



**VERİ MADENCİLİĞİ İLE TALEP TAHMİNİ VE BİR İŞLETMEDE
UYGULAMA**

Oktay Mesut ARSLAN

**YÜKSEL LİSANS TEZİ
YÖNETİM BİLİŞİM SİSTEMLERİ ANABİLİM DALI**

**GAZİ ÜNİVERSİTESİ
BİLİŞİM ENSTİTÜSÜ**

HAZİRAN 2020

Oktay Mesut ARSLAN tarafından hazırlanan “VERİ MADENCİLİĞİ İLE TALEP TAHMİNİ VE BİR İŞLETMEDE UYGULAMA” adlı tez çalışması aşağıdaki jüri tarafından OY BİRLİĞİ ile Gazi Üniversitesi Yönetim Bilişim Sistemleri Ana Bilim Dalında YÜKSEK LİSANS TEZİ olarak kabul edilmiştir.

Danışman: Dr. Öğr. Üyesi Bahar ÖZYÖRÜK

Endüstri Mühendisliği Bölümü, Gazi Üniversitesi

Bu tezin, kapsam ve kalite olarak Yüksek Lisans Tezi olduğunu onaylıyorum/onaylamıyorum.

Başkan: Prof. Dr. Cevriye GENCER

Endüstri Mühendisliği Bölümü, Gazi Üniversitesi

Bu tezin, kapsam ve kalite olarak Yüksek Lisans Tezi olduğunu onaylıyorum/onaylamıyorum.

Üye: Prof. Dr. Suat ÖZDEMİR

Bilgisayar Mühendisliği, Hacettepe Üniversitesi

Bu tezin, kapsam ve kalite olarak Yüksek Lisans Tezi olduğunu onaylıyorum/onaylamıyorum.

Tez Savunma Tarihi: 30 / 06 / 2020

Jüri tarafından kabul edilen bu tezin Yüksek Lisans Tezi olması için gerekli şartları yerine getirdiğini onaylıyorum.

.....

Doç. Dr. Aslıhan TÜFEKÇİ

Bilişim Enstitüsü Müdürü

ETİK BEYAN

Gazi Üniversitesi Bilişim Enstitüsü Tez Yazım Kurallarına uygun olarak hazırladığım bu tez çalışmada;

- Tez içinde sunduğum verileri, bilgileri ve dokümanları akademik ve etik kurallar çerçevesinde elde ettiğimi,
- Tüm bilgi, belge, değerlendirme ve sonuçları bilimsel etik ve ahlak kurallarına uygun olarak sunduğumu,
- Tez çalışmada yararlandığım eserlerin tümüne uygun atıfta bulunarak kaynak gösterdiğimi,
- Kullanılan verilerde herhangi bir değişiklik yapmadığımı,
- Bu tezde sunduğum çalışmanın özgün olduğunu,

bildirir, aksi bir durumda aleyhime doğabilecek tüm hak kayıplarını kabullendiğimi beyan ederim.

Mesut ARSLAN

...../...../.....

VERİ MADENCİLİĞİ İLE TALEP TAHMİNİ VE BİR İŞLETMEDE UYGULAMA
(Yüksek Lisans Tezi)

Oktay Mesut ARSLAN

GAZİ ÜNİVERSİTESİ
BİLİŞİM ENSTİTÜSÜ

Haziran 2020

ÖZET

Günümüz dünyasında bilişim teknolojisinin hızla gelişmesi birçok yeniliği beraberinde getirmektedir. Bu yeniliklerin en önemlilerinden biri ise büyük veri kavramı, bu verinin yönetilmesi ve kullanılmasıdır. Bu sebep ile büyük verilerden anlamlı ve kullanışlı bilgi edinimi yöntemi olarak ‘Veri Madenciliği’ ortaya çıkmıştır. Uzun yıllardır birçok alanda geleneksel yöntemler ile çözülmeye çalışılan karmaşık problemlerde bir çözüm arayışı olarak veri madenciliği yöntemleri kullanılmaya başlanmıştır. Veri madenciliği birçok alanda olduğu gibi talep tahmininde de başarılı bir şekilde kullanıldığı ve etkin sonuçlara ulaştığı gözlenmiştir. Bu tez çalışmasında da literatürde yer alan talep tahmini konusunda veri madenciliğini kullanan akademik çalışmalar detaylı bir şekilde incelenmiştir. Tez çalışması kapsamında Ankara da faaliyetlerine devam eden ve orta gerilim ekipmanları üreten bir firmanın ürettiği gaz yalıtımlı orta gerilim hücresi ürünü için talep tahmini yapılmaya çalışılmıştır. Firmanın ürününe olan talep birçok faktörden etkilendiği için istatistiksel olarak geliştirilen talep tahmini yöntemleri yerine veri madenciliği yöntemlerinden yararlanılmıştır. Bu amaçla firmanın 2016 – 2018 yılı satış verileri kullanılarak 2019 yılı için aylık talep tahminleri veri madenciliği yöntemleri kullanılarak yapılmıştır. Bu tahmin çalışması kapsamında merkez bankasının resmî sitesinden alınan aylık ‘Dolar Kurları’, ‘Enflasyon Oranları’ ve ‘Faiz Oranları’ nitelik olarak modele eklenmiştir. Çözüm WEKA (Waikato Environment for Knowledge Analysis) paket programının 3.9.4 sürümü (son sürüm) ile gerçekleştirilmiştir. Destek vektör makinaları (DVM) ve yapay sinir ağları (YSA) metotlarının kullanıldığı çalışmada elde edilen sonuçlar tartışılmıştır.

Bilim Kodu : 114611
Anahtar Kelimeler : Veri Madenciliği, Talep Tahmini, Üretim Sektörü
Sayfa Adedi : 49
Danışman : Dr. Öğr. Üyesi Bahar ÖZYÖRÜK

DEMAND ESTIMATION WITH DATA MINING AND IMPLEMENTATION IN AN ENTERPRISE

(M. Sc. Thesis)

Oktay Mesut ARSLAN

GAZİ UNIVERSITY
INFORMATICS INSTITUTE

June 2020

ABSTRACT

The rapid development of information technology in today's world brings with it many innovations. One of the most important of these innovations is the concept of big data and its management. For this reason, data mining has emerged as a method of obtaining meaningful and useful information from big data. Data mining methods have started to be used as a solution for complex problems which have been tried to be solved with traditional methods since past. It is observed that data mining is used successfully in demand forecasting as in many fields and reaches effective results. In this thesis study, academic studies using data mining on demand forecasting in the literature are examined in detail. Within the scope of the thesis, demand estimation has been made for the gas insulated medium voltage cell product produced by a company that continues its activities in Ankara and produces medium voltage equipment. Since the demand for the company's product is affected by many factors, data mining methods are used instead of statistically developed demand forecasting methods. For this purpose, monthly demand forecasts for 2019 were made using data mining methods, using the sales data of the company for 2016-2018. Within the scope of this estimation study, monthly 'Dollar Rates', 'Inflation Rates' and 'Interest Rates' taken from the official website of the central bank have been added to the model as qualifications. The solution was carried out in version 3.9.4 (latest version) of the WEKA (Waikato Environment for Knowledge Analysis) package program. The results obtained in the study using support vector machines (DVM) and artificial neural networks (ANN) methods are discussed.

Science Code : 114611

Key Words : Data Mining, Demand Estimation, Production Sector

Page Number : 49

Supervisor : Assist. Prof. Dr. Bahar ÖZYÖRÜK

TEŞEKKÜR

Bu çalışmamda bilgi, tecrübe, düşüncelerini benimle paylaşan ve aynı zamanda daha iyisini yapabilmem için beni motive eden, tezli yüksek lisans sürecimde beni her anlamda yönlendiren değerli hocam Sn. Dr. Öğr. Üyesi Bahar ÖZYÖRÜK’e, ‘Veri Ambarlama ve Veri Madenciliği’ dersini aldığım, tez konumu seçmemde ilham kaynağı olan, çalışmamın kritik zamanlarında yardımını esirgemeyen Sn. Prof. Dr. Cevriye GENCER hocama ve hayatımın her anında olduğu gibi çalışma sürecinde de desteğini esirgemedi ve yanımda olan sevgili eşime teşekkürü borç bilirim.



İÇİNDEKİLER

	Sayfa
ÖZET	iv
ABSTRACT.....	v
TEŞEKKÜR.....	vi
İÇİNDEKİLER	vii
ÇİZELGELERİN LİSTESİ.....	ix
ŞEKİLLERİN LİSTESİ.....	x
RESİMLERİN LİSTESİ	xi
SİMGELER VE KISALTMALAR.....	xii
1. GİRİŞ.....	1
2. VERİ MADENCİLİĞİ.....	3
2.1. Tanımlar	3
2.1.1. Veri.....	3
2.1.2. Bilgi.....	3
2.1.3. OLTP (OnLine Transaction Processing).....	3
2.1.4. OLAP (On-line Analytical Processing)	4
2.1.5. Veri tabanı.....	4
2.1.6. Veri ambarı	4
2.2. Veri Madenciliği Nedir?	5
2.3. Neden Veri Madenciliği?	5
2.4. Veri Madenciliğinin Tarihçesi	6
2.5. Veri Madenciliğinin Diğer Disiplinler ile İlişkisi	7
2.6. Veri Madenciliği Süreci	8
2.6.1. Problemin tanımlanması	8
2.6.2. Verilerin hazırlanması.....	9

	Sayfa
2.6.3. Modelin kurulması ve değerlendirilmesi	12
2.6.4. Modelin kullanılması	12
2.6.5. Modelin izlenmesi.....	13
2.7. Veri Madenciliği Yöntemleri	13
2.7.1. Tahmin edici modeller	13
2.7.2. Tanımlayıcı modeller	19
3. LİTERATÜR TARAMASI	25
4. UYGULAMA.....	33
4.1. Yapay Sinir Ağları	33
4.2. Destek Vektör Makinaları	37
4.3. Destek Vektör Makinaları ve Yapay Sinir Ağlarının Karşılaştırılması	41
5. SONUÇ	43
KAYNAKLAR	45
EKLER.....	47
EK-1. 2016 – 2019 yılları nitelikler veri tablosu	48
ÖZGEÇMİŞ	49

ÇİZELGELERİN LİSTESİ

Çizelge	Sayfa
Çizelge 4.1. Yapay sinir ağları ile MAPE değerleri	34
Çizelge 4.2. Destek vektör makinaları ile MAPE değerleri.....	37
Çizelge 4.3. DVM & YSA MAPE değerleri.....	41



ŞEKİLLERİN LİSTESİ

Şekil	Sayfa
Şekil 2.1. Veri indirgeme örneği.....	12
Şekil 4.1. GKNS 3-3-3 EDS 500 göre tahmin grafiği	34
Şekil 4.2. GKNS 3-3-3 EDS 400 göre tahmin grafiği	35
Şekil 4.3. GKNS 2-4-6 EDS 500 göre tahmin grafiği	35
Şekil 4.4. GKNS 2-4-6 EDS 400 göre tahmin grafiği	36
Şekil 4.5. GKNS 1 EDS 500 göre tahmin grafiği	36
Şekil 4.6. GKNS 1 EDS 400 göre tahmin grafiği	37
Şekil 4.7. C değeri 1 ve epsilon 0,001'e göre tahmin grafiği	38
Şekil 4.8. C değeri 1 ve epsilon 0,0001'e göre tahmin grafiği	39
Şekil 4.9. C değeri 2 ve epsilon 0,001'e göre tahmin grafiği	39
Şekil 4.10. C değeri 2 ve epsilon 0,0001'e göre tahmin grafiği	40
Şekil 4.11. C değeri 3 ve epsilon 0,001'e göre tahmin grafiği	40
Şekil 4.12. C değeri 3 ve epsilon 0,0001'e göre tahmin grafiği	41

RESİMLERİN LİSTESİ

Resim	Sayfa
Resim 2.1. Veri ambarı çalışma biçimi.....	5
Resim 2.2. Veri madenciliği ile ilişkili alanlar	7
Resim 2.3. Veri madenciliği süreci.....	8
Resim 2.4. YSA katman yapısı	19



SİMGELER VE KISALTMALAR

Bu çalışmada kullanılmış simgeler ve kısaltmalar, açıklamaları ile birlikte aşağıda sunulmuştur.

Simgeler

Açıklamalar

w	Ağırlık Vektörü
x	Düzlem Üzerindeki Herhangi Bir Nokta
$w^T x$	İki Vektörün İç Çarpımı
$h(x)$	Hiper Düzlem Denklemi
ed	Kesim Noktası
$k(x_i, x_j)$	Kernel Çekirdek Fonksiyonu
ε	Hata Toleransı
C	Ceza Katsayısı
γ	Çekirdek Fonksiyonu Parametresi

Kısaltmalar

Açıklamalar

BP-NN	Geri Yayılm Sinir Ağları (Backpropagation Neural Network)
DVM	Destek Vektör Makinaları
DVR, SVR	Destek Vektör Regresyonu (Support Vector Regressor)
EDS	Eğitilen Dönem Sayısı
GKA	Guguk Kuşu Araması
GKM	Gauss Karışım Modeli
GKNS	Gizli Katman Nöron Sayısı
GLM	Genelleştirilmiş Lineer Model
KNN	K – En Yakın Komşu (K Nearest Neighbor)
LR	Lineer Regresyon
MAPE	Ortalama Mutlak Yüzde Hata (Mean Absolute Percentage Error)

Kısaltmalar**Açıklamalar****RBF**

Radyal Temelli Fonksiyon
(Radial Basis Function)

TÜİK

Türkiye İstatistik Kurumu

WEKA

Bilgi Analizi için Waikato Programı
(Waikato Environment for Knowledge Analysis)

YSA

Yapay Sinir Ağları



1. GİRİŞ

Bilgisayar teknolojilerinin günümüzde hızlı gelişimi yeni problemleri, yeni fırsatları ve bunlar ile beraber yeni bilim dallarını, yeni bilimsel yöntemleri hayatımıza sokmaktadır. Günümüzde, toplumsal yaşamın her alanında kullanılmaya başlayan bilişim teknolojileri kişiler, işletmeler, kurum ve kuruluşlar için büyük veri depoları anlamına gelmektedir. Kişilerin telefon ile yaptıkları görüşmeleri, mesajlaşmaları veya internet üzerinden yaptıkları aramaları, seri üretim sektöründeki bir firma için malzeme hareketleri, satın alma ve satış işlemleri veya finansal hareketleri düşünürsek etrafımızda ne denli büyük bir veriye sahip olduğumuz anlaşılmaktadır. Bu boyutlarda büyük veriye sahip olmak, veriyi doğru şekilde saklama, doğru şekilde kullanma problemlerini getirmektedir. Dijitalleşmenin ve Endüstri 4.0'ın hızla ilerlediği günümüzde veri yönetimi başlı başına bir ana disiplin olarak karşımıza çıkmaktadır. Veri, bilgi olgusunun ham halidir. Bu açıdan bakıldığında, büyük boyutlardaki bu verilerin kişiler, işletmeler, kurum ve kuruluşlar için çok büyük bir bilgi madenine dönüştüğü aşikardır. Tamda bu perspektif ile verilerden bilgi edinmeyi amaçlayan 'Veri Madenciliği' kavramı son yıllarda trend konulardan biri haline gelmiştir. İşletmeler, gittikçe rekabetçi hale gelen ortamda veri tabanlarında yer alan verileri bir bilgi kaynağı olarak görmekte ve sürekli gelişmeyi hedefleyerek veri madenciliği yöntemleri ile daha doğru bir şekilde karar vermeye çalışmaktadır.

Bu çalışmada, veri madenciliğinin tanımı, tarihçesi ve algoritmaları detaylı bir şekilde açıklanmıştır. Ayrıca çalışma kapsamında veri madenciliği yöntemleri ile tahmin çalışması yapılan akademik yayınlar incelenmiştir. Tüm bu yayınlarından anlaşılan konu veri madenciliği yöntemlerinin tahmin çalışmalarında çok etkin sonuçlar vermesidir. Bunun sebebini özetlemek gerekirse, veri madenciliği yöntemleri çok büyük verilerin yönetilmesi için uygun olmasının yanı sıra geleneksel modellere göre daha esnek olmasıdır. Çünkü veri madenciliği yöntemleri çok fazla kriterin modele eklenmesine olanak sağlamaktadır. Üretim sektöründe faaliyet gösteren bir firmada talep tahmini hem stratejik kararlar hem de operasyonel kararlar üzerinde çok etkilidir. Talebin doğru tahmin edilmesi, geleceğe yönelik insan, malzeme, makine – ekipman ve tesis alanı gibi üretim sektörünün temel kaynaklarının doğru planlanmasını sağlamaktadır. Aksi durumda ise kaynakların verimsiz kullanıldığı ve ciddi maliyetlerin ortaya çıktığı durumlar ile karşılaşılabilir.

Talep tahmini konusu her sektörde karışık ve zor problemlerin başında gelmektedir. Bunun sebebi gerçek hayat problemlerinin, birçok girdisinin olması ve bu girdilerinde tahmin edilebilirliğinin düşük olmasıdır. Bu sebep ile ciddi avantajları olmasına ve çokça problemde etkin sonuç alınmasına rağmen, geleneksel yöntemler ile her zaman doğru tahminler yapılamamaktadır. Birçok girdiyi nitelik olarak ekleyebildiğimiz ve aynı zamanda bu girdilerin etkinliğini de ölçebildiğimiz veri madenciliği yöntemleri önemli bir alternatif olarak karşımıza çıkmaktadır.



2. VERİ MADENCİLİĞİ

Günümüzde geline nokta bilgisayar ve bilgisayara bağı tüm sistemler (internet, yazılımlar, donanımlar vs.) hayatımızın tüm alanında ciddi bir şekilde yer almaktadırlar. Bilişim sistemleri günlük yaşantının her alanında etkin bir şekilde kullanılmaktadır. Bu durum aynı zamanda çok sayıda veriyi hayatımıza dahil etmiştir. Sağlık alanında, işletmelerde, alışverişte toplanan büyük miktarda verilerin depolanması ve saklanması zor bir hal almıştır. Bu ihtiyaçtan dolayı çeşitli yazılım ve donanımlar zaman içinde gelişmiştir ve veri tabanı kavramı ortaya çıkmıştır. Oluşan veri tabanlarının içinde verilerin analiz edilip karar vermede yardımcı olabilmesi için bilgi keşfi kavramı önemli bir hal almıştır. Bilginin doğru ve güncel olarak keşfedilmesi için belli yöntemler geliştirilmiştir. Bütün bu gelişmeler ve yöntemler ‘Veri Madenciliği’ kavramını çok önemli hale getirmektedirler. (Han, 2012: 1,2; Gencer, 2019).

2.1. Tanımlar

Veri madenciliğini doğru anlamamız için bazı tanımların anlaşılması gerekmektedir. Bu kapsamda, veri, bilgi, OLTP, OLAP veri tabanı ve veri ambarı açıklanmıştır.

2.1.1. Veri

Herhangi bir işlemde geçmeden belli yollar ile ortamdan elde edilen tek başına bir anlam ifade etmeyen ve herhangi bir şekilde kullanılamayan değerdir. Daha kısa bir tanıma göre ise, veri gerçeklerin işlenmemiş halidir. (Ahi, 2015).

2.1.2. Bilgi

İşlenmiş veriye bilgi denir. Bir olay, konu veya durum için belirsizliği azaltan veya ortadan kaldıran olguya bilgi denir. (Ahi, 2015; Han, 2012: 1,2).

2.1.3. OLTP (OnLine Transaction Processing)

Bilişim tabanlı sistemlerde, kayıt ekleme, silme veya güncelleme gibi günlük işlemlerin yapıldığı sistemdir. Organizasyonların gün-içi hareketlerin oluşan bu sisteme ‘Canlı Sistemler’ de denilebilmektedir. (Han, 2012: 3,4).

2.1.4. OLAP (On-line Analytical Processing)

Çok boyutlu verilere çok boyutlu analiz yaparak karar vermeye yardımcı olan bir yapıdır. Çok boyutlu verilerde düzenleme, karşılaştırma ve sorgu yapılabilmesine olanak sağlamaktadır. (Han, 2012: 3,4).

2.1.5. Veri tabanı

Bir sistem içerisinde veriler düzensiz ve birbirleri ile ilişkisel bir bağ olmadan oluşurlar. Veriyi kontrol edebilmek için bu verilerin birbirleri ile ilişkisi çerçevesinde sistematik bir şekilde düzenlenmesi gerekmektedir. Bu yapıya veri tabanı denilmektedir. (Ahi, 2015; Klösgen, 2002: 73-89).

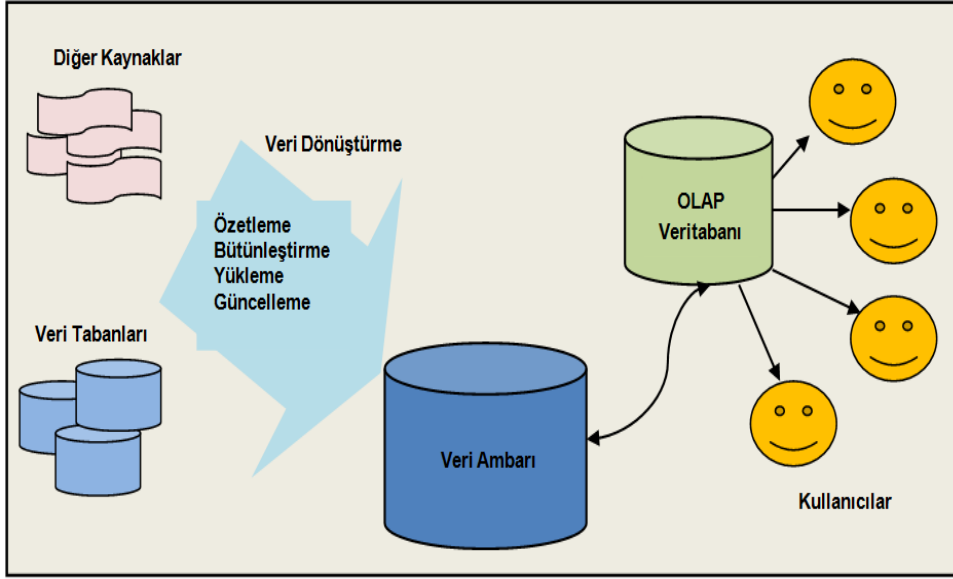
2.1.6. Veri ambarı

OLAP ile veri ambarı birbirlerinin tamamlayıcısı iki yapıdır. Veri ambarı, veriyi kullanıcılara doğru, tutarlı ve bütün bir şekilde ulaştırmayı sağlar. Bir veri ambarı aşağıdaki özelliklere sahip olmalıdır.

- Konuya yöneliktir,
- Bütünleşiktir,
- Durağandır ve
- Zaman boyutu mevcuttur.

Bu özellikler ile karar vermede etkili olan veri ambarı üzerine kullanıcı yazma işlemi yapamaz. Veri tabanına yazma işlemleri genel anlamda OLTP üzerinden gerçekleştirilir.

Veri ambarı çok boyut içerebildiğinden ‘çok boyutlu model’ de denilmektedir. Genel anlamda bakıldığında, Veri ambarları veri tabanlarından ve diğer dış kaynaklardan alınan verilerin dönüştürülmesi (özetleme, bütünleştirme vs.) ile oluşturulur ve son kullanıcıya aktarılmasını sağlar. (Han, 2012: 10; Klösgen, 2002: 198).



Resim 2.1. Veri ambarı çalışma biçimi

Şekil 2.1. de genel olarak veri ambarı mimarisini göstermektedir. Bazı uygulamalarda, son kullanıcıya veriler veri ambarı yoluyla bazılarında ise veri ambarına bağlı bir OLAP sistemi üzerinden ulaştırılabilmektedir.

2.2. Veri Madenciliği Nedir?

Veri ambarının ortaya çıkmasından sonra veri ambarında yer alan verilerin analiz edilerek faydalı ve daha önceden keşfedilmemiş bilgilerin ortaya çıkarılma süreci başlamıştır. Böylelikle ‘Veri Madenciliği’ kavramı ortaya çıkmıştır. Veri madenciliği veri ambarlarında yer alan çok sayıda veriden daha önce bulunmamış bilgi veya bilgilere ulaşmak ve bu bilgileri stratejik ve operasyonel karar almada kullanma sürecidir. Veri madenciliği için farklı kurum ve kuruluşların veya ilgili alanda çalışma yapanların farklı tanımları mevcuttur fakat hepsini ortak yönü aynıdır. Buda veri madenciliğinin bilgi keşfi süreci olmasıdır. (Han, 2012: 5; Gencer, 2019).

2.3. Neden Veri Madenciliği?

Bilişim sistemlerinin gelişmesi ile gittikçe rekabetçi bir hal alan üretim ve hizmet sektörleri daha çok veriye ulaşmaya başlamıştır. Bu ulaşılan büyük verilerden en faydalı bilgiye ulaşmanın öne çıkacağı bu rekabetçi ortamda veri madenciliği (bilgi keşfi) işletmeler için kaçınılmaz bir hal almaktadır. Teknolojik aletlerin ortaya çıkmasından bu yana veriyi

toplama bir sorun iken gelinen noktada toplanan veriden anlamlı ve karar vericiye yardımcı olacak bilginin edinilmesi önemli bir hal almıştır. Bu amaç ile birçok kurum ve kuruluş süregelen problemlerini çözmek için veri madenciliği algoritmalarını kullanmaya – geliştirmeye başlamışlardır. (Ahi, 2015; Han, 2012: 1; Gencer, 2019).

2.4. Veri Madenciliğinin Tarihçesi

Veri madenciliğinin ortaya çıkışını anlamak için önce ilk bilgisayardan bahsetmek gerekiyor. ENIAC (Electrical Numerical Integrator And Calculator) 2. Dünya savaşında geliştirilmiş ve ilk bilgisayar olarak kabul edilmektedir. Genel anlamda bilgisayarlar hesap yapmak üzerine geliştirilse de zaman için de ortaya çıkan verilerin saklanması içinde bir platform olmaya başlamışlardır. Bu süreçten sonra veri tabanları ortaya çıkmışlardır. Veri tabanları da zaman içerisinde büyümeye devam etmiş ve verinin yönetilmesi için veri ambarı kavramı ortaya çıkmıştır. Bu sayede veriler daha sistematik ve amaca yönelik kullanılabilirdi. Veri ambarı kavramı ile paralel olarak verinin modellenmesi kavramı ortaya çıkmıştır. Bu şekilde modellenerek büyük verilerin yönetilmesi sağlanmaya başlanmıştır. Aslında kavramsal olarak 1960’lı yıllarda ‘Veri Madenciliği’ ortaya çıkmıştır fakat o dönemde adına Veri taraması (data dredging), veri yakalaması (data fishing) gibi isimler verilmiştir. 1990’lı yıllara gelindiğinde ‘Veri Madenciliği’ ismi bilgisayar mühendisleri tarafından ortaya atılmıştır. Bu mühendisler göre geleneksel istatistiksel yöntemlere alternatif olarak bilgisayar kullanılarak veri analizini algoritmik yöntemler ile yapılabileceğini savunmuşlardır. Böylelikle veri madenciliği çalışmaları hız kazanmış ve hayatın her alanında çözüm olarak düşünölmeye başlanmıştır. (Ahi, 2015; Han, 2012: 4).

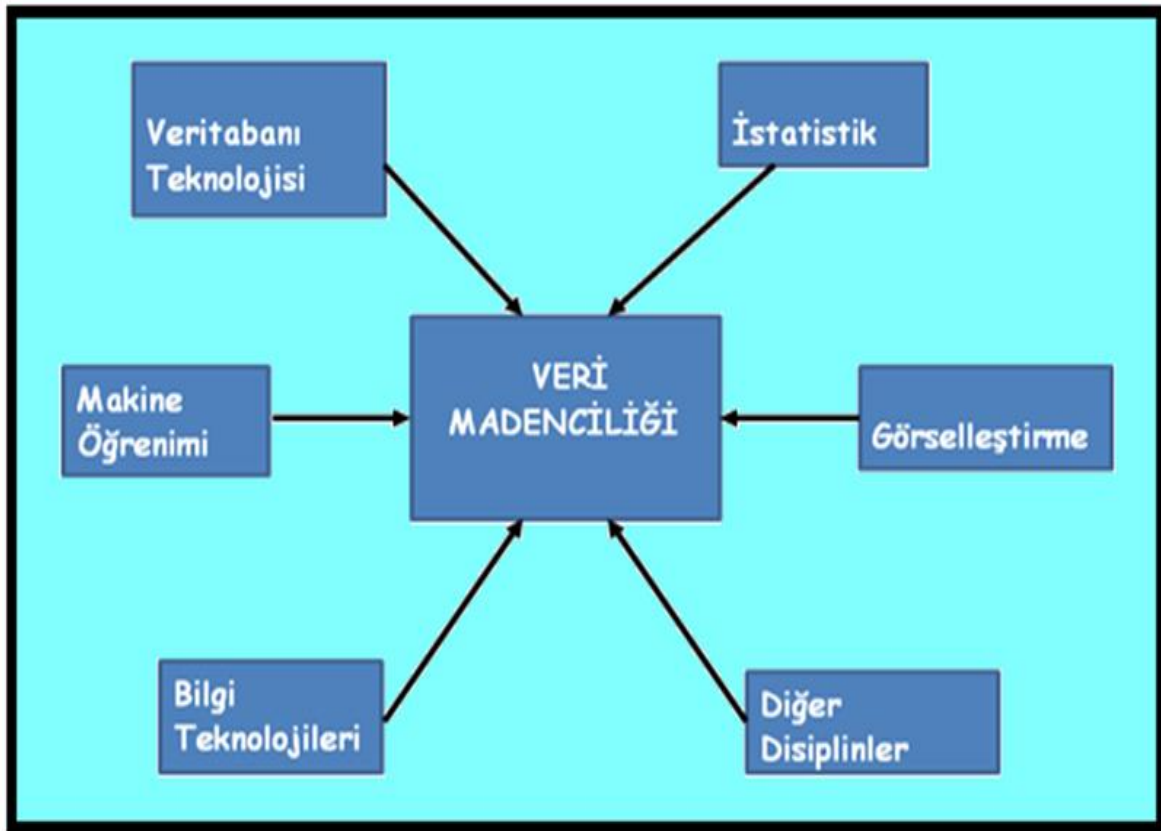
Günümüzde veri madenciliği aşağıdaki alanlarda kullanılabilmektedir.

- Pazarlama
- Bankacılık
- Sigortacılık
- Perakendecilik
- Borsa
- Telekomünikasyon
- Sağlık ve İlaç
- Endüstri

- Lojistik
- Bilim ve Mühendislik

2.5. Veri Madenciliğinin Diğer Disiplinler ile İlişkisi

Veri madenciliği kavramı birçok disiplin ile iç içe çalışan ve bu disiplinlerin gelişmesiyle ortaya çıkmış bir kavramdır. Özellikle makine öğrenimi, veri tabanları ve istatistik veri madenciliğini en çok besleyen bilim alanlarındandır. Hatta aynı hedefe odaklanıldığından istatistik ile veri madenciliği aynı disiplin olarak kabul edilebilmektedir. Fakat çok önemli ayrımlar mevcuttur. İstatistik daha çok teoriye dayalı ve bir hipotez üzerinden yol alırken veri madenciliği sezgisel algoritmalar ile yürütülür. Ayrıca veri madenciliğinde bir hipotez kurulmaz ve nereye ulaşılacağı çalışma bittiğinde yani sonuç verdiğinde bulunmuş olur. (Han, 2012: 5; Gencer, 2019). Veri madenciliğinin ilişkisi olan disiplinler şekil 2.2.de gösterilmektedir.

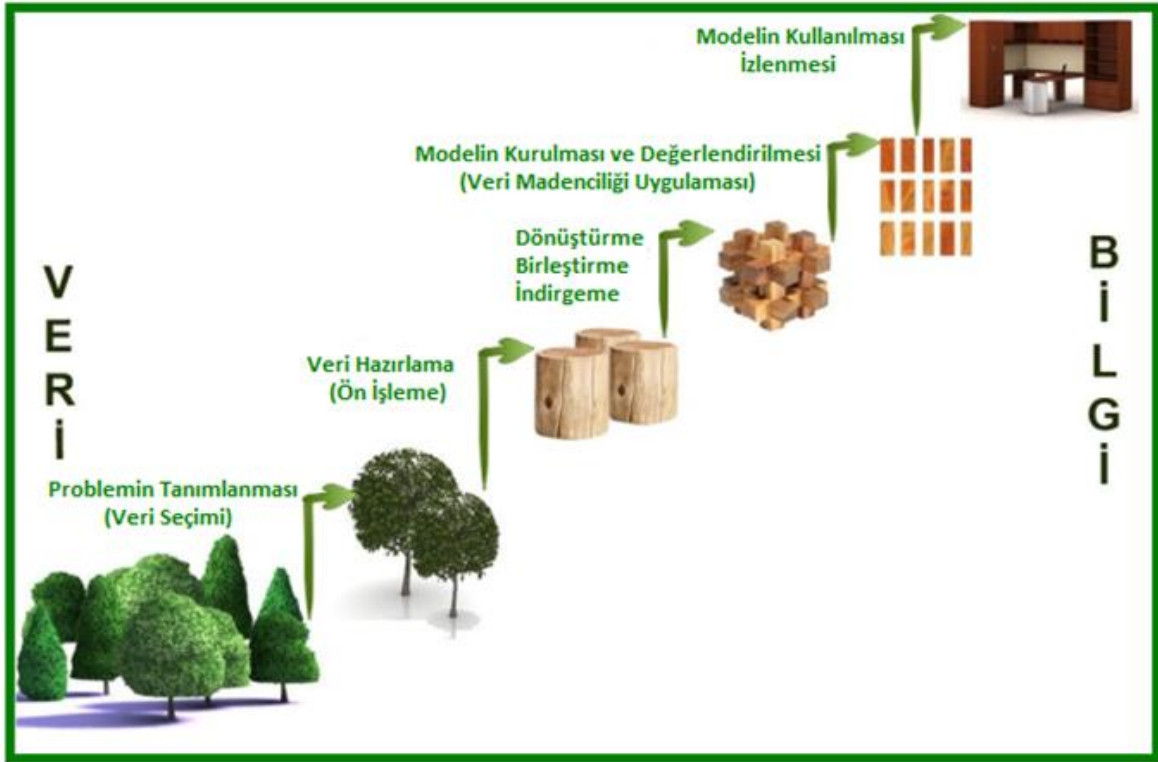


Resim 2.2. Veri madenciliği ile ilişkili alanlar

2.6. Veri Madenciliği Süreci

Veri madenciliği ile bilinmeyen ve daha önce keşfedilmemiş bilgiler elde edilebildiğinden veri madenciliği süreci tekrarlayan ve araştırmacı bir süreçtir. Veri madenciliği sürecinin başarılı bir şekilde ilerleyebilmesi için belli bir sistematik şekilde yürütmesi önemlidir. Başarılı bir veri madenciliği süreci 5 ana aşamadan oluşmaktadır. (Han, 2012: 6-8; Gencer, 2019).

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



Resim 2.3. Veri madenciliği süreci

2.6.1. Problemin tanımlanması

Veri madenciliğinin ilk ve kritik aşaması, problemin tanımlanmasıdır. Bu aşamada, asıl ihtiyacın doğru bir şekilde belirlenmesi ve ihtiyaca yönelik verilerin belirlenmesi

gerekmektedir. Sorunu tespit etmeden teşhis konulamayacağından bir veri madenciliği çalışmasında da problem anlaşılmadan süreç doğru bir şekilde yürütülemez. (Han, 2012: 6-8).

2.6.2. Verilerin hazırlanması

Problemin tanımlanması ve hangi verilerin kullanılması gerektiği belirlendikten sonra veri hazırlama aşamasına geçilir. Bu aşama veri madenciliğinin en kritik aşamasıdır. Bu aşama verilerin doğru bir şekilde toplanıp, seçilecek metoda uygun hale getirilmesi işlemi gerçekleştirilir. Aksi takdirde hiç sonuç alınamayacağı gibi sonuç alınsa dahi anlamsız veya yanıltıcı sonuçlar elde edilebilir. Bu sebep ile veri kalitesi sonuca direk etki etmektedir ve bir analiz sürecinde en uzun zaman (%80) genellikle verilerin hazırlanması için harcanmaktadır. Verilerin hazırlanması aşaması genel anlamda verilerin toplanması ve veri ön işleme aşamalarından oluşmaktadır. (Han, 2012: 6-8).

Verilerin toplanması

Veri toplama aşamasında bir işletme veya kuruluş genel anlamda kendi veri tabanlarını kullansa da dış kaynaklardan da faydalanmaktadırlar. Bu dış kaynaklar, nüfus sayımı, hava durumu gibi kaynaklar olabildiği gibi tedarikçi – müşteri veri tabanları, banka verileri veya veri pazarlayan şirketlerde olabilmektedir. Amaca uygun probleme katkı sağlayacak tüm verilerin toplama aşamasında elde edilmesi doğru çözüme ulaşmayı etkilemektedir. (Ahi, 2015; Han, 2012: 6-8).

Veri ön işleme

Veri ön işleme, veri hazırlığı aşamasının en büyük bölümünü oluşturmaktadır. Bu aşamada verinin kalitesi arttırılmaya ve model için en uygun hale getirilmeye çalışılmaktadır. Teknolojinin ve dolayısı ile bilişim sistemlerinin hızlı gelişimi çok büyük verileri ortaya çıkarması veri ön işlemenin önemini ve gerekliliğini arttırmaktadır. Veri ön işleme teknikleri genel anlamda 4 aşamadan oluşmaktadır. (Han, 2012: 6-8; Ahi, 2015; Klösger, 2002: 205).

- ✓ Veri temizleme (data cleaning)
- ✓ Veri birleştirme/bütünleştirme (data integration)

- ✓ Veri dönüştürme (data transformations)
- ✓ Veri indirgeme (data reduction)

Veri temizleme

Veriler çeşitli araçlar ile toplanırlar ve bu araçlar ile toplanan veriler birçok sebep ile yanlış, eksik veya tutarsız toplanmış olabilir. Bu sebepler, veri girişi yapan kişinin bilgi düzeyinden veya ilgili bilgiyi girmek istememesinden kaynaklandığı gibi teknolojik kısıtlardan da gerçekleşmiş olabilir. Veri temizleme aşaması ile farklı kaynaklardan toplanan veriler için eksik verilerin tamamlanması, tutarsız ve gürültülü verilerin düzeltilmesi sağlanmaktadır. Böylelikle veriler daha kullanışlı hale getirilmektedir. (Han, 2012: 6-8; Klösgen, 2002: 193).

Aşağıda veri temizlemede kullanılacak yöntemler listelenmiştir. (Ahi, 2015).

- Eksik veri içeren kısımlar silinebilir.
- Değişkenlerin ortalaması kayıp verilerin yerine kullanılabilir.
- Aynı sınıfa ait eksik veriler aynı sınıfın ortalama değerleri ile tamamlanabilir.
- Örneğin aynı kategoriye gren müşteriler için ortalama talep miktarı eksik verilerin yerine kullanılabilir.
- Var olan veriler kullanılarak eksik veriler tahmin edilebilir. Bu çözümde çeşitli tahmin algoritmaları, regresyon veya karar ağaçları kullanılabilir.

Ayrıca, gürültülü verileri tespit edip ayıklama için aşağıdaki yöntemler uygulanabilir.

- Sepetlere Ayırma
- Regresyon
- Kümeleme

Veri birleştirme

Veri madenciliği çalışmalarında genellikle farklı veri tabanlarından elde edilen veriler kullanılmaktadırlar. Bu verilerden veri madenciliğinde kullanılmak üzere veri ambarı oluşturulmaktadır. Fakat bu veri tabanlarında aynı şeyi ifade eden veriler farklı isimler ile

tutulmuş olabilir. Örneğin, bir üretim firmasında kullanılan malzemeleri ‘Stok ID’ ile tutarken farklı bir veri tabanında ‘Kalem Kodu’ ismiyle bulabiliriz. Bu tip verilerin entegre edilmesi sürecine veri birleştirme süreci denilmektedir. (Han, 2012: 6-8).

Veri dönüştürme

Veri madenciliği için hazırlanan veri setinde toplanan verilerin aynen kullanılması her zaman doğru olmayabilir. Değişkenler arasında çok büyük ortalama veya varyans farklılıkları olabileceğinden verilere dönüştürme işlemi yapılmalıdır. Bu dönüştürme işlemleri, birleştirme, genelleştirme ve normalleştirme. Bu tekniklerden bazen biri bazen ise birkaçı uygulanarak veriler uygun forma çevrilebilir. Veri dönüştürme işleminde en yaygın kullanılan yöntem normalleştirme işlemidir. Bu işlem için kullanılan yöntemler aşağıdaki gibidir. (Han, 2012: 6-8; Ahi, 2015; Gencer 2019).

Min - Max Normalleştirme: Min-max dönüştürme tekniği ile orijinal veriler üzerinde doğrusal dönüştürme yapılarak yeni bir veri aralığı elde edilir. Genellikle 0-1 aralığına uyacak şekilde dönüştürme yapılmaktadır. (Han, 2012: 83-87).

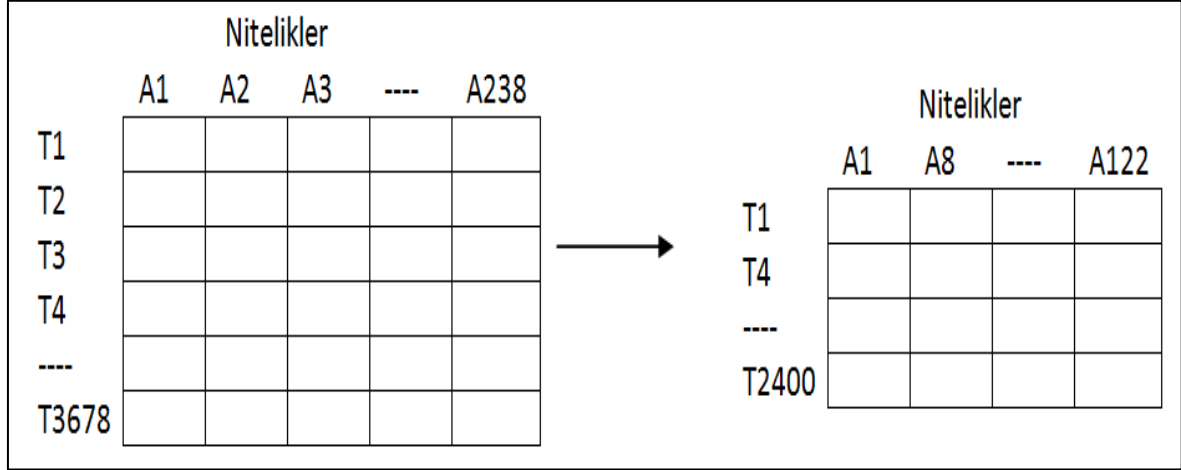
Z-Skor normalleştirme: Bu yöntem ile orijinal verilerin standart sapması ve ortalaması kullanılarak yeni veriler elde edilir. (Han, 2012: 83-87).

Ondalık ölçeklendirme: Orijinal verilerin ondalık kısmının hareket ettirilmesi ile uygulanmaktadır. Veri setindeki maksimum mutlak değere göre ondalık kısmın ne kadar hareket edeceği tespit edilmektedir. Örneğin veri setinde maksimum değer 575. Bu durumda ondalık kaydırma işlemi $n=3$ olacaktır ve yeni değer 0,575 olarak elde edilecektir. (Han, 2012: 83-87).

Veri indirgeme

Hazırlanan veri seti çok büyük olduğunda, veri madenciliği algoritmalarının uygulanması ve çalıştırması çok fazla vakit alabilmektedir. Bu durumlar da orijinal veri setinin özelliklerini bozmadan aynı özelliği koruyacak şekilde daha küçük sayıda bir veri seti elde edilmeye çalışılır. Bu tekniğe veri indirgeme tekniği denilmektedir. Veri indirgeme işlemi olarak 4 yöntem kullanılmaktadır. (Han, 2012: 83-87, 99; Ahi, 2015).

- ✓ Veri Birleştirme veya Veri Küpü (*Data Aggregation or Data Cube*)
- ✓ Boyut indirgeme (*Dimension Reduction*)
- ✓ Veri Sıkıştırma (*Data Compression*)
- ✓ Kesikli hale getirme (*Discretization*)



Şekil 2.1. Veri indirgeme örneği

2.6.3. Modelin kurulması ve değerlendirilmesi

Tanımlanan probleme yönelik verilerin toplanıp hazırlanmasından sonra probleme uygun modelin seçilmesi gerekmektedir. Veri madenciliği çalışmalarında genel olarak modele önceden karar verebilmek zor olduğundan probleme ve hazırlanan veri setine uygun olabilecek birçok model denenerek en uygun modele karar verilir. Bu aşamada denetimli ve denetimsiz öğrenme modellerinin hangisinin uygulandığı da kritik öneme sahiptir. Denetimli öğrenmede, verinin bir kısmı eğitim, bir kısmı test verisi olarak ayrılırken, denetimsiz öğrenmede verilere kümeleme analizi yapılır ve bu yol ile sınıflamaları yapılmaktadır. (Klösgen, 2002: 254).

2.6.4. Modelin kullanılması

Çeşitli yöntemler ile geliştirilen ve geçerliliği test edilen model bir çalışmanın ana modeli veya alt modeli olabilir. Bu aşamada model uygulanır ve veri madenciliğine uygun sonuçların elde edilmesi hedeflenmektedir. (Klösgen, 2002: 254).

2.6.5. Modelin izlenmesi

Geliştirilen ve sonuç alınan modelin zaman içinde gelişebilecek değişikliklere yönelik revize edilmesi veya tekrar en baştan geliştirilmesi gerekebilmektedir. Bu sebep ile hazırlanan modelin sürekli takip edilmesi ve yanlış sonuçlar vermemesi için düzenlenmesi gerekmektedir. (Han, 2012: 6-8; Ahi, 2015).

2.7. Veri Madenciliği Yöntemleri

Veri madenciliğinin amaçları farklı bilim insanları tarafından farklı şekilde kategorileştirilmiştir. Hepsinin ortak yönü ise veri madenciliğinde yer alan modellerin tanımlayıcı ve tahmin edici olmasıdır. (Han, 2012: 327-393; Gencer, 2019).

2.7.1. Tahmin edici modeller

Tahmin edici modeller ile sonuçları bilinen verilerden sonuçları bilinmeyen verilere yönelik tahmin çalışması yapılmaktadır. Bu sayede bilinmeyen eğilimler, örüntüler ve ilişkiler keşfedilmeye çalışılmaktadır. (Han, 2012: 327-393).

Sınıflandırma modelleri

Sınıflama modelleri veri sınıflarını ortaya çıkaran ve gelecek veri eğilimlerini tahmin etmek için kullanılan veri madenciliği modelleridir. Sınıflama modelleri 4 ana başlık ile kategorize edilebilir. (Han, 2012: 327-393; Klösgen, 2002: 258-298).

Bunlar;

- ✓ Karar ağaçları,
- ✓ Mesafeye dayalı yöntemler,
- ✓ İstatistiğe dayalı yöntemler,
- ✓ Diğer yöntemler.

Karar ağaçları

Karar ağaçları oluşturulmasının kolay olması, iyi güven düzeyine sahip olmaları, karar verme – yorumlama bakımından anlaşılır olması ve veri tabanları ile rahat entegre olabilmeleri açısından veri madenciliği sınıflandırma modelleri arasında en çok tercih edilen modellerdendir. Karar ağaçları, ağaç görünümünde olup tahmin edici bir modeldir. (Han, 2012: 330-332; Klösgen, 2002: 267).

Karar ağaçlarında 3 çeşit düğüm vardır. Bunlar;

Kök Düğüm: Kendisinden önce hiç dal olmayan ve üst düğüm olarak da adlandırılan düğümdür.

İç Düğüm: Kendisine sadece bir dal gelen ve kendisinden iki ya da daha fazla dal çıkan düğümdür.

Yaprak Düğüm: Kendisine sadece 1 dal gelen ve kendisinden dallanma olmayan ‘uç düğüm’, ‘kutup düğüm’ veya ‘terminal düğüm’ olarak isimlendirilen düğümdür.

Karar ağaçlarında yöntem temel olarak öğrenme ve sınıflama ikiye ayrılmaktadır. Öğrenme aşamasında sınıfı bilinen veriler karar ağacı algoritmaları ile belirlenir ve model oluşturulur. Model oluştuktan sonra sınıfı belli olmayan verilere model uygulanır ve hangi sınıfa ait olacakları tahmin edilir. Ağaç yapısında bu durum kök düğüm ile başlayıp yapraklara ulaşmaya kadar devam eden bir süreçtir. (Han, 2012: 330-332).

Karar ağaçlarında entropi, Gini indeksi ve Towing indeksi gibi yapılar ile dallanma oluşmaktadır. Bu indeksler ile dallanmanın doğru ve tutarlı oluşması sağlanmaktadır. (Han, 2012: 330-332).

En yaygın kullanılan karar ağaç algoritmaları aşağıdaki gibidir;

- ✓ ID3
- ✓ C4.5/C5
- ✓ CART
- ✓ CHAID

- ✓ SLIQ
- ✓ QUEST

Mesafeye dayalı yöntemler

Sınıflandırma yapmak için veriler arasındaki benzerlikleri ve mesafeyi dikkate alan yöntemlere mesafeye dayalı yöntemler denilmektedir. Bu sınıflandırma yönteminin en bilinen algoritması K-en yakın komşu algoritmasıdır.

K- En yakın komşu algoritması

Sınıfları bilinen verilerden yararlanılarak, yeni gelecek verilerin hangi sınıfa ait olacağını tahmin eden algoritmadır. Bu yöntem bilinmeyen verinin bilinen komşuları gibi davranacağı felsefesine dayanmaktadır. En yakın uzaklığa sahip k sayıda komşunun seçilmesine dayalı bir şekilde uygulanmaktadır. K-en yakın komşu algoritmasının 3 çeşit yaklaşımı bulunmaktadır. (Ahi, 2015; Han, 2012: 423).

Bunlar,

- ✓ k-en yakın komşu- klasik
- ✓ k-en yakın komşu- normalizasyon
- ✓ k-en yakın komşu- ağırlıklı oylama (Weighted Voting)'dir.

Bu yaklaşımlar temel olarak aynı şekilde uygulandığı gibi dikkate aldığı parametrelerin ağırlıkları veya değerleri yaklaşım bazlı değişmektedir. Hepsinde de yakın olanın bulunması hedeflenmektedir. Uzaklık için 3 yöntem yaygın olarak kullanılmaktadır. (Han, 2012: 423).

- ✓ Öklid uzaklık
- ✓ Minkowski uzaklık
- ✓ Manhattan uzaklık

İstatistiğe dayalı yöntemler

İstatistik, veri madenciliğinin en çok yararlandığı disiplinlerin başında gelmektedir. Birçok veri madenciliği algoritmasında istatistiksel yöntemler kullanılmaktadır. Bunlardan en yaygın kullanılanlarından bazıları regresyon ve Bayes/Naive Bayes algoritmalarıdır.

Regresyon

İstatistiksel veri madenciliği yöntemlerinin en belirginlerinden olan regresyon analizi, bağımlı değişkenler ile bağımsız değişken arasındaki ilişkinin bulunması, matematiksel şekilde ifade edilmesi ve bağımlı değişkenin ulaşacağı değeri tahmin etme yöntemidir. Sınıflandırma modellerinde veri tabanı özelliklerini bağımsız değişkenler, tahmin edilecek verileri bağımlı değişkenler ifade etmektedir. (Han, 2012: 18-19; Klösgen, 2002: 298).

Bayes/Naive bayes

İstatistiğe dayalı sınıflandırma yöntemlerinden bir diğeri de Bayes teoremidir. Bayes teoreminde Bayes sınıflayıcıları sınıf üyeliklerinin olasılıklarını tahmin eden yöntemdir. Bayes teoreminde en büyük önemli varsayım niteliklerin sınıflamada eşit etkileri olduğudur. Bu sebep ile niteliklerin birbirlerinden bağımsız oldukları düşünüldüğünde kullanılmaktadır. Naive Bayes algoritması, her kriterin sonuç üzerindeki etkilerinin olasılık üzerinden hesaplanması ile uygulanmaktadır. Bayes algoritması, büyük veri tabanlarında hızlı çalışması ve etkili olması sebebiyle veri madenciliği yöntemlerinden sık tercih edilenlerindendir. (Han, 2012: 350-351).

Diğer sınıflandırma yöntemleri

Karar ağacı algoritmaları, istatistiğe dayalı algoritmalar ve mesafeye dayalı algoritmalar haricinde literatürde farklı sınıflandırma metotları mevcuttur. Bunlardan yaygın olanlarından bazıları destek vektör makinaları, yapay sinir ağları ve genetik algoritmalarıdır.

Destek vektör makinaları

Destek vektör makinaları yöntemi istatistiksel öğrenme temeline dayanan kontrollü öğrenme algoritmasıdır. Bu algoritma kuramına göre belirli sayıda veri ile iyi bir genelleştirme yapılabilir. Ana hedefi, iki sınıf için en doğru ayrımın yapılmasıdır ve böylelikle hiperdüzlem tanımlanmaktadır. Bu yöntemde en doğru sınır, iki sınıf arasındaki en uzak düzlemdir. (Han, 2012: 408-413; Ahi, 2015).

Doğrusal veri setleri için destek vektör makinalarında hiper düzlem denklemi Eş. 2.1. de gösterilmektedir.

$$h(x) = v^T x + ed = 0 \quad (2.1)$$

Destek vektör makinaları doğrusal veriler üzerinde sahip olduğu verimli özelliklerin yanı sıra doğrusal olmayan problemlerdeki başarısı dolayısıyla da çok fazla tercih edilmektedir. DVM sınıflandırma yöntemi olarak kullanılmasıyla beraber regresyon analizi içinde kullanıldığından literatürde DVR (Destek Vektör Regresyonu) olarak da tanımlanmaktadır.

DVM tabanlı geliştirilen modellerin performansını DVM performansları yer almaktadır. Bunlar, Epsilon değeri (ϵ), ceza katsayısı (C) ve çekirdek fonksiyonu- parametresi (γ) doğru bir şekilde seçilmelidir. DVR'de yaygın olarak kullanılan çekirdek fonksiyonları genellikle doğrusal, polinom, radyal temelli fonksiyon (RBF) ve sigmoid olarak dört kategori altında toplanabilir. (Açıkkar ve Sivrikaya, 2020).

Doğrusal olmayan veri setlerinde destek vektör makinaları algoritmaları uygulanırken kernel fonksiyonu adı verilen yöntemler uygulanmaktadır. En sık kullanılan kernel fonksiyonları Eş. 2.2.deki gibidir.

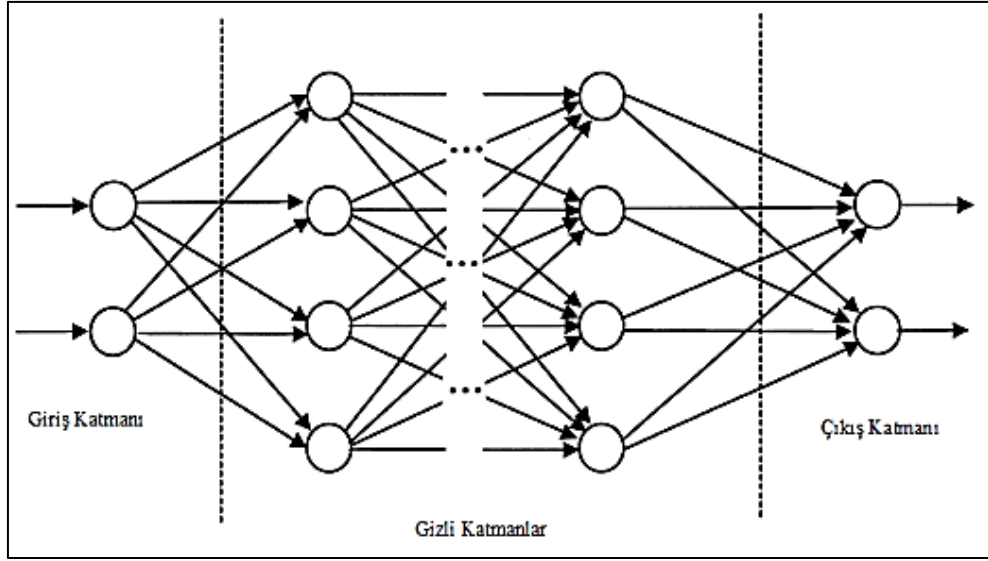
$$\begin{aligned} \text{Lineer fonksiyon: } K(x_i, x_j) &= x_i^T x_j \\ \text{Gaussian-RBF fonksiyon: } K(x_i, x_j) &= e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}} \\ \text{Polinomal: } K(x_i, x_j) &= (1 + x_i^T x_j)^p \end{aligned} \quad (2.2)$$

DVM parametrelerinin seçimi için belli bir formül veya bir matematiksel model olmadığından, bu parametreler belirlenirken arama algoritmaları kullanılabilir. Arama algoritması kullanılmadan ise, deneyerek doğru sonuçlara ulaşılması da bir yöntem olarak karşımıza çıkmaktadır. Bazı durumlar da kullanılan paket programların otomatik parametreleri ile de etkin sonuçlara ulaşılabilir.

Yapay sinir ağları

Yapay sinir ağları, insan beyninin öğrenme biçiminden yararlanılarak geliştirilmiş bir sınıflandırma algoritmasıdır. Bu yöntem ile, tanımlama, sınıflandırma ve tahmin yapılabilmektedir. İnsanlarda öğrenme, nöronlar arasındaki bağlantıların ayarlanması ile gerçekleşmektedir. Yapay sinir ağlarında ise öğrenme, girdi ve çıktı verilerinin işlenmesiyle, yani eğitme algoritmasının yakınsama sağlanana kadar tekrarlanarak ayarlanmasıyla gerçekleşmektedir. Yapay sinir ağlarındaki katman yapısı aşağıda gösterilmektedir. (Han, 2012: 398-400; Ahi, 2015; Klösgen, 2002: 304-310).

YSA metotlarında gizli katman sayısının ve nöron sayısının belirlenmesi sonuçlar için kritik öneme sahiptir. Nöron sayısı olması gerektiğinden fazla veya az olması durumunda model doğru çözüm veremez. Bu sebep ile doğru katman ve nöron sayısı tespit edilmelidir. Genellikle, nöron ve katman sayısının belirlenmesi çoklu denemeler ve çıkan sonuçların hataları kıyaslanarak bulunmaktadır. (Açıkkar ve Sivrikaya, 2020). Yapay sinir ağlarının katmanlı nöron yapısı resim 2.4.'de gösterilmiştir.



Resim 2.4. YSA katman yapısı

Genetik algoritmalar

Darwin'ın 'en iyi olan yaşar' prensibine dayanan sınıflandırma algoritmasıdır. Genetik algoritmalar, çözümlerin kodlanması, uygunlukların hesaplanması, çoğalma, çaprazlama ve mutasyon işlemlerinin uygulanması gibi aşamaları içermektedir. Bu algoritma öğrenme veri kümesi ile başlar ve bu küme üzerinden daha uygun olacağı tahmin edilen nesil oluşturmak için çalışmaya devam eder. Üretilen nesiller, bir önceki nesiller ile karşılaştırılır ve kötü bireyler popülasyondan elimine edilir. (Han, 2012: 426; Ahi, 2015).

2.7.2. Tanımlayıcı modeller

Tanımlayıcı modeller ile verilerin arasındaki gizli bilgiler bulunmaya çalışılır. Tanımlayıcı modeller tahmin için değil mevcut durum analizinin daha doğru yapılmasını sağlamaktadır.

Kümeleme modelleri

Veri tabanındaki verilerin benzerliklerine, ilişki durumlarına ve özelliklerine göre anlamlı ve kullanışlı bir şekilde gruplanmalarına kümeleme işlemi denilmektedir. Günümüzde, işletmeler, kurum ve kuruluşlar çok fazla veriye sahip olabildiğinden verileri doğru gruplamak için kümeleme algoritmalarını kullanmaktadırlar.

Kümeleme ile sınıfı olmayan verilere sınıf bulunma işlemi yapılmamaktadır. Bu veriler kümeleme yöntemleri ile gruplara ayrılırlar. Verilerin kaç kümeye ayrılacağı, hangi kümelerin oluşacağı elde edilen verilerin birbirlerine benzerliklerine göre belirlenmektedir. (Han, 2012: 444-448).

Kümelemede, tahmin veya sınıflandırma yöntemlerindeki gibi hedef seçilmez. Bunun yerine var olan veriler ile homojen gruplar elde edilmeye çalışılır. Bu yönüyle kümeleme denetimsiz öğrenmeye dayalı bir olgudur. (Han, 2012: 444-448).

Kümeler, verilerin birbirleri ile benzerlikleri ve uzaklıkları ile oluştuğundan kümelemede iki temel hedef vardır. Bunlar, küme için benzerliğin maksimize edilmesi ve kümeler arası benzerliğin minimize edilmesidir. Kümeleme çalışmalarında uzaklık belirleme teknikleri kullanılmaktadır. (Han, 2012: 444-448).

Bu tekniklerden yaygın olanlar aşağıda sıralanmıştır. (Han, 2012: 444-448).

- ✓ Öklid (Euclide) Uzaklığı,
- ✓ Manhattan (City-Blok) Uzaklığı,
- ✓ Minkowski Uzaklığı,
- ✓ Mahalanobis Uzaklığı
- ✓ Hotelling T2 Uzaklığı,
- ✓ Canberra Uzaklığı,
- ✓ Pearson Uzaklığı,

Kümeleme yöntemleri veri tipine, amaca ve kümelemenin oluşturulma şekline bağlı olarak farklı şekilde yapılabilir. Bunlar, Hiyerarşik ve hiyerarşik olmayan yöntemler olarak ikiye ayrılmaktadır.

Hiyerarşik kümeleme yöntemleri

Hiyerarşik kümeleme yöntemlerinde, verilere ait bir ana küme belirlenir ve aşamalı bir şekilde alt kümelere bölünmesi veya ayrı şekilde ele alınan kümelerin birleştirilerek bir küme edilmesi olarak tanımlanabilmektedir. Bu kümeleme biçiminde bir gözlem bir kümeye dahil edildikten sonra hep bu kümede yer alır ve değişiklik yapılmaz. (Han, 2012: 457).

Hiyerarşik kümelemede küme sayısı işleme başlamadan önce belirlenmektedir. Daha sonra işleme en yakın iki küme birleştirilerek devam edilir. Bu duruma göre küme sayısı 1 indirilerek yeni bir uzaklık matrisi oluşturulur. Bu işlemler $n-1$ defa tekrar edilir. (Han, 2012: 457).

Genel anlamda hiyerarşik kümeleme ikiye ayrılmaktadır;

- ✓ En yakın komşu algoritması (Tek bağlantı kümeleme yöntemi)
- ✓ En uzak komşu algoritması (Tam bağlantı kümeleme yöntemi)

En yakın komşu algoritması

En yakın komşu kümeleme algoritmasında gözlemler arasındaki uzaklıklar belirlenir ve en yakın 2 gözlem arsında ilk küme oluşturulur. Geri kalanların kendi aralarındaki mesafe ve oluşturulan kümeye mesafesi hesaplanır ve oluşturulan kümeye en yakın olan eklenir veya en yakın iki veriden yeni bir küme oluşturulur. Bu aşamadan sonra tekrar yeni uzaklıklar belirlenir. Bu aşamalar bütün kümelerin birleşip tek bir küme oluşturmasına kadar devam eder. (Han, 2012: 450).

En uzak komşu algoritması

En uzak komşu algoritması, en yakın komşu algoritması ile benzer şekilde çalışmaktadır. Bu algoritmada tek fark küme elemanları arasındaki verilerden en uzak olanlar belirlenir. Böylelikle kümeler arası uzunluklar belirlenir. Yine tek bir küme elde edilmeye çalışılmaktadır. (Han, 2012: 450).

Hiyerarşik olmayan kümeleme yöntemleri

Bu kümeleme yöntemi, küme sayısı ile alakalı bilgi varsa tercih edilmektedir. Bu yöntem genel anlamda n adet verinin k adet ($k < n$) kümeye ayrıştırılması işlemidir. (Gencer, 2019).

Bu yöntemde k adet küme için küme merkezleri rasgele seçilir. Daha sonra, veriler hangi küme merkezine yakın ise o kümeye dahil edilir ve yeni duruma göre küme merkezleri belirlenir. İşlemler, küme merkezlerinde hiçbir değişiklik olmayana kadar devam eder.

Hiyerarşik olmayan kümeleme, hiyerarşik yöntemlere göre daha hızlıdır ve büyük veri setleri için daha uygun bir kümeleme olmaktadır. Hiyerarşik olmayan kümelemede en çok 2 yöntem uygulanmaktadır. (Ahi, 2015).

- ✓ K-means (k-ortalamlar) algoritması,
- ✓ k-medoids algoritması.

K- Means (k-ortalamlar) algoritması

Bu algoritmanın amacı başlangıçta belli olan k adet küme için kareler toplamını minimize ederek n veriyi k adet kümeye yerleştirmektir.

Bu yöntemde, her kümenin merkezi hesaplanır ve her verinin küme merkezine mesafesi hesaplanır. Bu mesafeye göre veriler en yakın kümeye eklenir. Artan veya azalan gözlemlere göre ağırlıklı ortalamlar hesaplanır ve yeni küme merkezleri belirlenir. Bu işlemler değişiklik olmayana kadar devam etmektedir. Bu çalışma prensibi ile, k-means algoritması hızlı çalışan ve büyük veri setlerinde tercih edilen bir algoritmadır. (Ahi, 2015; Han, 2012: 451).

K-Medoids algoritması

K-medoids algoritması, k-means algoritmasının zayıf yönlerini ortadan kaldırmak için geliştirilmiştir. K-means algoritmasında değeri çok büyük bir gözlem küme merkezini çok büyük oranda değiştirebileceğinden kümelerin hassasiyeti bozulabilmektedir.

Bu algoritma, veri setini rasgele seçilen ve merkezi bir medoid olan k tane kümeye ayırır. Sonra, her gözlemin küme merkezlerine uzaklıkları belirlenir ve bu uzaklıklara göre veriler en yakın medoide yerleştirilir. Böylelikle yeni kümeler oluşturulmuş olur. Bir birim ile medoidin yer değiştirmesinin toplam maliyeti hesaplanır. Mevcut toplam maliyet, bir önceki ile karşılaştırılır. Bu işlemler, değiştirme maliyeti sıfırdan büyük veya sıfıra eşit çıkana kadar devam eder. Bu algoritma, gürültülü verilerden daha az etkilenmek ile beraber küçük veri setleri için daha uygundur. (Han, 2012: 454).

Birliktelik modelleri

Veri madenciliğinin ana konularından biri olan birliktelik, veri tabanlarından elde edilen geçmiş verilerin analiz edilerek, veriler arasındaki ilişkilerin tespit edilmesi ve bu tespiti göre stratejik kararlara yardımcı olan bir yaklaşımdır. Temel amaç, gizli ilişkileri bulmaktır.

Birliktelik analizinin en belirgin örneği ‘pazar sepeti analizi’dir. Bu analiz yöntemi, süpermarketlerde müşterilerin yaptıkları alışverişlerde satın almaları inceleyip hangi müşteri tipinin hangi ürünleri birlikte alacağını keşfetmeye çalışmaktadır. Bu bilgiye göre, market yöneticileri, pazarlama stratejilerini belirleyip, mevcut müşteriyi tutma ve yeni müşteriler kazanma yoluna gitmektedirler. (Ahi, 2015; Han, 2012: 243; Klösgen, 2002: 344-348).

Birliktelik analizi iki indeksten faydalanmaktadır. Destek ölçütü ile ürün grupları arasında ilişkinin ne kadar sık olduğu belirlenmektedir. Güven ölçütü ile de bir ürünün başka bir ürün ile hangi olasılıkla birlikte olacağı anlaşılmaktadır. Bu indekslerin olabildiğince yüksek olması gerekmektedir. (Han, 2012: 243).

Birliktelik algoritması olarak kullanılan çok sayıda yöntem mevcuttur. Bunlar aşağıda sıralanmıştır. (Han, 2012: 243).

- ✓ Apriori Algoritması
- ✓ AIS Algoritması
- ✓ SETM Algoritması
- ✓ HPA (Hash-based Parallel mining of Association rules) Algoritması
- ✓ PAR (Parallel Association Rules) Algoritması
- ✓ FP-Growth (Frequent Pattern-Growth) Algoritması
- ✓ DD (Data Distribution) Algoritması
- ✓ IDD (Intelligent Data Distribution) Algoritması



3. LİTERATÜR TARAMASI

Veri madenciliği yöntemleri ile talep tahmini yapılacak bu çalışmada çalışmanın derinliğinin artırılması ve doğru yöntemlerin uygulanması hedefi ile veri madenciliği yöntemleri ile tahmin çalışması yapılan akademik kaynaklar detaylı bir şekilde incelenmiştir. Bu inceleme, veri madenciliği kavramının günümüz trend konularından biri olması dolayısı ile 2014 yılı ve sonrası çalışmalar literatür taraması kapsamında incelenmiştir. 2014 yılı öncesinden sadece Han ve arkadaşlarının ‘Data Mining Concepts and Techniques’ isimli kitabı ve Willi Klösgen ve Jan M. Zytkow’un ‘Data Mining and Knowledge Discovery’ kitabı veri madenciliği açısından kaynak niteliğinde olduğundan incelmeye dahil edilmiştir. Derinlemesine incelenen akademik yayınlar aşağıda özetlenmiştir.

Klösgen, W., Zytkow, J.M. (2002), yazıldığı yıl itibari ile bakıldığında veri madenciliği açısından günümüze ışık tutan bir kitaptır. Çalışmada, veri madenciliğinin hangi kaynakları kullanarak gelişeceğini ve birçok alanda çözüm için kullanılabileceği vurgulanmıştır. Ayrıca veri madenciliği ile ilintili kavramların açıklandığı ve okuyucuyu veri madenciliğinin genel anlamda sınırlarının çizildiği bir çalışma olduğu görülmüştür.

Han J ve arkadaşları (2012), veri madenciliğini bütün yönleriyle ele alan bir kitap yayınlamışlardır. İlk baskısı 2000 yılında çıkan kitabın 2012 yılında yayınlanan 3. Baskısı incelenmiştir. Veri madenciliğinin genel anlamda hangi yönleri ele aldığı, hangi modellerin ne tip verilere nasıl uygulanacağı, sonuçların nasıl analiz edileceği yönünden veri madenciliğinde kaynak niteliğinde bir çalışma ortaya konulmuştur.

Ayhan S. ve Erdoğan Ş. (2014), veri madenciliğinin etkili yöntemlerinden Destek Vektör Makinaları yönteminde çekirdek fonksiyonu seçimine yönelik araştırma yapmışlardır. Bu çalışma kapsamında 12 veri seti ve 4 çeşit çekirdek fonksiyonu incelenmiştir. Destek vektör makinasının performansını veri setinin mi yoksa seçilen çekirdek fonksiyonunun mu etkilediği ANOVA kullanılarak belirlenmiştir. Daha sonra veri setlerinin etkileri çıkartılarak, çekirdek fonksiyonlarından hangisinin daha başarılı olduğu tespit edilmeye çalışılmıştır. İncelenen çekirdek fonksiyonlar Lineer, radyal tabanlı, polinomial ve sigmoid olup genel anlamda en başarılı çekirdek fonksiyonunun radyal tabanlı çekirdek fonksiyonu olduğu belirlenmiştir.

Ramos ve arkadaşları (2015), İspanya'nın güneyinde yer alan Jean şehrindeki ilaç talep tahmini çalışmasını gerçekleştirmişlerdir. Asıl amaçları, bölgesel değişkenlerin olduğu durumlarda talep tahminin veri madenciliği yöntemleri ile yapılabileceğini göstermektir. Veri madenciliği yöntemlerinden Genelleştirilmiş Lineer Model (GLM), Destek Vektör Makinaları (SVM) lineer kernel ve SVM- Gaussian Kernel (lineer ve lineer olmayan çekirdek fonksiyonlar) kullanılarak belli bir ilaca yönelik talep tahmini yapılmıştır. %5,78 hata oranı ve ortalama tüketime göre %61,44 daha doğru sonucu ile GLM metodu en iyi model seçilmiştir. Çıkan sonucun doğruluğu, bölgesel faktörlerin rol oynadığı ilaç talebinin tahmin edilmesinde veri madenciliğinin etkili olabileceğini göstermiştir.

İşlek İ. (2015), ulusal boyutta bir kuruyemiş sektörünü için talep tahmini çalışmasını veri madenciliği yöntemleri ile yapmaya çalışmıştır. Bu çalışma için firmanın 3 yıllık gerçek satış verileri, 98 adet deposu ve 70 ürünü veri seti olarak yer almaktadır. Çalışma kapsamında 4 model seçilmiş ve denenmiştir. Bu modeller, Lineer Regresyon, Çok Katmanlı Algılayıcı, Bayes Ağı ve Ardışık Minimal Optimizasyon'dur. Çalışmada ayrıca, ilgili modellerin 2,3, ve 4'lü kombinasyonları yığılmış genelleme yöntemi ile birleştirilip talep tahmini yapılmaya çalışılmıştır. Ayrıca iki parçalı çizge genelleme yöntemi ile de talep tahmini yapılmaya çalışılmıştır. Bu yöntem 2 aşamadan oluşmuş olup sadece 2. Aşamasında yöntemler birleştirilmiştir. Bütün bu denemelerin sonucunda, yığılmış genelleme metoduyla modellerin birleştirilmesinin model bazlı tahminden iyi sonuç verdiği görülse de iki parçalı çizge genelleme yöntemi ile gruplama yapılan modelin en iyi sonuca ulaştığı tespit edilmiştir.

Ahi L. (2015), amaca göre bireysel tüketimin sınıflandırılması (COICOP) ile 12 ana harcama grubu bazında harcama değişkenlerini tahmin etmeye çalışmıştır. Tahmin çalışmasının veri seti TÜİK kurumunun 2011 yılında yapmış olduğu Hanehalkı Bütçe Araştırması (HBA)'dan elde edilen 9918 hane halkı ve yine TÜİK'in 2012'de yaptığı Gelir ve Yaşam Koşulları Araştırması (GYKA)'dan elde edilen 17562 hane halkının birleştirilmesi ile oluşturmuştur. Tahmin çalışması için veri madenciliği yöntemlerinden faydalanılmış olup istatistiksel eşleştirme kullanılmıştır. İstenilen bütün değişkenleri aynı çatı altında toplama hedefiyle gerçekleştirilen istatistiksel eşleşme yöntemiyle, veri madenciliği çalışmalarına uygun veri seti hazırlanmıştır. Bu çalışma istatistik ile veri madenciliğinin farkını ve aynı zamanda uyumlu çalışmasını gösterir niteliktedir.

Bircan H. & Çam S. (2016), sağlık sektöründe hasta davranışlarının gruplanmasını veri madenciliği kümeleme yöntemleri ile yapmaya çalışmışlardır. Bu çalışma kapsamında, veri setini oluştururken 2006 ile 2011 yılları arasında hastaneye başvuran hasta kayıtları kullanılmıştır. 65 yaş üzeri ve 18 yaş altı hasta başvuruları dahil edilmeyerek veriler toplanmıştır ve toplam 78239 hasta verisi ile veri seti hazırlanmıştır. Kümeleme çalışması için k – means ve yoğunluk tabanlı kümeleme algoritmaları tercih edilmiştir. Ayrıca çalışmalarında hastalara ait sayısal veriler haricinde demografik verilerde veri setine dahil edilmiştir. Çalışmanın sonunda, yoğunluk tabanlı kümeleme analizinin k-means algoritmasına göre daha doğru sonuç verdiğini vurgulamışlardır.

Le Cam ve arkadaşları (2016), fan elektrik tüketimini veri madenciliği teknikleri ile tahmin etmeye çalışmışlardır. Bu kapsamda bina otomasyon sistemi üzerinde her 15 dakikada bir kaydedilen verileri kullanmışlardır. Kümeleme algoritmaları ile tipik olmayan veriler tespit edilmiştir. Geliştirilen karma (fiziksel ve kapalı devre lineer olmayan sinir ağı) model 23 günlük bir veri seti ile çalıştırılmış ve 6 saatlik bir tahmin yapılmıştır. Ayrıca, duyarlılık analizi için 4,8,12 günlük veri setleri ile aynı model çalıştırıldığında tahmini olumsuz etkilemediği ortaya çıkmıştır. Bu çalışma ile veri madenciliği temelli karma tahmin modelleri ile besleme fanının elektrik tüketiminin yapılabileceği gösterilmiştir.

Odabaş Ö. (2017), Telekom sektöründe ayrılma eğilimi olan müşterileri tahmin etmeye çalışmıştır. Tahmin çalışmasını veri madenciliği yöntemlerinden faydalanarak yapmaya çalışmıştır. Bu kapsamda, veri madenciliğinin sınıflandırma algoritmalarından ve nitelik seçimi tekniklerinden yararlanılmıştır. Bu teknikler Naïve Bayes, Lojistik Regresyon, SMO, J48, K-star'dır. Veri seti 18 aylık müşteri verileridir ve toplam 3300 müşteriyi kapsamaktadır. Çalışma sonunda en doğru sonuca %94,41 ile ham veriye uygulanan J48 algoritması ile ulaşılmıştır.

Öztürk ve arkadaşları (2017), Türkiye’de mobilya sektörünün geleceği için öngörü çalışması yapmışlardır. Mobilya talebini tahmin etmek için veri madenciliği yöntemlerinden faydalanılmıştır. Veri seti olarak TÜİK tarafından periyodik olarak yayınlanan 15 yıllık oturma izni alınan konut sayısı kullanılmıştır. 2002-2014 arası eğitim verisi olarak belirlenmiş olup 2015 ve 2016 verileri test verisi olarak kullanılmıştır. Çalışma veri madenciliğinin regresyon modeli üzerinden gerçekleşmiş olup, test verilerine yakın tahminler elde edilmiştir. Daha sonra aynı model ile 2023 yılı tahmini

gerçekleştirilmiş ve 2016 yılı değerlerinin yaklaşık 2 katı bir talep beklentisi ortaya çıkmıştır.

Miao ve arkadaşları (2017) veri madenciliği teknikleri ile yeni bir sürekli kan basıncı tahmini yaklaşımı geliştirmeye çalışmışlardır. Bu çalışmada yeni model beri madenciliği tekniklerinin geleneksel mekanik tabanlı model ile birleştirilmesi ile oluşturulmuştur. Bu model oluşturulurken çok değişkenli doğrusal regresyon ve destek vektör regresyon algoritmaları kullanılmıştır. Geliştirilen model 73 denek ile test edilmiştir. Önerilen yöntemi ile son teknoloji nabız geçiş süresi yöntemi ile karşılaştırıldığında farklı durum ve zaman aralıklarında önemli ölçüde daha az hata ile tahmin yaptığı tespit edilmiştir. Araştırmacılar, sonuç bölümünde daha etkili bir sonuç için çalışmaya devam edeceklerini belirtmişlerdir.

Karna ve arkadaşları (2018), yazılım projelerinde efor tahminini veri madenciliği yöntemleri yapmayı hedeflemişlerdir. Hırvat bir yazılım firmasının 42000 çalışma saatini aşan 27 projesi çalışmaya dahil edilmiş ve efor tahmini için kullanılmıştır. Bu 27 projeden 4 adedi test verisi olarak ayrılmış geri kalan 23 adedi ile model geliştirilmiştir. Geliştirilen model ile diğer 4 adet için efor tahmini yapılmış ve insanların tahminlerinden çok daha az hata ile tahmin yaptığı kanıtlanmıştır. Bu çalışma ile yazılım şirketleri gibi büyük veriye sahip yerlerde efor tahmini için veri madenciliğinin kullanılabileceğini göstermiştir. Bu yol ile tahmin hataları azaltılarak çalışmalar optimize edilebilmektedir.

Yang ve arkadaşları (2018), veri madenciliği yöntemleri ile öğrenci performans analizini, ilerleme analizini ve geleceğe yönelik öğrenci performans tahminini iyileştirilebileceğini gösteren bir çalışma yapmışlardır. Bu çalışma için öncelikle Öğrenci Öznitelik Matrisi oluşturmuşlardır. Böylelikle performans ile alakalı olmayan özniteliklerde ayıklanmıştır. Daha sonra, öğrencilerin önceki performansından analiz yapmak üzere sınıflandırma temelli Geri Yayılım Sinir Ağları (BP-NN) kullanılmıştır. Sonunda da, yapılan çalışmayı öğrencinin ilerleme analizini ve performans tahminini oluşturmak için kullanmışlardır. Bu aşamada hangi faktörün öğrenci nasıl etkiyeceğine yönelik öneriler oluşturulmaktadır. Çalışma 60 lise öğrencisinin akademik verilerinin toplanması ile gerçekleştirilmiştir. Özellikle performans tahmini için çıkan sonuçların doğruluğu veri madenciliği yöntemlerinin bu alanda kullanılmasının çok faydalı olabileceğini göstermiştir.

Ecemiş O. & Irmak S. (2018), paslanmaz çelik sektöründe distribütör firma için satış tahmini çalışması veri madenciliği yöntemlerinden yararlanarak yapmaya çalışmışlardır. Bu çalışma için firmanın veri tabanından 2008 ile 2016 yılları arasında gerçekleşmiş günlük satış verileri kullanılmıştır. Daha sonra, müşteri bilgileri ile eşleştirilerek aylık satış verileri elde etmişlerdir. Tahmin için veri madenciliği yöntemlerinden destek vektör makinaları ve yapay sinir ağları algoritmalarını kullanmışlardır. Çıkan sonuçlar karşılaştırıldığında veri madenciliği yöntemlerinin tahmin çalışması için başarılı sonuçlar verdiğini tespit etmişlerdir. Ayrıca destek vektör makinası algoritmasının yapay sinir ağlarına göre nispeten daha doğru sonuç verdiğini vurgulamışlardır.

Çakmakoglu A. (2018), finans sektöründe veri madenciliği yöntemleri ile kampanya modellenmesi çalışması yapmıştır. Çalışma ile, tele pazarlama yönteminde doğru müşterilerin aranması hedeflenmiştir. Çalışmada, veri madenciliği sınıflandırma yöntemlerinden aşırı öğrenme makinası kullanılmıştır. Ayrıca genetik algoritma kullanılarak modelin optimizasyonu sağlanmaya çalışılmıştır. Çalışma sonunda %92 düzeyinde sınıflandırma elde edilmiştir. Böylelikle, veri madenciliği yöntemlerinin finans alanında doğru sınıflama yapabileceği ve bu yöntemler ile daha az kaynak ile daha doğru kişilere ulaşılabilceği gösterilmiştir.

Türkoğlu ve arkadaşları (2019), soğuk dövme makinalarının arıza tahminini doğru bir şekilde gerçekleştirebilmek için veri madenciliğinden yararlanmışlardır. WEKA ve Rapidminer programlarında çeşitli yöntemler denenmiş ve WEKA’da J48 karar ağaçları ile en doğru (%77) sonuçlara ulaşılmıştır. J48 modeli 2014 ve 2017 yılları arasında gerçekleşen 28 arızayı önceden doğru bir şekilde tespit etmeyi başarmıştır.

Qui ve arkadaşları (2019), bir üretim tesisinde sıklıkla yaşanan ve üretim programını olumsuz etkileyen makine ve ekipman bozulması veya malzeme yokluğu gibi sorunları önceden tahmin edip üretim planının tekrar revize edilmesini sağlayacak bir model üzerine çalışmışlardır. Üretimden toplanan çok sayıda veriyi doğru sınıflandırmak için veri madenciliği yöntemlerinden Naive Bayes, karar ağaçları ve iki yöntemi birleştirerek karma bir sınıflandırma modeli oluşturmuşlardır. 3 model farklı boyutlardaki 3 veri setine uygulanmış ve hata oranları çıkarılmıştır. En düşük hata oranları hibrit (karma) model ile elde edilmiştir. Bu çalışma ile, üretimdeki aksaklıkların tahminin veri madenciliği

yöntemleri ile yapılacağı gösterildiği gibi Naive Bayes ve karar ağaçları yöntemleri birleştirildiğinde etkili olabildiği de gösterilmiştir.

Türk E. & Kiani F. (2019), beyaz eşya sektöründe için satış tahminlerini veri madenciliği yöntemlerinden faydalanarak yapmaya çalışmışlardır. Çalışma için veri madenciliği yöntemlerinden yapay sinir ağları ve regresyon yöntemleri kullanılmıştır. 2007 ve 2017 yılları arasındaki verileri çalışmaya dahil etmişlerdir. 2007-2015 verileri eğitim seti, 2016 ve 2017 verileri ise test verisi olarak belirlenmiştir. Yapılan çalışma ile gerçek verilere çok yakın tahminler elde edilmiş olup, veri madenciliği yöntemlerinin talep tahmininde çok önemli ölçüde fayda sağladığı gösterilmiştir.

Özcan ve arkadaşları (2019), veri madenciliği yöntemlerini kullanarak kalp hastalığı tahmini gerçekleştirmeye çalışmışlardır. Kalp hastalıklarında erken teşhisin hayati önem taşıdığı göz önünde bulundurularak önden tahmin etmek için ileri yayımlı yapay sinir ağları ve destek vektör makinaları kullanılarak model geliştirilmiştir. 13 adet nitelik belirlenmiş olup, 170 (90 adet eğitim, 80 adet test) kişiye ait veriler kullanılmıştır. YSA ile %83,33 DVM ile %91,67 oranında doğru tahmin başarılmıştır. Çalışmanın sonuç bölümünde DVM ile kalp hastalığı çeşitlerine yönelik tahmin çalışmasının yapılabileceği önerilmiştir.

Terzi M.B. & Arıkan O. (2019), koroner kalp hastalıklarını veri madenciliği yöntemlerini kullanarak tespit etmeye çalışmışlardır. Güncel sinyal işleme ve destek vektör makinası yöntem olarak seçilmiş olup, EKG verileri nitelik olarak model oluşturulurken kullanılmıştır. Ayrıca iskemik verilerin eksik olması durumunda gözetimsiz öğrenme metodu olarak Gauss Karışım modeli (GKM) kullanılmıştır. Çıkan sonuçlar değerlendirildiğinde DVM ile %90,47, GKM ile %81,25 lik doğruluk oranı ile tespit yapılabildiği görülmüştür.

Köse U. (2019), veri madenciliğinin destek vektör makinaları yöntemi ile diyabet teşhisine yönelik tahmin çalışması gerçekleştirmiştir. Veri seti 768 denekten (500 adedi diyabet negatif, 268 adedi diyabet pozitif) oluşturulmuştur. Çalışmaya diyabet teşhisinde etkili olduğu düşünülen 8 adet nitelik dahil edilmiştir. Çalışma kapsamında DVM ile zeki optimizasyon metotlarının hibrit edildiği bir yöntem ile gerçekleştirilmiştir. Zeki optimizasyon algoritmalarından 5 farklı yöntem ile DVM birleştirilmiş ve sonuçlar elde

edilmiştir. Sonuçlar incelendiğinde, Guguk kuşu araması (GKA) ile DVM'nin %94,27 ile en doğru sonucu verdiği görülmüştür.

Karadağ K. (2020), veri madenciliğinin makine öğrenmesi yöntemleri ile semen kalitesi tahmini gerçekleştirmeye çalışmıştır. Çalışma 18 – 36 yaş aralığından olan 100 farklı denekten alınan veriler ile gerçekleştirilmiştir. Veri setine nitelik olarak, alkol kullanımı, sigara kullanımı, günlük oturma süresi, mevsim gibi değerler normalize edilerek eklenmiştir. Çalışmada veri madenciliği yöntemlerinden destek vektör makinaları (DVM), lineer regresyon (LR), k – en yakın komşu (KNN) ve doğrusal ayırt edici (DAE) kullanılmıştır. Sonuçlar değerlendirildiğinde başarı oranları sırasıyla, DVM %91,7, KNN %87,5, DAE %70,8 ve LR %66,7 olarak tespit edilmiştir. Çalışma semen kalitesi analizinde veri madenciliği yöntemlerinin kullanılabileceğini gösterdiği gibi DVM'lerin tahmin çalışmalarında etkin sonuç vermesine de örnek teşkil etmektedir.

Açıkkar M. & Sivrikaya O. (2020), veri madenciliği yöntemleri ile yıkanmış Türk linyit kömürlerinin üst ısı değerlerini tahmin etmeye çalışmışlardır. Bu çalışmada veri kümesini, 3672 kömür numunesinin üst ısı değerleri, kül, kükürt ve uçucu madde değişkenleri ile oluşturmuşlardır. Çalışmada destek vektör makinaları ve yapay sinir ağları yöntem olarak uygulanmıştır. MAPE değerlerine göre sonuçlar kıyaslandığında iki yöntemle de doğru tahmin yapıldığı gösterilmiştir. DVM %1,19 hata değeri ile YSA'na 1,21 göre daha başarılı sonuç vermiştir.



4. UYGULAMA

Elektrik ekipmanlarının üretiminde faaliyet gösteren bir firmanın orta gerilim hücresi satışlarının tahmin edilmesine yönelik gerçekleştirilen bu çalışmada, veri seti ilgili firmanın 2016-2019 yılları arasında verileri alınarak oluşturulmuştur. 2016 – 2018 yılları arasındaki aylık satışlar eğitim verisi olup 2019 satışları test verisi olarak kullanılmıştır. Firmanın faaliyet gösterdiği ürünler yapılaşma ile doğru orantılı olduğundan ülkenin temel ekonomik parametrelerinden direk etkilendiği belirlenmiştir. Bu sebep ile satış miktarını direk etkileyen ‘dolar kuru’, ‘enflasyon oranı’ ve faiz oranları’ nitelik olarak oluşturulan modele eklenmiştir.

Oluşturulan veri seti incelendiğinde doğrusal olmadığı görülmektedir. Literatür tarama aşamasında gerek veri madenciliği açısından kaynak niteliğinde olan Han ve arkadaşlarının ‘Data Mining Concepts and Techniques’ isimli kitabı gerekse benzer veri setleri ile tahmin yapılan diğer bilimsel araştırmaların ışığında doğrusal olmayan verilerden yapılan tahmin çalışmalarında etkin sonuçlar elde edilebilen destek vektör makinaları ve yapay sinir ağları yöntemlerinin kullanılmasına karar verilmiştir. Oluşturulan modellerin çözümü WEKA (Waikato Environment for Knowledge Analysis) paket programının 3.9.4 sürümü (son sürüm) ile gerçekleştirilmiştir.

4.1. Yapay Sinir Ağları

Oluşturulan veri seti ile 2019 yılı tahmininde YSA Çok Katmanlı Algılayıcılar (Multilayer Perceptron) metodu uygulanmıştır. Bu yöntemde tahmin sonuçlarını gizli katman nöron sayısı (GKNS) ve eğitilen dönem sayısı (EDS) sonuçları direk etkilemektedir. 2 parametre için çok sayıda deneme yapılmış olup GKNS için {3-3-3, 2-4-6, 1} ve EDS için {400, 500} çalışma kapsamında değerlendirilmeye alınmıştır.

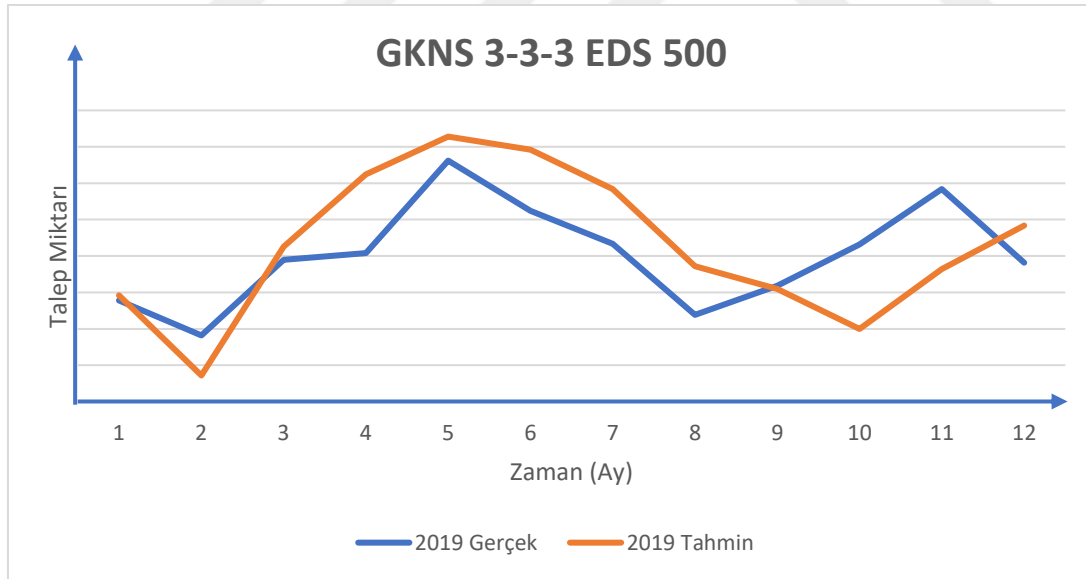
Çözümü gerçekleştirilip değerlendirmeye alınan 6 kombinasyonun MAPE hata değerlerine göre sonuçları çizelge 4.1.de sunulmuştur.

Çizelge 4.1. Yapay sinir ağları ile MAPE değerleri

MAPE		
GKNS \ EDS	400	500
3-3-3	%35,71	%31,76
2-4-6	%34,52	%29,26
1	%44,66	%36,51

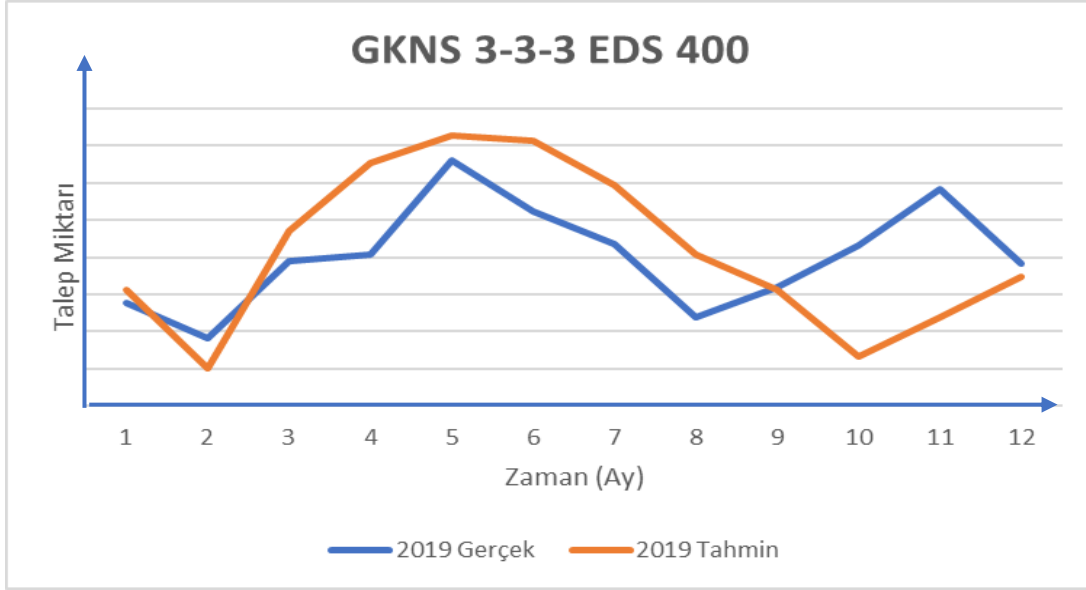
Çizelge 4.1. incelendiğinde parametreler GKNS 2-4-6 ve EDS 500 olduğunda en doğru tahmine %29,26 ile ulaşıldığı, GKNS 1 ve EDS 400 olduğunda en zayıf tahminin %44,66 ile yapıldığı görülmüştür. Bu sonuçlar ışığında katman sayısının ve eğitim dönem sayısının etkinliği de görülmüştür. Çözümü gerçekleştirilen 6 farklı kombinasyona 2019 yılı 12 ayı için gerçekleşen ve tahmin edilen veriler grafikler üzerinden incelenmiştir.

Şekil 4.1’de GKNS 3-3-3 ve EDS 500 parametrelerine göre 2019 yılı 12 aylık gerçekleşen ve tahmin edilen grafiği verilmiştir.



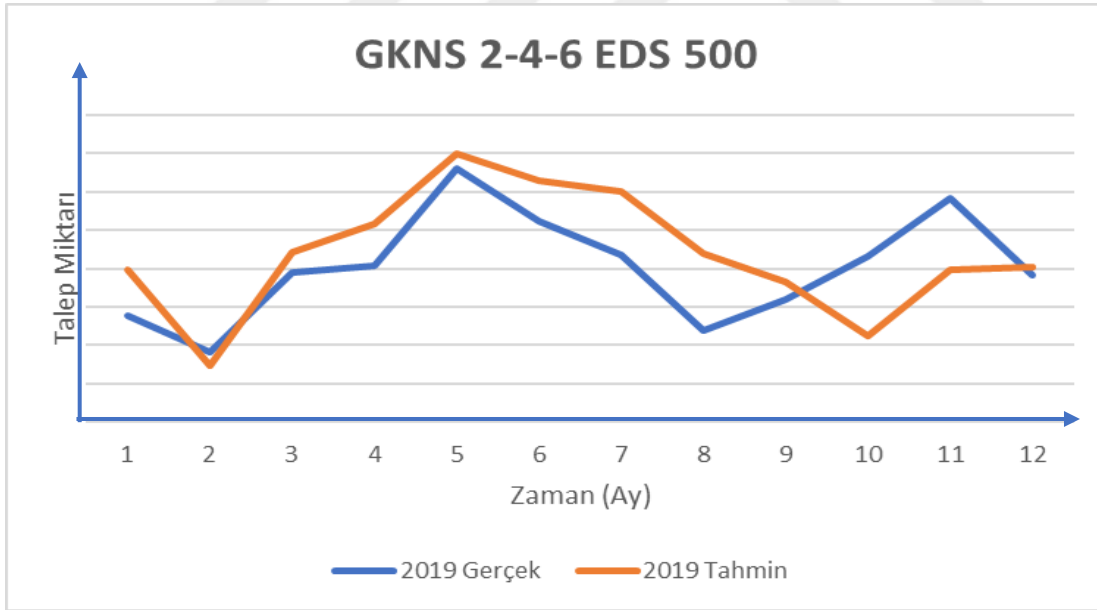
Şekil 4.1. GKNS 3-3-3 EDS 500 göre tahmin grafiği

Şekil 4.2’de GKNS 3-3-3 ve EDS 400 parametrelerine göre 2019 yılı 12 aylık gerçekleşen ve tahmin edilen grafiği verilmiştir.



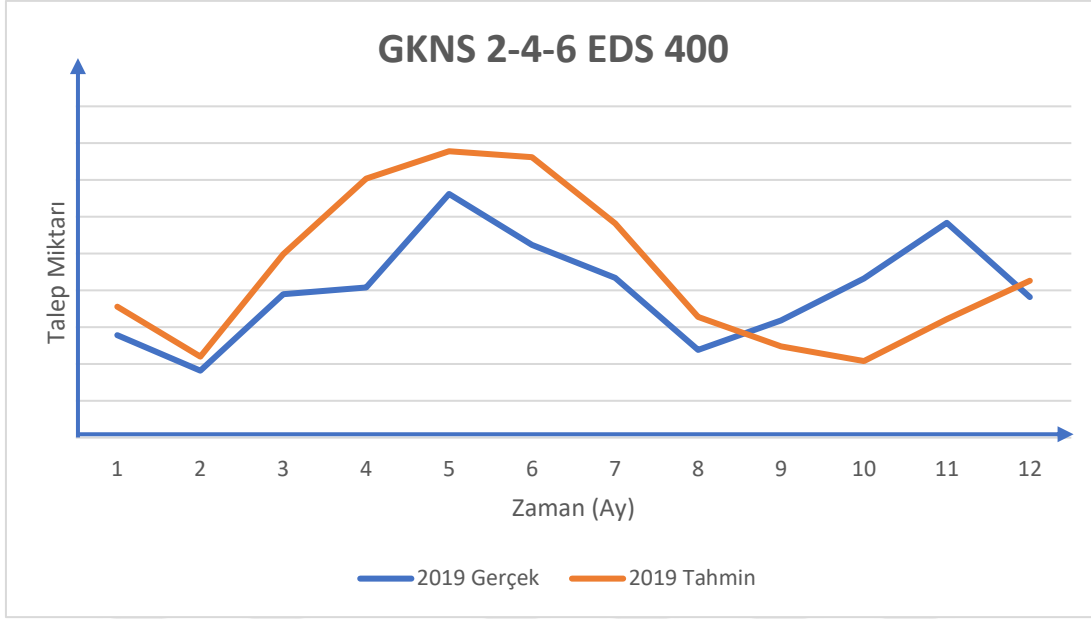
Şekil 4.2. GKNS 3-3-3 EDS 400 göre tahmin grafiği

Şekil 4.3’de GKNS 2-4-6 ve EDS 500 parametrelerine göre 2019 yılı 12 aylık gerçekleşen ve tahmin edilen grafiği verilmiştir.



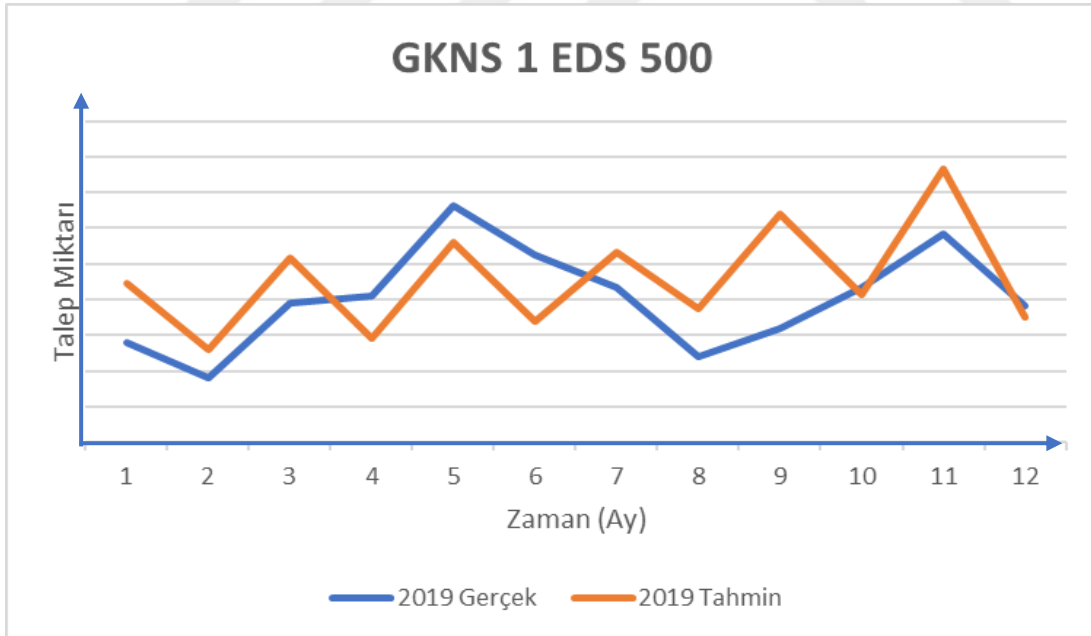
Şekil 4.3. GKNS 2-4-6 EDS 500 göre tahmin grafiği

Şekil 4.4’de GKNS 2-4-6 ve EDS 400 parametrelerine göre 2019 yılı 12 aylık gerçekleşen ve tahmin edilen grafiği verilmiştir.



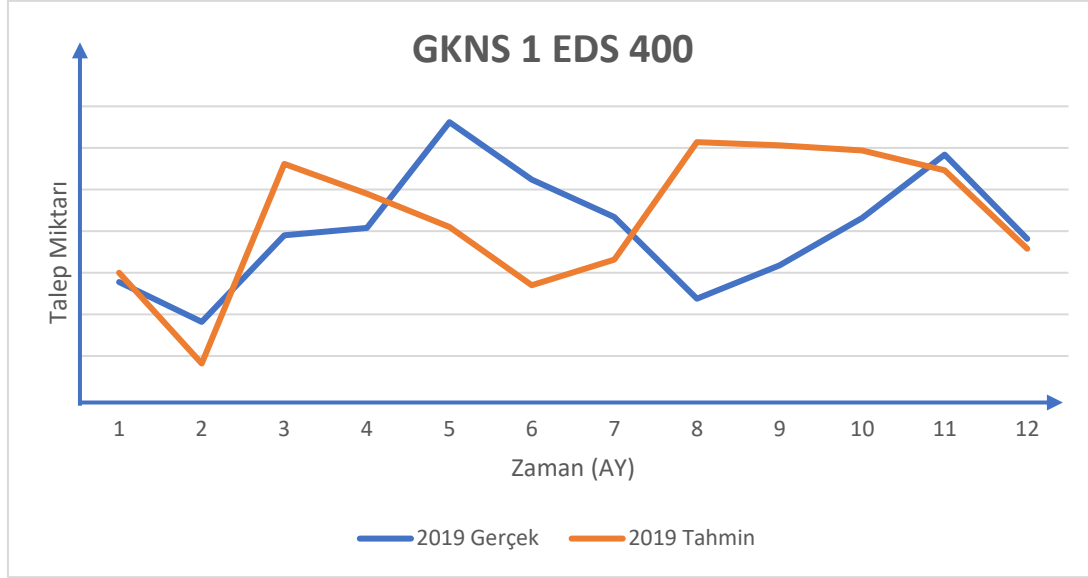
Şekil 4.4. GKNS 2-4-6 EDS 400 göre tahmin grafiği

Şekil 4.5’de GKNS 1 ve EDS 500 parametrelerine göre 2019 yılı 12 aylık gerçekleşen ve tahmin edilen grafiği verilmiştir.



Şekil 4.5. GKNS 1 EDS 500 göre tahmin grafiği

Şekil 4.6’de GKNS 1 ve EDS 400 parametrelerine göre 2019 yılı 12 aylık gerçekleşen ve tahmin edilen grafiği verilmiştir.



Şekil 4.6. GKNS 1 EDS 400 göre tahmin grafiği

4.2. Destek Vektör Makinaları

İlgili firmanın 2016 – 2018 yılları arasında satış verilerini kullanarak 2019 yılının doğru tahminine yönelik bu çalışmada destek vektör makinaları yöntemi uygulanmıştır. DVM’ler doğrusal olmayan veriler üzerinde etkinliği kanıtlanmış makine öğrenmesi yöntemlerindendir. Çalışmada DVM’nin Çok Terimli Kernel (Polinomial Kernel) metodu ile çözüm gerçekleştirilmiş olup, çıkan farklar incelenmiştir.

Destek vektör makinalarında C (ceza katsayısı) ve Epsilon (hata toleransı) değerlerinin seçimi oluşturulan modelin çıktıları üzerinde kritik etkiye sahiptir. Çalışma kapsamında C değeri için {1,2,3} ve epsilon değeri için {0,001,0,0001} olmak üzere toplam 6 kombinasyon için sonuçlar değerlendirmeye alınmıştır. Ulaşılan sonuçlar çizelge 4.2. de sunulmuştur.

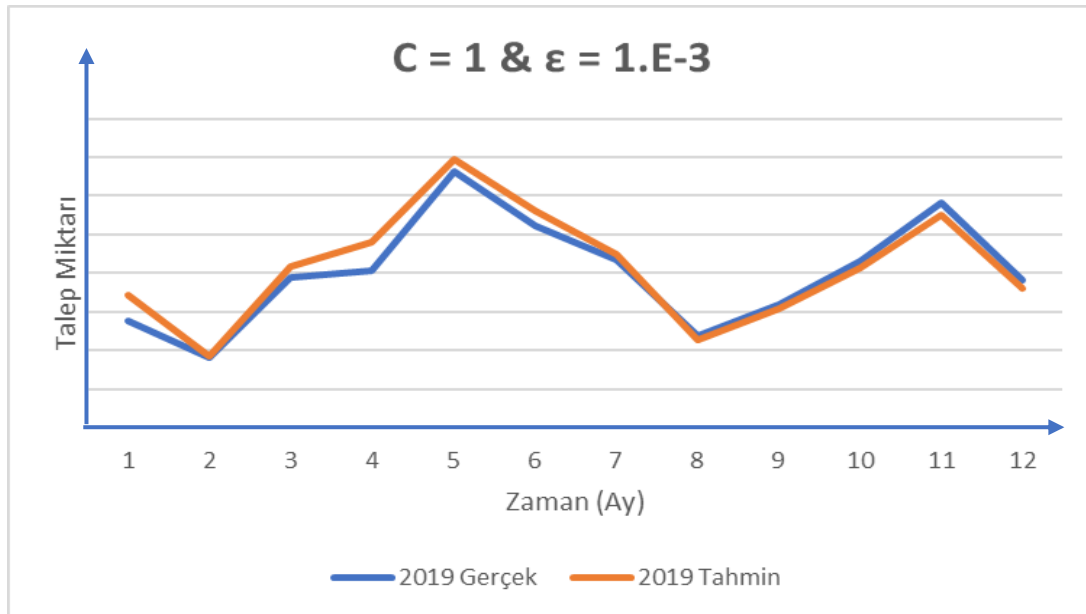
Çizelge 4.2. Destek vektör makinaları ile MAPE değerleri

C \ ε	MAPE	
	0,001	0,0001
1	%7,31	%7,22
2	%13,77	%13,37
3	%12,35	%12,52

Çizelge 4.2. de gösterildiği gibi, C değeri 1 ve epsilon değeri 0,001 için MAPE değeri %7,31, C değeri 1 ve epsilon değeri 0,0001 için MAPE değeri %7,22, C değeri 2 ve epsilon değeri 0,001 için MAPE değeri %13,77, C değeri 2 ve epsilon değeri 0,0001 için MAPE değeri %13,37, C değeri 3 ve epsilon değeri 0,001 için MAPE değeri %12,35 ve C değeri 3 ve epsilon değeri 0,0001 için MAPE değeri %12,52 oranlarında hata tespit edilmiştir.

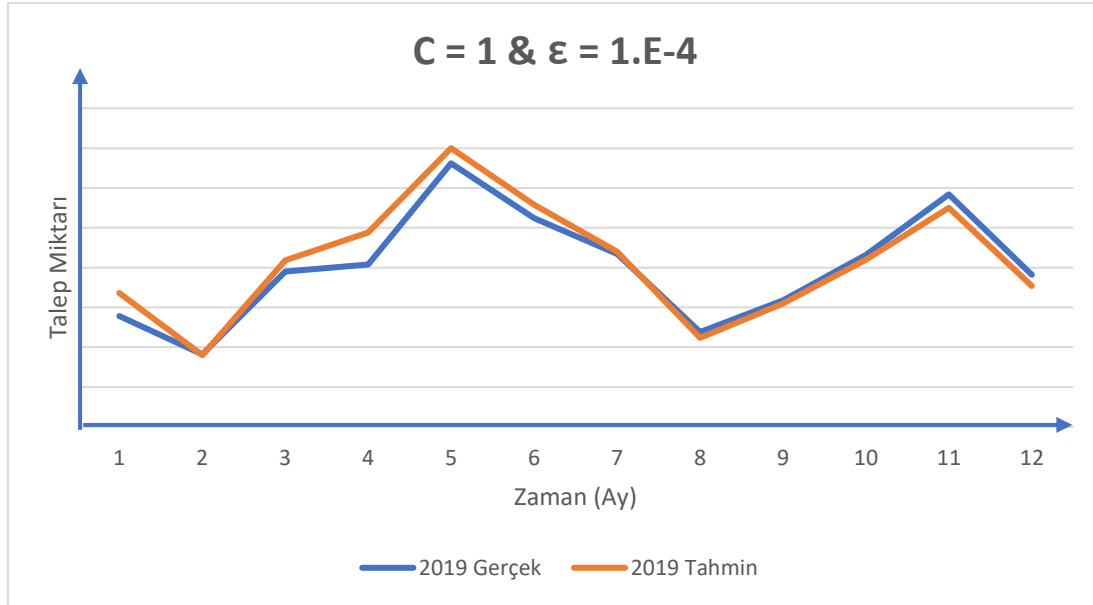
Destek vektör makinalarının çok terimli kernel metodu ile çözümü gerçekleştirilen model çıktıları değerlendirildiğinde yüksek oranda başarı sağladığı görülmektedir. Değerlendirilmeye alınan 6 kombinasyon içerisinde en düşük doğruluk seviyesi %13,77 MAPE değeri ile C değeri 2 ve epsilon değeri 0,001 ile çalıştırılan çözüm olurken, en yüksek doğruluk C değeri 1 ve epsilon değeri 0,0001 iken %7,22 olarak tespit edilmiştir. Çözümü gerçekleştirilen 6 farklı kombinasyona 2019 yılı 12 ayı için gerçekleşen ve tahmin edilen veriler grafikler üzerinden incelenmiştir.

Şekil 4.7.'de C değeri 1 ve epsilon 0,001 parametrelerine göre 2019 yılı 12 aylık gerçekleşen ve tahmin edilen grafiği verilmiştir.



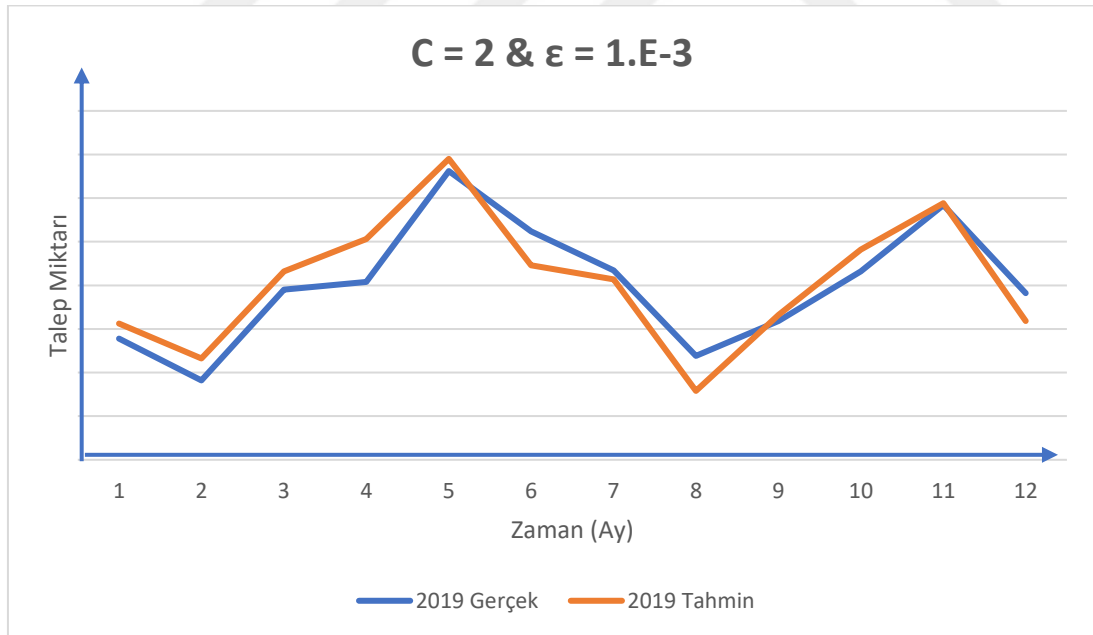
Şekil 4.7. C değeri 1 ve epsilon 0,001'e göre tahmin grafiği

Şekil 4.8.'de C değeri 1 ve epsilon 0,0001 parametrelerine göre 2019 yılı 12 aylık gerçekleşen ve tahmin edilen grafiği verilmiştir.



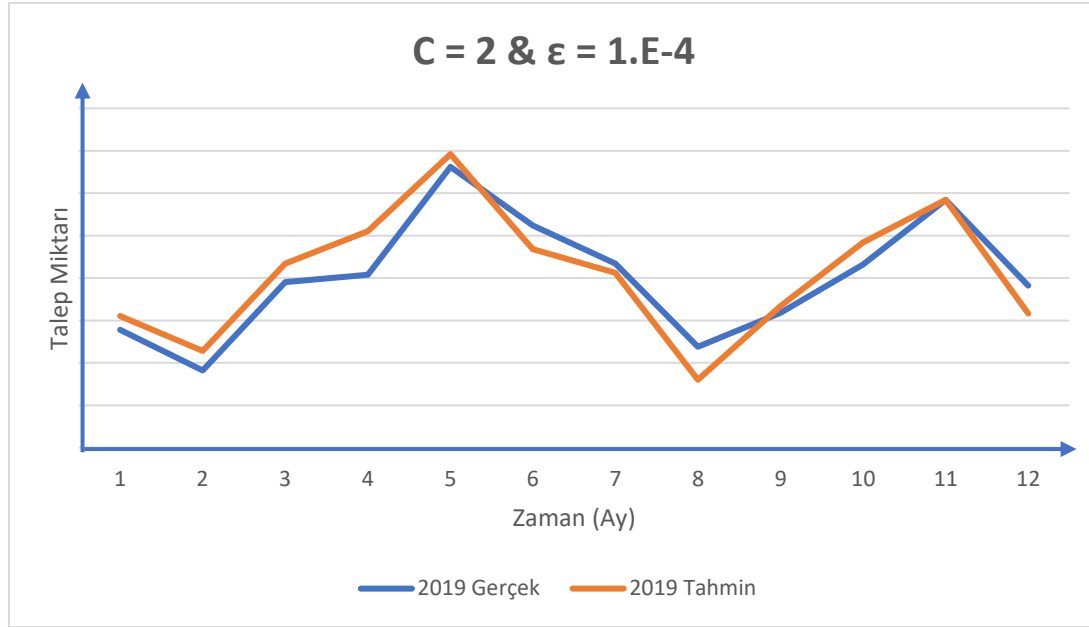
Şekil 4.8. C değeri 1 ve epsilon 0,0001'e göre tahmin grafiği

Şekil 4.9.'de C değeri 2 ve epsilon 0,001 parametrelerine göre 2019 yılı 12 aylık gerçekleşen ve tahmin edilen grafiği verilmiştir.



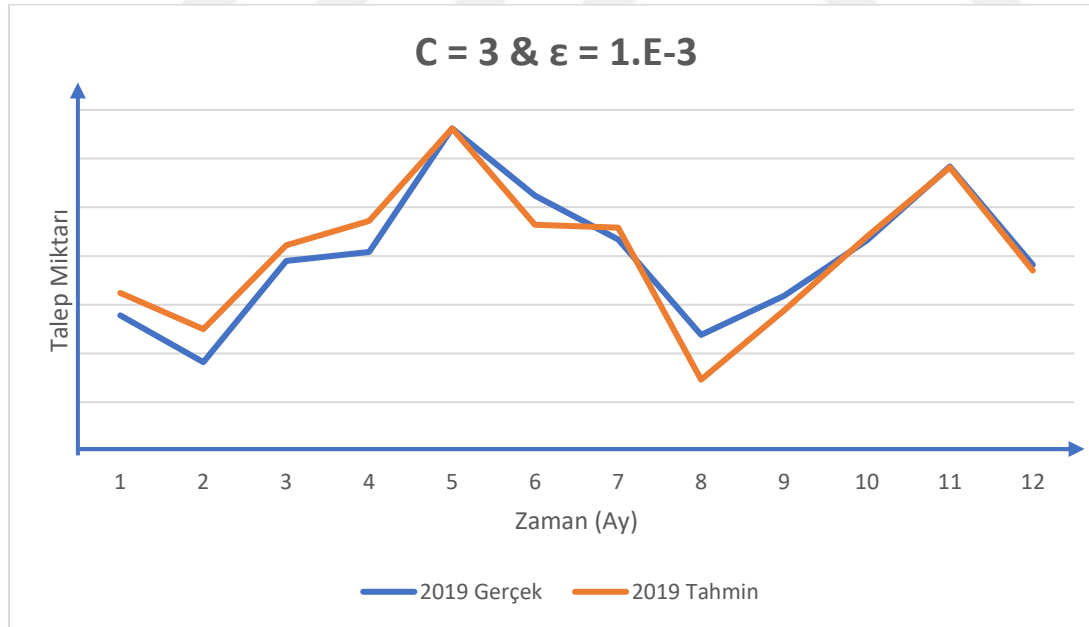
Şekil 4.9. C değeri 2 ve epsilon 0,001'e göre tahmin grafiği

Şekil 4.10.'de C değeri 2 ve epsilon 0,0001 parametrelerine göre 2019 yılı 12 aylık gerçekleşen ve tahmin edilen grafiği verilmiştir.



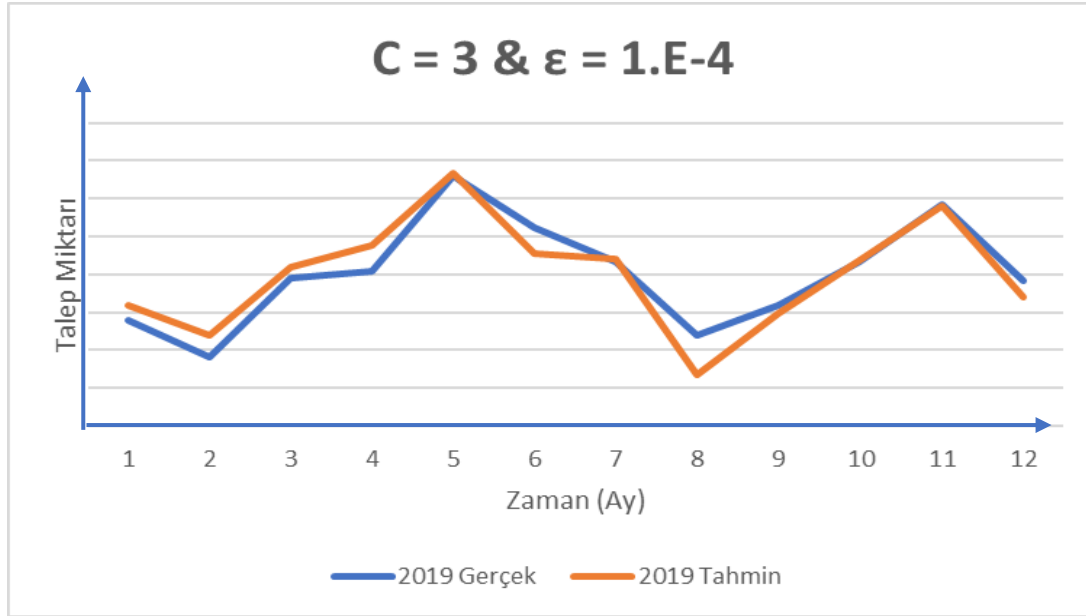
Şekil 4.10. C değeri 2 ve epsilon 0,0001'e göre tahmin grafiği

Şekil 4.11.'de C değeri 3 ve epsilon 0,001 parametrelerine göre 2019 yılı 12 aylık gerçekleşen ve tahmin edilen grafiği verilmiştir.



Şekil 4.11. C değeri 3 ve epsilon 0,001'e göre tahmin grafiği

Şekil 4.12.'de C değeri 3 ve epsilon 0,0001 parametrelerine göre 2019 yılı 12 aylık gerçekleşen ve tahmin edilen grafiği verilmiştir.



Şekil 4.12. C değeri 3 ve epsilon 0,0001'e göre tahmin grafiği

4.3. Destek Vektör Makinaları ve Yapay Sinir Ağlarının Karşılaştırılması

2016- 2018 yılları satış verilerinin kullanılarak 2019 yılı satış verilerini tahmin etmek için incelenen bilimsel çalışmalarda da saptandığı üzere etkin sonuçlara ulaşılabilecek iki ana veri madenciliği yöntemlerinden DVM ve YSA kullanılmıştır. Destek vektör makinalarının çok terimli kernel metodu ve yapay sinir ağlarının çok katmanlı algılayıcılar metodu çözüm için kullanılmıştır. İki metot içinde çalışma sürecinde çok fazla kombinasyon denenmiş olup en doğru sonuca ulaşılabilen 6'şar adet kombinasyon değerlendirmeye alınmıştır. Çıkan sonuçlar çizelge 4.3.de verilmiştir.

Çizelge 4.3. DVM & YSA MAPE değerleri

MAPE HATA DEĞERLERİ			
DVM Parametreleri	DVM MAPE	YSA MAPE	YSA Parametreleri
C =1 & ε = 1.E-3	%7,31	%31,76	GKNS 3-3-3 & EDS 500
C =1 & ε = 1.E-4	%7,22	%35,71	GKNS 3-3-3 & EDS 400
C =2 & ε = 1.E-3	%13,77	%29,26	GKNS 2-4-6 & EDS 500
C =2 & ε = 1.E-4	%13,37	%34,52	GKNS 2-4-6 & EDS 400
C = 3 & ε = 1.E-3	%12,35	%36,51	GKNS 1 & EDS 500
C = 3 & ε = 1.E-4	%12,52	%44,66	GKNS 1 & EDS 400

Çizelge 4.3 incelendiğinde DVM yöntemiyle en düşük MAPE %7,31 olarak tespit edilmiştir. Yine aynı tablodan YSA yöntemiyle en düşük MAPE değerinin %29,26 olduğu sunulmaktadır. Veri madenciliği ile yapılan çalışmalar veri madenciliğinin yapısı gereği optimum sonuç vermemektedir ve literatürde veri madenciliği ile talep tahmini kapsamında incelenen bilimsel araştırmalardan da gözlemlendiği üzere YSA yöntemiyle elde edilen en düşük hata değeri (%29,26) kabul edilebilir sınırlardadır ve stratejik kararlar almada yardımcı olabilecek düzeydedir. Fakat DVM yöntemi ile elde edilen en yüksek %13,77 ve en düşük %7,22 MAPE sonuçları ile YSA yöntemi ile elde edilen en düşük %29,26 MAPE değerleri karşılaştırıldığında DVM yönteminin ilgili veri setiyle çalışıldığında YSA yöntemine göre daha iyi sonuç verdiği ve hem stratejik kararlar da hem de operasyonel kararlarda etkin bir şekilde kullanılabileceği tespit edilmiştir.

5. SONUÇ

Veri madenciliği, günümüz teknolojisinde vazgeçilmez bir bilgi keşfi aracıdır. Büyük verilerden oluşan günümüz dünyasında verilerden faydalı bilgi edinmek amacıyla veri madenciliği yöntemleri etkin bir şekilde kullanılmaya başlanmıştır. Temel anlamda sınıflandırma, kümeleme ve birliktelik üzerine çalışmaya uygun algoritmalar içeren veri madenciliği kavramı farklı sektörlerde farklı problemlere yönelik verilerin analizinde ve verilerden anlamlı bilgi edinilmesinde etkin bir şekilde kullanıldığı görülmektedir.

Orta gerilim ekipmanları üreten bir firmada satış tahmini yapılan bu çalışmada veri madenciliği yöntemlerinden destek vektör makinaları ve yapay sinir ağları kullanılmıştır. Bu yöntemler, doğrusal olmayan veri setlerinde çokça kullanıldıklarından ve etkili sonuç alabildiklerinden ilgili tahmin çalışmasında çözüm metodu olarak seçilmişlerdir. İki yöntem içinde, veri seti hazırlanırken ilgili firmanın 2016 – 2018 yılı satışları eğitim verisi, 2019 yılı satışları test verisi olarak kullanılmıştır. İki yöntemde, satışlar üzerinde etkisi bilinen ‘dolar kuru’, ‘enflasyon oranı’ ve ‘faiz oranı’ nitelik olarak eklenmiştir. İki yöntemde seçilen parametrelerin sonucu etkilediği bilindiğinden birçok parametrik kombinasyon denenmiş olup 6’şar kombinasyonun sonucu değerlendirmeye alınmıştır.

Model çözümleri incelendiğinde, YSA metodu ile en yüksek %44,66 ve en düşük %29,26 hata oranı ile tahmin yapıldığı tespit edilirken, DVM metodu ile en yüksek %13,77 en düşük %7,22 hata oranı ile tahmin yapılmıştır. İki yöntemin sonuçları karşılaştırıldığında en iyi tahmine %7,22 hata oranı ile destek vektör makinaları yöntemleriyle ulaşıldığı görülmüştür. Bu sonuçlara göre çalışma kapsamındaki veri seti ile DVM’nin YSA yöntemine göre çok daha doğru tahmin yaptığı tespit edilmiş olup, etkin bir şekilde kullanılabileceği bulunmuştur.

Bilgi kavramının önemini sürekli arttırdığı günümüzde üst yönetimin ve ara yönetim personellerinin aldığı stratejik ve operasyonel kararlar için çok önemli bir girdi olan talep tahmini konusunda veri madenciliği yöntemlerin önemli ölçüde doğru sonuçlara ulaşılabilirdiği çalışma çıktılarından açıkça görülmektedir. Bu durum bilişim teknolojisinin yönetimler açısından ne denli önem kazandığını ve faydalı olabileceğini göstermektedir.

Çok fazla veriden ‘Nitelikli Bilgi’ edinme sanatı olarak tanımlayabileceğimiz veri madenciliğinin benzer çalışmalarda geliştirilerek kullanılması gerekliliği, çalışmanın bir diğer önemli sonucu olmuştur. Bir sonraki çalışmalarda çalışma kapsamında kullanılan ‘dolar kuru’, ‘faiz oranı’ ve ‘enflasyon oranı’ niteliklerinin başka nitelikler ile güçlendirilerek modele eklenmesi daha etkin sonuçlara ulaştırabilir. Çalışma kapsamında 3 yıllık eğitim verisi kullanılmış olup, eğitim verisinin oranı artırıldığında sonuçların gerçeğe daha çok yaklaşacağı öngörülmektedir. Çalışmada DVM ve YSA ile geliştirilen modellere ek olarak benzer çalışmalarda kullanılacak çok fazla veri madenciliği yöntemi olduğu düşünüldüğünde talep tahmini konusunda diğer yöntemlerinde doğru veri setlerinde çözüm olabileceği ve bu yönlü çalışmaların faydalı olacağı düşünülmektedir.



KAYNAKLAR

1. Klösgen, W., Zytkow, J.M. (2002). *Data mining and knowledge discovery*. Newyork, USA: Oxford University Press. 73-89, 193-198, 205, 254, 258-298, 304-310, 344-348.
2. Ahi, L. (2015). *Veri madenciliği yöntemleri ile ana harcama gruplarının paylarının tahmini*. Yüksek Lisans Tezi, Hacettepe Üniversitesi İstatistik Ana Bilim Dalı, Ankara.
3. Ayhan, S., Erdoğan, Ş. (2014). Destek vektör makineleriyle sınıflandırma problemlerinin çözümü için çekirdek fonksiyonu seçimi. *Eskişehir Osmangazi Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, 9(1), 175- 198
4. Bircan, H., Çam, S. (2016). Veri madenciliğinde kümeleme analizi ve sağlık sektöründe bir uygulaması. *Cumhuriyet Üniversitesi. İktisadi ve İdari Bilimler Dergisi*, 17(2), 85-96.
5. Çakmakoglu A. (2018). *Finans sektöründe veri madenciliği teknikleri kullanılarak kampanya modelleme*. Yüksek Lisans Tezi, İstanbul Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.
6. Ecemiş, O., Irmak, S. (2018). Paslanmaz çelik sektörü satış tahmininde veri madenciliği yöntemlerinin karşılaştırılması. *Sosyal Bilimler Dergisi*, 8(15), 148-169.
7. Han, J., Kamber, M., Pei J. (2012). *Data mining concepts and techniques*. USA: Morgan Kaufmann Publishers. 1-5, 6-8, 10, 18-19, 83-87, 99, 243, 327-393, 398-400, 408-413, 423, 426, 444-448, 450-457.
8. Gencer, C. (2019). *Veri ambarlama ve veri madenciliği. 2018 – 2019 bahar yarıyılı ders notları*. Ankara, Gazi Üniversitesi, Bilişim Enstitüsü.
9. İşlek, İ. (2015). *Talep tahmini için model topluluklarının kullanılması*. Yüksek Lisans Tezi, İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.
10. Karna, H., Vickovic, L., Gotovac, S. (2018). Application of data mining methods for effort estimation of software projects. *Softw: Pract Exper*, 49, 171–191.
11. Le Cam, M., Daoud, A., Zmeureanu, R. (2016). Forecasting electric demand of supply fan using data mining techniques. *Energy*, 101, 541-557.
12. Miao, F., Fu, N., Yuan, T., Zhang, Fellow, IEEE, Ding, X. R., Hong, X., He, Q., Li, Y. (2017). A Novel Continuous Blood Pressure Estimation Approach Based on Data Mining Techniques. *IEEE Journal of Biomedical and Health Informatics*, 21, 6.
13. Odabaş, Ö. (2017). *Veri madenciliği teknikleri ile telekom sektöründe ayrılan müşteri analizi*. Yüksek Lisans Tezi, İstanbul Ticaret Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.
14. Öztürk, Y., Yılmaz, K., Burdurlu, E. (2017). Veri madenciliği ile Türkiye mobilya sektörü için 2023 öngörülleri. *İleri Teknoloji Bilimleri Dergisi*, 757-763.

15. Qiu, Y., Rapinder, S., Chaoyang, Z., Shao, C., Tao, Z., Vahid, G. L., Kaibo, J., Ji, W. (2019). Data mining-based disturbances prediction for job shop scheduling. *Advances in Mechanical Engineering*, 11(3) 1–14.
16. Ramos, M. I., Cubillas, J. J., Feito, F. R. (2015). Improvement of the prediction of drugs demand using spatial data mining tools. *Journal of Medical Systems*, 40, 6.
17. Türkoğlu, B., Kömesli, M., Ünlütürk, M. S. (2019). Application of data mining in failure estimation of cold forging machines: An industrial research. *Studies in Informatics and Control*, 28(1) 87-94.
18. Türk, E., Kiani, F. (2019). Yapay sinir ağları ile talep tahmini yapma: Beyaz eşya üretim planlaması için YSA uygulaması. *İstanbul Sabahattin Zaim Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, 1(1), 30-37.
19. Yang, F., Li, F. W. B. (2018). Study on student performance estimation, student progress analysis, and student potential prediction based on data mining. *Computers & Education*, 123, 97–108.
20. Özcan, İ., Taşar, B., Tatar, B. Yakut, O. (2019). Destek vektör makinası algoritması ile kalp hastalıklarının tahmini. *Anatolian Journal of Computer Science*, 4(2), 74-79.
21. Terzi, M. B., Arıkan, O. (2019). *Koroner Arter hastalığının destek vektör makineleri ve Gauss karışım modeli ile tespiti*. IEEE 2019 Medical Technologies Congress, Bilkent Üniversitesi, Ankara.
22. Açıkkar, M., Sivrikaya, O. (2020). Yıkanmış Türk linyit kömürlerinin üst ısı değeri için destek vektör regresyonu ile tahmini. *Avrupa Bilim ve Teknoloji Dergisi*, 18, 16-24.
23. Karadağ, K. (2020). Makine öğrenme yöntemleri ile semen kalitesi tahmini. *Avrupa Bilim ve Teknoloji Dergisi*, 18, 306-311.
24. Köse, U. (2019). Zeki optimizasyon tabanlı destek vektör makineleri ile diyabet teşhisi. *Politeknik Dergisi*, 22, 557-566.



EK-1. 2016 – 2019 yılları nitelikler veri tablosu

Yıl	Ay	Dolar Kuru	Faiz Oranı	Enflasyon Oranı
2016	1	3.01	11.10	9.58
	2	2.94	11.51	8.78
	3	2.89	11.65	7.46
	4	2.83	11.63	6.57
	5	2.93	11.57	6.58
	6	2.92	11.43	7.64
	7	2.96	11.13	8.79
	8	2.96	10.91	8.05
	9	2.96	10.57	7.28
	10	3.07	10.21	7.16
	11	3.27	10.17	7.00
	12	3.49	10.33	8.53
2017	1	3.73	10.34	9.22
	2	3.67	10.42	10.13
	3	3.67	10.71	11.29
	4	3.65	11.12	11.87
	5	3.56	11.71	11.72
	6	3.52	12.39	10.90
	7	3.56	12.90	9.79
	8	3.51	13.21	10.68
	9	3.47	13.37	11.20
	10	3.66	13.50	11.90
	11	3.88	13.53	12.98
	12	3.85	13.53	11.92
2018	1	3.77	13.62	10.35
	2	3.78	13.67	10.26
	3	3.88	13.70	10.23
	4	4.05	13.80	10.85
	5	4.41	13.91	12.15
	6	4.63	14.86	15.39
	7	4.75	15.19	15.85
	8	5.73	15.65	17.90
	9	6.37	16.77	24.52
	10	5.86	18.99	25.24
	11	5.37	21.34	21.62
	12	5.31	22.31	20.30
2019	1	5.37	23.24	20.35
	2	5.26	23.73	19.67
	3	5.44	23.77	19.71
	4	5.74	22.96	19.50
	5	6.05	22.23	18.71
	6	5.81	21.74	15.72
	7	5.67	21.42	16.65
	8	5.62	21.10	15.01
	9	5.71	20.13	9.26
	10	5.78	18.48	8.55
	11	5.73	17.42	10.56
	12	5.84	14.56	11.84

ÖZGEÇMİŞ

Kişisel Bilgiler

Soyadı, adı : ARSLAN, Oktay Mesut
 Uyruğu : T.C.
 Doğum tarihi ve yeri :
 Medeni hali :
 Telefon :
 e-mail :

Eğitim

Derece	Eğitim Birimi	Mezuniyet Tarihi
Yüksek lisans	Gazi Üniversitesi / Yönetim Bilişim Sistemleri	Devam ediyor
Lisans	Çankaya Üniversitesi / Endüstri Mühendisliği	2014
Lise	Erzincan Fen Lisesi	2008

İş Deneyimi

Yıl	Yer	Görev
2019-Halen	Ulusoy Elektrik	Kıdemli Değişiklik Yönetimi Mühendisi
2018-2019	KSB Pompa	Üretim ve Malzeme Planlama Mühendisi
2017-2018	Tübitak	Bilimsel Prog. Uzm. Yrd.
2015-2017	Ulusoy Elektrik	Üretim ve Malzeme Planlama Mühendisi

Yabancı Dil

İngilizce

Yayınlar

Arslan, O.M., Özyörük, B. (2020). *Veri Madenciliği ile Talep Tahmini ve Bir İşletmede Uygulama*. 7. Uluslararası Bilimsel Araştırmalar Kongresi, Ankara,

Hobiler

Seyahat Etmek, Sinema, Kitap Okumak



GAZİLİ OLMAK AYRICALIKTIR.