226

| | | | | | | | Open <= 9729.0161: 1 (5.0)

| | | | | | | | Open > 9729.0161

| | | | | | | | | Open <= 20825.87: 0 (2.0)

| | | | | | | | | Open > 20825.87: 1 (2.0)

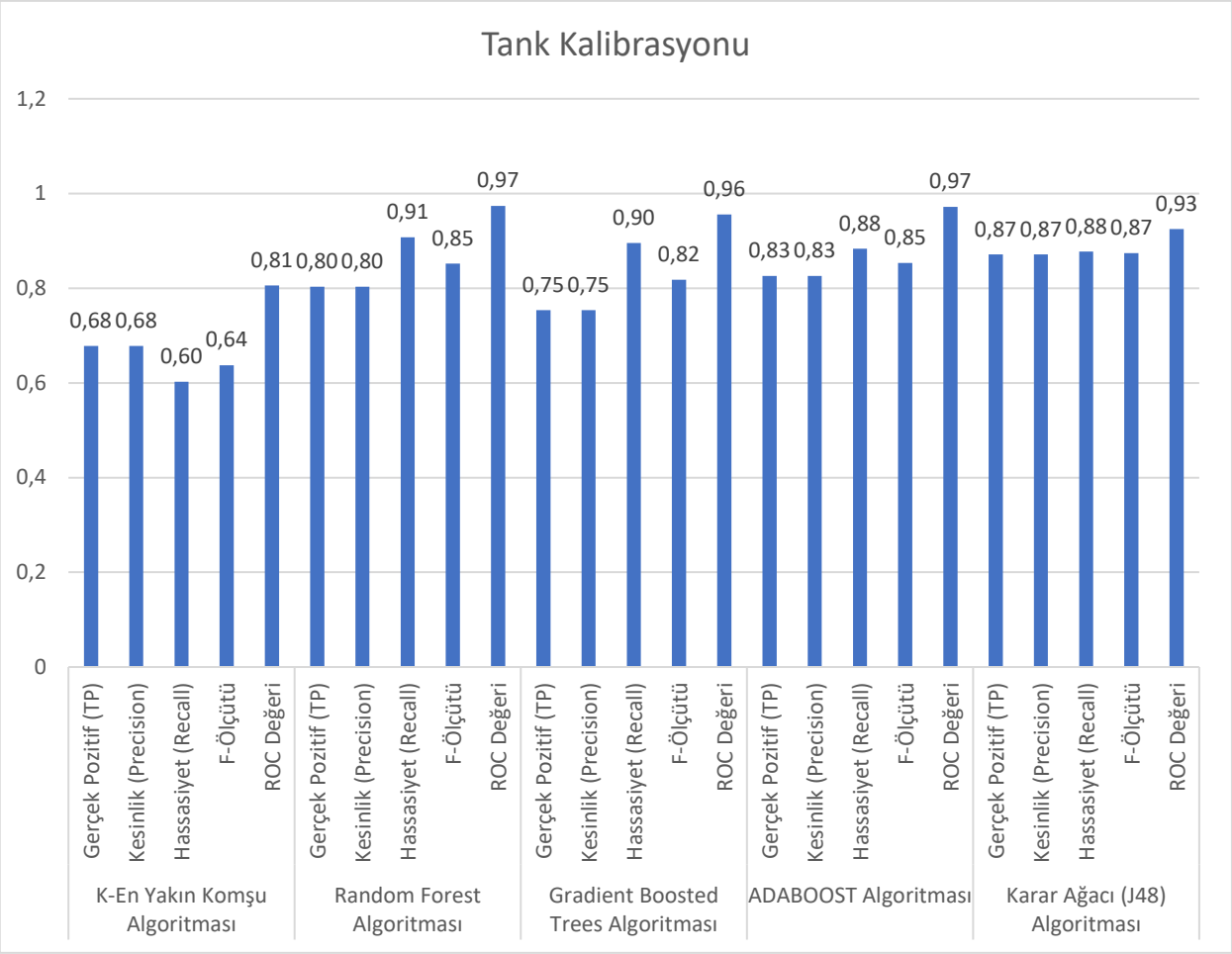
| | | | | | Rate (%) > 48.505226: 0 (2.0)

Şekil 40. Karar Ağacı (J48) Arıza Modeli Karar Çıktısı

Kaynak: "yazar tarafından oluşturulmuş/derlenmiştir"

4.5. Kullanılan Sınıflandırma Algoritmalarının Karşılaştırılması

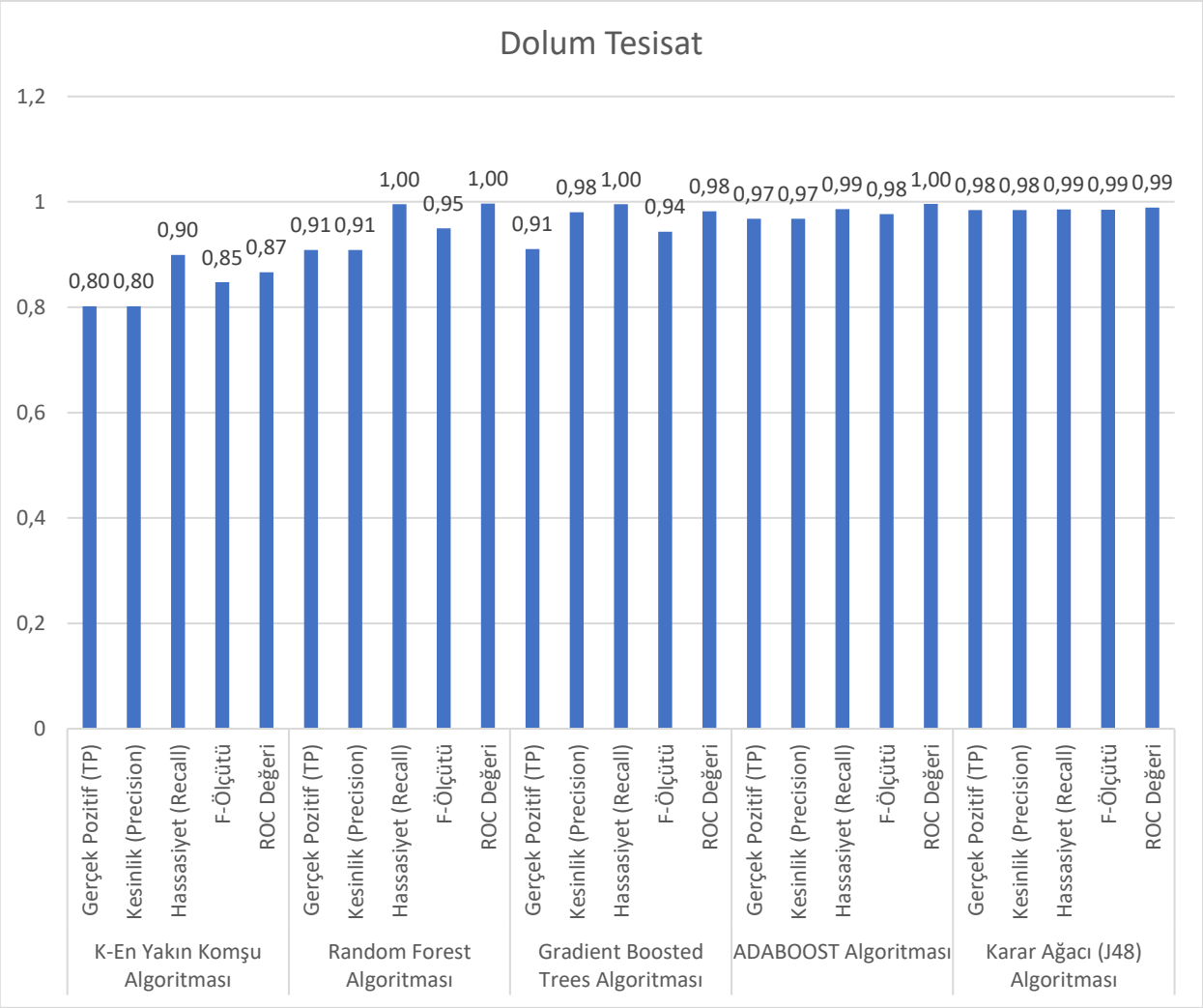
Uygulamamızda dört ayrı kategoriye ayrı ayrı olarak K-En Yakın Komşu algoritması, Gradient Boosted Trees Algoritması, ADABOOST Algoritması, Rastgele Orman Algoritması ve Karar Ağacı (J48) Algoritması yöntemleri uygulanmıştır. Her modelin başarımını ölçmek için farklı sınıflandırma yöntemleri başarım ölçütlerinden yararlanılmıştır. Şekil 41'de tank kalibrasyonu dört farklı modele göre sınıflandırma başarı performansları gösterilmiştir. Tank kalibrasyonu modelinde gerçek pozitif, kesinlik, duyarlılık ve F-Ölçütü başarım değerlerine göre yapılan karşılaştırmada karar ağaçları (J48) algoritması diğer yöntemlerden daha iyi olsa da random forest algoritması ve ADABOOST algoritmasına yakın değerler göstermektedir. ROC değerine göre karşılaştırma sağlandığında en iyi sonucu Random Forest algoritması ve ADABOOST algoritmalarının verdiği görülmüştür.



Şekil 41. Tank Kalibrasyonun Sınıflandırma Kriterlerine Göre Performanslarının Karşılaştırılması

Kaynak: "yazar tarafından oluşturulmuş/derlenmiştir"

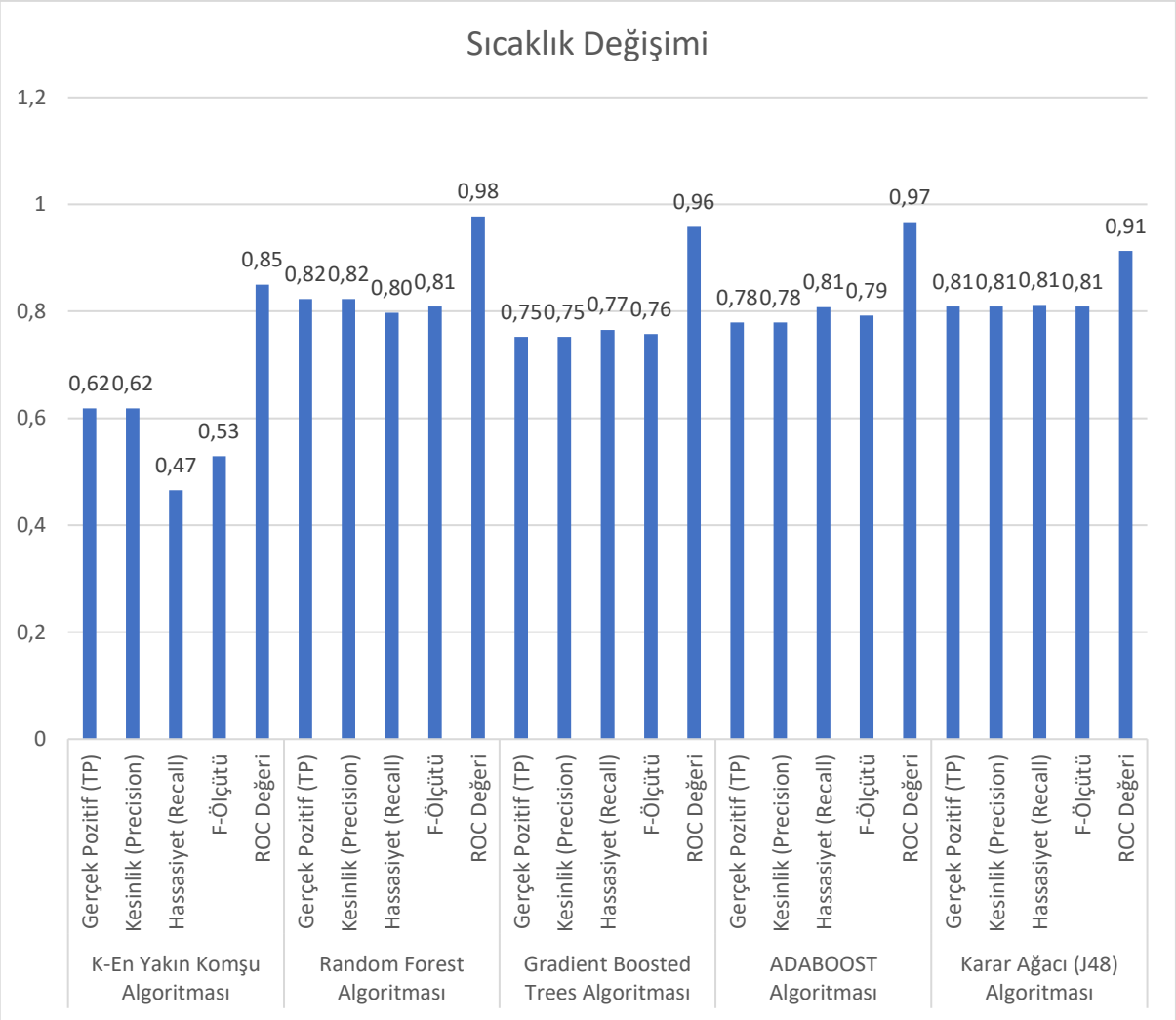
Dolum-Tesisat modelinde gerçek pozitif, kesinlik, duyarlılık ve F-Ölçütü başarım değerlerine göre yapılan karşılaştırmada karar ağaçları (J48) algoritması diğer yöntemlerden daha iyi olsa da random forest algoritması, ADABOOST algoritması ve Gradient Boosted Trees Algoritması yakın değerler göstermektedir. ROC değerine göre karşılaştırma sağlandığında en iyi sonucu Random Forest algoritması ve ADABOOST algoritmalarının verdiği görülmüştür.



Şekil 42. Dolum-Tesisat Sınıflandırma Kriterlerine Göre Performanslarının Karşılaştırılması

Kaynak: "yazar tarafından oluşturulmuş/derlenmiştir"

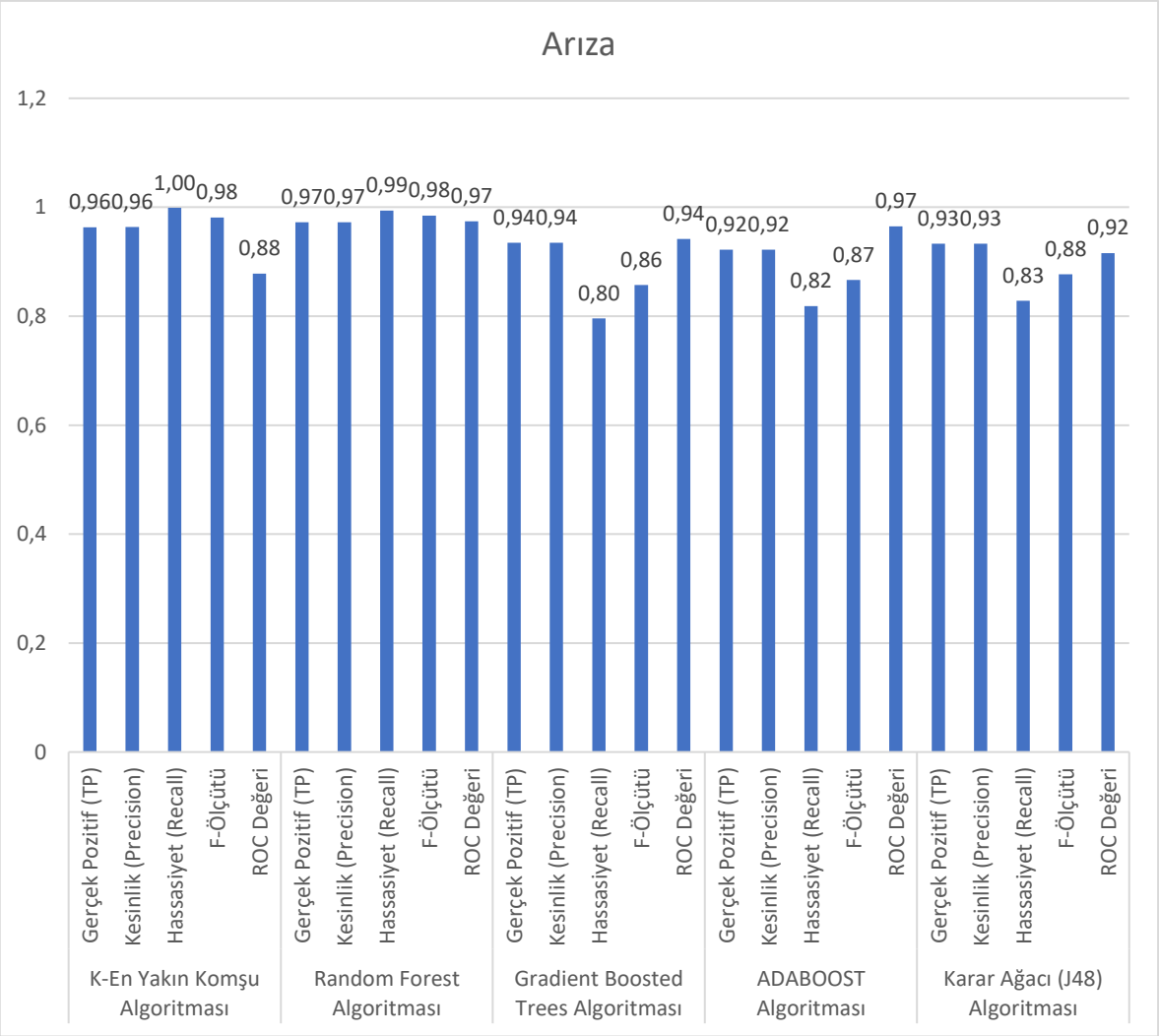
Sıcaklık Değişimi modelinde gerçek pozitif, kesinlik, duyarlılık ve F-Ölçütü başarım değerlerine göre yapılan karşılaştırmada Random Forest Algoritması diğer yöntemlerden daha iyi olsa da karar ağacı (J48) Algoritması, ADABOOST algoritması ve Gradient Boosted Trees Algoritması yakın değerler göstermektedir. Diğer modellere göre sıcaklık değişim modelinin daha az başarılı sonuçlar vermesinin sebebi ne kadarlık sıcaklık değişiminin ne kadar fark yaratacağının tam olarak bilinmemesidir. ROC değerine göre karşılaştırma sağlandığında en iyi sonucu Random Forest algoritması ve ADABOOST algoritmalarının verdiği görülmüştür.



Şekil 43. Sıcaklık Değişimi Sınıflandırma Kriterlerine Göre Performanslarının Karşılaştırılması

Kaynak: "yazar tarafından oluşturulmuş/derlenmiştir"

Arıza modelinde gerçek pozitif, kesinlik, duyarlılık ve F-Ölçütü başarım değerlerine göre yapılan karşılaştırmada K-En Yakın Komşu algoritmasının diğer yöntemlerden daha iyi olsa da random forest algoritması, ADABOOST algoritması ve Gradient Boosted Trees Algoritması yakın değerler göstermektedir. ROC değerine göre karşılaştırma sağlandığında en iyi sonucu Random Forest algoritması ve ADABOOST algoritmalarının verdiği görülmüştür.



Şekil 44. Arıza Değişimi Sınıflandırma Kriterlerine Göre Performanslarının Karşılaştırılması

Kaynak: "yazar tarafından oluşturulmuş/derlenmiştir"

BULGULAR

Bu çalışmada kullanılacak olan sınıflandırma algoritmaları belirlenirken literatürde en çok tercih edilen beş algoritma belirlenmiştir. Bu algoritmalar K-En Yakın Komşu Algoritması, Rastgele Orman Algoritması, Gradient Boosted Tress Algoritması, ADABOOST Algoritması ve Karar Ağacı (J48) Algoritması olarak belirlenmiştir. Çalışmada sınıflandırma yöntemlerinin karşılaştırılmasının yanı sıra sınıflandırma ölçütleri de karşılaştırılmıştır

Yöntemlerin modeller bazında sınıflandırma başarılarına bakıldığında Sınıflandırma yapılan veri seti dengesiz olmasına rağmen sınıflandırma başarılarının tablo.13'te olduğu gibi başarılı olduğu görülmektedir. En iyi sınıflandırma yöntemlerinin karar ağacı (J48) algoritması, ADABOOST Algoritması, Gradient Boosted Trees Algoritması ve Rastgele Orman Algoritması olduğu görülmüştür. K-en yakın komşu algoritmasının ise sınıflandırma için yeterli olmadığı görülmüştür.

Gelecek çalışmalarda, geliştirilecek yazılım programları sayesinde veri tabanlarına yansıyan verilerin bilgisayar ortamında sorunsuz ve hızlı bir şekilde analizlerin gerçekleştirilerek daha hızlı aksiyonların alınması sağlanılabilir. Hızlı alınan aksiyonlar sayesinde çevre kirliliğinin ve maliyetlerin azalmasına katkı sağlayacaktır.

Tablo 13. Algoritmaların Başarı Oranları

Sınıflandırma Algoritması	Kalibrasyon Doğruluk Başarı Oranı (%)	Dolum Doğruluk Başarı Oranı (%)	Sıcaklık Değişimi Doğruluk Başarı Oranı (%)	Arıza Doğruluk Başarı Oranı (%)
K En Yakın Komşu Algoritması	0,677	0,801	0,618	0,961
Rastgele Orman Algoritması	0,803	0,908	0,822	0,972
Gradient Boosted Tress Algoritması	0,753	0,914	0,752	0,935
ADABOOST Algoritması	0,826	0,968	0,779	0,922
Karar Ağacı (J48) Algoritması	0,871	0,984	0,808	0,933

Kaynak: "yazar tarafından oluşturulmuş/derlenmiştir

SONUÇ

Akaryakıt sektöründe veri madenciliği uygulamaları gün geçtikçe gelişmekte ve uygulamalar yaygınlaşmaktadır. Akaryakıt sektöründe kullanılan yöntemler ve çeşitli analizler ile beraber akaryakıt hırsızlığı, sızıntı, dolum anında miktar aşımları, aşırı dolum sonrası taşma gibi önemli konular takip edilerek aksiyonlar alınmaktadır. Ülkemizde de son yıllar ile birlikte veri madenciliği yöntemleri sayesinde akaryakıt veriler üzerinden birçok akaryakıt hırsızlığı ve sızıntı gibi önemli tespit yapılmış ve aksiyonlar alınarak çeşitli yaptırımlar gerçekleştirilmiştir. Akaryakıt sızıntıları çevreye zarar verebileceği gibi akaryakıt firmalarını da maliyet yönünden zarara uğratabilmektedir. Ülkemizde akaryakıt firmalarının otomasyon sistemlerinin teknik ve veri tabanı kapsamında takibini ve mutabakatını yaparak EPDK tarafından görevlendirerek denetleyen birkaç firma bulunmaktadır. Bu işletmelerde anlık olarak veriler veri tabanına yansıdığından sorunları veri analizi ile tespit ederek kategori bazında sınıflandırmak zor bir süreçtir. Bu çalışmada bir otomasyon takip firmasından bir petrol şirketine ait 2022-2023 yıllarına ait veri seti kullanılarak daha önce analiz edilmiş ve kategori bazında sınıflandırılmış olan veri setlerinin çeşitli veri madenciliği yöntemleri kullanılarak sınıflandırılmış ve yöntemlerin sınıflandırma başarısı ve yapılan analizlerin doğruluğu değerlendirilmiştir. Veri setinde daha önceden analizleri gerçekleştirilmiş akaryakıt sektörü için önemli olan 4 ana kategori olan dolum-tesisat, kalibrasyon, arıza ve sıcaklık değişimi kategorileri seçilmiştir. Sınıflandırma işlemleri öncesinde veri setleri uygun hale getirilmek için çeşitli adımlar izlenmiştir. Sınıflandırma yapılan veri seti dengesiz olmasına rağmen sınıflandırma başarılarının tablo.13'te olduğu gibi başarılı olduğu görülmektedir.

KAYNAKÇA

Agrawal, R., Imielinski, T. ve Swami, A. (1993). Database mining: A performance perspective. *IEEE Transactions on Knowledge and Data Engineering, 5*(6), 914-925.

Akpınar, H. (2017). *Data veri madenciliği veri analizi* (2. b.). Papatya Yayıncılık.

Aktan, E. (2018). Büyük veri: Uygulama alanları, analitiği ve güvenlik boyutu. *Bilgi Yönetimi Dergisi, 1*(1), 1-22.

Alan, B. (2016). *Veri madenciliği ve market veri tabanında birliktelik kurallarının belirlenmesi.* [Yüksek lisans tezi] Recep Tayyip Erdoğan Üniversitesi.

Almuttar, A. ve Alkhafaji, S. (2023). Using data mining techniques deep analysis and theoretical investigation of COVID-19 pandemic. *Measurement: Sensors*(27), 1-5.

Artsın, M. (2020). Bir metin madenciliği uygulaması: VOSVIEWER. *Eskişehir Teknik Üniversitesi Bilim ve Teknoloji Dergisi B - Teorik Bilimler, 8*(2), 344-354.

Aydın, C. (2018). Makine öğrenmesi algoritmaları kullanılarak itfaiye istasyonu ihtiyacının sınıflandırılması. *Avrupa Bilim ve Teknoloji Dergisi, 14*(1), 169-175.

Aydoğan, E. ve Gencer, C. (2007). Kaba küme yaklaşımı kullanarak veri madenciliği problemlerinde sınıflandırma amaçlı yapılmış olan çalışmalar. *Savunma Bilimleri Dergisi, 6*(2), 17-32.

Ayşe, T. G. (2007). *Forecasting Simulated Retail Demand Using Statistical and Data Mining Techniques.* [Yüksek lisans tezi] Koç Üniversitesi.

Boland, J., Baumann, D. ve Dziegielewski, B. (1981). *An Assessment of Municipal and Industrial Water Use Forecasting Approaches.* Institute for Water Resources (U.S.).

Ceran, G. (2006). *Esnek Akış Tipi Çizelgeleme Problemlerinin Veri Madenciliği ve Genetik Algoritma Kullanarak Çözülmesi* [Yüksek lisans tezi]. Selçuk Üniversitesi.

Chen, M., Huang, C. ve Wu, P. (2005). Aggregation of orders in distribution centers using data mining. *Expert Systems with Applications, 28*(3), 453-460.

Chen, R. S., Hu, Y. C. ve Tzeng, G. H. (2003). Finding fuzzy classification rules using. *Pattern Recognition Letters, 24*(1-3), 509-519.

Coşkun, C. ve Baykal, A. (2011, 2-4 Şubat). *Veri madenciliğinde sınıflandırma algoritmalarının bir örnek üzerinde karşılaştırılması* [Bildiri sunumu]. *XII. Akademik Bilişim Konferansı,* Malatya.

Çataloluk, H. (2012). *Gerçek tıbbi veriler üzerinden veri madenciliği yöntemlerini kullanarak hastalık teşhisi* [Yüksek lisans tezi]. Bilecik Üniversitesi.

Çelik, M. (2009). *Veri Madenciliğinde Kullanılan Sınıflandırma Yöntemleri ve Bir Uygulama* [Yüksek lisans tezi]. İstanbul Üniversitesi.

Çelik, U., Akçetin, E. ve Gök, M. (2017). *RAPIDMINER ile veri madenciliği* (1. b.). Pusula 20 Teknoloji ve Yayıncılık.

D, D. ve Dagbui, A. (2014). Dealing with construction cost overruns using data mining. *Construction Management and Economics, 32*(7-8), 682-694.

Dalman, A. (2017). *A wet-stock management and leak detecıtıon system for fuel tanks* [Yüksek lisans tezi]. Yeditepe Üniversitesi.

Damle, C. ve Yalçın, A. (2007). Flood prediction using time series data mining. *Journal of Hydrology, 333*(2-4), 305-316.

Demirel, Ş. ve Yakut, G. (2019). Karar ağacı algoritmaları ve çocuk işçiliği üzerine bir uygulama. *Sosyal Bilimler Araştırma Dergisi, 8*(4), 52-65.

Deniz, S. (2016). *Otomotiv Sektöründe Yakıt Tüketimi ve Emisyon Seviyelerinin Veri Madenciliği ile Analizi* [Yüksek lisans tezi]. Gazi Üniversitesi.

Doğan , K. ve Arslantekin, S. (2016). Büyük veri: Önemi,yapısı ve günümüzdeki durum. *DTFC Dergisi, 56*(1), 15-36.

Dos santos, S., Steiner, T., Fenerich, A. ve Lima, R. (2019). Data mining and machine learning techniques applied to public health problems: A bibliometric analysis from 2009 to 2018 *Computers & Industrial Engineering*, *138(2019),* 1-10.

Dönmez, Z. (2008). *Bayi performans değerlendirilmesinde bir veri madenciliği uygulaması.* [Yüksek lisans tezi]. İstanbul Teknik Üniversitesi.

Dudas, C., Ng, A., Pehrsson, L. ve Boström, H. (2013). Integration of data mining and multi-objective optimisation for decision support in production systems development. *International Journal of Computer Integrated Manufacturing, 27*(9), 824-839.

Dünder, M. (2021). *Diyabet hastalarına ait verilerin istatistiksel yöntemler ve verimadenciliği teknikleri kullanılarak incelenmesi* [Yüksek lisans tezi]. Ondokuz Mayıs Üniversitesi.

Elmas, Ç. (2016). *Yapay Zeka Uygulamaları* (3. b.). Seçkin Yayıncılık.

Fu, C. (2011). A review on time series data mining. *Engineering Applications of Artificial Intelligence, 24*(1), 164-181.

Gemici, B. (2012). *Veri madenciliği ve bir uygulaması* [Yüksek lisans tezi]. Dokuz Eylül Üniversitesi.

Gezer, M., Erol, Ç. ve Gülseçen, S. (2007, 31 Ocak-2 Şubat). *Bir web sayfasının web madenciliği ile analizi* [Bildirimi sunumu]. IX. Akademik Bilişim Konferans Bildirileri, Kütahya.

Girija, N. ve Srivatsa, S. (2006). A research study: Using data mining in knowledge base business strategies. *Information Technology Journal, 5*(3), 590-600.

Gün, N. (2008). *Assortment Planning Using Data Mining Algorithms* [Yüksek lisans tezi]. Boğaziçi Üniversitesi.

Gürbüz, F. ve Turna, F. (2018). Rule extraction for tram faults via data mining for safe transportation. *Transportation Research Part A: Policy and Practice, 116*, 568-579.

Güvenç, E. (2001). *Student Performance Assessment in Higher Education Using Data Mining* [Yüksek lisans tezi]. Boğaziçi Üniversitesi.

Hui, S. ve Jha, G. (2000). Data mining for customer service support. *Information and Management, 38*(1), 1-13.

Irmak, S., Köksal, D. ve Asilkan, Ö. (2012). Hastanelerin gelecekteki hasta yoğunluklarının veri madenciliği yöntemleri ile tahmin Edilmesi. *Uluslararası Alanya İşletme Fakültesi Dergisi, 4*(1), 101-114.

Jothi, N., Rashid, N. ve Hüseyin, V. (2015). Data mining in healthcare – A review. *Procedia Computer Science, 72*, 306-313.

Kalaycı, Ş. (2018). *SPSS uygulamalı çok değişkenli istatistik teknikleri.*Dinamik Akademi.

Kamath, P. ve Patil, P. (2021). Crop yield forecasting using data mining. *Global Transitions Proceedings, 2*(2), 402-407.

Kantardziç, M. (2003). *Data mining: Concepts, models, methods, and algorithms.* SPİ Global.

Karaboğa, T. (2020). *Büyük veri analitiği yönetsel kabiliyetlerinin firma performansına etkisi: Veri odaklı kültür ve büyük veri strateji uyumunun aracılık etkisi* [Doktora Tezi]. Yıldız Teknik Üniversitesi.

Kass, G. (1980). An exploratory technique for ınvestigating large quantities of categorical data. *Journal of the Royal Statistical Society., 29*(2), 119-127.

Kılıç, S. (2013). Klinik karar vermede ROC analizi. *Journal of Mood Disorders, 3*(3), 135-140.

Kiremitçi, B. (2005). *Veri ambarlarında veri madenciliği ve ulaştırma-lojistik sektöründe bir uygulama* [Yüksek lisans tezi]. İstanbul Üniversitesi.

Koonce, D. ve Tsai, C. (2000). Using data mining to find patterns in genetic algorithm solutions to a job shop schedule. *Computers & Industrial Engineering, 38*(3), 361-374.

Köksal, G., Batmaz, İ., Testik, C. ve Güntürkün, F. (2010). İmalat sektöründe kalite iyileştirmede veri madenciliği tekniklerinin kullanımı. *Verimlilik Dergisi*(2), 47-65.

Kuşaksızoğlu, B. (2006). *Veri madenciliği yardımıyla mobil telekomünikasyon şebekelerinde sahtekarlık tespiti* [Yüksek lisans tezi]. Bahçeşehir Üniversitesi.

Loh, Y. ve Shih, S. (1997). Splıt selectıon methods for classıfıcatıon trees. *Statistica Sinica, 7*(4), 815-840.

Lu, X., Dong, Z. ve Li, X. (2005). Electricity market price spike forecast with data mining techniques. *Electric Power Systems Research, 73*(1), 19-29.

Martin, L., Baena, L., Garach, L., Lopez, G. ve Ona, J. (2014). Using data mining techniques to road rafety ımprovement in Spanish Roads. *Procedia - Social and Behavioral Sciences, 160*, 607-614.

Mastrogiannis, N., Boutsinas, B. ve Giannikos, I. (2009). A method for improving the accuracy of data mining classification algorithms. *Computers & Operations Research, 36*(10), 2829-2839.

Mitroshin, P., Shitova, Y., Shitov, Y., Vlasov, D. ve Mitroshin, A. (2022). Big data and data mining technologies application at road transport logistics. *Transportation Research Procedia, 61*, 462-466.

Ngai, E., Xui, L. ve Chau, D. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications, 36*(2), 2592-2602.

Olafsson, S., Li, X. ve Wu, S. (2008). Operations research and data mining. *European Journal of Operational Research 1, 187*(3), 1429-1448.

Ouali, A., Şerif, A. ve Krebs, M. (2006). Data mining based bayesian networks for best classification. *Computational Statistics & Data Analysis, 51*(2), 1278-1292.

Özekes, S. (2003). Veri madenciliği modelleri ve uygulama alanları. *İstanbul Ticaret Üniversitesi Dergisi, 2*(3), 65-82.

Özseven, T. ve Düğenci, M. (2011, 2-4 Şubat*). LOG preprocessing: Web kullanım madenciliği ön İşlem aşaması uygulma yazılımı* [Bildiri sunumu]. XIII. Akademik Bilişim Konferansı Bildirileri, Malatya.

Rupnik, R., Kukar, M. ve Krisper, M. (2007). Integrating data mining and decision support through data mining based decision support system. *Journal of Computer Information Systems, 47*(3), 89-104.

Sağlam Hadık, E. (2015). *An integrated methodology with data mining techniques for retail industry* [Yüksek lisans tezi]. Yıldız Teknik Üniversitesi.

(2017). Adaboost. In: Sammut, C., Webb, GI (ed.). *Makine öğrenmesi ve veri madenciliği ansiklopedisi* (s. 19-20). Springer.

Savaş, S., Topaloğlu, N. ve Yılmaz, M. (2012). Veri madenciliği ve Türkiye'deki uygulama örnekleri. *İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi*(21), 1-23.

Schuh, G., Prote, P. ve Hunnekes, P. (2020). Data mining methods for macro level process planning. *Procedia CIRP, 88*, 48-53.

Sforna, M. (2000). Data mining in a power company customer database. *Electric Power Systems Research, 55*(3), 201-209.

Han, J., Kamber, M. ve Pei, J. (2012). *Data mınıng concepts and Techniques*. Morgan Kaufmann.

Şahin, D. (2010). *Rotalama problemlerinin veri madenciliği ile çözümü* [Yüksek lisans tezi]. Gebze Yüksek Teknoloji Enstitüsü.

Şeker, E. (2015). Metin Madenciliği (Text Mining). *YBS Ansiklopedi, 3*(2), 30-32.

Şekeroğlu, S. (2010). *Hizmet sektöründe veri madenciliği uygulaması* [Yüksek lisans tezi]. İstanbul Teknik Üniversitesi.

Şengür, D. (2013). *Öğrencilerin akademik başarılarının veri madenciliği metotları ile tahmini* [Yüksek lisans tezi]. Fırat Üniversitesi.

Şimşek, F. (2019). *Veri madenciliği sınıflandırma tekniklerini kullanarak üretim sistemlerinde hatalı ürünlerin tespit edilmesi* [Yüksek lisans tezi]. Kırıkkale Üniversitesi.

Tuzcu, S. (2018). *Ders yönetim sistemi tabanlı veri madenciliği ve öğrenme analitiği* [Yüksek lisans Tezi]. Eskişehir Osmangazi Üniversitesi.

Ulaş, A. (2001). *Market basket analysıs for data mınıng* [Yüksek lisans tezi]. Boğaziçi Üniversitesi.

Uyumaz, Ö. (2017). *Bankacılık sektöründe pazarlama stratejilerinin belirlenmesinde sınıflandırma ve veri madenciliği* [Yüksek lisans tezi]. Beykent Üniversitesi.

Ünsal, Ç. (2023). *Veri madenciliği teknikleri ile Türkiye'de yaşam memnuniyeti düzeyleri üzerine Bir Uygulama* [Doktora tezi]. Ankara Hacı Bayram Veli Üniversitesi.

Üre , N. (2021). *Veri madenciliği sınıflandırma algoritmalarının performans karşılaştırılması: Tiroid hastalığının tahmini üzerinde bir uygulama* [Yüksek lisans Tezi]. Kafkas Üniversitesi.

Vezhnevets, A. ve Vezhnevets, V. (2005). 'Modest adaboost' – teaching adaboost to generalize better. *Graphicon, 12*(5), 987-997.

Zengin, K., Esgi, N., Erginer, E. ve Aksoy, E. (2011). A sample study on applying data mining research techniques in educational science: Developing a more meaning of data. *Procedia - Social and Behavioral Sciences, 15*, 4028-4032.