KIRIKKALE ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

ENDÜSTRİ MÜHENDİSLİĞİ ANABİLİM DALI
YÜKSEK LİSANS TEZİ

VERİ MADENCİLİĞİ SINIFLANDIRMA TEKNİKLERİNİ KULLANARAK ÜRETİM SİSTEMLERİNDE HATALI ÜRÜNLERİN TESPİT EDİLMESİ

MUHAMMED FATİH ŞİMŞEK

AĞUSTOS 2019

Endüstri Mühendisliği Anabilim Dalında Muhammed Fatih ŞİMŞEK tarafından hazırlanan VERİ MADENCİLİĞİ SINIFLANDIRMA TEKNİKLERİNİ KULLANARAK ÜRETİM SİSTEMLERİNDE HATALI ÜRÜNLERİN TESPİT EDİLMESİ adlı Yüksek Lisans Tezinin Anabilim Dalı standartlarına uygun olduğunu onaylarım.

Prof. Dr. Süleyman ERSÖZ
Anabilim Dalı Başkanı

Bu tezi okuduğumu ve tezin **Yüksek Lisans Tezi** olarak bütün gereklilikleri yerine getirdiğini onaylarım.

Dr. Öğr. Üyesi Adnan AKTEPE
Danışman

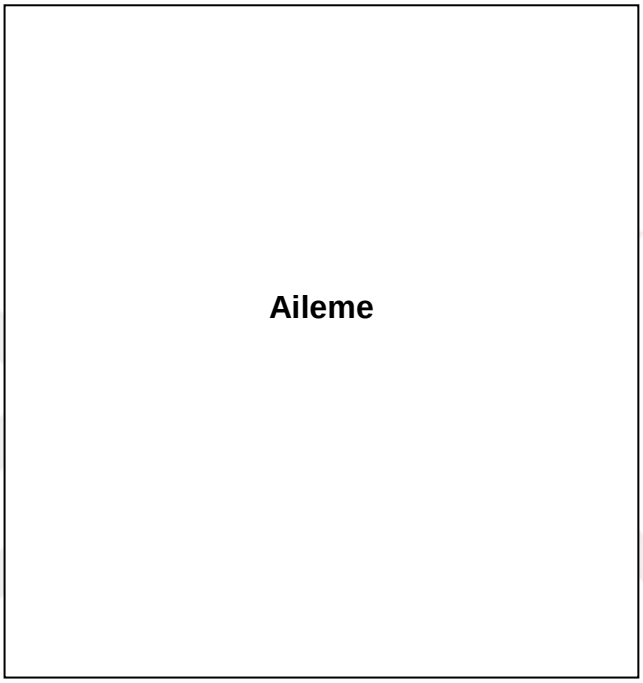
Jüri Üyeleri

Başkan	: Prof. Dr. Süleyman ERSÖZ	_____
Üye (Danışman)	: Dr. Öğr. Üyesi Adnan AKTEPE	_____
Üye	: Dr. Öğr. Üyesi Ahmet YÜCEL	_____

……/…../…….

Bu tez ile Kırıkkale Üniversitesi Fen Bilimleri Enstitüsü Yönetim Kurulu Yüksek Lisans derecesini onaylamıştır.

Prof. Dr. Recep ÇALIN
Fen Bilimleri Enstitüsü Müdürü


Aileme

ÖZET

VERİ MADENCİLİĞİ SINIFLANDIRMA TEKNİKLERİNİ KULLANARAK ÜRETİM SİSTEMLERİNDE HATALI ÜRÜNLERİN TESPİT EDİLMESİ

ŞİMŞEK, Muhammed Fatih
Kırıkkale Üniversitesi
Fen Bilimleri Enstitüsü
Endüstri Mühendisliği Anabilim Dalı, Yüksek Lisans Tezi
Danışman: Dr. Öğr. Üyesi Adnan AKTEPE
Ağustos 2019, 88 sayfa

Günümüzde yaşanan teknolojik gelişmeler sayesinde, hemen her alanda çeşitli veriler kolaylıkla elde edilebilmekte ve depolanabilmektedir. Fakat elde edilen bu veriler çoğu zaman ham olarak bir anlam ifade etmemektedir. Verilerin kullanılabilir hale gelmeleri için çeşitli süreçlerden geçmeleri gerekmektedir. Veri madenciliği teknikleri sayesinde ham verilerden kullanılabilir bilgiler elde edilmektedir.

Birçok alanda veri madenciliği tekniklerinden yoğun bir biçimde yararlanılmaktadır. Fakat veri madenciliği tekniklerinin üretim alanında kullanımı diğer alanlara kıyasla daha azdır. Bu çalışma ile, üretim alanında az sayıda olan veri madenciliği uygulamalarına bir örnek daha eklenmiştir.

Bu çalışmada, bir üretim atölyesinden elde edilen fiili üretim verileri WEKA yazılımı üzerinde bulunan çeşitli sınıflandırma algoritmalarında kullanılmıştır. Algoritmalardan elde edilen sonuçlar ile ürünlerin hatalı ya da sağlam olmasına etki eden koşullar için STATİSTİKA yazılımı kullanılarak karar ağaçları ve kurallar tespit edilmiştir.

Anahtar kelimeler: Veri Madenciliği, Üretim Verileri, Sınıflandırma

ABSTRACT

DETERMINING FAILED PRODUCTS IN PRODUCTION SYSTEMS USING DATA MINING CLASSIFICATION TECHNIQUES

ŞİMŞEK, Muhammed Fatih
Kırıkkale University
Graduate School of Natural and Applied Sciences
Department of Industrial Engineering, Master Science Thesis
Supervisor: Asst. Prof. Dr. Adnan AKTEPE
August 2019, 88 Pages

Today, thanks to the technological developments, miscellaneous data sets from many different domains can easily be obtained and stored. However, these data sets are mostly in raw format and do not provide so much information. The data should go through various processes to make them ready for information extraction. Data mining techniques enable us to obtain the information from raw data sets.

Data mining techniques are used extensively in many areas. However, the use of data mining techniques in the production area is less compared to other fields. With this study, a new example of data mining applications in the production area was added.

In this study, a real production data set from a production workshop were used in various classification algorithms on WEKA software. Using STATISTIKA software, decision trees and rules have been determined for the conditions that affect the results of the algorithms and the faulty or robust products.

Key Words: Data Mining, Production Data, Classification

TEŞEKKÜR

Tez konumun belirlenmesinden başlayarak tez yazım sürecinde yönlendirmeleriyle bana destek olan değerli tez danışmanım Sayın Dr. Öğr. Üyesi Adnan AKTEPE'ye ve değerli hocam Sayın Prof. Dr. Süleyman ERSÖZ'e; ihtiyaç duyduğum zamanlarda akademik tecrübesini ve hayat tecrübesini paylaşmaktan çekinmeyen değerli hocam Sayın Dr. Öğr. Üyesi Ahmet YÜCEL'e; son olarak da tüm hayatım boyunca hiçbir fedakarlıktan kaçınmayan, yardımlarını ve desteklerini her konuda yanımda hissettiğim, benim için her şeyden daha kıymetli olan aileme sonsuz teşekkür ederim.

İÇİNDEKİLER DİZİNİ

Sayfa

ŞEKİLLER DİZİNİ

ŞEKİL Sayfa

ÇİZELGELER DİZİNİ

ÇİZELGE Sayfa

SİMGELER VE KISALTMALAR DİZİNİ

SİMGELER DİZİNİ

e	Hata Terimi
$P(c \mid x)$	Koşullu Olasılık

KISALTMALAR DİZİNİ

VABK	Veri Ambarlarından Bilgi Keşfi
CRISP-DM	Cross-Industry Standard Process for Data Mining
SEMMA	Sample, Explore, Modify, Model, Assess

1. GİRİŞ

Günümüzde teknolojide yaşanan hızlı gelişmeler sayesinde artık hemen her alanda veri elde etmek ve bu verileri depolamak oldukça kolay ve ucuz bir şekilde yapılabilmektedir. Fakat bu veri yığınları işlenerek anlamlı bilgilere dönüşmedikleri takdirde bir anlam ifade etmemekte, kullanılabilmeleri için içlerinden faydalı anlamlı bilgilerin elde edilmesi gerekmektedir. Bu ham anlamsız verileri işleyip anlamlı bilgiler çıkarma sürecine veri madenciliği ismi verilmektedir.

Veriye ulaşmanın ve depolamanın kolaylaşması sayesinde hemen her alanda veri madenciliği uygulamaları yapılmaktadır. Veri madenciliği ile ilgili yapılan çalışmalar incelendiğinde, çalışmaların büyük çoğunlunun hizmet ve servis sistemlerinin verileri kullanılarak yapılmış olduğu görülmektedir. Hizmet ve servis sistemlerine kıyas edildiğinde, üretim sistemlerinin verileri kullanılarak yapılan çalışmaların sayısının oldukça az olduğu görülmektedir. Bu nedenden dolayı bu çalışmada veri madenciliği yöntemleri ile bir üretim sisteminin verilerini kullanılmış ve az sayıda olan çalışmalar arasına bir yenisinin eklenmesi ve literatüre bu alanda katkı sunulması amaçlanmıştır.

Bu çalışmada üretim sistemlerinde veri madenciliği ile ilgili yapılan yayınlar incelenmiş, veri madenciliği ve kullanılan yöntemler hakkında bilgiler verilmiş ve son olarak da bir üretim işletmesinin atölyesinden elde edilmiş fiili üretim verileri kullanılarak veri madenciliği uygulaması yapılmıştır. Çalışma sekiz ana bölümden oluşmaktadır ve bölümlerde bahsedilen çeşitli konular hakkında bilgiler aşağıda belirtildiği şekildedir.

Çalışmanın giriş bölümünü izleyen ikinci bölümünde öncelikle bir literatür taraması yapılarak üretim sistemlerinde veri madenciliği ile ilgili yayınlanmış olan çalışmaların özeti sunulmuştur.

Çalışmanın üçüncü bölümünde veri ambarlarından bilginin keşfedilme süreci ve veri madenciliği kavramları açıklanmıştır. Veri madenciliğinin tarihsel gelişim sürecinden ve günümüzde kullanıldığı alanlardan bahsedilmiştir. Veriden bilgiye

ulaşma sürecinde literatürde yaygın olarak kullanılan yaklaşımlar hakkında bilgi verilmiştir.

Dördüncü bölümde veri ambarlarında anlamsız halde duran verilerin anlamlı bilgiler haline gelinceye kadar gerçekleştirilen süreçler ve bu süreçlerin teknik detayları anlatılmıştır.

Beşinci bölümde tahmin edici ve tanımlayıcı metotlar açıklanarak, denetimli ve denetimsiz öğrenme kavramları hakkında bilgiler verilmiştir. Farklı veri madenciliği metotları tanıtılarak sınıflandırma, regresyon, kümeleme ve birliktelik analizi gibi veri madenciliği metotlarından bahsedilmiştir.

Altıncı bölümde bu çalışmanın da konusu olan veri madenciliği metotlarından sınıflandırma metodu uygulanırken yaygın olarak kullanılan yöntemler ayrıntılı bir şekilde incelenmiştir. Karar ağaçları, Bayes Sınıflandırma Teknikleri, K-En Yakın Komşu Algoritması, Yapay Sinir Ağları, Genetik Algoritmalar ve Destek Vektör Makineleri gibi sınıflandırma işlemlerinde yaygın olarak kullanılan teknikler hakkında bilgiler verilmiştir.

Yedinci bölümde, önceki bölümlerde bahsedilen veri madenciliği sınıflandırma teknikleri kullanılarak fiili üretim verileriyle bir uygulama yapılmıştır. Fiili üretim verileri çeşitli sınıflandırma algoritmalarında kullanılmış ve algoritmaların başarı oranları karşılaştırılmıştır.

Sekizinci bölümde, başarı oranı en yüksek çıkan algoritma ile bir karar ağacı oluşturularak ürünlerin sağlam veya hatalı olma durumlarına etki eden değişken değerleri belirlenmiş ve karar kuralları oluşturulmuştur. Böylelikle sadece tespit edilen değişkenlerin değerlerine bakarak ürünlerin hatalı olup olmayacağı tahmin edilmiştir.

2. LİTERATÜR ARAŞTIRMASI

Veri madenciliği günümüzde hemen her alanda yoğun şekilde kullanılmaktadır. Fakat yapılan çalışmalar incelendiğinde çalışmaların daha çok hizmet sektörü üzerinde yoğunlaştığı, üretim sektöründe yapılan çalışmaların hizmet sektörüne kıyasla oldukça az olduğu görülmektedir. Bu çalışmanın konusunu ise fiili atölye ortamının verileri oluşturmaktadır. Bu sebeple bu bölümde literatürde yer alan üretim sistemlerinin verilerini kullanarak yapılan çalışmalar derlenmiştir.

Özden ve Chen (1999), plastik sanayi için enjeksiyon kalıplama işleminde ürün ve süreç tasarımı, malzeme ve işlemeye yönelik konular, süreci izleme ve ürün sağlamlığını tanımlama gibi çoklu kalite özellikleri için sinir ağ modeli örneği sunmuşlardır. Üç girdisi, bir gizli katmanı ve beş çıkışı olan bir sinir ağı modeli ile aynı anda beş kritik kalite değişkenini yüksek doğrulukla eş zamanlı olarak modellemişlerdir.

Mieno vd. (1999), yarı iletken üretimi sürecinde veri madenciliği tekniklerini uygulamışlardır. Regresyon ağacı analizi ile proses üzerinde bile tespit edilmesi zor olan arıza nedenlerini önceden tespit etmişlerdir ve doğrulama yöntemleri ile hata nedeninin doğruluğunu kanıtlamışlardır.

Skinner vd. (2002), levha üretiminde veri madenciliği tekniklerini uygulayarak levhaların kalitesi ve verimi üzerindeki etkileri incelemişlerdir. Düşük verimli ürünlerin nedenlerini belirleyerek gelecekte üretimi artırabilecek bilgileri elde etmeye çalışmışlardır. Kümeleme ve temel bileşenler yöntemini bir sınıflandırma ve regresyon ağacı (CART) yöntemiyle karşılaştırmışlardır. Karşılaştırma sonucu CART yönteminin diğer yöntemlerden daha kullanışlı olduğunu ifade etmişlerdir.

Li vd. (2003), cam imalatında üretim sürecini iyileştirmek için sinir ağı ve CART methodu kullanarak makine ölçümleriyle, kalite ölçümleri arasındaki ilişkiyi modellemeye çalışmışlardır. Her iki yöntemle sundukları modelde kalite spesifikasyon tolerans aralığında bir hataya ulaşmışlardır.

Kowalski ve Kowalska (2003), endüksiyon motorlarda rotor, stator ve rulman hatalarının teşhisi için yapay sinir ağlarını kullanmışlardır. Sinir ağlarını stator akımı ve mekanik titreşim spektrumlarının ölçüm verilerini kullanılarak eğitmiş ve test etmişlerdir. Yapay sinir ağlarının tüm hata türlerinin tespiti için tatmin edici sonuçlar verdiğini öne sürmüşlerdir.

Baykasoğlu (2005), bir çeşit kompozit çimentonun 28 gün sonundaki oluşacak basma mukavemetini önceden tahmin etmek için gen denklem programlama, yapay sinir ağları ve regresyon analizi yöntemlerini kullanmış ve bu yöntemlerin karşılaştırmalarını yapmıştır.

Çoban (2006), üretim sektöründe faaliyet gösteren bir işletmenin gerçek verilerini kullanarak işletmenin tedarikçileri ile olan ilişkilerine etki edecek anlamlı sonuçlar elde etmiştir.

Chen vd. (2006), LCD sürücü IC paketleme fabrikasında ürünlerin yeterlilik yüzdesini arttırmak ve ürünün kalite problemlerine bir analiz sistemi oluşturmak için veri madenciliği yöntemlerinden karar ağaçlarını kullanılmışlardır. Karar ağaçlarının sonuçlarını daha önce kullanılmış olan sinir ağları ile kıyaslamışlardır. Araştırma sonuçlarının, karar ağacı algoritmasının kullanımının kalite problemi sınıflandırma ve LCD sürücü IC ambalaj endüstrisinin analizinde sinir ağlarından daha uygun olduğunu ifade etmişlerdir.

Dal vd. (2006), makinelerin titreşim standartları ile ilgili verileri Yapay Sinir Ağları modelinde eğitim seti olarak kullanmışlardır. Gizli katman sayısı 5, 10, 15, 25, 50 ve 75 olan farklı ağları karşılaştırıp en iyi çözümü veren ağı bulmaya çalışmışlardır. Modellerin eğitilmesi sonucu elde ettikleri çıktı değerlerini, tablodaki gerçek değerlerle karşılaştırdıklarında, eğittikleri modelin titreşim analizi için kullanılabileceğini tespit etmişlerdir. Çalışmanın farklı standartlar üzerinde uygulanabileceğini ve önceden bulunmuş ölçüm bilgileri kullanılarak kestirimci bakım programları oluşturulabileceğini öne sürmüşlerdir.

Kayaalp (2007), endüstride yaygın olarak kullanılan asenkron motorlarda hataların önceden tespit edilebilmesi için sınıflandırma tekniklerinden karar ağacı algoritmalarını kullanmıştır. Motorlarda meydana gelebilecek kısa devre, yalıtım bozuklukları veya motor milinde oluşabilecek mekanik dengesizliklerin tespitini amaçlamıştır. WEKA yazılımının sınıflandırma algoritmalarını kullanmış ve RepTree karar ağacının ürettiği kuralların geçerliliğini ispatlamıştır.

Bakır vd. (2008), bir döküm şirketi tarafından üretilen ürünlerde kusura neden olan en etkili değişkenleri belirlemek için yaptıkları çalışmada, işlem parametreleri ve hata tipleri arasındaki ilişkiyi modellemek için regresyon analizini ve karar ağaçlarını kullanmışlardır. Lojistik regresyon modelleri başarısız olmasına rağmen, karar ağacı yaklaşımı tatmin edici sonuçlar vermiştir. Sonuçları şirketin kalite ekibine sunduklarında, modeldeki bazı parametrelerin ve eşik değerlerinin anlamlı bulunduğunu, bazılarının ise beklenmeyen ilginç bulunduğunu belirtmişlerdir.

Çelik (2009), bir otomotiv yan sanayi firmasının kesim bölümünde hataların en aza indirgenmesi amacıyla hata tanımlamalarını yapıp Ana Bileşenler Analizi, Kanonik Korelasyon Analizi, Çoklu Regresyon Analizi yöntemlerini kullanmış ve mevcut durumla karşılaştırmıştır. En ve boy değerleri için istenilen sonuçlara yüzde yüz ulaştığını ve delik çapı için ise yüzde elli beş oranında bir iyileşme sağladığını ifade etmiştir.

Çetin (2009), bir üretim işletmesinde ürünlerin uygunsuz olarak ayrılmasının nedenlerini analiz etmek için SPSS Clementin 11.1 yazılımını kullanarak karar ağaçları ve yapay sinir ağları ile bir model geliştirerek uygunsuz ürünlerin sayısını azaltmaya çalışmıştır.

Gürbüz vd. (2009), Türkiye'de faaliyet göstermekte olan bir hava yolu işletmesinin parça söküm raporları verileri üzerinde bir veri madenciliği çalışması yapmışlardır. Yaptıkları çalışma ile uçaklarda kullanılan parçaların, düzeltici ve önleyici faaliyetlerin yapılabilmesi için, parça üzerinde herhangi bir arıza oluşmadan önce uygun ikaz seviyelerini tespit etmeyi amaçlamışlardır. Geliştirdikleri kuralları

doğrulukları ve güvenilirlikleri ile test ederek anlamlı kurallar oluşturduklarını ifade etmişlerdir.

Köksal vd. (2009), yaptıkları TÜBİTAK projesinde sanayi kuruluşlarında ürün ve süreçlerin kalitesini artırmaya yönelik kullanılabilecek veri madenciliği yaklaşımları belirlemeye ve yeni yaklaşımlar sunmaya çalışmışlardır. Kalitenin tanımlanması, tahmin edilmesi, sınıflandırılması ve parametrelerinin optimizasyonu problemlerini ele almışlardır ve bu problemlerin çözümü için veri hazırlama ve ön işlemenin yanısıra kümeleme, tahmin etme, sınıflandırma, birliktelik analizi ve optimizasyon gibi veri madenciliği işlevlerinin gerekli olabileceğini belirlemişlerdir. Ayrıca kalite iyileştirme alanında etkinliği artırabilecek çeşitli yöntem ve metotlar geliştirmişlerdir.

Kahya Özyirmidokuz (2009), halı imalat sürecini ve ürün kalitesini iyileştirmek amaçlı veri madenciliği uygulamaları yapmıştır. Halı hatalarının ve makine duruşlarının sebeplerini bilgi kazancı tekniği ile bir standarda kavuşturmuştur. Modelleme süreci sonunda, karar ağaçları ve sinir ağları model çıktıları elde etmiştir.

Bilekdemir (2010), sınıflandırma tekniklerinden karar ağaçları ile bir su sayacı üretim tesisinde makinelerin üretim sürelerini tahmin etmeye çalışmıştır. Karar ağaçları ile sürekli değişkenlerin tahmini yapılamadığından dolayı, bu durumlarda yapay sinir ağları, genetik algoritmalar gibi veri madenciliği tekniklerinin kullanılabileceğini öne sürmüştür.

Gu vd. (2011), bir çeşit karbon çeliğinin yüzey sertleştirme işlemini etkileyen nitelikleri belirlemek amacıyla yaptıkları deney sonuçlarını kullanarak genetik algoritma geri yayılımlı sinir ağı (GA-BP) kullanılarak tahmin ve optimizasyon modeli geliştirilmişlerdir. Sertleşen tabakaların özellikleri ve işlem parametreleri arasındaki doğrusal olmayan bir ilişki tespit etmişlerdir. Tahmin sonuçlarının ölçülen sonuçlarla oldukça iyi bir uyum içinde olmasından dolayı GA-BP tahmin modelinin güvenilir olduğunu belirtmişlerdir.

Abhang ve Hameedullah (2012), bir tür çelik alaşımın yüzey pürüzlülüğü parametlerinin tahmin modelinin geliştirilmesi için çoklu regresyon ve yapay sinir ağlarını kullanmışlardır. Sonuçlara göre, geri yayılım modeline sahip yapay sinir ağı modelinin çoklu regresyon modellerine kıyasla yüksek doğrulukta olduğunu belirlemişlerdir.

Yalçın Pirinççiler ve Şen (2012), sürekli iyileştirme ve bu amacı benimseyen Altı Sigma metodolojisi teknikleri için kullanılan DMAIC döngüsündeki aşamalarda veri madenciliği tekniklerine yer vermişlerdir. Verilerin toplanması, ölçülmesi ve analiz edilmesi sürecinde veri madenciliği tekniklerinden faydalanmışlardır.

Ordu (2013), demir-çelik sektöründe veri madenciliği tekniklerini kullanarak geçmiş yıllardaki üretim verileri ile uzun ürünler olarak nitelendirilen ürün grubu üzerinde üretime ilişkin değişkenleri incelemiş ve veri madenciliğinde sınıflandırma temelli teknikler ile analizler yaparak üretim miktarına ilişkin tahmin modeli ve üretimi etkileyen en önemli değişkenleri tespit etmiştir.

Öncel Çekim ve Karasoy (2013), karar ağacı yöntemi ile elde ettikleri düğümleri Cox regresyon modeline adım fonksiyonu olarak ekleyerek, Cox regresyon yöntemi ve karar ağaçları yöntemleriyle karma bir model oluşturmuşlardır. Oluşturdukları modeli kömür işletmesinde kullanılan makinelere ait veriler üzerinde kullanarak makinelerdeki lastik ömürlerinin üzerinde etkili olan değişkenleri belirlemeye çalışmışlardır. Oluşturdukları karma modelin klasik Cox regresyonundan daha iyi sonuçlar verdiğini belirtmişlerdir.

Yıldız (2014), yapmış olduğu doktora tezi çalışmasında kumaş hata denetimi ile ilgili iki farklı çalışma gerçekleştirmiştir. İlk çalışmasında hatalı kumaş görüntülerine, görüntü işleme algoritmaları uygulamıştır. Hatalı alanın tespitini yaptıktan sonra şekilsel ve histogram özelliklerini çıkarmış, Fuzzy C-Means Algoritması ile yapılan uygulamada ortalama %87 kümeleme başarısı elde etmiştir. İkinci çalışmasında ise 3 farklı kumaş türü üzerinde farklı hatalara sahip olan video üzerinde anlık hata tespiti, sınıflandırma ve yer tespiti gerçekleştirmiş ve K En Yakın Komşu, Bayes Ağları, Karar Ağaçları kullanarak sınıflandırma yapmış ve K En Yakın Komşu Algoritması

kullanarak yaptığı sınıflandırmada diğer yöntemlere göre daha iyi sonuçlar elde etmiştir.

Erden ve Nazarov (2017), tekstil endüstrisinde ipliklerin sürtünme özelliklerini Kaba kümeler teorisi kullanarak analiz etmişlerdir. Gerçek veri setleri ile çalışarak dört özellik sınıfı arasından hammadde özelliğinin, iplik sürtünmesine etki eden en önemli özellik olduğunu tespit etmişlerdir. Tekstil endüstrisinde kullanılan verilerin belirsizlik içermesi nedeniyle, az ve belirsiz veriler ile çalışabilen Kaba kümeler teorisi kullanılarak verilerin analizinde ve yorumlanmasında karar vericilerin işinin kolaylaşacağını söylemişlerdir.

Şeker ve Yüksek (2017), bir makine öğrenme metodu olan derin öğrenmeyi görüntü işleme ile teknikleri ile birlikte kumaş hatalarının tespiti için kullanmışlardır. Derin öğrenme modelinin hiper parametlerinin ince ayarları ile oynamalar yaparak öznitelik çıkarımı başarısını artırmayı amaçlamışlardır. Kendi veri setleri üzerinde %96'lık bir başarı oranı sağladıklarını belirtmişlerdir.

Tapkan ve Özmen (2018), iplik üreten bir tesiste nitelik seçimi ve sınıflandırma ile iplik kalitesini belirlemişlerdir. Önce Taguchi deneysel tasarım yöntemi kullanılarak iplik kalitesini etkileyen etkin nitelikler tespit edilmiştir. Daha sonra ise veri madenciliği yöntemlerinden sınıflandırma yöntemiyle maliyete-duyarlı ve maliyete-duyarsız olarak iki farklı şekilde kural çıkarımları yapmışlardır.

Karadağ (2018), yaptığı tez çalışmasında veri madenciliği teknikleri ile bir ambalaj firmasının üretim sürecinde çıkan fire miktarını tahmin etmeyi hedeflemiştir. Üretim değişkenlerinin fire üzerindeki etkisini belirlemek için on farklı versiyon oluşturmuş ve farklı algoritmalar ile denemeler yapmıştır. Denemeleri sonunda hangi üretim girdilerinin farklı versiyonlarda çıktıları ne derece etkilediğini tespit etmeye çalışmıştır.

Türkoğlu vd. (2018), soğuk dövme makinelerinin duraksama sayısının azaltılmasına yönelik yaptıkları çalışmada veri madenciliği tekniklerini kullanmışlardır. Soğuk

dövme makinelerinden gelen bilgileri veri madenciliği algoritmalarında kullanarak analizler sonucunda umut verici sonuçlar bulduklarını ifade etmişlerdir.

Canlı ve Toklu (2019), üretim verileriyle otomotiv sanayisinde kalite kontrol sürecinde veri madenciliği sınıflandırma algoritmalarını kullanarak bir karar destek sistemi oluşturmayı hedeflemişlerdir. Sınıflandırma algoritmalarından C4.5, Rastgele Orman, Sıralı Minimal Optimizasyon ve Naive Bayes algoritmalarını kullanarak modelleri farklı oranlarda hold-out ve cross-validation performans yöntemleriyle değerlendirmişlerdir. En iyi performans gösteren algoritma olarak C4.5 algoritmasını bulmuşlardır. Algoritmaların, işlem tamamlanmadan önce ürünün arızalı olduğunu tespit ederek kalite analizini çok hızlı ve kolay hale getirdiğini ve bu sayede işçilik ve malzeme maliyetinin azaldığını belirlemişlerdir.

Tunçkaya (2019), fosil yakıtlı bir enerji santralinde kritik operasyon parametrelerinden 19 adetini seçerek Yapay Sinir Ağlarıyla (YSA) santralin modellemesini gerçekleştirmiştir. Prosesin en kritik çıkış değişkenlerinden biri olan ve elektrik üretim miktarını önemli bir şekilde etkileyen ana buhar basıncı değerinin tahmin edildiği bir kestirim çalışması sunmuştur. Elde ettiği sonuçları, istatistiksel kestirim yöntemlerinden Çoklu Doğrusal Regresyon yöntemi ile karşılaştırmıştır. Model çıktılarının performansları, kök ortalama karesel hata ve determinasyon katsayısı yaklaşımları ile karşılaştırmıştır. Yapay Sinir Ağı modelinden elde ettiği değerlerin, Çoklu Doğrusal Regresyon modelinden bulduğu değerlere göre daha başarılı olduğunu tespit etmiştir. En büyük mutlak hata değeri için model başarımını % 99,83 bulmuş ve giriş parametrelerinin değişimi ile ana buhar basıncı değerlerini operatörlerin kolay ve gerçeğe yakın şekilde önceden tahmin ederek takip edebileceğini öne sürmüştür.

3. VERİ AMBARLARINDAN BİLGİ KEŞFİ VE VERİ MADENCİLİĞİ KAVRAMLARI

Günümüzde artan rekabet koşulları ve gelişen teknolojik hayat, birçok alan için bilgiye sahip olmanın önemini ve bilgiye duyulan ihtiyacı artırmıştır. Bu sebeple bilgiye ulaşmak için hemen her alanda çok çeşitli verilere gereksinim duyulmaktadır. Bilişim teknolojilerinde yaşanan hızlı gelişim ve donanım maliyetlerinin düşmesi ile de veri ambarlarında oldukça büyük boyutlarda veriler depolanabilmekte ve verilere olan gereksinim karşılanmaya çalışılmaktadır. Fakat çoğu zaman bu büyük karmaşık veriler bilgi açısından bir anlam ifade etmemekte ve bu verilerin anlamlı ve kullanılabilir hale gelmeleri için işlenmeleri gerekmektedir.

Geleneksel sorgu veya raporlama tekniklerinin büyük veri boyutları karşısında yetersiz kalması, Veri Ambarlarından Bilgi Keşfi (VABK) kavramının ortaya çıkmasına neden olmuştur (Akpınar, 2000). Veri yığınları arasında fark edilmeyen fakat bizim için gerekli olan bilginin elde edilmesi için çeşitli süreçlerden geçmesi gerekmektedir.

Veri Ambarlarından Bilgi Keşfi (VABK) ve Veri Madenciliği (Data Mining) kavramları bazı araştırmacılar tarafından karıştırılmaktadır. Veri Madenciliği VABK'nın bir alt süreci olduğu halde bu iki kavramı karıştıran araştırmacılar kavramları benzer biçimde kullanmaktadır (Özdemir vd., 2009). Bu kavramların birbiriyle karıştırılmasının sebebi, Veri Madenciliği adımının, VABK aşamalarında modelin oluşturulması ve değerlendirilmesi gibi sürecinin en önemli basamağını oluşturmasıdır (Akpınar, 2000). Literatürde Veri Madenciliği VABK sürecinin bir parçası olmasına karşın, pratikte Veri Madenciliği, VABK sürecinin tamamını içeren bir kavram olarak kullanılmaktadır (Talan,2016).

Veri madenciliği geniş anlamıyla; var olan sorunları çözmek, önemli kararları almak ve geleceği tahmin için gereken ama henüz keşfedilmemiş olan potansiyel bilgiyi kullanışlı hale getiren analiz teknikleri bütünü olarak tanımlanabilir (Aydemir, 2017). Klasik istatistikte özet ve aşırı düzenlenmiş verilerle çalışılırken, veri madenciliğinde çok daha fazla veri ve değişken ile çalışılır. Veri madenciliğini diğer metotlardan

ayıran en önemli yön, önceden akla gelmeyecek ve tahmin edilemeyecek bilgiyi ortaya çıkarmasıdır (Şekeroğlu, 2010). Veri madenciliği teknikleri veriler arasında bulunan, insan gözlem ve hesaplamaları ile fark edilemeyecek bağıntıların ortaya çıkarılmasını sağlarlar. Bunun sonucunda ise karmaşık veriler arasından kullanılabilir ve faydalı bilgiler keşfedilip elde edilir.

3.1. Veri Madenciliğinin Tarihsel Gelişimi

Veri madenciliğinin temelleri 1950'lerde üretilen ilk bilgisayarlara kadar dayanmaktadır. Üretilen bu ilk bilgisayarlar daha çok sayım amaçlı kullanılmışlardır (Canlı, 2017). 1960'lı yıllarda ise veri ambarı sistemleri geliştirilmiştir. Bu sayede bilgi ve dokümanların veri ambarlarında saklanabilmesi mümkün hale gelmiştir (Odabaş, 2017).

1970'lere gelindiğinde ilişkisel veri ambarı yönetim sistemi uygulamaları kullanılmaya başlanmıştır. Basit kurallar içeren uzman sistemler geliştirilmiş ve böylece basit anlamda makine öğrenimi kavramına adım atılmıştır. 1980'lerde ise ilişkisel veri ambarı yönetim sistemleri yaygınlaşmış çeşitli alanlarda kullanılmaya başlanmıştır. Şirketler tarafından müşterileri, rakipleri ve ürünleriyle ilgili verileri içeren veri ambarları oluşturulmaya başlanmıştır. Bu yıllarda geleneksel algoritmalara dayalı istatistiksel araçlar kullanılmış ve oldukça güzel sonuçlar da alınmıştır. Fakat veri boyutu arttıkça güvenilirliğin azaldığı gözlemlenmiştir. 1990'larda katlanarak artan bu veri boyutları karşısında veri ambarlarından gerekli olan faydalı bilginin nasıl elde edilebileceği araştırılmaya başlanmıştır (Savaş vd, 2012). Bilgisayar mühendisleri, büyük miktardaki veri içinden geleneksel istatistiksel yöntemlerinin yerine algoritmik bilgisayar modülleri kullanılarak veri analizinin yapılabileceğini belirterek veri madenciliği ismini kullanmışlardır (Boyacı, 2017). 1992 yılına gelindiğinde ise ilk veri madenciliği yazılımı oluşturulmuştur (Canlı, 2017).

2000'li yıllarda veri madenciliği sürekli gelişmiş ve neredeyse tüm alanlarda kullanılmaya başlanmıştır. Elde edilen sonuçların faydaları görüldükçe, bu alana karşı gösterilen ilgi artmıştır (Fakı, 2015). Gelişen teknoloji sayesinde bilgisayarların

hem hızlı bir şekilde işlem yapması hem de büyük boyutlarda veri depolamaya olanak sağlamalarından dolayı veri madenciliği çalışmaları daha da yaygınlaşarak farklı metot ve algoritmaların geliştirilmesine neden olmuştur. Bu sayede birçok alanda keşfedilmemiş faydalı bilgiler değerlendirilebilmektedir (Yıldırım, 2016).

3.2. Veri Madenciliğinin Kullanıldığı Alanlar

Günümüzde veri madenciliği hemen her alanda yaygın bir şekilde kullanılmaktadır. Veri madenciliği uygulamalarının yoğun bir şekilde kullanıldığı alanlar ve bu alanlarda yapılan çalışmalara ait çeşitli örnekler aşağıda belirtilmiştir (Yurdakul, 2015; Özarslan, 2014):

Pazarlama Alanında;

- Müşterilerin satın alma alışkanlıklarının tespit edilmesinde,
- Mevcut müşterilerin elde tutulmasında ve yeni müşterilerin bulunmasında,
- Kaybedilen müşterilerin benzer özelliklerinin tespit edilmesinde,
- Market sepet analizi ile ürünlerin satışları arasındaki ilişkinin tespit edilmesinde,
- Müşterilerin değerlendirilmesinde ve özellikleri birbirine benzeyen müşterilerin belirlenmesinde,
- Satış tahminlerinin yapılmasında

Bankacılık Alanında;

- Müşteri davranışları arasındaki benzerliklerin bulunmasında ve müşterilerin gruplandırılmasında,
- Aldıkları hizmeti iptal etme riski bulunan müşterilerin tespit edilmesinde,
- Farklı finansal göstergeler arasındaki gizli korelasyonların bulunmasında,
- Kredi kartı dolandırıcılıklarının tespit edilmesinde,
- Müşterilerin kredi taleplerinin değerlendirilmesinde.

Sigortacılık Alanında;

- Yeni poliçe alacak müşterilerin ve poliçelerini yenilemeyecek müşterilerin tahmininde,
- Sigorta dolandırıcılıklarının tespit edilmesinde,

- Riskli müşteri gruplarının tespiti edilmesinde

İnternet alanında;

- Web sayfalarında gezinen kullanıcıların profilinin belirlenmesinde,
- İnternet alışveriş siteleri kullanıcılarının satın alma profillerinin belirlenmesinde,
- Web sayfalarını kullanan ziyaretçilerin sayfa içerisindeki davranışlarının analiz edilmesinde
- Bilişimsel saldırıların çözümlenmesinde,
- Kullanıcı davranışlarına göre web sayfalarının önerilmesinde,

Telekomünikasyon Alanında;

- Kaynak kullanımının iyileştirilmesinde,
- Müşteri davranışlarına göre yeni hizmetlerin sunulmasında
- Dolandırıcılık yapan müşteriler için geçmiş veriler kullanılarak model oluşturma ve benzeri davranışları yapanları belirlemede,
- Arama zamanı, mekânı, süresi, aranılan bilgiler gibi verilerden çeşitli örüntüleri tespit edilmesinde,

Sağlık alanında;

- Hastalık haritalarının hazırlanmasında,
- Hastalık tanılarının belirlenmesinde ve tıbbi teşhis konulmasında,
- Hastalıkları etkileyen faktörlerin ortaya çıkartılmasında,
- İlaç kullanımında olası sahtekarlıkların belirlenmesinde,
- Test sonuçlarının tahmin edilmesinde,
- Hastalara ait tıbbi verilerden hastanın risklerinin tahmin edilmesinde,

Eğitim alanında;

- Öğrenci profillerine göre başarı durumlarının tahmin edilmesinde,
- Benzer özellikleri gösteren öğrencilerin belirlenmesinde,
- Ölçme ve değerlendirme sistemlerinin geliştirilmesinde,
- Öğrenme ortamlarının geliştirilmesine yönelik araştırma-geliştirme çalışmalarının yapılmasında,

Biyomedikal ve DNA alanında;

- DNA dizilimindeki benzerliklerin karşılaştırılmasında,
- Zengin genetik veri ambarlarının meydana getirilmesinde,
- Genler arasındaki ilişkilerin belirlenmesinde,
- Genlerin hastalıkların farklı seviyelerindeki etkilerinin belirlenmesinde,
- Biyomedikal verilerin anlaşılmasında görsel araçlardan faydalanılmasında,

İmalat alanında ve mühendislik uygulamalarında;

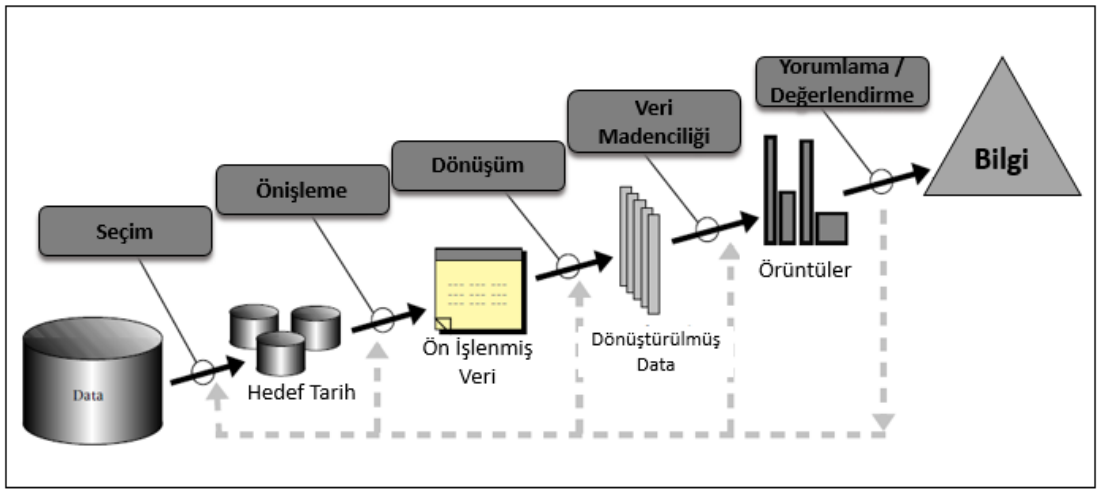
- Araştırma ve geliştirme faaliyetlerinde,
- Ürün hatalarındaki sapmaların belirlenmesinde,
- Çeşitli örüntülerin tanımlanmasında,
- Simülasyon uygulamalarında veri madenciliği teknikleri kullanılmaktadır.

3.3. Veri Ambarlarında Bilgi Keşfi Sürecinde Kullanılan Yaklaşımlar

Büyük veriler içeren, analizlerin klasik hesaplama yöntemleri yürütülemeyeceği, temel istatistik yöntemlerinin tek başına aşamayacağı problemler karşısında veri madenciliği yöntemlerinin kullanılması kaçınılmaz olmuştur. Küçük verilere kolaylıkla uygulanabilen geleneksel analizler, veri büyüklüğünün artması ve ayrıca farklı veri türlerinin oluşması ile yetersiz duruma gelmiştir. Bu verilerin geleneksel metotlarla analiz edilmeye çalışılması, zaman ve maliyet kaybına neden olmaktadır. Bu kaybın önlenmesine, veri madenciliğinin ve bilgisayarların devreye girmesiyle çözüm bulunmuştur. Böylece büyük veriler arasından faydalı olan bilginin kolaylıkla bulunması sağlanarak zaman ve maliyet kaybının önüne geçilmiştir (Çığşar, 2017).
Veri Ambarlarından Bilgi Keşfi sürecinin farklı kaynaklarda farklı şekilde ele alındığı görülmektedir (Emre, 2017). Bu bölümde literatürde yaygın olarak kullanılan yaklaşımların birkaçından bahsedilecektir.

3.3.1.Fayyad Yaklaşımı

Fayyad vd. (1996), Veri Ambarlarında Bilgi Keşfi sürecini Şekil 3.1.de görüldüğü gibi tanımlamışlar ve sürecinin adımlarını aşağıdaki gibi detaylandırmışlardır:



Şekil 3.1. VABK Sürecini Oluşturan Adımlara Genel Bir Bakış (Fayyad vd., 1996)

- Birinci adım: VABK sürecinin hedefinin, müşteri bakış açısıyla tanımlanarak uygulama alanının ve uygulama alanıyla ilgili ön bilginin anlaşılması gerekmektedir.

- İkinci adım: Bir hedef veri seti oluşturulmasıdır.

- Üçüncü adım: Veri temizleme ve veri ön işlemedir. Eğer mümkünse gürültülü veriler tespit edilmeli ve eksik veri alanlarına yönelik stratejiler belirlenmelidir.

- Dördüncü adım: Amaca bağlı olarak boyut azaltma veya dönüştürme yöntemleri ile değişkenlerin sayısı azaltılabilir veya veriler arasındaki örüntüyü değiştirmeyen ve veriyi temsil eden küçük veri setleri bulunabilir.

- Beşinci adım: VABK sürecinin birinci adımdaki hedeflerini gerçekleştirmek için özetleme, sınıflandırma, regresyon, kümeleme gibi veri madenciliği tekniğinin seçilmesidir.

- Altıncı adım: Veri madenciliği metot ve algoritmasının seçildiği aşamadır.

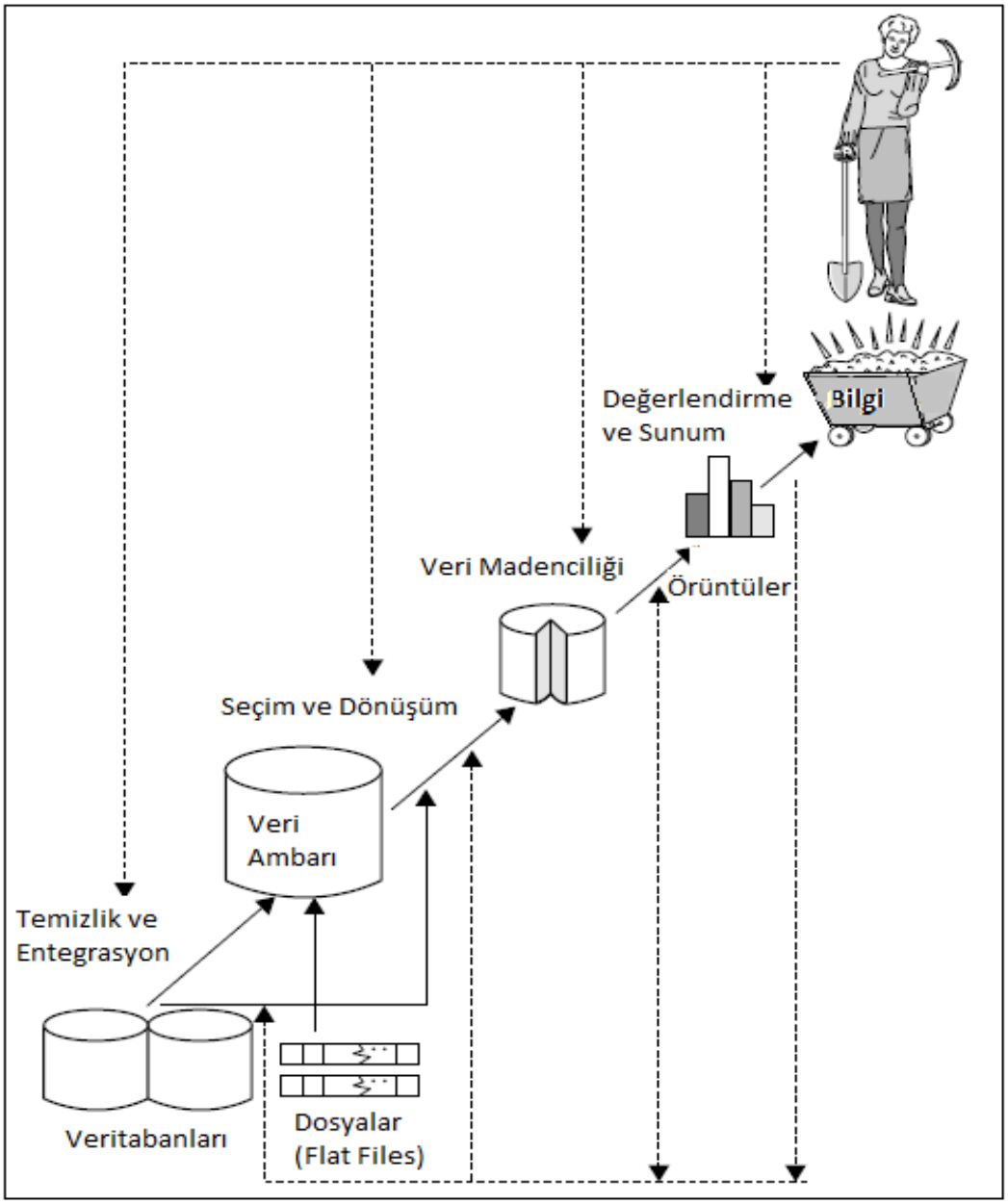
- Yedinci adım: Veri madenciliği aşamasıdır. Örüntülerin, sınıflandırma kuralları veya ağaçların, regresyon ve kümelenme gibi temsillerin oluşturulduğu adımdır.

- Sekizinci adım: Modellerin yorumlandığı adımdır. Bu adım, çıkarılan modellerin veya modellerin verilerinin görselleştirilmesini içerebilir. Daha önceki adımlardan birine tekrar dönülmesini gerektirebilir.

- Dokuzuncu adım: Keşfedilen bilgi üzerinde hareket etmektir: Bilgiyi doğrudan kullanmak, bilgiyi başka bir sisteme dahil etmek veya sadece belgeleyerek ilgili taraflara rapor etmektir. Bu süreç aynı zamanda önceden inanıldığı (veya çıkarılan) bilgiyle olası çelişkileri kontrol etmeyi ve çözmeyi de içerir.

3.3.2. Han Yaklaşımı

Han vd. (2012), veri madenciliği alanında önemli çalışmalardan biri olan Data Mining Concepts and Techniques isimli kitapta, VABK sürecini Şekil 3.2.'de görüldüğü gibi görselleştirmişler ve adımların detaylarını aşağıda belirtildiği gibi açıklamışlardır:



Şekil 3.2. Bilgi Keşfi Sürecinde Bir Adım Olarak Veri Madenciliği. (Han vd., 2012)

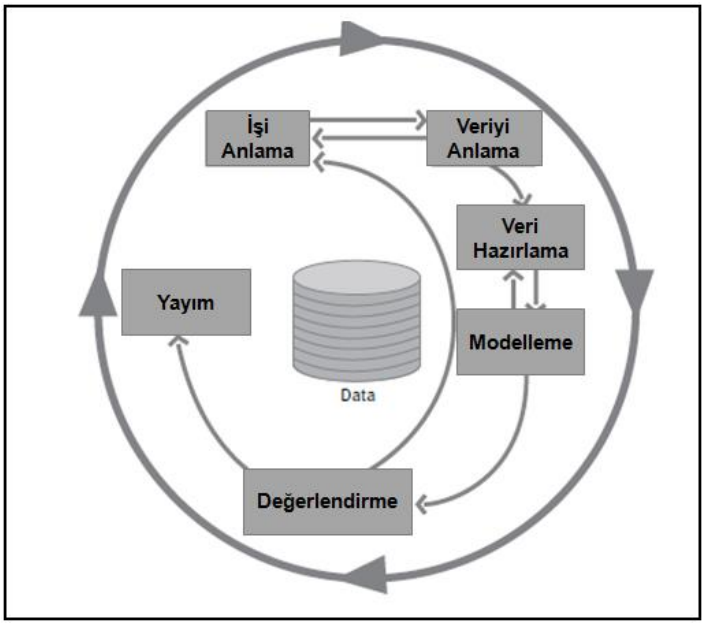
1. Veri temizleme: Gürültülü ve tutarsız verileri belirleyip çıkarma adımıdır.
2. Veri entegrasyonu: Birden fazla veri kaynağının birleştirildiği adımdır.
3. Veri seçimi: Analizle ilişki olan verilerin veri ambarından seçildiği aşamadır.
4. Veri dönüşümü: Verilerin madencilikte kullanılabilecek uygun formlara dönüştürüldüğü aşamadır.
5. Veri madenciliği: Örüntüleri çıkarmak için veriye akıllı yöntemlerin uygulandığı aşamadır.

6. Örüntü değerlendirme: Bilgiyi temsil eden ilginç örüntülerin tanımlandığı aşamadır.
7. Bilgi sunumu: Keşfedilen bilgiyi kullanıcılara iletmek amacıyla görselleştirme ve bilgi sunum metotlarının uygulandığı aşamadır.

3.3.3. CRISP-DM (Cross-Industry Standard Process for Data Mining)

SPSS, NCR, Daimler Chrysler gibi firmalar, endüstride kullanılabilecek şekilde VABK sürecini standart hale getirmek için bir konsorsiyum oluşturmuşlardır. Çalışmalarının sonucunda oldukça geniş bir alanda farklı sektörlerde kullanılabilecek şekilde CRISP-DM (Cross-Industry Standard Process for Data Mining) adını verdikleri metodolojiyi üretmişlerdir (Chapman vd., 2000).

Chapman vd. (2000), veri madenciliği projesinin yaşam döngüsünün altı aşamadan oluştuğunu ve aşamalarının katı olmadığını, farklı aşamalar arasında ileri geri hareket etmenin her zaman gerekli olduğunu belirtmişlerdir. CRISP-DM (Cross-Industry Standard Process for Data Mining) sürecini Şekil 3.3.'de görüldüğü gibi şematize etmişlerdir.



Şekil 3.3. CRISP-DM Süreci (Chapman vd., 2000)

Dış daire, veri madenciliğinin kendisinin döngüsel doğasını simgelemektedir ve bir çözüm yayımlandıktan sonra da veri madenciliği sona ermez. Daha sonraki veri madenciliği süreçleri, öncekilerin deneyimlerinden faydalanacaktır. CRISP-DM sürecinin altı aşamasının detayları ise aşağıda bahsedildiği gibidir (Chapman vd., 2000)

İşi Anlama: Bu başlangıç aşaması, proje hedeflerini ve gereksinimlerinin iş perspektifinden değerlendirildiği, ardından bu bilgiyi bir veri madenciliği problemi tanımına dönüştürüldüğü ve hedeflere ulaşmak için bir ön planın tasarlandığı aşamadır.

Veriyi Anlama: Veri toplama ile başlayan veri anlama aşaması; veri ile ilgili bilgi sahibi olunduğu, veri kalitesinin sorunlarının tanımlandığı, verilerle ilgili ilk bilgilerin keşfedildiği ve ilginç alt kümeleri fark etmeyi sağlayan etkinliklerin yer aldığı aşamadır.

Veri Hazırlama: Veri hazırlama aşaması, ilk ham verilerden başlayarak modelleme araçlarına beslenecek olan nihai veri setini oluşturana kadar gerekli olan tüm faaliyetleri kapsar.

Modelleme: Modelleme tekniklerinin seçildiği ve uygulandığı aşamadır. Bazı tekniklerde veri formları için belirli gereksinimler vardır. Bu nedenle genellikle veri hazırlama aşamasına dönmek gerekebilir.

Değerlendirme: Modelin iş hedeflerini düzgün bir şekilde yerine getirdiğinden emin olmak için, modeli tam olarak değerlendirmek ve bunu oluşturmak için atılan adımların gözden geçirildiği aşamadır. Önemli bir amacın ya da yeterince dikkate alınmamış önemli bir iş konusunun olup olmadığı belirlenir.

Yayım: Modelin oluşturulması genellikle projenin sonu değildir. Genellikle, bir kuruluşun karar verme süreçlerinde canlı modeller uygulanmasını gerektirir.

3.3.4. SEMMA Yaklaşımı

Bir yazılım şirketi olan SAS firması tarafından geliştirilen ve bilgi keşfi sürecini standart bir şekle sokmayı amaçlayan diğer bir yaklaşım SEMMA metodudur. SEMMA kelimesi sample (örnekle), explore (araştır), modify (düzenle), model (modelle), assess (değerlendir) kelimelerinin baş harflerinden oluşturulmuştur. Süreç beş aşamalı bir döngüden oluşur. SEMMA adımları aşağıda belirtildiği gibidir (Olson ve Delen, 2008; Azevedo ve Sanoz, 2008; Akdemir,2016):

- **S**ample (Örnekle): Büyük bir veri kümesi içinden, hızlı bir şekilde işlemek için yeterince küçük olan fakat önemli ve anlamlı bilgileri içerecek kadar da büyük olan bir kısmının çıkarılmasıdır. Örneklenen küçük veri tüm veri setini istatistiksel açıdan temsil etmelidir. Verilerin tamamına yayılmış bir örüntü varsa örneklenen küçük kısımda da yer alacak şekilde seçilmelidir.

- **E**xplore (Araştır): Veri kümesini daha iyi anlamak için, beklenmedik eğilimlerin ve anormalliklerin aradığı aşamadır. Verileri örnekledikten sonra, bir sonraki adım, doğal eğilimler veya gruplamalar için bunları görsel veya sayısal olarak araştırmaktır. Keşif sürecini hassaslaştırmaya ve yönlendirmeye yardımcı olur. Görsel keşif net eğilimleri ortaya çıkarmazsa, istatistiksel teknikler aracılığıyla veriler araştırabilir.

- **M**odify (Düzelenle): Kullanıcının modelleme süreci için değişkenleri belirlediği, seçtiği ve dönüştürdüğü aşamadır. Bu aşamada değişken sayısının azaltılması veya yeni değişken eklenmesi gibi işlemler yapılabilir, aykırı ve gürültülü veriler tespit edilip gerekli düzenlemeler gerçekleştirilebilir. Aşama sonunda veriler analiz yapmak için uygun hale getirilmiş olur.

- **M**odel (Modelle): Analiz yapmak için hazır hale gelmiş olan verilere çeşitli veri madenciliği algoritmaların uygulanarak verilerde bulunan bilgilerin keşfedildiği adımdır.

- **A**ssess (Değerlendir): Veri keşfi sürecindeki bulguların yararlılığının ve güvenilirliğinin değerlendirdiği adımdır. Bu son adımda, modellerin ne kadar iyi performans gösterdiği değerlendirilir. Bir modeli değerlendirmenin bir yolu, verilerin bir kısmını örneklem oluştururken bir kenara koymak ve model oluştuktan sonra model yapımı sırasında kullanılmayan bu veriler üzerinde modeli test etmektir.

Çalışmada bahsedilen örneklerden de görüldüğü gibi veri ambarlarından bilgi keşfi süreci için birbirinden farklı çeşitli yaklaşımlar mevcuttur. Fakat genel olarak her birinde ortak olan yönler dikkate alınarak özetlemek gerekirse, veri keşfi sürecinin aşağıda belirtilen adımlardan oluştuğu söylenebilir:

- Problemin Tanımlanması
- Verilerin Hazırlanması
- Modelin Kurulması ve Değerlendirilmesi
- Modelin Uygulanması ve İzlenmesi

Belirtilen adımlarda ve bu adımların alt başlıklarında yapılması gereken işlemler çalışmanın ilerleyen bölümünde detaylı olarak açıklanmıştır.

4.VERİ AMBARLARINDAN BİLGİ KEŞFİ SÜRECİ AŞAMALARI

4.1.Problemin Tanımlanması

Problemin tanımlanması adımı, bilgi keşfi sürecinin ilk adımıdır ve uygulamanın hangi amaç için yapılacağı belirlenir. Ayrıca bu adımda uygulama sürecinin ne şekilde ilerleyeceği planın da yapılması gerekir (Aydemir, 2017). Problemin belirlenmesi aşaması veriden bilgi keşfine giden sürecinin en önemli aşamasıdır. Belirlenen amaç, problem ile tam uyuşmaz ise çalışmadan doğru sonuç alınamamakla birlikte hatalı kararlar verilmesine neden olarak başka yanlışlara da yol açabilir (Odabaş, 2017). Yani veri madenciliği uygulamalarında başarılı olmanın ilk koşulu, çalışmanın hangi amaç için yapılacağının açık ve net bir şekilde belirlenmiş olmasına bağlıdır. Başarılı olmak için probleme odaklanılmalı, problem açık ve net bir şekilde belirlenmeli, çıktıların başarı seviyelerinin nasıl ölçüleceği doğru bir şekilde tanımlanmalı, yanlış tahminlerin getirebileceği kayıplar ve doğru tahminlerin sağlayacağı faydalar da bu adımda tahmin edilmeye çalışılmalıdır (Çetin, 2009).

Amaçların açık net ve anlaşılır olması, veri madenciliği yöntem ve algoritmalarının seçimini belirleyen ana unsurdur. Beklentileri net olmayan amaçlar, uygun olmayan yöntem ve algoritmalarının tercih edilmesine neden olurlar. Uygun olmayan yöntem ve algoritmalarının tercih edilmesi ise başarısız bir modellemenin gerçekleştirilmesine yol açar (Talan, 2016).

4.2.Verilerin Hazırlanması

Verilerin hazırlanması aşaması, ham veriden başlayıp işlenmeye hazır olan nihai veriyi elde edene kadar yapılması gereken bütün işlemlerin yapıldığı oldukça yoğun emek gerektiren bir adımdır (Fakı, 2015). VABK sürecinin en önemli ve en çok vakit alan adımı olarak, verilerin model için hazırlanması adımı gösterilebilir. Veriler elde edilirken; insanların sebep olduğu sorunlardan, verinin bir yerden başka bir yere aktarılırken oluşan sorunlardan, donanımdan kaynaklı meydana gelebilecek sorunlardan vb. sebeplerden dolayı elde edilen verilerde çeşitli hatalar meydana

gelebilir. Ham şekliyle verilerde; eksik değerlere, yanlış değerlere ya da olağandan çok çok büyük ya da küçük değere sahip olma gibi sorunlar görülebilir. Bundan dolayı modellemeden önce verilerdeki bu sorunların giderilerek verilerin kullanıma uygun bir biçime çevrilmesi gerekmektedir. Aksi takdirde verilerden hatalı sonuçlar elde edilecektir (Kılınç, 2015).

Verilerin model için hazırlandığı bu aşama kendi içerisinde veri toplama, veri bütünleştirme, veri temizleme, veri dönüştürme, veri indirgeme gibi adımlarından oluşmaktadır. Bu adımlar sonucunda veriler modelleme için uygun hale gelmiş olacaklardır.

Veri hazırlama süreci adımlarının belirli bir sırası veya tekrar sayısı yoktur. Süreçte meydana gelecek problemlerden dolayı sık sık geri dönülüp veri hazırlama sürecinin baştan yapılması durumu meydana gelebilir (Yıldırım, 2016).

4.2.1. Veri Toplama

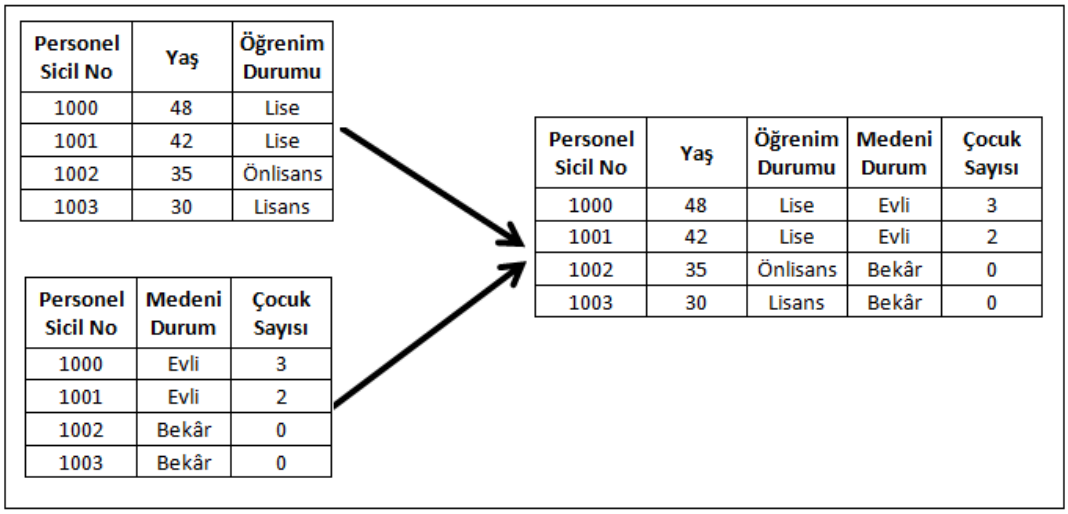
Tanımlanmış olan problem için gerekli olacak veriler toplanmadan önce bu verilerin toplanacağı kaynakların belirlenmesi gerekmektedir. Verilerin nereden, ne şekilde toplandığı büyük önem taşımaktadır. Çünkü ihtiyaç olandan az verinin bulunması, uygulama sürecini aksatabileceği gibi ihtiyaç olandan çok verinin bulunması ise uygulama sürecinin uzamasına neden olabilir. Böylece zaman ve emek gibi kaynakların kaybı ya da veri kirliliğine neden olacak durumlar ortaya çıkabilir (Uyumaz, 2017).

Elde edilen verilerin çalışmanın ihtiyaçlarını karşılayıp karşılamadığını anlamak için; verilerin formatının ne olduğu, ne kadar kayıt sayısı elde edildiği, kayıtların kaç özellikten oluştuğu gibi verilerin diğer özelliklerinin incelenmesi ve değerlendirilmesi gerekmektedir. Ayrıca ileride aynı projenin tekrardan veya projeye benzer projelerin yapılması durumunda yardımcı olması için verilerin elde edildiği kaynakların, veriyi toplarken kullanılan yöntemlerin, yüz yüze gelinen sorunların ve gerçekleştirilen çözümlerin veri toplama adımında kayıtlarının tutulması önemlidir (Şık, 2014).

4.2.2.Veri Bütünleştirme

Veriler toplanırken birden fazla veri kaynağında yer alabilir. Değişik kaynaklardan (veri ambarları, veri küpleri, metin dosyaları, vb.) sağlanan veriler arasındaki uyumu sağlamak için veri bütünleştirme işleminin gerçekleştirilmesi gerekir. Bütünleştirme işlemi yapılırken verilerin karakteristik özellikleri, toplanma biçimi vb. özelliklerinin göz önüne alınarak tutarlılığın sağlanması gereklidir (Dondurmacı ve Çınar, 2014).

Veri bütünleştirme aşamasında, aynı niteliğe ait sabit bir değişken baz alınarak bütünleştirme işlemi yapılır. Personel sicil numarası, öğrenci numarası gibi, farklı veri kaynaklarından toplanan verilerde aynı sabit değeri taşıyan değişkenler kullanılarak yapılır. Şekil 4.1.'de iki farklı tabloda yer alan bilgilerin personel sicil numarası gibi sabit değer yardımıyla bütünleştirilerek tek bir veri kaynağına dönüştürülmesi bu işleme örnek olarak gösterilebilir.



Şekil 4.1. Bütünleştirilmiş Veri Örneği

Farklı kaynaklardan toplanan verilerde değişkenler farklı şekillerde ve isimlerde tutulmuş olabilirler. Yukarıda verilen personel sicil numarası örneği için, kaynaklardan birinde personel_sicil_no ismiyle veri tutulmuşken diğerinde per_sicil_id şekliyle veri tutulmuş olabilir. Veri birleştirme aşamasında bu tarz durumlara dikkat edilerek işlemlerin gerçekleştirilmesi gerekmektedir.

Veri bütünleştirme işleminde dikkat edilmesi gereken bir diğer nokta ise aynı değişkene ait verilerin farklı formatlarda tutuluyor olabileceğidir. Örneğin cinsiyet değişkenine ait veriler, bir veri kaynağında kadın-erkek şeklinde tutuluyor iken diğer bir veri kaynağında 0-1 şeklinde tutuluyor olabilir. Bu tarz farklı formatlarda tutulan veriler için de gerekli düzenlemelerin yapılarak veri bütünleşmesinin doğru bir şekilde gerçekleştirilmesi sağlanmalıdır.

4.2.3.Veri Temizleme

Bazen çalışılan veride beklenilen özelliklerin olmadığı görülebilir. Verilerde eksik kısımlar veya olağandan oldukça uzak değerler yer alabilir. Bu tarz istenmeyen durumlarda eksik ve aykırı verilerin temizlenmesi gerekmektedir. Eksik verilerin düzenlemesi için aşağıda belirtilen metotlar kullanılabilir (Akın, 2012):

- Eksik değerlerin olduğu kayıtlar veri kümesinden çıkartılabilir.

- Eksik değerlerin yerine genel bir sabit yazılabilir. Örnek verilirse tüm eksik değerler için "bilinmiyor" gibi bir değer yazılabilir. Fakat bu durum çeşitli kayıpların ve yanlışların oluşmasına neden olabilir.

- Değişkenin tüm verilerinin ortalaması alınarak eksik değer yerine yazılabilir.

- Değişkenin tüm verilerinin ortalamasının yerine, eksik veri ile aynı sınıfa ait değerlerin ortalaması alınarak eksik olan değer yerine yazılabilir.

- Veriler arasındaki ilişkiler tespit edilerek veya makul bir model kurularak eksik değerlerin tahmin edilmesi sağlanarak eksik değer yerine yazılabilir.

Veri temizleme adımında temizlenmesi gereken diğer veriler ise gürültülü verilerdir. Veri girişi yapan kişilerin dalgınlığı, veri aktarırken oluşan format değişiklikleri veya veri kaybı gibi nedenlerden ötürü gürültülü veriler oluşabilmektedir. Örneğin bir kişinin yaşının 360 olması veya kilosunun 650 olması ya da doğum tarihinin 4986 olması gibi gürültülü veri içeren verilerin de temizlenmesi gerekmektedir.

Veri temizleme işleminin kullanılması gerekli olan bir başka yer ise olağandan oldukça farklı olan verilerdir (outliers). Örneğin maaş ortalamasının 2000 olduğu bir kaynakta bir adet değerin maaşının 54500 olması durumu. Bu verilerin belirlenmesi için histogram, kümeleme analizi, regresyon analizi gibi metotlardan yararlanılabilir (Oğuzlar, 2003).

Temizleme adımında yapılan hatalar sürecin başa dönmesine sebep olabilir. Bu nedenle veri temizleme adımında çalışmanın amacına ulaşmasında gerekli olan verilerin kayba uğramaması için dikkatli ve titiz davranılmalı, sadece çıkartılması gerçekten gerekli olan verilerin çıkartılmasına dikkat edilmelidir (Özbay, 2015).

4.2.4. Veri Dönüştürme

Tüm veri madenciliği algoritmalarının kendine özgü özellikleri mevcuttur ve bu özelliklerin birisi de işleyebildikleri veri türleridir. Her algoritma her veri türü ile çalışamaz. Dolayısıyla algoritmaların kullanımına uygun hale getirmek için veriler üzerinde dönüştürme işlemlerinin yapılması gerekli olabilir. Kullanılacak algoritmalar hangi veri tipiyle çalışıyor ise (sayısal, kategorik, 0-1 değerleri gibi) dönüştürme işleminin o doğrultuda uygulanması gerekmektedir (Kılınç, 2015).

En yaygın kullanılan veri dönüştürme işlemlerinden birisi veri normalleştirmedir. Bazı veri normalleştirme işlemleri aşağıdaki gibi sıralanabilir (Tahminciler, 2014):

- **Min-Max:** Orijinal veriler, hedeflenen aralığına doğrusal dönüşüm ile dönüşürler. Bazı veri madenciliği algoritmaları sadece 0-1 aralığında çalıştığından dönüştürülmüş aralık genellikle 0-1 aralığı olmaktadır.

- **Z Skor:** Değişkenin herhangi bir değeri için, bilinen Z dönüşümü kullanılarak değişkenin ortalama ve standart sapmasına bağlı olarak hesaplanır.

- **Ondalık Ölçekleme.** Değişkenin ondalık kısımları ile oynanarak yapılan normalleştirmedir. Örneğin 825 sayısı 0,825 gibi bir değer ile normalleştirilebilir.

4.2.5.Veri İndirgeme

Büyük veri kümeleri üzerinde yapılan uygulamalarda işlem süreleri doğal olarak uzun olacaktır. Aynı çıktıları verecek daha ufak veri kümeleriyle çalışmak, uzun işlem sürelerinin düşmesine neden olacaktır. Bu sebeple veri boyutlarının azaltılması için örneklemeler yapılabilir. Büyük veri kümesinin girdi dağılımını koruyarak daha küçük örneklerin oluşturulması mümkündür. Ayrıca veride gereksiz olan özellikler elenerek de veri boyutlarını azaltma yoluna gidilebilir. Sonuca olumlu ya da olumsuz bir etki etmeyen veya algoritmanın başarılı sonuç almasını engelleyecek değişkenlerin atılması bilgi keşfi sürecini olumlu şekilde etkileyecektir (Kılınç, 2015).

Veri indirgeme adımında aşağıdaki metotlar uygulanabilir (Oğuzlar, 2003):

- **Veri Birleştirme veya Veri Küpü:** Veriler tek tek kendi başlarına değil, benzer veriler gruplanarak genel kavramlarla ifade edilir. Örneğin elimizde 2000-2003 yılları için dört faklı çeyrek dönemin satış tutarlarının yer aldığı bir veri kümesi olsun. Çeyrek dönemlik satış tutarlarını yıllık satış tutarı şekline çevirdiğimizde veri birleştirmesi yapmış oluruz. Satış tutarlarını yıllık şeklinde tuttuğumuzda veri kümesinin hacmi küçülmüş olur fakat herhangi bir bilgi kaybı yaşanmış olmaz.

  Veri küpleri ise çok değişkenin yer aldığı birleştirilmiş bilginin saklandığı yerlerdir. Örneğin bir satış tutarlarının; yıllar, satışı gerçekleştirilen mallar ve firmanın farklı satış noktaları aynı küp üzerinde gösterilebilir. Veri küpleri özet bilgi hızlıca ulaşılmasını sağlayan yapılardır.

- **Boyut indirgeme:** Veri kümeleri bazen gereksiz şekilde oldukça fazla değişkene sahip olmaktadır. Örneğin bir ürünün satışına ait çalışılan bir veri kümesinde müşterilerin telefon numaraları gereksiz bir değişken olarak yer alabilir. Bu tarz gereksiz değişkenler hem bilgi keşfi sürecini yavaşlatacak hem de yanlış örüntülerin elde edilmesine yol açacaktır. Bu tarz gereksiz verilerin tespit edilip, veri kümesinden atılması çalışmanın hızı ve doğruluğu açısından faydalı olacaktır.

- **Veri sıkıştırma:** Şifreleme ve veri dönüştürme işlemleri yardımıyla boyut sıkıştırması yapılarak asıl veriyi temsil eden ve olduğundan daha az alan kaplayan veri kümesinin elde edilmesi işlemidir.

- **Kesikli Hale Getirme:** Bazı veri madenciliği algoritmaları sadece kategorik değerler ile çalışırlar. Bu algoritmalarda kullanılacak sürekli verilerin kesikli verilere dönüştürülmesi gerekir. Örneğin kazanılan aylık ücret gibi sürekli bir değişken 0-2000, 2000-4000, 4000+ gibi kategorik bir şekilde ifade edilebilir. Bu şekilde bir veri indirgeme yapıldığı zaman detay bazı bilgiler kayıp olsa da genelleştirilmiş veriler daha anlamlı sonuçlar verecektir.

4.3. Modelin Kurulması ve Değerlendirilmesi

Bu aşama modelleme tekniklerinin seçimini, test tasarımının oluşumunu, modelin oluşumunu ve modelin değerlendirilmesini içerir. Adımlarda bir önceki adıma geri dönme ihtiyacı duyulabilir (Mocan, 2016).

4.3.1.Modelleme Tekniklerinin Seçimi

Tanımlanmış olan probleme en uygun veri madenciliği modelinin seçilebilmesi için verilerin çok sayıda model üzerinde denenmesi gerekmektedir. Bundan dolayı veri hazırlama ve model kurma adımları, en uygun olan modele karar verilinceye kadar yenilenen süreçlerdir (Boyacı, 2017). Oldukça çeşitli algoritmalar var olmasına karşın seçilecek olan algoritmaların, uygulanacak problemin yapısına göre belirlenmesi gerekmektedir. Kullanılan algoritmaların doğruluk, kesinlik, özgünlük, duyarlılık gibi ölçütleri göz önüne alınarak algoritmaların başarı ölçütleri belirlenir. Kullanılan verilerden en iyi başarıyı sağlayan algoritma tespit edilerek seçim yapılır (Aydın, 2017).

4.3.2.Test Tasarımının Oluşumu

Veri madenciliği genellikle oldukça büyük veri kümelerinin analizi için kullanılır. Standart veri madenciliği sürecinde bu büyük veriler bölünerek, bir kısmı modelin

geliştirilmesi için (training set) kullanılır, ayrılan diğer bir kısmı ise inşa edilen modelin test edilmesi için (test set) kullanılır. Verileri bu şekilde bölerek bir parçasını modelin geliştirilmesi için kullanmak ve ayrı bir parçası üzerinde test etmek daha kesin sonuçlar elde etmemizi sağlayacaktır. Bazı uygulamalarda parametlerin tahmini için doğrulama kümesi (validation set) adında üçüncü bir veri bölümü daha yapılabilmektedir (Olson ve Delen, 2008).

Bir modelin doğruluğunun testi için aşağıda belirtilen yöntemler kullanılmaktadır (Akpınar, 2000):

- **Basit Geçerlilik/Doğrulama Testi (Simple Validation Test):** Bu yöntemde verilerin tercihe göre %5 ile %33 arasındaki bir kısmı test verisi için ayrılır ve diğer kısmı ile model eğitilir. Model eğitildikten sonra test için ayrılan veriler üzerinde uygulanarak modelin doğruluğu test edilmiş olur.

- **Çapraz Geçerlilik/Doğrulama Testi (Cross Validation Test):** Veri kümesi rastgele olacak şekilde ikiye ayrılır. Ayrılan veri kümelerinden biri eğitim işlemi için diğer kalan kısım ise test işlemi için kullanılır. Daha sonra model ikinci kez uygulanarak eğitim için kullanılan veriler test için, test için kullanılan veriler de eğitim için kullanılır. Hata oranlarının ortalaması alınarak modelin hata oranı olarak kullanılır.

- **N Katlı Çapraz Geçerlilik/Doğrulama Testi (N-Fold Cross Validation Test):** Veri kümesi rastgele N adet parçaya bölünür ve her defasında bir parçası test ve kalan diğer parçalar eğitim verisi olarak kullanılır. Tüm parçalar test kümesi olacak biçimde N defa model uygulanır ve elde edilen N denemenin ortalama hata oranı modelin hata oranı olacaktır. Örneğin yaygın olarak kullanılan 10 Kat Çapraz Geçerlilik/Doğrulama Test Şematiği Şekil 4.2.'de görüldüğü gibidir.

	PARÇA 1	PARÇA 2	PARÇA 3	PARÇA 4	PARÇA 5	PARÇA 6	PARÇA 7	PARÇA 8	PARÇA 9	PARÇA 10
TEST 1	EĞİTİM	EĞİTİM	EĞİTİM	EĞİTİM	EĞİTİM	EĞİTİM	EĞİTİM	EĞİTİM	EĞİTİM	TEST
TEST 2	EĞİTİM	EĞİTİM	EĞİTİM	EĞİTİM	EĞİTİM	EĞİTİM	EĞİTİM	EĞİTİM	TEST	EĞİTİM
TEST 3	EĞİTİM	EĞİTİM	EĞİTİM	EĞİTİM	EĞİTİM	EĞİTİM	EĞİTİM	TEST	EĞİTİM	EĞİTİM
TEST 4	EĞİTİM	EĞİTİM	EĞİTİM	EĞİTİM	EĞİTİM	EĞİTİM	TEST	EĞİTİM	EĞİTİM	EĞİTİM
TEST 5	EĞİTİM	EĞİTİM	EĞİTİM	EĞİTİM	EĞİTİM	TEST	EĞİTİM	EĞİTİM	EĞİTİM	EĞİTİM
TEST 6	EĞİTİM	EĞİTİM	EĞİTİM	EĞİTİM	TEST	EĞİTİM	EĞİTİM	EĞİTİM	EĞİTİM	EĞİTİM
TEST 7	EĞİTİM	EĞİTİM	EĞİTİM	TEST	EĞİTİM	EĞİTİM	EĞİTİM	EĞİTİM	EĞİTİM	EĞİTİM
TEST 8	EĞİTİM	EĞİTİM	TEST	EĞİTİM	EĞİTİM	EĞİTİM	EĞİTİM	EĞİTİM	EĞİTİM	EĞİTİM
TEST 9	EĞİTİM	TEST	EĞİTİM	EĞİTİM	EĞİTİM	EĞİTİM	EĞİTİM	EĞİTİM	EĞİTİM	EĞİTİM
TEST 10	TEST	EĞİTİM	EĞİTİM	EĞİTİM	EĞİTİM	EĞİTİM	EĞİTİM	EĞİTİM	EĞİTİM	EĞİTİM

Şekil 4.2. 10 Kat Çapraz Geçerlilik/Doğrulama Test Şematiği

- **Bootstrapping:** Küçük veri kümelerinin bulunduğu örneklerde modelin hata tahmininde kullanılan bir diğer tekniktir. Önce model tüm veri kümesi üzerine kurulur ve sonra tekrarlı örneklerle en az 200 hatta bazen binin üzerinde adetlerle oldukça fazla sayıda öğrenim kümesi oluşturularak veri kümesi içinde öğrenim kümesinden geriye kalan örnekler test verisi olarak kullanılarak hata oranı hesaplanır.

4.3.3.Modelin Oluşumu

Modelin kuruluş sürecinde, denetimli ve denetimsiz öğrenmenin kullanıldığı modellere göre farklılıklar görülür. Denetimli öğrenmede, ilgili sınıflar önceden belirli olan ölçütlere göre ayrıştırılır ve her bir sınıf için örnekler verilir. Böylece verilen örneklerden yola çıkılarak modeller her bir sınıf için özellikleri keşfeder ve kural çıkarımları elde eder. Denetimsiz öğrenmede ise, modellerin gözlemlerden yola çıkarak örneklerin arasındaki benzerlikleri ve ilişkileri keşfetmesi ve örnek kümelerini tanımlamaya çalışması beklenir (Uyumaz, 2017).

4.3.4.Modelin Değerlendirilmesi

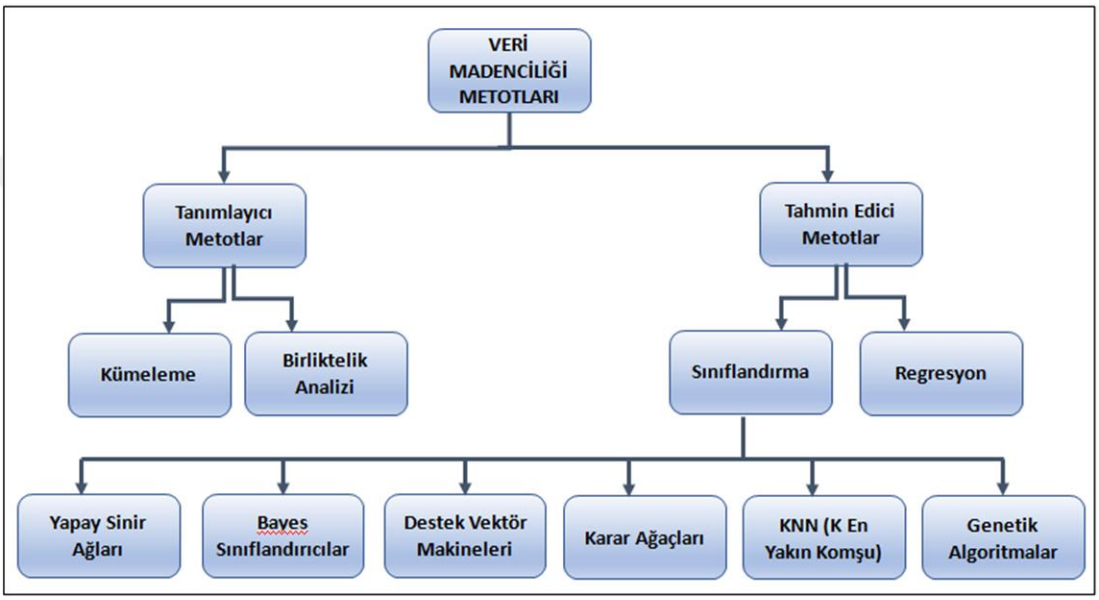
Bu aşama içerisinde modelden elde edilen sonuçların değerlendirilmesi, model oluşturulurken geçen süreçlerin değerlendirilmesi, bir sonraki aşamanın ne olacağına karar verilmesi gibi çeşitli değerlendirmeler yapılır. Hangi modelin çalışmanın amacına ne oranda katkı sağladığı belirlenmeye çalışılır. Modelleme süreci sonunda istenilen hedeflere ulaşılamadıysa, başarısızlığa neden olan etmenlerin belirlenmeye çalışılması ve başarısızlık durumunun etkilerinin değerlendirmelerinin de yapılması gerekmektedir (Levent, 2016).

4.4. Modelin Uygulanması ve İzlenmesi

Değerlendirme süreci sonunda en başarılı bulunan model, başlangıçta belirlenen amaca yönelik tek başına kullanılabileceği gibi herhangi bir sistemin parçası olarak farklı bir sisteme entegre edilerek de kullanılabilir. Modelin kullanılmaya başlaması ile birlikte model izlenmeye başlanır. Çalışmanın niteliğinin değişmesi, veri depolama sistemlerinde yaşanabilecek farklılıklar, zamanla yeni kayıt değişkenlerinin oluşması veya bazı değişkenlerin zamanla anlamını kaybetmesi modelin başarısını olumsuz olarak etkileyebilmektedir. Bu tarz değişikliklerin analiz edilerek model üzerinde oluşan olumsuz durumların giderilmesi gerekir. Ayrıca zamanla modelde kullanılan algoritmaların veya yöntemlerin değiştirilmesi gibi işlemler de yapılabilir (Talan, 2016).

5. VERİ MADENCİLİĞİ METOTLARI

Veri madenciliği fonksiyonları açısından genel olarak, tahmin edici (predictive) metotlar ve tanımlayıcı (descriptive) metotlar olarak iki temel kategoriye ayrılabilir (Han vd., 2012). Veri madenciliği metotlarının genel hiyerarşisi Şekil 5.1.'de görüldüğü gibidir. Bu yöntemlerin detayları başlıklar halinde ilerleyen bölümde açıklanmıştır.



Şekil 5.1. Veri Madenciliği Metotları Gösterimi

5.1.Tahmin Edici Metotlar ve Denetimli/Gözetimli (Supervised) Öğrenme

Tahmin edici metotlar, eldeki verileri kullanarak geleceğe ilişkin bir olayın sonuçlarını tahmin etmek için kullanılırlar. Mevcut ve önceden tanımlanmış bir sınıfın özelliklerini inceleyerek bir eğitim süreci sonunda sonuç sınıfı ile ilgili gerekli çıkarımları yaparlar. Yeni bir örnek geldiğinde, bu yeni örneği önceden tanımlanmış ve özellik çıkarımları yapılmış olan uygun sınıfa atama işlemini gerçekleştirirler. Sınıflar önceden belirli olduğu için yeni örneğin sonucunun kategorik veya rakamsal olarak alabileceği değer kümeleri bellidir. Tahmin edilen sonuçların kalitesi yani ne doğrulukta tahmin edildiği de sonucun ne olduğu kadar önemlidir. Dolayısıyla tahmini sonuç ile birlikte tahminin kalitesine yönelik güven aralığı, olasılık vb.

tahminin doğruluğunu gösteren istatistiki değerlerinde belirlenmesi önemlidir (Argüden ve Erşahin, 2008).

Tahmin edici metotlar, bilinmeyen sonuçlara erişirken sonuçları bilinen bir öğrenme kümesini kullandığından dolayı denetimli/gözetimli (supervised) öğrenme olarak da adlandırılırlar. Test verisi üzerindeki özelliklerin öğrenme verisine uyumu, başarı ölçütü olarak ifade edilmektedir. Denetimli öğrenme metotlarına örnek olarak sınıflandırma ve regresyon analizi metotları verilebilir (Erduran, 2017).

Veri madenciliği metotları arasında sınıflandırma ve regresyon analizi en yaygın olarak kullanılan metotlardır. Sınıflama ve regresyon modelleri arasındaki temel fark, tahmin edilen bağımlı değişkenin (sonuç değişkeni) kategorik ya da süreklilik değere sahip olması ile ilgilidir (Ergün, 2008). Eğer bulunması hedeflenen sonuç değişkeni kesikli ise tahmin modeli için sınıflandırma metodu, eğer sonuç değişkeni sürekli ise tahmin modeli için regresyon metodu kullanılır (Şık, 2014).

5.1.1. Sınıflandırma Metodu

Sınıflandırma metodu, veri madenciliği metotları içinde en çok kullanılanıdır. Verilerin sınıflandırılması işleminde, test verisi ve öğrenme verisi olarak iki veri kümesi, öğrenme adımı ve sınıflama adımı olarak da iki adımı mevcuttur. Öğrenme adımında, veri madenciliği algoritmaları öğrenme verisini kullanarak analizler yapar ve çeşitli kural çıkarımları oluştururlar. Amaç bir niteliğin değerini diğer nitelikler yardımıyla tespit etmektir. Yeni bir veri geldiğinde oluşturulan bu kurallar kullanılarak yeni veri hakkında nasıl karar verileceği belirlenir. Sınıflama adımında ise, kuralların doğruluğunu ölçmek için test verisi üzerinde öğrenilen kurallar denenir. Kuralların doğruluğu kabul edilebilir bir seviyede ise, o zaman kurallar yeni verinin sınıflandırılmasında kullanılabilirler (Türkoğlu, 2016).

Sınıflandırılma modellerinde yaygın olarak kullanılan teknikler aşağıda belirtilmiş ve bu teknikler ilerleyen bölümlerde detaylı olarak açıklanmıştır. Yaygın olarak kullanılan sınıflandırma teknikleri aşağıda belirtilen tekniklerdir:

- Karar Ağaçları (Decision Trees)
- Bayes Sınıflandırması (Bayesian Classification)
- K-En Yakın Komşu (K-Nearest Neighbour)
- Yapay Sinir Ağları (Artificial Neural Networks)
- Genetik Algoritmalar (Genetic Algorithms)
- Destek Vektör Makineleri (Support Vector Machines)

5.1.2. Regresyon Analizi Metotları

Regresyon modelleri, süreklilik gösteren değerlerin tahmini için kullanılırlar. Amaçları, girdiler ile çıktı arasındaki ilişkileri tespit ederek buna göre bir model oluşturmak ve en doğru tahmini bulmaya çalışmaktır. Sonuç değişkeni 'bağımlı değişken' olarak girdi değişkenleri ise 'bağımsız değişken' olarak tanımlanır. Bağımlı değişkenin (sonuç değişkeni) alacağı değer genellikle bir güven aralığı içinde tespit edilir. Girdiler duruma göre bir veya daha çok sayıda olabilir. Gerçek hayattaki problemlerin neredeyse tamamında doğru tahmini bulabilmek için birden fazla girdi değişkeninden yararlanmak gerekir. Girdi değişkenlerinin sonucun doğru tahmin edilmesine ne oranda katkı yaptıkları önemlidir. Bazen sonuca katkısı çok küçük oranda olan girdi değişkenlerini modelden atmak, daha verimli bir model oluşturulmasını sağlayabilir (Yurdakul, 2015).

Regresyon modelleri matematiksel olarak da Eşitlik 5.1.'de görüldüğü gibi gösterilebilir: Bağımsız değişkenin veya değişkenlerin (girdi değişkenlerinin) katsayıları olan parametreler ($q_{1,\ldots,}q_n$) ve bağımlı değişkenin (sonuç değişkeni) üzerindeki etkisini tespit etmek amacıyla, bağımlı değişken ile tahmin edilen değerin arasındaki fark hata terimi (e) olmak üzere,

$$y=F(x,q) + e \qquad (5.1)$$

eşitliğini sağlayan en F fonksiyonu için q değerinin tespit edilmesi süreci regresyon analizi olarak tanımlanır (Şık, 2014). Regresyon modeli tek bir adet bağımsız değişkenden oluşuyorsa basit doğrusal regresyon, birden fazla bağımsız değişkenden oluşuyor ise çoklu regresyon modeli olarak adlandırılır (Erduran, 2017).

Sınıflandırma modelleri ve regresyon modelleri arasındaki en temel fark; sınıflandırma modelleri kategorik değerleri tahmin etmek için kullanılırken, regresyon modelleri ise süreklilik gösteren değerleri tahmin etmek için kullanılmaktadır. Fakat çok terimli lojistik regresyon gibi kategorik değerlerin tahminine imkan sağlayan regresyon tekniklerinin geliştirilmesi gibi nedenlerden dolayı regresyon ve sınıflandırma modelleri gittikçe birbirine yaklaşmaktadır (Kılıç, 2014).

5.2.Tanımlayıcı Metotlar ve Denetimsiz/Gözetimsiz (Unsupervised) Öğrenme

Tanımlayıcı metotların amacı veri kümesinde bulunan veriler arasındaki ilişkileri, bağlantıları ve örüntüleri keşfetmektir. Tahmin edici metotlarda olduğu gibi bir öğrenme süreci sonucunda, veriyi önceden belirli bir sonuç değişkenine atamazlar. Eldeki mevcut verileri kendi içlerinde değerlendirerek davranış biçimlerini tespit etmeyi ve aynı davranış özelliklerini gösteren alt veri setlerini belirlemeye çalışırlar (Argüden ve Erşahin, 2008). Yani kısacası sonuç değişkeninin özelliklerinden yola çıkarak bir öğrenme metodolojisi geliştirmezler, verilerin davranış biçimlerinin tüm özelliklerini dikkate alarak kendileri oluşturmaya çalışırlar.

Tanımlayıcı metotlar, denetimsiz/gözetimsiz (unsupervised) metotlar olarak da adlandırılırlar. Herhangi bir bilgi başlangıç bilgisi verilmediğinden dolayı denetimsiz öğrenmenin sonuçları kesin doğruluk içermeyebilir. Kümeleme ve birliktelik analizi metotları denetimsiz öğrenme yöntemlerine örnek olarak verilebilir (Erduran, 2017).

5.2.1.Kümeleme Metotları

Kümeleme metodu ilk kez Londra'da ortaya çıkan kolera salgınında çok sayıda can kaybının yaşanması üzerine, bu durumun çözümü için kullanılmıştır. John Snow adında bir kişi harita üzerinde ölen kişilerin yerlerini işaretlediğinde bazı bölgelerde işaretlerin yoğunlaştığını fark ediyor. O bölgelerdeki su pompaları incelendiğinde ana sokaklardan birinde atık su tesisindeki problem fark edilmiştir. Su tesisindeki problemin çözülmesi kolera salgının sonlaması için yeterli olmuş ve koleradan

kaynaklı ölümler engellenmiştir. Günümüzde de kümeleme analizi ile istatistik, makine öğrenmesi ve örüntü tanıma gibi birçok alanda araştırmalar yapılmakta ve ilginç sonuçlar keşfedilmektedir (Türkoğlu, 2016).

Kümeleme metotları, denetimsiz ve tanımlayıcı metotlar içerisinde yer almaktadır. Bu metotların sınıflandırma ve regresyon metotlardan farkı, bulunan mevcut verileri daha önceden belirli olan bir sınıflandırmaya göre değil, belirli olmayan ve kendi keşfettiği bir bölümlemeye göre gruplandırmasıdır. Kümeleme analizi ile, veri setinde doğal şekilde meydana gelen alt gruplar keşfedilmeye çalışılır (Özcan, 2014). Yani sınıflandırmada sınıflar önceden belli iken, kümelemede oluşan gruplar önceden belli değildir ve kümeleme işlemi sonrasında oluşur. Grupları oluşturma aşamasında aynı gruba atanacak verilerin birbirleri ile benzerliklerinin artırılmasına, gruplar arasında ise benzerliklerin azaltılmasına çalışılır (Paçaman, 2014).

Grupların ayrımı benzerlik ve uzaklık matrisleri kullanılarak yapılır. Benzerlik, iki nesne arasındaki ilişkinin kuvveti biçiminde tanımlanabilir. Uzaklık ise, iki nesne arasındaki zıtlık veya farklılık olarak tanımlanabilir. Nesneler benzerlik ve uzaklık ilişkilerine göre birbirlerinden ayırt edilir ve böylece veri kümeleri alt gruplara ayrılmış olur. Nicel veriler için bu ölçümler çeşitli matematiksel yöntemler ile yapılmaktadır (Uğurlu, 2015).

Uzaklık veya benzerlik ölçüleri veri kümesinde bulunan değişkenlerin ölçü birimlerine göre değişiklik göstermektedir. Değişkenlerin oransal, kesikli, sayısal ya da ikili (binary) değişken olmasına göre literatürde yer alan çeşitli benzerlik ve farklılık ölçülerinden kullanılacak olan uzaklık ölçüsü farklılık gösterecektir (Ergün,2008).

Kümeleme metodu denetimsiz öğrenme uygulaması olmasından dolayı küme sayısının önceden belirlenmemesi gerektiği halde kümeleme algoritmalarının birçoğu küme sayısının analizden önce belirlenmesini istemektedir. Dolayısıyla kullanıcılar, farklı küme sayılarını denemekte ve sonucun doğruluğunu test ederek optimum küme sayısını bulmaya çalışmaktadırlar. Ayrıca optimum küme sayısını bulmaya yarayan literatürde kullanılan bazı indeksler de bulunmaktadır (Erduran, 2017).

Literatürde çok sayıda kümeleme metodu mevcuttur. Kullanılacak olan algoritmanın seçimi, çalışmanın amacına ve verinin tipine göre değişecektir (Ergün, 2008). Kümeleme analizi, gruplar içi ön tahmin yapılması, veri yapısının belirlenmesi, verilerin indirgenmesi ve ayrık değerlerin tespiti gibi amaçlar için de kullanılmaktadır (Erduran, 2017).

5.2.2.Birliktelik Analizi Metotları

Veri kümesi incelenerek, birlikte gerçekleşen olayların belirli olasılıklarla tespit edildiği veri madenciliği yöntemlerine birliktelik analizi adı verilir (Canlı, 2017). Bu yöntemlerde kurallar değerlendirilirken destek ve güven ölçütleri adında iki ölçüt kullanılır. Destek ölçütü; veri kümesinde birlikte bulunan nesnelerin sayısının veri kümesinde bulunan olay sayısına oranıdır. Güvenilirlik ölçütü ise; X seçeneği gerçekleştiğinde Y seçeneğinin de gerçekleşme olasılığıdır. X ve Y seçeneklerinin birlikte bulunduğu olay sayısının X'lerin bulunduğu olay sayısına bölünmesi ile elde edilir (Çığşar, 2017). Destek ve güven ölçütlerinin minimum eşik değerleri belirlenir ve bu eşik değerlerinden büyük olan değerlerin yer aldığı birliktelik kuralları dikkate alınır, eşik değerlerinin altında kalan kurallar ise dikkate alınmaz (Argüden ve Erşahin, 2008). Destek ve güven değerleri ne kadar büyük olursa aradaki birlikteliğin o derece kuvvetli olduğu ifade edilebilir.

Birliktelik analizi kuralları çoğunlukla pazarlama alanında kullanılmaktadır. Birliktelik kurallarından hareketle pazar sepet analizi (market sepet analizi) olarak bilinen veri madenciliği uygulamalarıyla müşterilerin alışveriş davranışları belirlenmeye çalışılır. Bir müşteri bir ürün aldığında, bu ürünle birlikte diğer ürünlerden başka hangisi ya da hangilerini de aldığının belirli bir olasılığa göre tespit edilmesi amaçlanır. Bu ürünler tespit edildiğinde, mağazaların rafları bu sonuçlara göre düzenlenerek bu ürünleri müşterilere daha kolay ulaştırılması hedeflenir (Köse, 2015). Ayrıca bu veriler kampanya, promosyon düzenleme gibi stratejilerde kullanılabilirler (Can, 2017).

Birliktelik kurallarını tek boyutlu birliktelik kuralları ve çok boyutlu birliktelik kuralları şeklinde ikiye ayırabiliriz. Tek boyutlu birliktelik kuralına iki ürünün

arasındaki satın alınma ilişkisini örnek verilebilirken, çok boyutlu birliktelik kuralına hava durumu, yer ve gün gibi birden fazla özelliğe göre değişebilen satın alma ilişkisinin incelenmesi örnek olarak verilebilir (Erkuş, 2015).

Ardışık zamanlı örüntüler adı verilen ve birliktelik analizine benzeyen bir diğer veri madenciliği metodunda ise birbirini izleyen dönemlerde gerçekleşen olayların birbirleri arasındaki ilişkiler incelenir. Birliktelik analizinde aynı zamanlı olayların arasındaki birliktelikler tespit edilirken, ardışık zamanlı örüntüler de ise birbirini izleyen dönemlerde gerçekleşen birliktelikler göz önünü alınır. "X ve Y ürünlerini alan müşteriler %90 ihtimalle Z ürününü alır" biçimindeki çıkarımlar birliktelik analizi kurallarına örnek verilebilirken "K ameliyatı olan bir kişinin on gün içinde %50 ihtimalle L enfeksiyonu oluşacaktır" biçimindeki zaman faktörünü de göz önüne alan çıkarımlar ise ardışık zamanlı örüntüler kurallarına örnek olarak verilebilirler (Ciga, 2015).

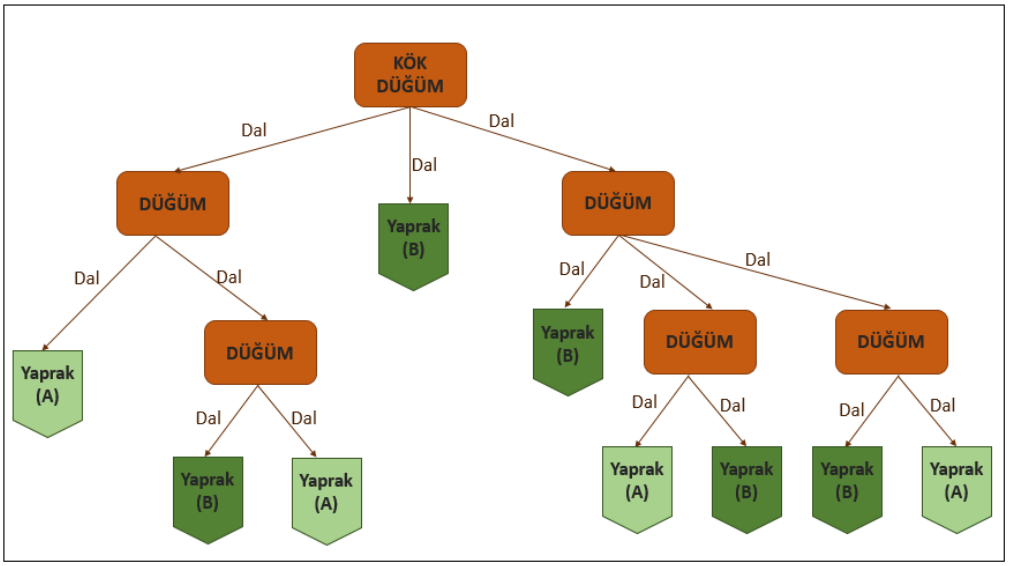
İçerisinde kategorik değişken olan her yerde birliktelik kuralları uygulanabilir. Birliktelik kurallarında en bilinen ve yaygın olarak kullanılan algoritma "Appriori Algoritmasıdır" (Diler, 2016).

6. VERİ MADENCİLİĞİNDE SINIFLANDIRMA TEKNİKLERİ

Bu bölümde, bu tezin de konusunu oluşturan sınıflandırma metodu için literatürde yoğun olarak kullanılan tekniklerden bahsedilecektir.

6.1. Karar Ağaçları

Adından da anlaşıldığı üzere tıpkı bir ağacın dal ve yaprakları gibi açılımlara sahip olan, karar düğümlerinden, dal ve yapraklardan oluşan yapılara karar ağaçları denilmektedir. Karar ağaçlarında sınıflandırma işlemi kök düğüm ile başlar, daha sonra düğümler alt dallarına ayrılırlar ve bu işlem yaprak elde edilinceye kadar devam eder. Burada düğümler karar işleminin uygulandığı adımlardır, düğümlerin sonuçlarına göre dallar meydana gelir ve elde edilen dallar ile tahmin sonuçlarının bulunduğu yapraklara (karar sınıfına) gidilir. Fakat dalların sonucunda bir sınıflama tamamlanmıyorsa tekrardan düğüm oluşturulur ve yaprağa yani karar sınıfına ulaşana kadar işlem tekrarlanır (Diler, 2016). Şekil 6.1.'de örnek bir karar ağacı görülmektedir.



Şekil 6.1. Karar Ağacı Örneği

Karar ağaçları hem sınıflandırma hem de regresyon modelleri için kullanılabilmektedirler (Armutlu, 2018). Karar ağaçları, kolay modellenebilmeleri, kolay yorumlanabilmeleri ve güvenilirliklerinin yüksek olmaları nedeniyle veri madenciliği yöntemleri arasında en fazla kullanılan sınıflandırma metodudur (Levent, 2016).

Karar ağacı algoritmalarına örnek olarak ID3 (Iterative Dichotomiser 3), C4.5 Algoritmaları, CART (Classification And Regression Trees) Algoritmaları, Rastgale Orman (Random Forest) Algoritmaları örnek olarak gösterilebilir.

Kök düğümün ne olacağı, dallanmaların neye göre yapılacağı kullanılan karar ağacı algoritmasına göre farklılık göstermektedir (Atasoy, 2015). Karar ağacı oluşturma işleminde kullanılan algoritma değişirse, ağacın yapısı da değişir ve dolayısıyla sınıflama işleminin sonuçları da değişmiş olacaktır (Can, 2017).

Entropi adı verilen ölçüm, hangi özelliklerin hangi sırada kullanılarak karar ağacının oluşturulacağının belirlenmesinde kullanılan bir ölçüttür ve bu amaçla kullanılan ölçütlerin en yaygın olanıdır. Entropi ölçümü ne kadar fazla ise, elde edilecek olan sonuçlarda o oranda belirsiz ve kararsızdır. Dolayısıyla özellikler için entropi hesabı yapıldığında entropi değeri en az olan özellik kök düğüm olarak kullanılır (Uyumaz, 2017). Diğer alt dallanmalarda da entropi hesabı göz önüne alınarak açılım yapılır.

Karar ağaçları oluşturulduktan sonra ise karar kuralları oluşturulur ve oluşturulan kurallar veri kümesine yeni gelen verilerin sınıflarının tahmini için kullanılır. Örneğin kredi durumunun sınıfını tahmin eden bir modelde “yaş =41-50 arası ve gelir=yüksek ise kredi durumu=mükemmel” şeklinde bir kural çıkarımı yapılmış olsun. Bu veri kümesine geliri yüksek olan 45 yaşında yeni bir kişi geldiğinde model bu kişinin kredi durumu tahminini mükemmel olarak sınıflandıracaktır (Yurdakul, 2015).

6.2. Bayes Sınıflandırıcılar

Bayes sınıflandırıcılar, mevcut sınıflandırılmış olan verileri kullanarak yeni gelen verinin hangi bir sınıfa dahil olabileceğinin ihtimalini tahmin eden istatistik tabanlı yöntemlerdir. Öncelikle mevcut öğrenme kümesindeki değerlerin ve sınıfların bulunma sıklığını hesaplayarak yeni gelen veri için bu hesaplara göre hangi sınıfa dahil olabileceğinin olasılığını tahmin ederler (Fakı, 2015).

Bayes sınıflandırıcıların temeli Bayes Teoremine dayanır. Bayes sınıflandırıcılar, kolay uygulanabilirliği ve hızlı hesaplama yeteneği ile veri madenciliği sınıflandırma algoritmaları içinde araştırmacılar tarafından tercih edilen algoritmalar arasında yer alırlar (Uyumaz, 2017).

Bayes yöntemi koşullu olasılığı temel alır. Eşitlik 6.1'de görülen $P(C \mid X)$, "X" in olması durumunda "C" nin olması olasılığı ve P(C) sınıfın olasılığıdır (Altun, 2018). Eşitlik 6.1'de paydada görülen P(X), Eşitlik 6.2'de görüldüğü gibi yazılabilir ve Bayes Teoremi Eşitlik 6.3'de görüldüğü gibi ifade edilebilir (Türker, 2013).

$$P(C \mid X) = \frac{P(X|C) \cdot P(C)}{P(X)} \quad (6.1)$$

$$P(X) = P(C_1) \cdot P(X \mid C_1) + \ldots + P(C_k) \cdot P(X \mid C_k) = \sum_{j=1}^{k} P(Cj) \bullet P(X \mid Cj) \quad (6.2)$$

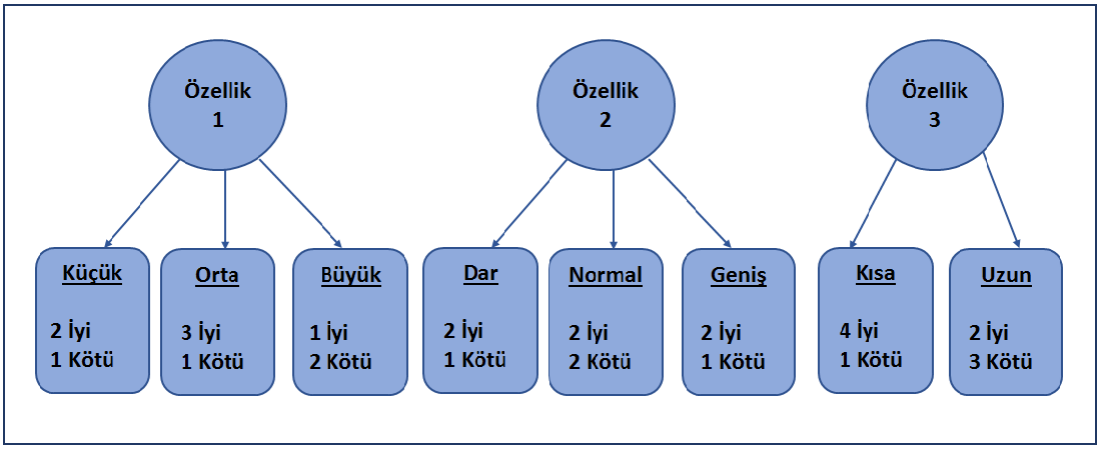
$$P(C_i \mid X) = P(X \mid C_1) \bullet P(C_1) / [\sum_{j=1}^{k} P(Cj) \bullet P(X \mid Cj)], \quad (i=1,2,\ldots,k) \quad (6.3)$$

Bayes sınıflandırmanın nasıl yapıldığı, Çizelge 6.1.'de görülen veri seti üzerinde yapılan örnek bir hesaplama ile aşağıda belirtilmiştir.

Çizelge 6.1. Naive Bayes Algoritması Hesabı İçin Örnek Veri Seti

ÖZELLİK 1	ÖZELLİK 2	ÖZELLİK 3	SINIF
Küçük	Dar	Kısa	İyi
Büyük	Normal	Uzun	Kötü
Orta	Normal	Kısa	İyi
Orta	Dar	Kısa	İyi
Orta	Geniş	Kısa	İyi
Büyük	Normal	Kısa	Kötü
Küçük	Geniş	Uzun	İyi
Orta	Dar	Uzun	Kötü
Küçük	Geniş	Uzun	Kötü
Büyük	Normal	Uzun	İyi

Örnek veri setimizde bulunan 10 adet örnekten, 6 adet örnek iyi sınıfına ait 4 adet örnek ise kötü sınıfına aittir. Dolayısı ile iyi sınıfına ait olma ihtimali P(İyi) = 0,6 iken, kötü sınıfına ait olma ihtimali ise P(Kötü) = 0,4 olarak bulunur. Verilerin özelliklerine göre dağılımı ise Şekil 6.2.'de görülmektedir.



Şekil 6.2. Naive Bayes Hesaplaması İçin Verilerin Özelliklere Göre Dağılımı

Özellik 1 için hesaplanan olasılıklar Çizelge 6.2.'de, Özellik 2 için hesaplanan olasılıklar Çizelge 6.3.'de, Özellik 3 için hesaplanan olasılıklar Çizelge 6.4.'de gösterilmektedir.

Çizelge 6.2. Naive Bayes Örneği Özellik 1 İçin Olasılıklar

	İyi	Kötü	Toplam
Küçük	2/6	1/4	3/10
Orta	3/6	1/4	4/10
Büyük	1/6	2/4	3/10

Çizelge 6.3. Naive Bayes Örneği Özellik 2 İçin Olasılıklar

	İyi	Kötü	Toplam
Dar	2/6	1/4	3/10
Normal	2/6	2/4	4/10
Geniş	2/6	1/4	3/10

Çizelge 6.4. Naive Bayes Örneği Özellik 3 İçin Olasılıklar

	İyi	Kötü	Toplam
Kısa	4/6	1/4	5/10
Uzun	2/6	3/4	5/10

Veri setimize yeni gelen ve nitelikleri "Özellik 1: Büyük, Özellik 2: Dar, Özellik 3: Kısa" olan bir örnek için, Bayes Teoreminin örneği hangi sınıfa dahil edeceği aşağıda bahsedilen şekilde hesaplanabilir. Bahsedilen yeni gelen örneğin Bayes Teoremine göre iyi sınıfına atanma olasılığı Eşitlik 6.4' de, kötü sınıfına atama olasılığı Eşitlik 6.5'de gösterildiği gibidir.

$$P(\text{İyi} \mid X) = \frac{P(X \mid \text{İyi}) \bullet P(\text{İyi})}{P(X)} \tag{6.4}$$

$$P(\text{Kötü} \mid X) = \frac{P(X \mid \text{Kötü}) \bullet P(\text{Kötü})}{P(X)} \tag{6.5}$$

Eşitlik 6.4 ve Eşitlik 6.5'yer alan de formüllerin içinde bulunan olasıklar, Eşitlik 6.6, Eşitlik 6.7, Eşitlik 6.8, Eşitlik 6.9'da görüldüğü gibi hesaplanabilir.

$$P(X \mid \text{İyi}) = P(\text{Büyük} \mid \text{İyi}) \bullet P(\text{Dar} \mid \text{İyi}) \bullet P(\text{Kısa} \mid \text{İyi}) = 1/6 \bullet 2/6 \bullet 4/6 \quad (6.6)$$

$$P(X \mid \text{İyi}) = 1/6 \bullet 2/6 \bullet 4/6 = 0{,}037 \quad (6.7)$$

$$P(X \mid \text{Kötü}) = P(\text{Büyük} \mid \text{Kötü}) \bullet P(\text{Dar} \mid \text{Kötü}) \bullet P(\text{Kısa} \mid \text{Kötü}) \quad (6.8)$$

$$P(X \mid \text{Kötü}) = 2/4 \bullet 1/4 \bullet 1/4 = 0{,}031 \quad (6.9)$$

Eşitlik 6.4 ve Eşitlik 6.5'in paydasında görülen P(X) olasılığı ise Eşitlik 6.10 ve 6.11'de gösterildiği gibi hesaplanabilir.

$$P(X) = P(\text{İyi}) \bullet P(X \mid \text{İyi}) + P(\text{Kötü}) \bullet P(X \mid \text{Kötü}) \quad (6.10)$$

$$P(X) = 0{,}6 \bullet 0{,}037 + 0{,}4 \bullet 0{,}031 = 0{,}0346 \quad (6.11)$$

Bulunan değerler Bayes Teoremi formülünde yerlerine konulduğunda yeni gelen örneğin iyi sınıfına atanma olasılığı Eşitlik 6.12'de görüldüğü gibi, kötü sınıfına atanma olasılığı ise Eşitlik 6.13'de görüldüğü gibi bulunur.

$$P(\text{İyi} \mid X) = \frac{P(X \mid \text{İyi}) \bullet P(\text{İyi})}{P(X)} = \frac{0{,}037 \bullet 0{,}6}{0{,}0346} = 0{,}642 \quad (6.12)$$

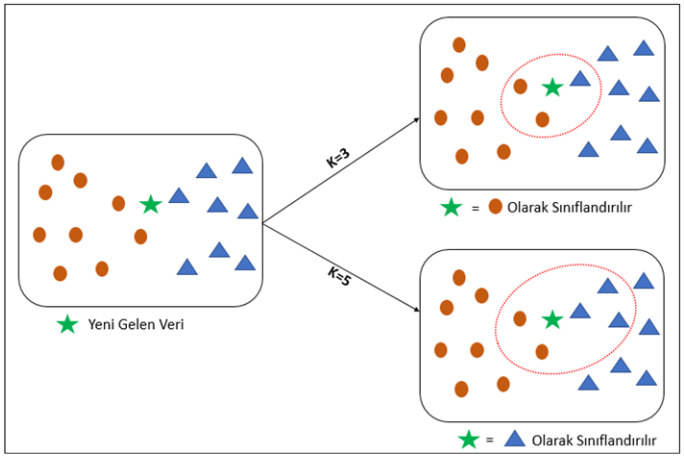
$$P(\text{Kötü} \mid X) = \frac{P(X \mid \text{Kötü}) \bullet P(\text{Kötü})}{P(X)} = \frac{0{,}031 \bullet 0{,}4}{0{,}0346} = 0{,}358 \quad (6.13)$$

Eşitlik 6.12'de bulunan iyi sınıfına atanma olma olasılığı, Eşitlik 6.13'de bulunan kötü sınıfına atanma olasılığından daha yüksek bulunduğundan dolayı, örnekte nitelikleri bahsedilen yeni bir veri geldiği zaman, Bayes Teoremi bu yeni örneğin sınıfını "iyi" olarak belirleyecektir.

6.3. K-En Yakın Komşu (K-Nearest Neighborhood, KNN)

K-En Yakın Komşu Algoritması ile yeni gelen bir veri örneği sınıflandırılmak istenildiğinde, o veri örneğinin etrafında bulunan ve daha önceden sınıflandırılmış olan k adet en yakın komşusuna olan uzaklığını ölçerek yeni örneği adet olarak fazla olan komşularının bulunduğu sınıfın içine dahil eder. Burada önce bir k değeri belirlenmesi gerekir. Algoritma, belirlenen k adet en yakın komşusuna bakarak çalışacağı için belirlenecek k değerine göre algoritmanın başarısı değişecektir. Yeni örnek ile eğitim seti içerisindeki uzaklıkları ölçmek için çeşitli uzaklık yöntemleri kullanılmaktadır. Öklid uzaklığı, Minkowski uzaklığı, Manhattan uzaklığı, Chebyschev uzaklığı kullanılan uzaklık ölçme yöntemlerine örnek olarak gösterilebilir (Yakupoğlu, 2018).

K-En Yakın Komşu Algoritması, büyük boyutlu veri kümelerinde oldukça yüksek performansta sonuçlar vermektedir. Öğrenme kümesinde yeterli sayıda veri bulunması, öznitelik sayısı, belirlenen k sayısı ile kaç en yakın komşuya bakılacağı, kullanılan uzaklık ölçütü ve ağırlıklandırma yönteminin kriterleri gibi etkenler yöntemin performansını etkilemektedir (Güzel, 2018). K sayısının farklı alınması durumunda sınıflandırmanın nasıl etkilendiği Şekil 6.3.'de gösterilmiştir. Şeklin sol tarafında yer alan veri setine, yıldız ile temsil edilen yeni veri örneği geldiğinde, k=3 alınan durumda, belirlenmek istenen yeni örnek A sınıfına atanırken, k=5 durumunda ise belirlenmek istenen yeni örnek B sınıfına atanmaktadır. Veri seti için en uygun olan k değeri ise farklı denemeler yapılarak tespit edilebilir.

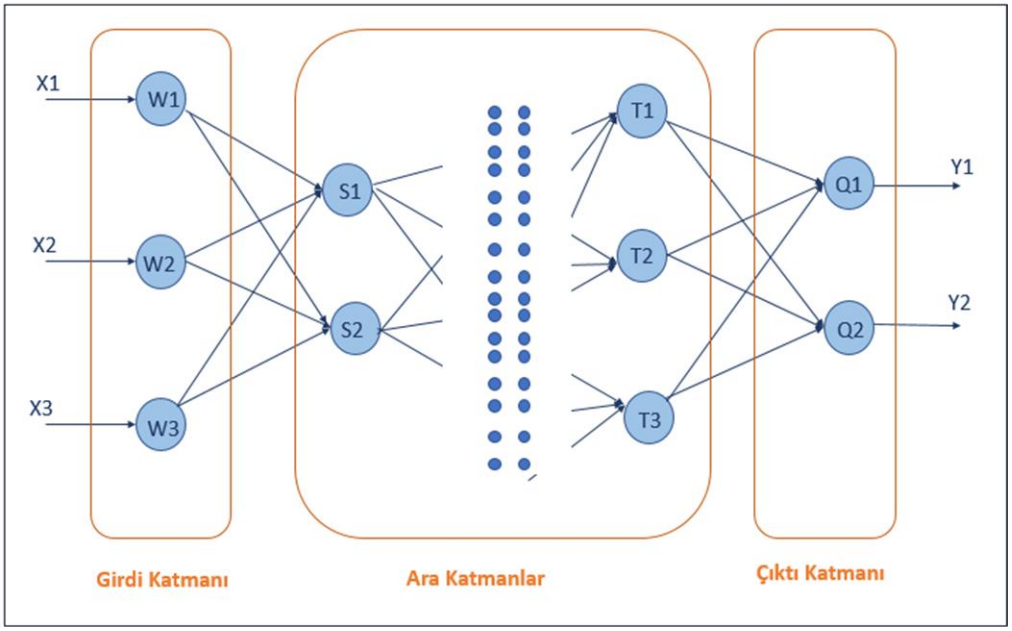


Şekil 6.3. K Sayısının Sınıflandırmaya Etkisi

6.4. Yapay Sinir Ağları

Yapay sinir ağlarının en genel tanımı; "İnsan beyni ve biyolojik sinir sisteminin işleyiş ve öğrenme mekanizmasını, elektronik ortamlar üzerinde taklit etmeye çalışan sistemlerdir" şeklinde yapılabilir (Yakut, 2012). Biyolojik sinir sistemlerinde öğrenme nöronlar arasında meydana gelen iletişim ile olurken, yapay sinir ağlarında öğrenme ise katmanlar arasında gelişen bir dizi döngülerin tekrar etmesiyle gerçekleşir (Yıldız, 2019).

Yapay sinir ağları yapısı genel olarak birkaç katmandan oluşur. İlk katman olan girdi katmanı verilerin analiz edilmek üzere fonksiyona girmesi için kullanılır. Son katman olan çıktı katmanı ise öğrenme sonucunda elde edilen sonuçların ortaya konduğu katmandır. Girdi ve çıktı katmanları arasında ise işlemlerin gerçekleştirildiği sayısı birden fazla olabilen gizli katmanlar vardır. Çok katmanlı yapılarda ilk gizli katmandan elde edilen sonuçlar bir sonraki katmanın girdileri olacak şekilde kademeli olarak ilerler (Yalçın, 2019). Örnek bir yapay sinir ağı yapısı Şekil 6.4.'de görüldüğü gibidir.



Şekil 6.4. Yapay Sinir Ağı Örneği

Yapay sinir ağlarında ana etken ağı kontrol edecek ya da yönetecek kişinin değişkenlere atadığı rollerdir. Her bir değişken bir ağırlık vektörü ile ifade edilir ve çıktı değişkenine ağırlığı ölçüsünde etki eder (Yıldız, 2019). Yapay sinir ağlarında öğrenme süreleri uzun olduğu için, süre problemi olmayan veya verilerin boyutunun çok büyük olmadığı durumlarda kullanılmaları daha uygundur (Yalçın, 2019).

6.5. Genetik Algoritmalar

Genetik algoritmalar biyolojiden esinlenerek çaprazlama, mutasyon ve gen gibi terimlerden yararlanıp en iyi çözüme ulaşmayı hedefler. Genetik algoritma çalıştırıldığında, problem öncelikle bir başlangıç çözümü oluşturmak için kodlanır. Daha sonra biyolojide olduğu gibi seçim, çaprazlama ve mutasyon yoluyla yeni değerler üretilir. Sonlandırma kriterine uyan çözümü bulana kadar döngü bu şekilde devam eder (Özaltındiş, 2018).

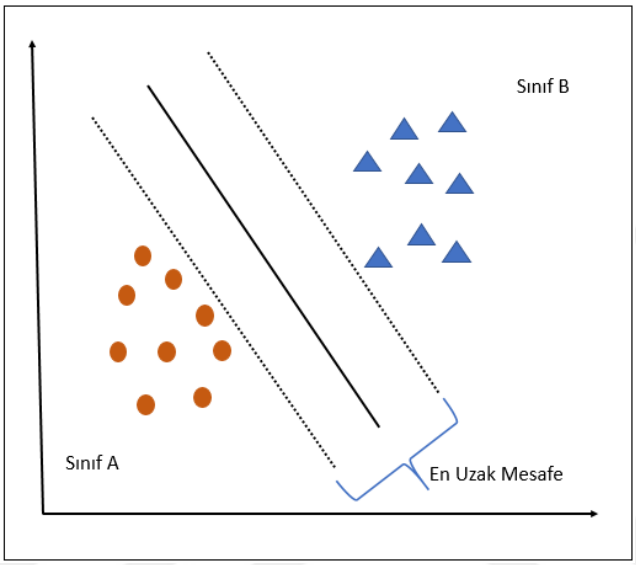
Gen, popülasyonu oluşturan bireylerin kod yapısındaki en küçük birimdir. Bir genin yapısında genelde ikili tabandaki sayılar kullanılmakla birlikte tamsayılar, karakterler, ağaç biçimi vb. farklı semboller de kullanılabilmektedir. Kromozom ise birden çok sayıda genin bir araya gelmesiyle oluşan yapıdır ve her bir kromozom bir bireyi temsil eder. Bireyler aşağıda göründüğü şekilde 0-1 gibi ikili bitlerle, tamsayılarla, rasyonel değerlerle ya da karakterlerle kromozom yapılarına göre farklı şekillerde kodlanabilirler (Gitmez, 2018):

- Kromozom A= 101100110
- Kromozom B=43552346
- Kromozom C= 4.1 5.3 4.2 1.5
- Kromozom D=CBDAABCDC

6.6. Destek Vektör Makineleri

Destek vektör makineleri sınıfları birbirlerinden ayırmak için kullanılan bir denetimli öğrenme algoritmasıdır. İki sınıfı birbirlerinden ayırabilmek için, optimum şekilde sınıfların aralarında bulunan en uzak mesafeyi tespit etmeye çalışır. En uzak

mesafeyi tespit ettikten sonra, buraya bir hiperdüzlem çizer. Çizilen bu hiperdüzlem sayesinde sınıflar birbirinden ayrılmış olur ve böylece sınıflandırma işlemini gerçekleştirilir (Saylan, 2018). Destek vektör makinelerinin bir örneği Şekil 6.5.'de gösterilmektedir. En uzak mesafenin ortasından geçen doğru ile sınıflar arasına bir sınır çizilmekte ve yeni gelen veri hangi sınıf tarafında kalıyor ise o sınıfın tarafına atama yapılarak sınıflar belirlenmiş olmaktadır.



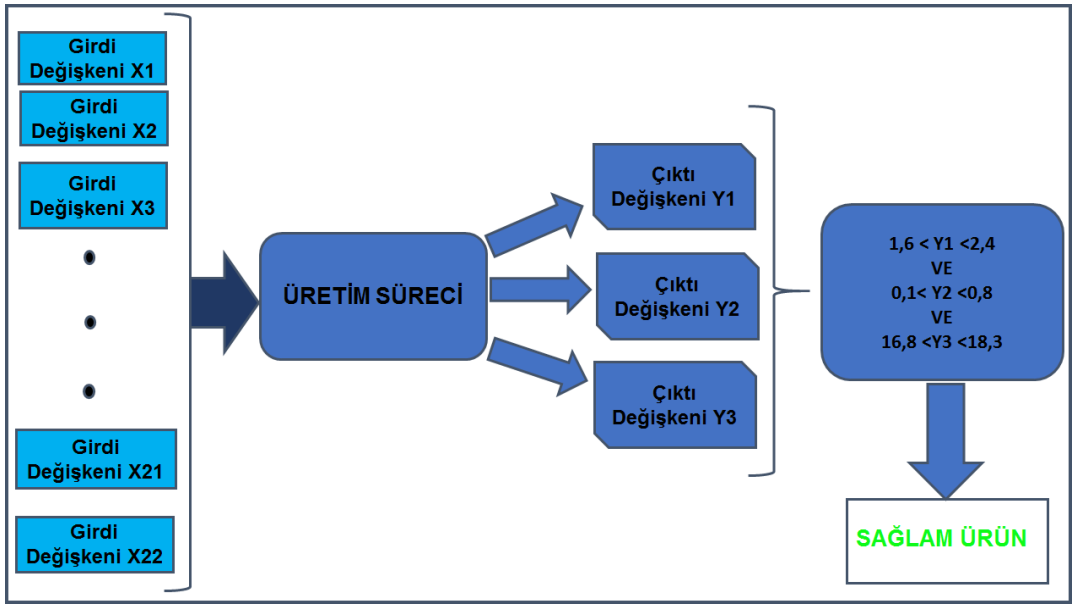
Şekil 6.5. Destek Vektör Makineleri Örneği

7. VERİ MADENCİLİĞİ UYGULAMASI

Uygulama, bir üretim işletmesinin kaynak atölyesinde üretilen bir ürünün fiili üretim verileri kullanılarak gerçekleştirilmiştir. Ürün kalitesini belirleyen üç adet çıktı değişkenine etki eden yirmi iki adet girdi değişkeni ile, WEKA adlı makine öğrenmesi yazılımı kullanılarak veri madenciliği uygulaması yapılmıştır. Firma gizliği nedeniyle, tüm veriler belirli bir katsayı ile değiştirilmiş ve firma ismi ile değişkenlerin isimleri çalışmada paylaşılmamıştır. Veri setinin örnek bir kesiti Çizelge 7.1.'de, ürün kalitesini tespit etmede kullanılan metodolojinin şeması ise Şekil 7.1.'de gösterilmektedir.

Çizelge 7.1. Veri Seti Kesiti Örneği

No	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16	X17	X18	X19	X20	X21	X22		Y1	Y2	Y3
1	0,076	0,18	19,73	4,244	15,75	1	1	15,8	21	21,2	7,8	72,4	14,4	23,4	228,6	21,8	2,4	2,6	2,6	0,2	77,2	117,8		2,190	0,214	17,490
18	0,088	0,204	19,69	4,244	15,75	1	1	16,2	21,6	20,6	7,2	72,6	15	23	229,8	21,8	2,8	2,2	2,6	0,4	61,2	115,6		2,162	0,250	17,478
19	0,088	0,204	19,69	4,244	15,75	1	1	17,2	20,2	20,8	8	73	14,8	24	229,8	22	2	2,4	3,2	0,4	80	74,8		2,078	0,068	17,644
34	0,12	0,182	19,65	4,244	15,75	1,4	1,2	16,2	21	20,4	7,2	72,4	15	23,8	228,4	22,2	2,4	2	1,8	0,2	103,2	94,2		2,190	0,228	17,596
42	0,12	0,182	19,65	4,244	15,75	1,4	1	18	20,8	22	7	72,8	15,2	23	230,4	22,6	2,6	3,2	2	0,2	67,6	62,4		1,852	0,900	17,126
50	0,08	0,186	19,68	4,244	15,75	1,4	1,4	15	20,6	20,4	7	72,4	14,4	23,4	228,8	21,6	2,4	1,8	2,8	0,2	109	119,6		2,100	0,452	17,292
65	0,08	0,186	19,68	4,184	15,78	1,4	1	15,8	21,4	20,4	7,6	72,2	15	23	230,2	22,4	2,8	2,6	3	0,4	85,6	100,4		2,256	0,118	17,524
83	0,082	0,142	19,74	4,184	15,78	1	1,2	16,8	21,6	20,8	8	71,4	14,6	23	229,8	22,6	2,4	2,8	2,2	0,2	101	115,8		1,480	0,542	17,874
91	0,082	0,142	19,74	4,184	15,78	1,4	1,2	15,4	21,8	21,8	7,2	71	14,4	23,8	229	22,6	2,4	2,6	2	0,2	62,2	65,2		1,866	0,274	17,730

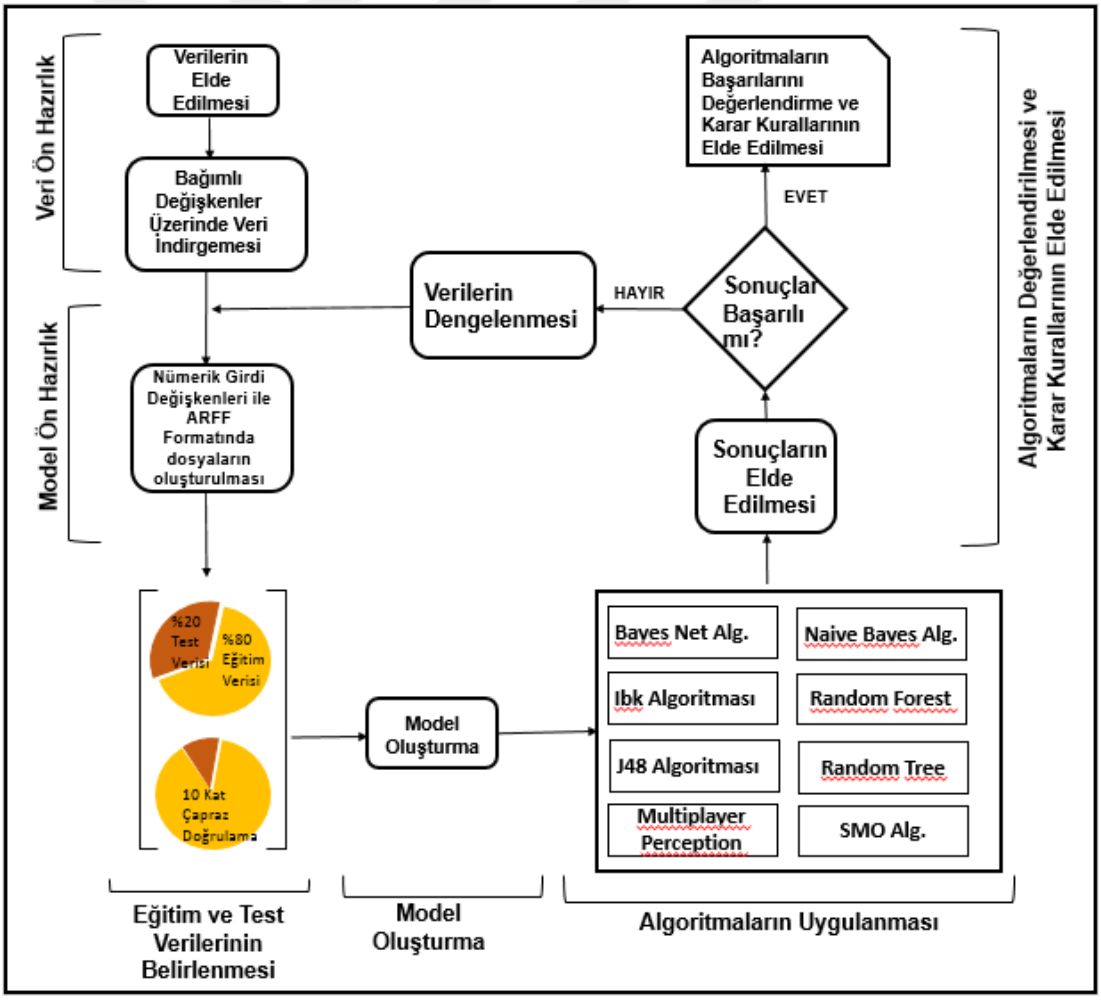


Şekil 7.1. Ürün Kalitesini Tespit Etme Şematiği

Bir ürünün sağlam olabilmesi için Y1, Y2, Y3 değişken değerlerinin Şekil 7.1.'de belirtilen aralıkları sağlaması gerekmektedir. Eğer Y1, Y2, Y3 değerlerinden herhangi birisi olması gereken aralığın dışında kalır ise o veriye ait ürün hatalı ürün olarak nitelendirilmektedir. Çizelge 7.1'de sarı ile işaretli olan satırlar hatalı ürünleri temsil etmektedir, temsili örneklerde görülen 19 numaralı satırda Y2 değeri alt aralıktan daha düşük olduğu için, 42 numaralı satırda Y2 değeri üst aralıktan daha yüksek olduğu için ve 83 numaralı satırda Y1 değeri alt aralıktan daha düşük olduğu için bu ürünler hatalı ürün olarak sınıflandırılmaktadır.

7.1.Çalışmada Uygulanan Veri Madenciliği Metodolojisi

Çalışmada uygulanan veri madenciliği metodolojisi Şekil 7.2.'de görülmektedir.



Şekil 7.2. Çalışmada Kullanılan Veri Madenciliği Metodolojisi

Çalışmada ilk aşama olarak, kullanılan veri seti üzerinde veri ön işleme adımları uygulanmıştır. Bağımlı değişkenler olan üç adet çıktı değişkeni veri indirgenmesi ile hatalı veya sağlam olacak şekilde tek değişken haline getirilmiştir. Oluşturulan tek çıktı değişkeninde "1" ile ifade edilen değerler sağlam parçaları, "0" ile ifade edilen değerler ise hatalı parçaları göstermektedir.

Çıktı değişkeni tek değişkene dönüştürüldükten sonra, WEKA yazılımına Excel formatında veri tanımlanamadığı için bulunan veriler WEKA yazılımının tanıdığı arff formatına dönüştürülmüştür. Veriler arff formatına dönüştürülürken girdi değişkenlerini nümerik olarak tanımlanmıştır.

Veriler algoritmalar tarafından çözümlenmeden önce eğitim ve test verisi olarak ayrılmıştır. Modellerin eğitim seti üzerinde öğrenme yapması ve test verisi üzerinde başarısını ölçmesi sağlanmıştır. Eğitim ve test verileri belirlenirken yaygın olarak kullanılan 10 kat çapraz doğrulama test yöntemi ve verilerin %80eğitim-%20test olarak ayırma metodu uygulanmıştır. Her iki metot da WEKA yazılımında yer alan farklı algoritmalara uygulanmış ve algoritmaların sonuçları karşılaştırılmıştır.

93 adeti sağlam ürün verisi, 8 adeti hatalı ürün verisi olan veri seti çeşitli sınıflandırma algoritmalarına uygulanmıştır. Veri setinin algoritmalara on kat çapraz doğrulama test yöntemi ile uygulandığında elde edilen karışıklık matrisleri (confusion matrix) Çizelge 7.2.'de, %80eğitim-%20test yöntemi ile uygulandığında elde edilen 20 adet test verisinin karışıklık matrisleri (confusion matrix) ise Çizelge 7.3.'de görüldüğü gibidir. Karışıklık matrislerinde satırlarda yer alan değerler verilerin gerçekteki adetlerini, sütunlarda yer alan değerler ise algoritmaların tahminleri sonucunda bulduğu adetleri göstermektedir.

Çizelge 7.2. Dengeleme Öncesi Karışıklık Matrisleri (10 Kat Çapraz Test Yöntemi)

10 Kat Çapraz Test Sonucu Karışıklık Matrisleri		**Tahmin**	
	Reel	**0**	**1**
Bayes Net Algoritması	**0**	0	8
	1	0	93
Ibk Algoritması	**0**	0	8
	1	5	88
J48 Algoritması	**0**	0	8
	1	1	92
Multiplayer Perception	**0**	1	7
	1	6	87
Naive Bayes Algoritması	**0**	0	8
	1	2	91
Random Forest Alg.	**0**	0	8
	1	0	93
Random Tree Alg.	**0**	1	7
	1	11	82
SMO Algoritması	**0**	0	8
	1	0	93

Çizelge 7.3. Dengeleme Öncesi Karışıklık Matrisleri (%80 Eğitim- %20 Test)

%80 Eğitim, %20 Test Sonucu Karışıklık Matrisleri	**Reel**	**Tahmin**	
		0	**1**
Bayes Net Algoritması	**0**	0	3
	1	0	17
Ibk Algoritması	**0**	1	2
	1	2	15
J48 Algoritması	**0**	0	3
	1	0	17
Multiplayer Perception	**0**	1	2
	1	5	12
Naive Bayes Algoritması	**0**	0	3
	1	3	14
Random Forest Alg.	**0**	0	3
	1	0	17
Random Tree Alg.	**0**	1	2
	1	5	12
SMO Algoritması	**0**	0	3
	1	0	17

Çizelge 7.2. ve Çizelge 7.3.'te görülen matrisler incelendiğinde, doğru sınıflandırılan öğe sayının yüksek olmasına rağmen özellikle hatalı ürünleri (sıfır etiketli) sınıflandırmada oldukça başarısız oldukları görülmektedir.

Hatalı ürünlerin niçin yanlış sınıflandırıldığının nedeni araştırıldığında; veri setinde sağlam ürün adetinin hatalı ürün adetine oranla oldukça fazla olmasından kaynaklı, veri dengesizliği oluştuğu ve algoritmaların az miktarda bulunan hatalı ürün sınıfı hakkında veri üzerinde tam öğrenme gerçekleştiremediğinden dolayı, verileri sağlam ürün sınıfına atamaya meyilli oldukları tespit edilmiştir.

Veri setinde 93 adet sağlam ürün verisi bulunurken sadece 8 adet hatalı ürün verisi bulunmaktadır. Dolayısı ile modeller hatalı ürün verilerini öğrenmekte yetersiz kalmakta ve örnekleri veri setinde baskın olarak bulunan sağlam ürün verisi sınıfına atamaya eğilim göstermektedirler.

Bu aşamada veri ön işleme adımına dönülerek unbalanced (dengesiz dağılmış) veriler, balanced (dengeli dağılmış) hale getirilerek veri dengesizliği ortadan kaldırılmıştır. Veri dengeleme işlemi sonucunda sağlam ve hatalı ürün sayısı eşitlenerek bu sorun çözülmüştür.

7.2. Verilerin Dengeli (Balanced) Hale Getirilmesi

Algoritmaların doğru sonuçlar verebilmesi amacıyla, çıktı değişkenlerinin veri setinde dengeli bir şekilde dağılması için veri ön işleme adımına geri dönülmüştür. WEKA Yazılımında veri ön işleme prosesinde yer alan sentetik veri üretme modülü Synthetic Minority Oversampling Technique (SMOTE) ile sağlam ürünlerin verileri ile hatalı ürünlerin verileri arasında veri dengeleme işlemi yapılmıştır. Programın içindeki bu modül sayesinde verilerin içindeki bağıntıları bozmayacak şekilde, en yakın komşularına bakarak veri örüntüleri korunmuş bir biçimde, sadece adet olarak çoğaltım işleminin yapılması sağlanmıştır. Veri dengeleme işlemi ile hatalı ürün ve sağlam ürün miktarları eşit hale getirilmiş ve algoritmaların dengeli bir öğrenme yapabilmesine olanak sağlanmıştır. Dengeleme işleminden sonra 93'ü sağlam 93'ü hatalı olacak şekilde 186 veri elde edilmiş ve elde edilen bu veriler algoritmalara yeniden uygulanmıştır.

Veri dengeleme adımı uygulanmadan önce ve uygulandıktan sonra verilere ait temel istatistiksel özellikler Çizelge 7.4.'de görüldüğü gibidir. Dengeleme işlemi sonrasında, birkaç ufak değişiklik dışında, dengeleme işleminden önce verilerde mevcut olan minimum maksimum değerler, standart sapma, ortalama gibi temel istatistiksel özelliklerin korunduğu görülmektedir.

Çizelge 7.4. Girdi Değişkenlerinin Temel İstatistiki Özellikleri

	Veri Dengeleme İşleminden Önce				Veri Dengeleme İşleminden Sonra			
	Minimum Değer	**Maximum Değer**	**Ortalama**	**Standart Sapma**	**Minimum Değer**	**Maximum Değer**	**Ortalama**	**Standart Sapma**
X1	0,076	0,120	0,088	0,014	0,076	0,120	0,088	0,012
X2	0,142	0,204	0,172	0,024	0,142	0,204	0,174	0,023
X3	19,650	19,736	19,704	0,032	19,650	19,736	19,701	0,028
X4	4,184	4,244	4,217	0,030	4,184	4,244	4,217	0,027
X5	15,746	15,782	15,762	0,012	15,746	15,782	15,762	0,016
X6	1	1,400	1,200	0,162	1	1,400	1,167	0,155
X7	1	4,400	1,182	0,170	1	1,400	1,167	0,149
X8	14	18	15,865	1,144	14	18	16,053	1,004
X9	20	22	21,091	0,602	20	22	21,243	0,579
X10	20	22	20,966	0,628	20	22	20,960	0,538
X11	7	8	7,469	0,312	7	8	7,509	0,287
X12	71	73	71,923	0,634	71	73	72,055	0,609
X13	14,400	15,200	17,798	0,304	14,400	15,200	14,795	0,282
X14	23	24	23,517	0,359	23	24	23,471	0,334
X15	228	232	230,022	1,227	228	232	230,171	0,960
X16	21,600	22,600	22,103	0,351	21,600	22,600	22,167	0,326
X17	1,800	3,200	2,558	0,440	1,800	3,200	2,493	0,395
X18	1,800	3,200	2,497	0,493	1,800	3,200	2,531	0,436
X19	1,800	3,200	2,535	0,472	1,800	3,200	2,639	0,440
X20	0,200	0,400	0,307	0,100	0,200	0,400	0,294	0,092
X21	60,600	119	90,347	17,831	60,600	119	91,468	15,421
X22	60	119,800	90,859	18,444	60	119,800	91,403	17,847

7.3. Algoritmaların Uygulanması ve Başarılarının Karşılaştırılması

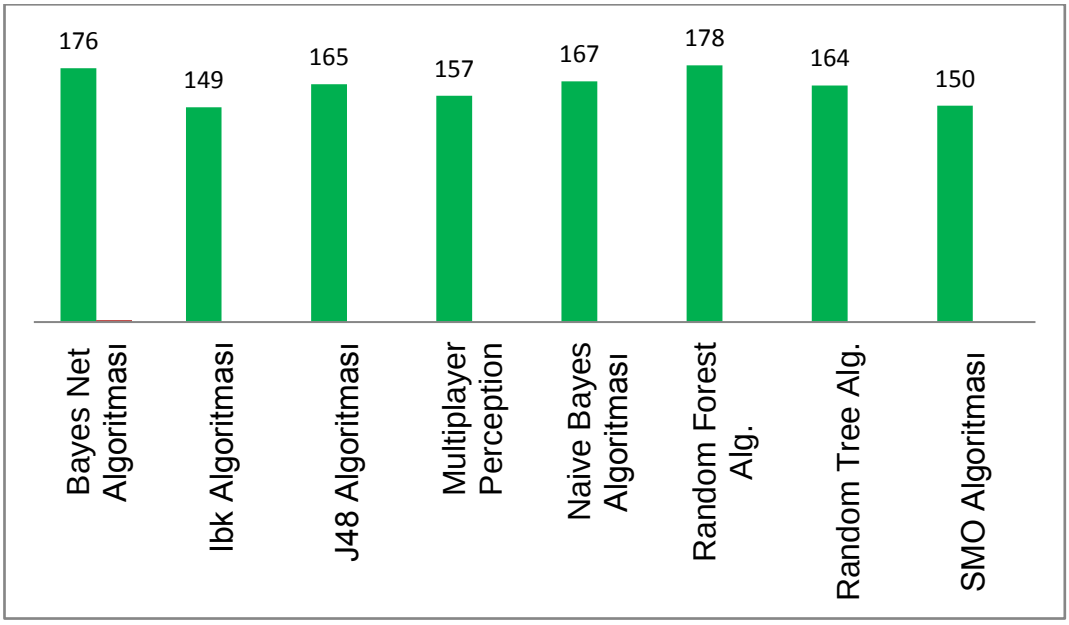
Veri dengeleme işleminden sonra elde edilen veri seti, on kat çapraz doğrulama test yöntemi ve %80eğitim-%20test yöntemi ile eğitilmiş ve algoritmalara tekrardan uygulanarak yeniden test edildiğinde anlamlı sonuçlar elde edildiği görülmüştür.

7.3.1. Algoritmaların On Kat Çapraz Doğrulama Test Yöntemi İle Uygulanması

Elde edilen 186 veri çeşitli sınıflandırma algoritmalarına uygulanmış ve on kat çapraz doğrulama test yöntemi ile test edilmiştir. Test sonucunda algoritmaların doğru sınıflandırdığı öğe sayısı Çizelge 7.5.'de, grafiği ise Şekil 7.3.'de gösterilmektedir.

Çizelge 7.5. Doğru Sınıflandırılan Öğe Sayısı (On Kat Çapraz Test Yöntemi)

10 Kat Çapraz Test Sonuçları	Adet
Bayes Net Algoritması	176
Ibk Algoritması	149
J48 Algoritması	165
Multiplayer Perception	157
Naive Bayes Algoritması	167
Random Forest Alg.	178
Random Tree Alg.	164
SMO Algoritması	150



Şekil 7.3. Doğru Sınıflandırılan Öğe Sayısı Grafiği (On Kat Çapraz Test Yöntemi)

On kat çapraz doğrulama yöntemi ile test edilen algoritmaların doğru sınıflandırdığı öğe sayısının tüm veriler içindeki yüzdesi Çizelge 7.6.'da gösterilmektedir.

Çizelge 7.6. Doğru Sınıflandırılan Öğe Yüzdesi (On Kat Çapraz Test Yöntemi)

10 Kat Çapraz Test Sonuçları	**Yüzde**
Bayes Net Algoritması	94,62
Ibk Algoritması	80,11
J48 Algoritması	88,71
Multiplayer Perception	84,41
Naive Bayes Algoritması	89,78
Random Forest Alg.	95,70
Random Tree Alg.	88,17
SMO Algoritması	80,65

Dengelenmiş veriler kullanılarak on kat çapraz doğrulama test yöntemi ile uygulanan algoritmaların sonuçları incelendiğinde; oluşturulan modeller arasında en başarılı olan modelin %95,70 başarı oranı ile 186 veri arasından 178 veriyi doğru sınıflandıran Random Forest Algoritması olduğu görünmektedir.

On kat çapraz doğrulama yöntemi ile test olan algoritma sonuçlarının karışıklık matrisleri (confusion matrix) Çizelge 7.7.'de göründüğü gibidir. Karışıklık matrislerinde satırlarda yer alan değerler verilerin gerçekte mevcut olan adetlerini, sütunlarda yer alan değerler ise algoritmaların tahminleri sonucunda bulduğu adetleri göstermektedir.

Çizelge 7.7. Dengeleme Sonrası Karışıklık Matrisleri (10 Kat Çapraz Test Yöntemi)

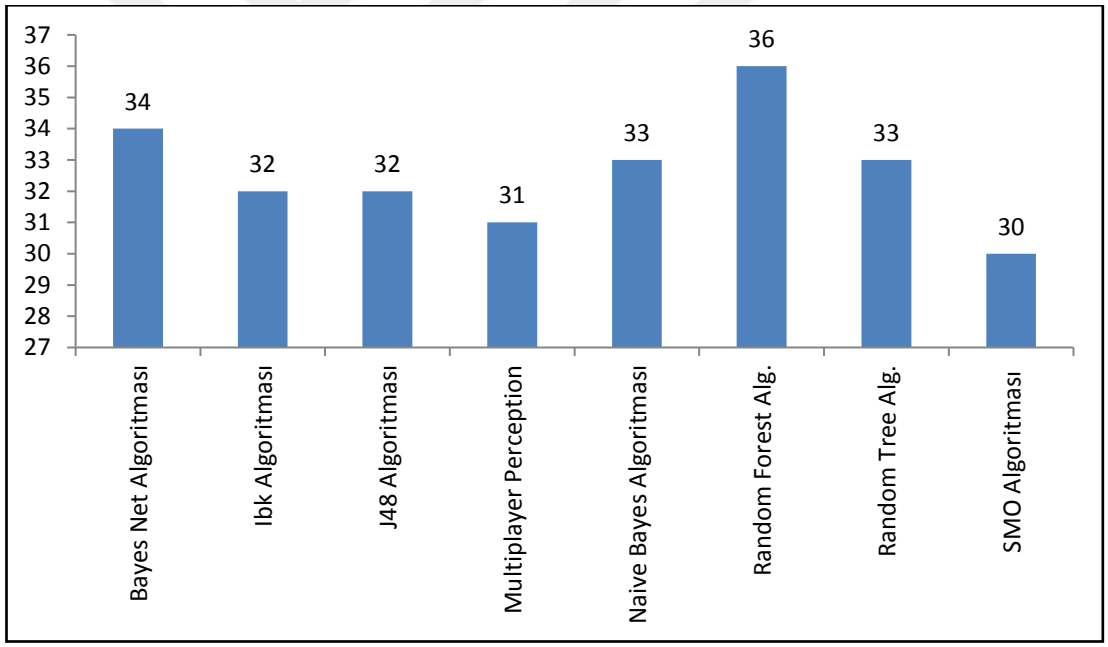
10 Kat Çapraz Test Sonucu Karışıklık Matrisleri	**Reel**	**Tahmin**	
		0	**1**
Bayes Net Algoritması	**0**	85	8
	1	2	91
Ibk Algoritması	**0**	92	1
	1	36	57
J48 Algoritması	**0**	83	10
	1	11	82
Multiplayer Perception	**0**	86	7
	1	22	71
Naive Bayes Algoritması	**0**	81	12
	1	7	86
Random Forest Alg.	**0**	88	5
	1	3	90
Random Tree Alg.	**0**	80	13
	1	9	84
SMO Algoritması	**0**	81	12
	1	24	69

7.3.2. Algoritmaların %80 Eğitim- %20 Test Verileri İle Uygulanması

Bu kısımda verilerin farklı bir test yöntemi test edilmesi ve başarılı olan algoritmaların başarı oranlarını devam ettirip ettiremeyecekleri denenmek istenmiştir. Verilerin %80'i eğitim verisi olarak kullanılarak modelin bu verilerle öğrenme yapması ve daha sonra ayrılan %20'lik veri kısmı üzerinde modeli test etmesi sağlanmıştır. Algoritmalar modeli öğrendikten sonra tüm verinin %20'si olan 37 veri üzerinden modelin doğruluğunu test etmişlerdir. Algoritmaların doğru sınıflandırdığı öğe sayısı Çizelge 7.8.'de grafik üzerinde gösterimi ise Şekil 7.4.'de görüldüğü gibidir.

Çizelge 7.8. Doğru Sınıflandırılan Öğe Sayısı (%80 Eğitim- %20 Test Yöntemi)

%80 Eğitim, %20 Test Yöntemi Sonuçları	Adet
Bayes Net Algoritması	34
Ibk Algoritması	32
J48 Algoritması	32
Multiplayer Perception	31
Naive Bayes Algoritması	33
Random Forest Alg.	36
Random Tree Alg.	33
SMO Algoritması	30



Şekil 7.4. Doğru Sınıflandırılan Öğe Sayısı Grafiği (%80 Eğitim- %20 Test)

%80 Eğitim - %20 Test verisi yöntemi ile uygulanan algoritmaların doğru sınıflandırdığı öğe sayısının, tüm veriler içindeki yüzdesi Çizelge 7.9.'da gösterilmektedir.

Çizelge 7.9. Doğru Sınıflandırılan Öğe Yüzdesi (%80 Eğitim- %20 Test Yöntemi)

%80 Eğitim, %20 Test Yöntemi Sonuçları	**Yüzde**
Bayes Net Algoritması	91,89
Ibk Algoritması	86,49
J48 Algoritması	86,49
Multiplayer Perception	83,78
Naive Bayes Algoritması	89,19
Random Forest Alg.	97,30
Random Tree Alg.	89,19
SMO Algoritması	81,08

Dengelenmiş verilerin algoritmalara %80eğitim-%20 test yöntemi ile uygulanan sonuçları incelendiğinde; modeller arasında en başarılı olan modelin %97,30 başarı oranı ile 37 veri arasından 36 veriyi doğru sınıflandıran Random Forest Algoritması olduğu görünmektedir. Böylece test yöntemi değiştirildiğinde de en başarılı algoritma değişmemiş yine Random Forest Algoritması olmuştur.

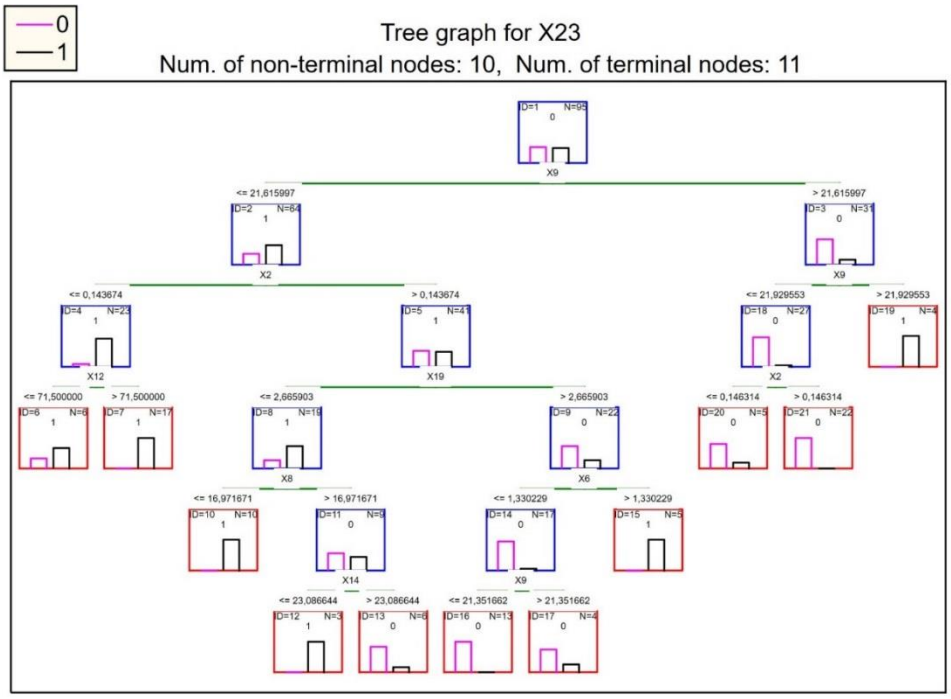
Verilerin %80eğitim-%20test yöntemi ile uygulandığı algoritma sonuçlarının karışıklık matrisleri (confusion matrix) Çizelge 7.10.'da göründüğü gibidir. Karışıklık matrislerinde satırlarda yer alan değerler verilerin gerçekteki adetlerini, sütunlarda yer alan değerler ise algoritmaların tahminleri sonucunda bulduğu adetleri göstermektedir.

Çizelge 7.10. Dengeleme Sonrası Karışıklık Matrisleri (%80 Eğitim-%20 Test)

%80 Eğitim, %20 Test Yöntemi Karışıklık Matrisleri	**Reel**	**Tahmin**	
		0	**1**
Bayes Net Algoritması	**0**	14	3
	1	0	20
Ibk Algoritması	**0**	17	0
	1	5	15
J48 Algoritması	**0**	12	5
	1	0	20
Multiplayer Perception	**0**	15	2
	1	4	16
Naive Bayes Algoritması	**0**	14	3
	1	1	19
Random Forest Alg.	**0**	16	1
	1	0	20
Random Tree Alg.	**0**	15	2
	1	2	18
SMO Algoritması	**0**	13	4
	1	3	17

8. SONUÇLAR VE ÖNERİLER

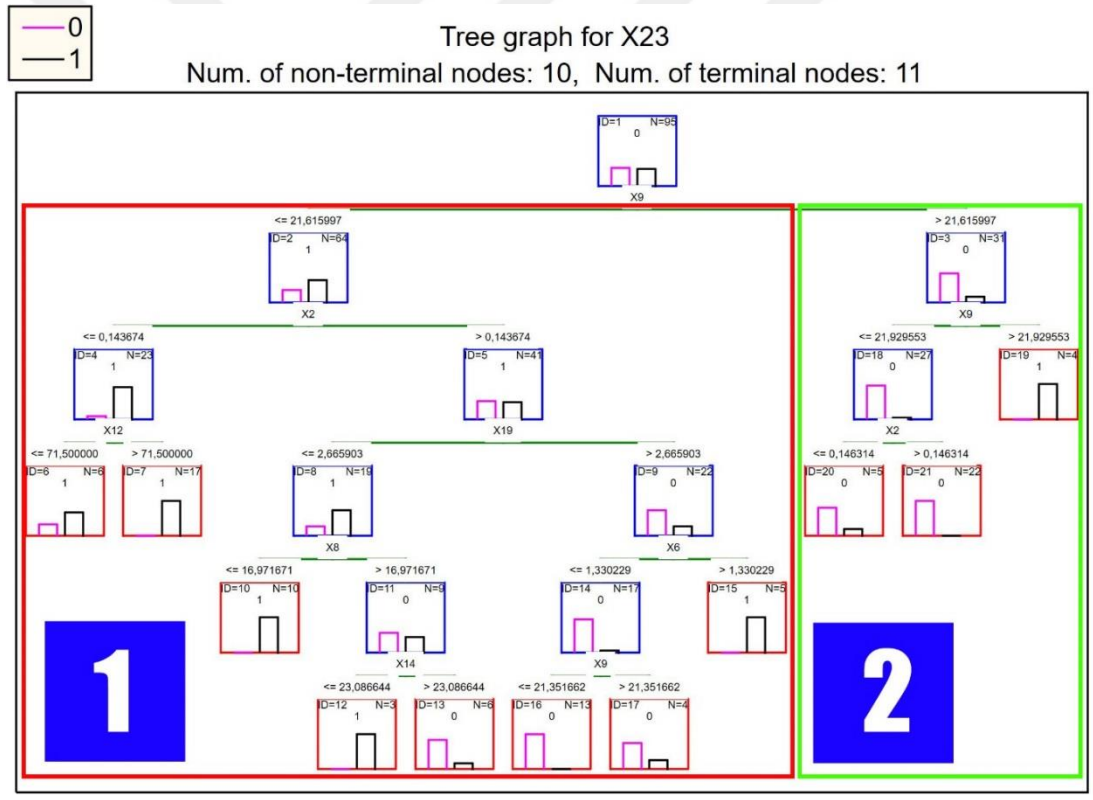
Hatalı ve sağlam ürünlerin tespit edilmesi için, yedinci bölümde uygulanan her iki test yönteminde de en fazla doğruluk oranına sahip olan Random Forest Algoritmasının sonuçları kullanılmıştır. Algoritmanın ağaç yapısı elde edilmiş ve hangi dallarda sağlam ürünlerin hangi dallarda hatalı ürünlerin yer aldığı keşfedilmiştir. Dallara ulaşmak için gerekli olan yollar tespit edilerek karar kuralları çıkarılmıştır. Random Forest Algoritmasının görsel ağaç yapısını ve kural çıktılarını WEKA yazılımı sonuç olarak vermeyip sadece istatistiki değerleri verdiğinden dolayı görsel ağaç yapısı için STATİSTİKA isimli istatistik ve veri madenciliği yazılımı kullanılmıştır. STATİSTİKA yazılımında Random Forest Algoritmasına girdi olarak WEKA ile aynı veriler verilmiş ve WEKA'nın elde ettiği sonuçlara uyumlu sonuçlar elde edilmiştir. Random Forest Algoritması ile elde edilen ağaç yapısı Şekil 8.1.'de görüldüğü gibidir.



Şekil 8.1. Random Forest Algoritmasından Elde Edilen Ağaç Yapısı

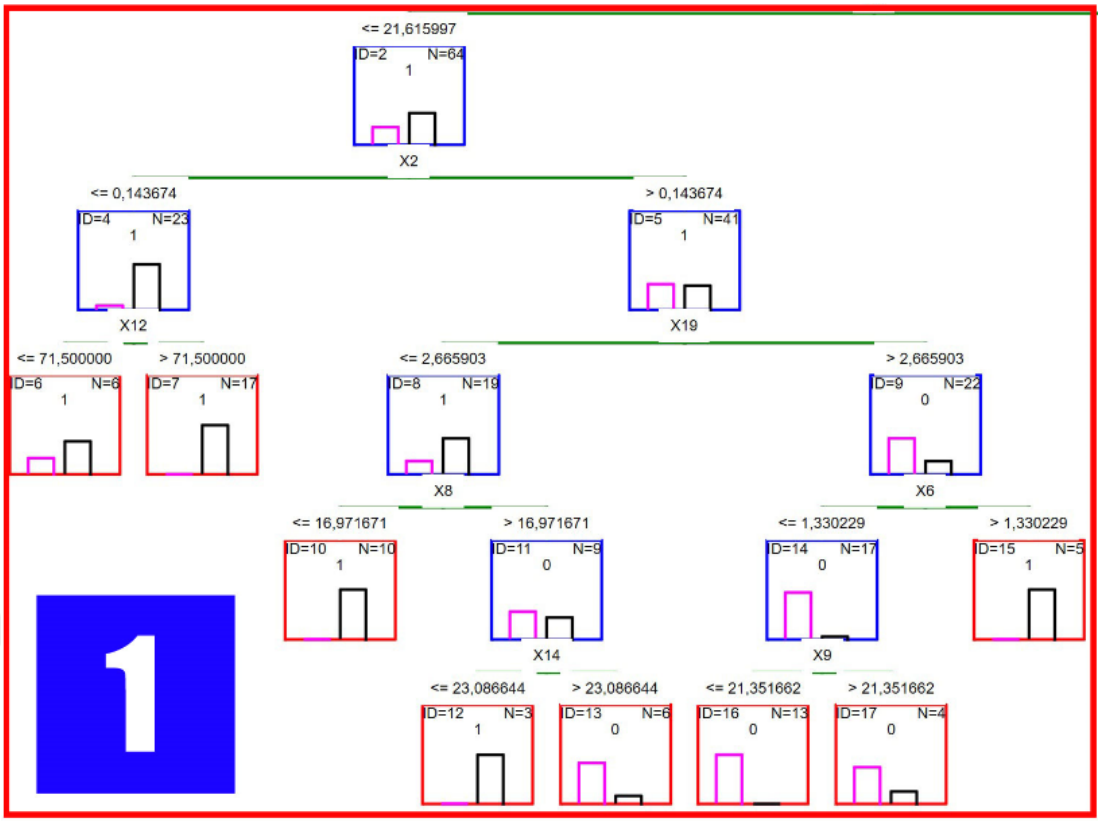
Sonuçlar incelendiğinde mavi renkli düğümler karar aşamasının bitmediği ve dallanmanın devam ettiği düğümleri göstermektedir. Kırmızı renkli düğümler ise algoritmanın dallanmayı bitirdiği ve bir sonuca ulaştığı yaprakları temsil etmektedir.

Kök düğüm olarak algoritmanın X9 değişkenini belirlediği görülmektedir. X9 değişkenin değerinin 21,616'dan küçük veya eşit olması durumu ile 21,616'dan büyük olması durumuna göre ağaç yapısı dallanmaktadır. Ağaç yapısının daha detaylı görünebilmesi için bu iki durum büyük ölçekli olarak Şekil 8.2.'de görüldüğü gibi iki farklı parçaya ayrılmış ve parçalar detaylı olarak açıklanmıştır. Değerlerin daha net görünmesi ve açıklamalarda karmaşıklık olmaması açısından, değerler virgülden sonra yuvarlanarak üç basamak olarak ifade edilmiştir.



Şekil 8.2. Random Forest Algoritmasından Elde Edilen Ağaç Yapısı (Parçalı)

X9 değerinin 21,616'dan küçük veya eşit olması durumunda yapmış olduğu dallanmanın büyütülmüş ölçekli detaylı görünümü Şekil 8.3.'de görüldüğü gibidir.



Şekil 8.3. Random Forest Algoritmasından Elde Edilen Ağaç Yapısı (Parça 1)

X9 değişkenin değerinin 21,616'dan küçük veya eşit olması durumunda karar ağacı X2 değişkenine bakarak yeniden dallanma yapmaktadır. X2 değişkeninin değerinin 0,144'den küçük veya eşit olması durumuna göre ve 0,144'den büyük olması durumuna göre ağaç dallanmaktadır.

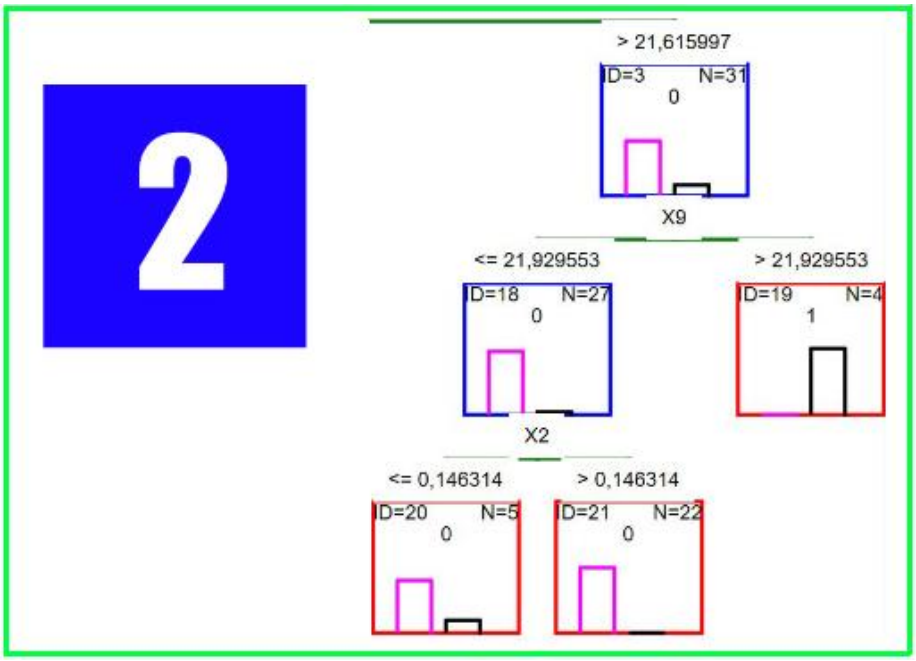
X2 değişkenin değerinin 0,144'den küçük veya eşit olması durumunda X12 değişkenin değerine bakarak sınıflandırma yapmaktadır. X12 değişkeninin değerinin 71,5'den küçük veya eşit olması durumu ile 71,5'den büyük olması durumlarında farklı doğruluk oranları ile ürünü SAĞLAM sınıfına atamaktadır.

X2 değişkenin değerinin 0,144'den büyük olması durumunda X19 değişkenin değerine bakarak ağaç yeniden dallanmaktadır. X19 değişkenin değerinin 2,666'dan küçük veya eşit olması durumunda X8 değişkenine bakarak, X19 değişkenin değerinin 2,666'dan büyük olması durumunda ise X6 değişkenine bakarak yeniden dallanma yapmaktadır.

- X8 değişkeninin değerinin 16,972'den küçük veya eşit olması durumunda karar ağacı ürünü SAĞLAM olarak sınıflandırmıştır. X8 değişkeninin değerinin 16,972'den büyük olması durumunda ise X14 değişkenine bakarak karar vermiştir. X14 değişkeninin değerinin 23,087'den küçük veya eşit olması durumunda ürünü SAĞLAM sınıfına, 23,087'den büyük olması durumunda ise HATALI sınıfına atamaktadır.

- X6 değişkeninin değerinin 1,330'dan büyük olması durumunda karar ağacı ürünü SAĞLAM olarak sınıflandırmıştır. X6 değişkeninin değerinin 1,330'dan küçük veya eşit olması durumunda ise X9 değişkeninin değerine bakarak karar vermiştir. X9 değişkeninin değerinin 21,352'den küçük veya eşit olması durumu ile 21,352'den büyük olması durumlarında farklı doğruluk oranları ile ürünü HATALI sınıfına atamaktadır.

X9 değerinin 21,616'dan büyük olması durumunda yapmış olduğu dallanmanın büyütülmüş ölçekli detaylı görünümü Şekil 8.4.'de görüldüğü gibidir.



Şekil 8.4. Random Forest Algoritmasından Elde Edilen Ağaç Yapısı (Parça 2)

X9 değişkenin değerinin 21,616'dan büyük olması durumunda karar ağacı X9 değişkenin farklı bir değerine bakarak yeniden dallanma yapmaktadır. X9 değişkeninin değeri 21,930'dan büyükse ürünü SAĞLAM sınıfına atamaktadır. X9 değişkeninin 21,930'dan küçük veya eşitse X2 değişkenine bakarak karar vermektedir.

X2 değişkenin değerinin 0,147'den küçük veya eşit olması durumu ile 0,147'den büyük olması durumunda farklı doğruluk oranları ile ürünü HATALI sınıfına dahil etmektedir.

Ağaç Yapısından Elde Edilen Kurallar Özetle Aşağıdaki Gibidir:

- X9 değişkeni 21,616'dan küçük-eşit VE X2 değişkeni 0,144'den küçük-eşit ise ürün SAĞLAM

- X9 değişkeni 21,616'dan küçük-eşit VE X2 değişkeni 0,144'den büyük VE X19 değişkenin değeri 2,666'dan küçük-eşit VE X8 değişkeni 16,972'den küçük-eşit ise ürün SAĞLAM

- X9 değişkeni 21,616'dan küçük-eşit VE X2 değişkeni 0,144'den büyük VE X19 değişkenin değeri 2,666'dan küçük-eşit VE X8 değişkeni 16,972'den büyük VE X14 değişkeni 23,087'den küçük-eşitse ürün SAĞLAM

- X9 değişkeni 21,616'dan küçük-eşit VE X2 değişkeni 0,144'den büyük VE X19 değişkenin değeri 2,666'dan küçük-eşit VE X8 değişkeni 16,972'den büyük VE X14 değişkeni 23,087'den büyükse ürün HATALI

- X9 değişkeni 21,616'dan küçük-eşit VE X2 değişkeni 0,144'den büyük VE X19 değişkenin değeri 2,666'dan büyük VE X6 değişkeni 1,330'dan büyükse ürün SAĞLAM

- X9 değişkeni 21,616'dan küçük-eşit VE X2 değişkeni 0,144'den büyük VE X19 değişkenin değeri 2,666'dan büyük VE X6 değişkeni 1,330'dan küçük-eşitse ürün HATALI

- X9 değişkeni 21,616'dan büyük ve 21,930'a küçük-eşit ise ürün HATALI

- X9 değişkeni 21,930'dan büyükse ürün SAĞLAM şeklinde karar kuralları elde edilmiştir.

Çalışmada bir üretim sisteminden elde edilen veriler kullanılarak WEKA üzerinde veri dengelemesi yapılmış ve dengelenmiş veriler ile algoritmalardan oldukça başarılı sonuçlar elde edilmiştir. On kat çapraz doğrulama test yöntemi ve %80eğitim-%20test yöntemi gibi her iki test yönteminde de en başarılı sonucu Random Forest Algoritması vermiştir. Algoritma on kat çapraz doğrulama test yönteminde %95,70 gibi bir başarı oranı yakalarken, veriler %80 eğitim-%20 test verisi olarak kullanıldığında % 97,30 gibi yüksek bir başarı oranı elde etmiştir.

Algoritmaların yakaladıkları yüksek başarı oranları, üretimde hatalı ürün ve sağlam ürün verisi gibi dengesiz dağılım gösteren veri setleri ile çalışılırken, veri setlerinin dengeli hale getirildiğinde algoritmalardan başarılı sınıflandırma sonuçları alınabileceğini göstermektedir.

En başarılı sonucu veren Random Forest Algoritması'nın ağaç yapısı çıkarılarak ürünlerin hatalı ya da sağlam olmalarını etkileyen değişkenler tespit edilmiş ve karar kuralları belirlenmiştir.

Sonuç olarak, çıkarılan karar kuralları sayesinde çok fazla sayıda olan girdi değişkenlerinden sadece birkaç tanesinin değerine bakarak ürünün sağlam mı hatalı mı olacağı belirlenebilmektedir. Böylece ürün kalitesini önemli derecede etkileyen birkaç değişkenin değeri, üretimin erken aşamalarında gözlemlenerek ürünlerde daha hata meydana gelmeden önce gerekli önlemler alınabilir. Bu sayede hatalı ürün üretiminin neden olacağı işgücü, zaman, maliyet gibi kayıplar önlenebilir veya azaltılabilir.

9. KAYNAKLAR

Abhang, L.B., Hameedullah, M., Modeling and Analysis of Surface Roughness in Steel Turning Using Regression and Neural Networks. Advances in Engineering, Science and Management (ICAESM), International Conference on. IEEE, p. 317-322, 2012.

Akdemir, Ç., Hilenin Veri Madenciliği ile Ortaya Çıkartılması ve Perakende Sektöründe Bir Uygulama. Doktora Tezi. Marmara Üniversitesi, İstanbul, 2016.

Akın, M., Kanserli Hücrelerin Mikroarray Gen İfadelerinin İncelenmesi ve Veri Madenciliği Yöntemleri Kullanarak Sınıflandırılması. Yüksek Lisans Tezi. Gazi Üniversitesi, Ankara, 2012.

Akpınar, H., Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği. İstanbul Üniversitesi İşletme Fakültesi Dergisi, 29(1), 1-22, 2000.

Altun, S., MR Spektroskopi Temelli Beyin Tümörü Teşhisinde Veri Madenciliği Uygulamaları. Yüksek Lisans Tezi. Kahramanmaraş Sütçü İmam Üniversitesi, Kahramanmaraş, 2018.

Argüden, Y., Erşahin B., Veri madenciliği: veriden bilgiye, masraftan değere.37-42. ARGE Danışmanlık Yayınları, İstanbul, 2008.

Armutlu, Ş., Veri Madenciliği İle Kütüphane Kullanımı Ve Ders Baiarısı Arasındaki İliikinin İncelenmesi. Yüksek Lisans Tezi. Uşak Üniversitesi, Uşak, 2018.

Atasoy, Y., Veri Madenciliği Yöntemleri İle Ankilozan Spondilit Hastalığında Radyografik Progresyona Etkili Faktörlerin Analizi. Yüksek Lisans Tezi. İstanbul Üniversitesi, İstanbul, 2015.

Aydemir, B., Veri Madenciliği Yöntemleri Kullanarak Meslek Yüksek Okulu Öğrencilerinin Akademik Başarı Tahmini. Yüksek Lisans Tezi. Pamukkale Üniversitesi, Denizli, 2017.

Aydın, Ö., Elektronik Harp İle Toplanan Verilerin Veri Madenciliği Yöntemleri İle Analiz Edilmesi. Yüksek Lisans Tezi. Bahçeşehir Üniversitesi, İstanbul, 2017.

Azevedo, A. I. R. L., Santos, M. F., KDD, SEMMA and CRISP-DM: a parallel overview. IADS-DM, 2008.

Bakır, B., Batmaz, İ., Güntürkün, F. A., İpekçi, İ. A., Köksal, G., Özdemirel, N. E., Defect Cause Modeling With Decision Tree And Regression Analysis. World Acad Sci Eng Technoly, International Journal of Mechanical, Aerospace, Industrial, Mechatronic and Manufacturing Engineering. Vol:2, No:12, 2008.

Baykasoğlu, A., Veri Madenciliği ve Çimento Sektöründe Bir Uygulama. Akademik Bilişim Konferansı, 2005.

Bilekdemir, G., Veri Madenciliği Tekniklerini Kullanarak Üretim Süresi Tahmini ve Bir Uygulama. Yüksek Lisans Tezi. Dokuz Eylül Üniversitesi, İzmir, 2010.

Boyacı, A., Öğretmenlerin Algılanan Örgütsel Destek ve Örgütsel Özdeşleme Düzeylerinin Veri Madenciliği İle Analizi. Yüksek Lisans Tezi. Hitit Üniversitesi, Çorum, 2017.

Can, O., Türkiye Sağlık Araştırmasının Veri Madenciliği Teknikleri İle İncelenmesi. Yüksek Lisans Tezi. Kafkas Üniversitesi, Kars, 2017.

Canlı, H., Otomotiv Sektöründe Kalite Kontol Sürecinde Veri Madenciliği Yöntemleri İle Karar Destek Sistemi Uygulaması. Yüksek Lisans Tezi. Düzce Üniversitesi, Düzce, 2017.

Canlı, H., Toklu, S., İmplementation of Decision Support System with Data Mining Methods in the Quality Control Process of the Automotive Sector. Düzce University Journal of Science & Technology. 7, 102-114, 2019.

Chapman, P., Clinton J., Kerber R., Khabaza T., Reinartz T., Shearer C., Wirth R., CRISP-DM 1.0 Step-by-step data mining guide.10-11. SPSS Şirketi, 2000.

Chen, R.S., Yeh, K.C., Chang, C.C., Chien, H.H., Using Data Mining Technology to Improve Manufacturing Quality –A Case Study of LCD Driver IC Packaging Industry. Seventh ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing (SNPD'06)., p. 115-119, 2006.

Ciga, A.B., Gerçek Laboratuvar Verilerinin Veri Madenciliği Teknikleri İle Analizi. Yüksek Lisans Tezi. Afyon Kocatepe Üniversitesi, Afyonkarahisar, 2015.

Çelik, M., Bir Otomotiv Yan Sanayi Kuruluşunda Veri Madenciliği Uygulaması. Yüksek Lisans Tezi. Uludağ Üniversitesi, Bursa, 2009.

Çetin, M., Bir Üretim İşletmesinde Veri Madenciliği Uygulaması. Yüksek Lisans Tezi. Sakarya Üniversitesi, Sakarya, 2009.

Çığşar, B., Kredi Risklerinde Veri Madenciliği Sınıflandırma Algoritmaları. Yüksek Lisans Tezi. Çukurova Üniversitesi, Adana, 2017.

Çoban, A., İmalat Sanayinde Veri Madenciliği Destekli Tedarikçi Seçimi Uygulaması. Doktora Tezi. Sakarya Üniversitesi, Sakarya, 2006.

Dal, H., Morgül, Ö.K., Şahin, İ., Yapay Sinir Agı (YSA) Kullanarak Titreşim Tabanlı Makina Durum İzlemesi ve Hata Teşhisi. SAÜ Fen Bilimleri Enstitüsü Dergisi. 1O. Cilt, 2. Sayı, 45-50, 2006.

Diler, S., Veri Madenciliği Süreçleri Ve Karar Ağaçları Algoritmaları İle Bir Uygulama. Yüksek Lisans Tezi. Yüzüncü Yıl Üniversitesi, Van, 2016.

Dondurmacı, G.A., Çınar A., Finans Sektöründe Veri Madenciliği Uygulaması. Akademik Sosyal Araştırmalar Dergisi. Yıl:2, Sayı:2/1, 258-271, 2014.

Emre, İ.E., Veri Madenciliği İle Çocukluk Çağındaki Akut Romatizmal Ateşin Kalp Hastalığına Etkilerinin Analizi. Yüksek Lisans Tezi. İstanbul Üniversitesi, İstanbul, 2017.

Erden, N., Nazarov, M., İplik Sürtünme Özelliklerinin İncelemesinde Kaba Kümeler Yaklaşımı. Erzincan Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 9(1), 75-86, 2016.

Erduran, G.Y., Online Müşteri Şikayetlerinin Veri Madenciliği İle İncelenmesi. Doktora Tezi. Trakya Üniversitesi, Edirne, 2017.

Ergün, E., Ürün Kategorileri Arasındaki Satış İlişkisinin Birliktelik Kuralları Ve Kümeleme Analizi İle Belirlenmesi Ve Perakende Sektöründe Bir Uygulama. Doktora Tezi. Afyon Kocatepe Üniversitesi, Afyonkarahisar, 2008.

Erkuş, S., Veri Madenciliği Yöntemleri İle Kardiyovasküler Hastalık Tahmini Yapılması. Yüksek Lisans Tezi. Bahçeşehir Üniversitesi, İstanbul, 2015.

Fakı, B.M., Veri Madenciliği Yöntemlerini Kullanarak Anemi Sınıflandırılmasına Yönelik Bir Uygulama. Yüksek Lisans Tezi. İstanbul Teknik Üniversitesi, İstanbul, 2015.

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

Gitmez, M., Metasezgisel Algoritmalar İle Veri Madenciliğinde Aykırı Değerlerin Tespiti Uygulamaları. Yüksek Lisans Tezi. Harran Üniversitesi, Şanlıurfa, 2018.

Gu, L., Liu-ying, W., Gui-ming, C., Shao-chun, H., Parameters Optimization of Plasma Hardening Process Using Genetic Algorithm and Neural Network. Journal of Iron and Steel Research International, 18(12), 57-64, 2011.

Gürbüz, F., Özbakır, L., Yapıcı, H., Türkiye'de Bir Havayolu İşletmesine Ait Parça Söküm Raporlarına İlişkin Veri Madenciliği Uygulaması. Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi. Cilt 24, No 1, 73-78, 2009.

Güzel, S., Veri Madenciliğinde Sınıflandırma Algoritmaları Kullanılarak Hepatit Hastalığının Tespiti. Yüksek Lisans Tezi. Kahramanmaraş Sütçü İmam Üniversitesi, Kahramanmaraş, 2018.

Han, J., Pei J., Kamber M., Data Mining Concepts and Techniques. Elsevier, United States of America, 2012.

Kahya Özyirmidokuz, E., Veri Madenciliği Tekniklerini Kullanarak İmalat Verilerinin Modellenmesi ve Analizi. Doktora Tezi. Erciyes Üniversitesi, Kayseri, 2009.

Karadağ, G., Prediction of Production Wastage Via Data Mining. Yüksek Lisans Tezi. Yaşar Üniversitesi, İzmir, 2018.

Kayaalp, K., Asenkron Motorlarda Veri Madenciliği İle Hata Tespiti. Yüksek Lisans Tezi. Süleyman Demirel Üniversitesi, Isparta, 2007.

Kılıç, B., Öğrencilerin Sınav Kaygısını Etkileyen Faktörlerin Veri Madenciliği İle İrdelenmesi. Yüksek Lisans Tezi. İstanbul Aydın Üniversitesi, İstanbul, 2014.

Kılınç, Ç., Üniversite Öğrenci Başarısı Üzerine Etki Eden Faktörlerin Veri Madenciliği Yöntemleri İle İncelenmesi. Yüksek Lisans Tezi. Eskişehir Osmangazi Üniversitesi, Eskişehir, 2015.

Kowalski, C.T., Kowalska, T.O., Neural networks application for induction motor faults diagnosis. Mathematics and computers in simulation, 63(3-5), 435-448, 2003.

Köksal, G., Batmaz, İ., Karasözen,B., Kayalıgil, S., Testik, M.C., Özdemirel, N.E., Weber, G.W., Bakır, B., Öztürk, B., Kalite İyileştirmede Veri Madenciliği Kullanımı ve Geliştirilmesi, TÜBİTAK Destekli Projeler Veri Tabanı, 2009-486, TÜBİTAK MAG Proje 105M138, 2009: 1-89, 2009.

Köse, Y., Değerli Müşterilerde Ürün Kategorileri Arasındaki Satış İlişkilerinin Veri Madenciliği Yöntemlerinden Birliktelik Kuralları Ve Kümeleme Analizi İle Belirlenmesi Ve Ulusal Bir Perakendecide Örnek Uygulama. Yüksek Lisans Tezi. Selçuk Üniversitesi, Konya, 2015.

Levent, E.B., Veri Madenciliği Ve Havacılık Sektöründe Bir Uygulama. Yüksek Lisans Tezi. Yıldız Teknik Üniversitesi, İstanbul, 2016.

Li, M., Feng, S., Sethi I.K., Luciow, J., Wagner, K., Mining Production Data with Neural Network & CART. Third IEEE International Conference on Data Mining (ICDM'03), 731-734, 2003.

Mieno, F., Sato, T., Slubnya, Y., Odagiri, K., Tsuda, H., Take, R., Yield Improvement Using Data Mining System. IEEE International Symposium on Semiconductor Manufacturing Conference Proceedings. 391-394, 1999.

Mocan, G., Perakendecilikte Veri Madenciliği Uygulamaları Ve Sorunları. Yüksek Lisans Tezi. Yıldız Teknik Üniversitesi, İstanbul, 2016.

Odabaş, Ö., Veri Madenciliği Teknikleri İle Telekom Sektöründe Ayrılan Müşteri Analizi. Yüksek Lisans Tezi. İstanbul Ticaret Üniversitesi, İstanbul, 2017.

Oğuzlar, A., Veri Ön İşleme. Erciyes Üniversitesi İktisadi ve İdari Bilimler Dergisi. 21, 67-76, 2003.

Olson, D.L., Delen D., Advanced Data Mining Techniques. 16-23. Springer Science & Business Media, Berlin, 2008.

Ordu, B., Veri Madenciliğinde Sınıflayıcı Teknikler İle Demir Çelik Sektöründe Uzun Ürünlerin Üretimine İlişkin Bir Tahmin Modellemesi. Yüksek Lisans Tezi. Karabük Üniversitesi, Karabük, 2013.

Öncel Çekim, H., Karasoy, D., Karar Ağacı ile Cox Karma Modeli ve Lastik Verileri Üzerine Bir Uygulama. İstatistikçiler Dergisi: İstatistik ve Aktüerya, 6(1), 41-50, 2013.

Özaltındiş, T., Mekansal-Zamansal Veri Madenciliğinde Kümeleme Analizi. Yüksek Lisans Tezi. Mimar Sinan Güzel Sanatlar Üniversitesi, İstanbul, 2018.

Özarslan, S., Öğrenci Performansının Veri Madenciliği İle Belirlenmesi. Yüksek Lisans Tezi. Kırıkkale Üniversitesi, Kırıkkale, 2014.

Özbay, Ö., Öğretim Yönetim Sistemi Üzerinde Üniversite (Lisans) Düzeyindeki Öğrenci Hareketliliğinin Veri Madenciliği Yöntemleriyle Analizi. Yüksek Lisans Tezi. Başkent Üniversitesi, Ankara, 2015.

Özcan, C., Veri Madenciliğinin Güvenlik Uygulama Alanları ve Veri Madenciliği ile Sahtekârlık Analizi. Yüksek Lisans Tezi. İstanbul Bilgi Üniversitesi, 2014.

Özdemir, A., Aslay F.Y., Çam H., Veri Tabanında Bilgi Keşfi Süreci: Gümüşhane Devlet Hastanesi Uygulaması. Sosyal Ekonomik Araştırmalar Dergisi. 10(20), 347-366, 2009.

Özden, G.A., Chen, Y.T., Design Quality and Robustness with Neural Networks. IEEE Transactıons On Neural Networks. Vol. 10, No. 6, 1518-1527, 1999.

Paçaman, N., Mobil Cihazlarda Veri Madenciliği Sonuçlarının Gösterilmesine Yönelik Uygulama Geliştirilmesi. Yüksek Lisans Tezi. Ege Üniversitesi, İzmir, 2014.

Savaş, S., Topaloğlu N., Yılmaz M., Veri Madenciliği ve Türkiye'deki Uygulama Örnekleri, İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi, Yıl:11 Sayı:21, 1-23, 2012.

Saylan, S., Veri Madenciliği Teknikleri İle İstenmeyen Türkçe E-Postaların Önlenmesi Üzerine Bir Uygulama. Yüksek Lisans Tezi. Marmara Üniversitesi, İstanbul, 2018.

Skinner, K.R., Montgomery, D.C., Runger G.C., Fowler, J.W., McCarville, D.R., Rhoads T.R., Stanley, J.D., Multivariate Statistical Methods for Modeling and Analysis of Wafer Probe Test Data. IEEE Transactıons On Semıconductor Manufacturıng. Vol. 15, No. 4, 523-530, 2002.

Şeker, A., Yüksek, A.G., Stacked Autoencoder Method for Fabric Defect Detection. Cumhuriyet Üniversitesi Fen Fakültesi Fen Bilimleri Dergisi (CFD), Cilt 38, No. 2, 2017.

Şekeroğlu, S., Hizmet Sektöründe Bir Veri Madenciliği Uygulaması. Yüksek Lisans Tezi. İstanbul Teknik Üniversitesi, İstanbul, 2010.

Şık, M.Ş., Veri Madenciliği Ve Kanser Erken Teşhisinde Kullanımı. Yüksek Lisans Tezi. İnönü Üniversitesi, Malatya, 2014.

Tahminciler, E., Erythromcın İlacının Yan Etkilerinin Araştırılması Üzerine Veri Madenciliği Çalışması. Yüksek Lisans Tezi. Okan Üniversitesi, İstanbul, 2014.

Talan, M.İ., Veri Madenciliği İle Karpal Tünel Sendromuna Yönelik Ön Tanı Destek Ve Hasta Takip Sisteminin Geliştirilmesi. Yüksek Lisans Tezi. Gazi Üniversitesi, Ankara, 2016.

Tapkan,, P.Z., Özmen, T., Bir iplik üretim tesisinde nitelik seçimi ve sınıflandırma ile iplik kalitesinin belirlenmesi. Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi. 24(4), 713-719, 2018.

Tunçkaya, Y., Fosil Yakıtlı Bir Enerji Santrali Prosesinin Modellenmesi ve Ana Buhar Basıncı Parametresinin Kestirim Başarımı Analizi. Düzce Üniversitesi Bilim ve Teknoloji Dergisi, 7, 488-504, 2019.

Türker, N., Lineer Regresyon Modelinde Bayes Tahmin Ediciler. Yüksek Lisans Tezi. Çukurova Üniversitesi, Adana, 2013.

Türkoğlu, B., Komesli, M., Ünlütürk, M.S. Veri Madenciliği Üzerine Endüstriyel Bir Durum Çalışması. 5th International Management Information Systems Conference, 2018.

Türkoğlu, T., Çoklu Ölçüt Oy Değerleri Üzerinden Veri Madenciliği. Yüksek Lisans Tezi. Anadolu Üniversitesi, Eskişehir, 2016.

Uğurlu, T., Veri Madenciliği Teknikleri İle Konut Fiyatı Belirleme. Yüksek Lisans Tezi. Beykent Üniversitesi, İstanbul, 2015.

Uyumaz, Ö., Bankacılık Sektöründe Pazarlama Stratejilerinin Belirlenmesinde Sınıflandırma Ve Veri Madenciliği. Yüksek Lisans Tezi. Beykent Üniversitesi, İstanbul, 2017.

Yakupoğlu, Y., Eğitimsel Veri Madenciliği Ve Bir Uygulaması. Yüksek Lisans Tezi. İstanbul Teknik Üniversitesi, İstanbul, 2018.

Yakut, E., Veri Madenciliği Tekniklerinden C5.0 Algoritması Ve Destek Vektör Makineleri İle Yapay Sinir Ağlarının Sınıflandırma Başarılarının Karşılaştırılması: İmalat Sektöründe Bir Uygulama. Doktora Tezi. Atatürk Üniversitesi, Erzurum, 2012.

Yalçın, L., Sağlık Sektöründe Veri Madenciliği. Yüksek Lisans Tezi. Milli Savunma Üniversitesi, İstanbul, 2019.

Yalçın Pirinççiler, E.C., Şen, A., Süreç İyileştirme Çalışmalarında Veri Madenciliği Yaklaşımının Kullanılması Üzerine Bir Çalışma, Muğla Sıtkı Koçman Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, Sayı 29, 57-77, Güz 2012.

Yangın, A. Yapay Sinir Ağı Teknikleri Kullanarak Eğitim Yayıncılığı Sektöründe Veri Madenciliği. Yüksek Lisans Tezi. İstanbul Aydın Üniversitesi, İstanbul, 2017.

Yıldırım, M., İldeki Kurumlar Arası Çalışma Performansının Arttırılmasında Veri Madenciliği Tekniklerinin Kullanılması. Yüksek Lisans Tezi. Fırat Üniversitesi, Elazığ, 2016.

Yıldız, H., Finans Sektöründe Veri Madenciliği Kredi Skorlama. Yüksek Lisans Tezi. Milli Savunma Üniversitesi, İstanbul, 2019.

Yıldız, K., Kumaş Hatalarının Isıl Görüntüleme ve Görüntü İşleme Teknikleri İle Tespit Edilmesi. Doktora Tezi. Marmara Üniversitesi, İstanbul, 2014.

Yurdakul, S., Veri Madenciliği ile Lise Öğrenci Performanslarının Değerlendirilmesi. Yüksek Lisans Tezi. Kırıkkale Üniversitesi, Kırıkkale, 2015.