T.C.
PAMUKKALE ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

NANOFİBER KAPLI FİLTRE MALZEMELERİNİN KALİTE STANDARTLARININ BELİRLENMESİNDE VERİ MADENCİLİĞİ

YÜKSEK LİSANS TEZİ

AYLİN SABANCI

DENİZLİ, HAZİRAN - 2019

T.C.
PAMUKKALE ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

NANOFİBER KAPLI FİLTRE MALZEMELERİNİN KALİTE STANDARTLARININ BELİRLENMESİNDE VERİ MADENCİLİĞİ

YÜKSEK LİSANS TEZİ

AYLİN SABANCI

DENİZLİ, HAZİRAN - 2019

KABUL VE ONAY SAYFASI

Aylin SABANCI tarafından hazırlanan "**Nanofiber Kaplı Filtre Malzemelerinin Kalite Standartlarının Belirlenmesinde Veri Madenciliği**" adlı tez çalışmasının savunma sınavı 17.06.2019 tarihinde yapılmış olup aşağıda verilen jüri tarafından oy birliği / oy çokluğu ile Pamukkale Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı Yüksek Lisans Tezi olarak kabul edilmiştir.

Jüri Üyeleri	İmza
Danışman Prof. Dr. Sezai TOKAT	.....
Üye Prof. Dr. Ecir Uğur KÜÇÜKSİLLE	.....
Üye Dr. Öğr. Ü. Elif HAYTAOĞLU	.....

Pamukkale Üniversitesi Fen Bilimleri Enstitüsü Yönetim Kurulu'nun 17/07/2019 tarih ve 29/11-02 sayılı kararıyla onaylanmıştır.

.....

Prof. Dr. Uğur YÜCEL

Fen Bilimleri Enstitüsü Müdürü

Bu tezin tasarımı, hazırlanması, yürütülmesi, araştırmalarının yapılması ve bulgularının analizlerinde bilimsel etiğe ve akademik kurallara özenle riayet edildiğini; bu çalışmanın doğrudan birincil ürünü olmayan bulguların, verilerin ve materyallerin bilimsel etiğe uygun olarak kaynak gösterildiğini ve alıntı yapılan çalışmalara atfedildiğine beyan ederim.

Aylin SABANCI

ÖZET

NANOFİBER KAPLI FİLTRE MALZEMELERİNİN KALİTE STANDARTLARININ BELİRLENMESİNDE VERİ MADENCİLİĞİ
YÜKSEK LİSANS TEZİ
AYLİN SABANCI
PAMUKKALE ÜNİVERSİTESİ FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
(TEZ DANIŞMANI:PROF.DR. SEZAİ TOKAT)

DENİZLİ, HAZİRAN - 2019

Günümüzde teknolojinin gelişmesiyle birlikte bilgiye verilen önem ve veri miktarı artmaktadır. Büyük miktarlardaki verinin araştırılması, analiz edilmesi ve veri yığınları arasından anlamlı bilgiler ortaya çıkarılması için veri madenciliği tekniklerinin kullanılması gerekmektedir.

Bu tez çalışmasında, veri madenciliği tekniklerinin endüstriyel bir probleme uygulanması üzerinde durulmuştur. Çalışmada kullanılan veri seti nanoteknoloji alanında faaliyet gösteren özel bir şirketten alınmıştır. EN 779:2012 kalite standartlarına göre filtre sınıfının belirlenmesi amacıyla nanofiber kaplı filtre malzemelerinin laboratuvar ölçümleri veri madenciliği yöntemleriyle analiz edilmiştir. C4.5 Karar Ağacı, Rastgele Orman, Yapay Sinir Ağları, Naive Bayes sınıflandırma algoritmaları ile k-means ve bulanık c-ortalama kümeleme algoritmaları kullanılmıştır. Veri analizleri RStudio geliştirme ortamında R programlama dili ile gerçekleştirilmiş.

Verilerin normalizasyonu aşamasında; minimum-maksimum, ondalık ölçeklendirme, z-değeri ve sigmoid normalizasyon yöntemleri karşılaştırılmıştır. Yapılan analizler sonucunda k-en yakın komşu algoritması kullanılarak elde edilen 0.8902 doğruluk değeri ile en başarılı yöntemin sigmoid normalizasyon yöntemi olduğu tespit edilmiştir. Verilerin sınıflandırılması aşamasında; model performans değerlendirme yöntemi olarak hold-out performans değerlendirme yöntemi ve k-kat çapraz geçerleme yöntemi uygulanmıştır. Sınıflandırma algoritması olarak C4.5 Karar Ağacı, Rastgele Orman, Naive Bayes ve Yapay Sinir Ağları algoritmaları kullanılarak modeller oluşturulmuştur. Modellerin performansları doğruluk, kesinlik, duyarlılık, F-ölçütü ve Kappa değerine göre karşılaştırılmıştır. Bilgi kazancına dayalı özellik seçim algoritması veri setine uygulanarak özellikler önem derecesine göre sıralanmıştır. Verilerin kümeleme aşamasında; k-means ve bulanık c-ortalama algoritmaları kullanılarak modeller oluşturulmuştur. Modellerin performansları entropi ve saflık başarı ölçütlerine göre karşılaştırılmıştır.

ANAHTAR KELİMELER: Veri Madenciliği, Sınıflandırma, Kümeleme, Nanoteknoloji, Nanofiber malzeme, EN 779:2012

ABSTRACT

DATA MINING IN DETERMINING THE QUALITY STANDARDS OF NANOFIBER COATED FILTER MATERIALS

MSC THESIS

AYLİN SABANCI

PAMUKKALE UNIVERSITY INSTITUTE OF SCIENCE

COMPUTER ENGİNEERİNG

(SUPERVISOR:PROF.DR. SEZAİ TOKAT)

DENİZLİ, JUNE 2019

Nowadays with the evolution of technology, the importance of information and the amount of data increases. Data mining techniques should be used to investigate and analyze large amounts of data and extract meaningful information from among the data stacks.

In this thesis study, the application of data mining techniques to an industrial problem is emphasized. The data set used in the study was obtained from a private company operating in the field of nanotechnology. In order to determine the filter class according to EN 779: 2012 quality standards, laboratory measurements of nanofiber coated filter materials were analyzed by data mining methods. C4.5 Decision Tree, Random Forest, Artificial Neural Networks, Naive Bayes classification algorithms and k-means and fuzzy c-means clustering algorithms were used as data mining methods. Data analysis was performed with R programming language in RStudio development environment..

During the normalization of data; the minimum-maximum, decimal scaling, z-value and sigmoid normalization methods were compared. The most successful method was found to be the sigmoid normalization method with the accuracy value of 0.8902 obtained using the nearest neighbor algorithm. In the classification stage of the data; hold-out performance evaluation method and k-fold cross validation method were applied as model performance evaluation method. Models were created using C45 Decision Tree, Random Forest, Naive Bayes and Artificial Neural Networks algorithms as the classification algorithm. The performances of the models were compared according to accuracy, precision, recall, F-criterion and Kappa value. In the clustering stage of the data; models were created using k-means and fuzzy c-mean algorithms. The performances of the models were compared according to the success criteria of entropy and purity.

KEYWORDS: Data Mining, Classification, Clustering, Nanotechnology, Nanofiber metarials, EN 779:2012

İÇİNDEKİLER

Sayfa

ŞEKİL LİSTESİ

Sayfa

TABLO LİSTESİ

Sayfa

SEMBOL LİSTESİ

IBM	**:**	International Business Machines
VGCNF / VE	**:**	Vapor-Grown Carbon Nanofiber/Vinyl Ester
FCM	**:**	Fuzzy c-means - Bulanık c-ortalamalar
SOM	**:**	Self Organizing Maps
PLC	**:**	Programmable Logic Controller
VTYS	**:**	Veri Tabanı Yönetim Sistemi
SQL	**:**	Structured Query Language
IMS	**:**	Information Management System
CRISP-DM	**:**	Cross-Industry Standard Process for Data Mining
KDD	**:**	Knowledge and Database Discovery
CART	**:**	Classification and Regresion Trees
RSM	**:**	Response Surface Methodology - Yüzey Tepki Yöntemi
YSA	**:**	Yapay Sinir Ağları

ÖNSÖZ

Veri madenciliği kavramı, büyük veri yığınları arasından ham veriyi araştırıp, işleyerek faydalı bilgiye dönüştürme süreci olarak tanımlanabilir. Her geçen gün tarihte daha önce olmadığı kadar büyük bir hızla veri üretilmektedir. Verilerin bilgiye dönüştürülmesi aşamasında kullanıcıların ya da iş analistlerinin aklına gelmeyen sorgular olabilmektedir. Bunun için de veri madenciliği ortaya çıkmıştır. Veriler arasında gizli kalmış bilgiler, desenler veri madenciliği teknikleri ile ortaya çıkarılabilmektedir. Günümüzde neredeyse bilim, iş ve mühendislik alanlarının tümünde büyük, karmaşık ve bilgi bakımından önemli verilerin anlaşılması ortak ihtiyaç haline gelmiştir. Verilerden anlamlı bilgiler çıkartmak ve bu bilgilerin işlenmesi büyük öneme sahiptir.

Bu tez çalışmasında nanofiber kaplı filtre malzemelere ait endüstriyel ortamdan elde edilen ölçümler veri madenciliği yöntemleri ile sınıflandırılmıştır. Veri madenciliği yöntemleri olarak C4.5 karar ağacı, rastgele orman sınıflandırma algoritması ile yapay sinir ağları, Naive Bayes teknikleri ve k-means, bulanık c-ortalamalar kümeleme algoritmaları kullanılmıştır.

Bu tez çalışmasının yürütülmesinde değerli fikir ve önerileriyle bana yol göstererek ilgi ve desteğini esirgemeyen danışman hocam Prof. Dr. Sezai TOKAT'a teşekkürlerimi sunarım. Eğitim ve çalışma hayatım boyunca her zaman yanımda olan maddi ve manevi desteklerini en önemlisi de sevgilerini hiç esirgemeyen aileme içtenlikle teşekkür ederim.

1. GİRİŞ

Günümüzün hızla gelişen bilgi teknolojilerine paralel olarak endüstri alanında toplanan veri miktarında da artış olmuştur. Artan bu verilerin analiz edilmesi ve yönetilmesi gün geçtikçe ihtiyaç haline gelmektedir. Bu ihtiyaçlar doğrultusunda hem akademide hem iş dünyasında veri madenciliği uygulama alanları gelişmiştir.

Veri madenciliği, sahip olduğu istatistik ve matematik gibi disiplinlerin bir arada kullanıldığı yöntemlerle verilerden birtakım örüntüler veya kurallar elde edilmesini, karar destek sürecine katkı sağlayacak bilginin ortaya çıkarılmasını ve araştırılan konuya yönelik tahminlerde bulunulmasını sağlar. Veri madenciliği yöntemleri ve algoritmalarının yazılımlarla bilgisayar ortamında kolayca uygulanabilmesi, bu kavramın oldukça geniş bir kullanıcı kesimine ve araştırma alanına erişmesini sağlamıştır. Böylece bilginin artan önemine paralel olarak veri madenciliğinin öneminin de arttığı söylenebilir (Erken 2017).

Veri tabanı sistemlerinde meydana gelen gelişmeler tıp, bankacılık, ekonomi, finans gibi birçok alanda veri madenciliği tekniklerinin de gelişmesini sağlamıştır. Tahminleme, modelleme ve yapay zeka teknikleriyle yapılan uygulamalarda amaç maliyetleri düşürmek, satışları arttırmak ve araştırma-geliştirme çalışmalarını daha etkin kılmaktır (Gürsoy 2011).

Birçok alanda uygulanan veri madenciliği farklı işlevleri yerine getirmek için farklı algoritmalar içerir. Algoritmalar verileri inceler ve incelenen verinin özelliklerine en yakın olan modeli belirler. Veri madenciliği ana teknik olarak sınıflandırma ve tahmin etme, kümeleme, ilişkilendirme kuralları, zaman serisi analizi ve metin madenciliği ile sosyal ağ analizi ve duyarlılık analizi gibi bazı yeni teknikleri içerir.

Bilim ve teknolojinin hızla gelişmesi, yeni ve çok disiplinli teknolojilerin doğmasına da öncülük etmektedir. Nanoteknoloji son yıllarda önemli gelişim kaydederek birçok ülke tarafından stratejik alan olarak belirlenmiştir (UNESCO, 2015). Avrupa Komisyonu tarafından nanoteknoloji, daha yeşil bir ekonomiye geçişi

destekleyen birkaç endüstriyel uygulama alanında sürdürülebilir rekabet edebilirliğe ve büyümeye katkıda bulunan altı "Anahtar Etkinleştirme Teknolojisinden" biri olarak kabul edilmiştir (EC, 2012). Yeni, yüksek katma değerli ve üstün özellikli ürünlerin üretilmesine olanak sağladığı için nanoteknoloji rekabet gücüne dolaylı olarak etki edebilecek bir teknoloji olarak görülmektedir (Sevinç 2017). Nanoteknoloji uygulamalarını kimya, biyoloji, fizik, tıp, enerji, bilişim, malzeme, elektronik, uzay, tarım gibi birçok alanda görmek mümkündür (Sevinç 2017).

Birçok endüstri alanında devrim niteliğinde olan nanoteknoloji, başta yapay sinir ağları olmak üzere veri madenciliği tekniklerine konu olmuştur. Diğer veri madenciliği tekniklerinde sınırlı sayıda araştırmalar yapıldığı görülmüştür. Bu çalışmada nano boyuttaki malzemelerden olan nanoliflere sınıflandırma ve kümeleme veri madenciliği teknikleri uygulanmıştır. Filtre standartlarına göre sınıflandırılan nanofiber kaplı filtre malzemelerinin filtre sınıfını tahmin etmede hangi tekniklerin daha başarılı olduğu ve filtre numunelerine kümeleme tekniklerinin uygulanması konularında literatüre katkı sağlayabilir.

Bu tez çalışmasında nanolif kaplı filtre malzemelerinin Avrupa Standardı olan EN779:2012 standardına göre sınıflandırılmasında kullanılmak üzere TSI 8130 cihazından alınan ölçümler veri madenciliği yöntemleri ile sınıflandırılmıştır. Veri madenciliği yöntemleri olarak C4.5 karar ağacı, rastgele orman, yapay sinir ağları, Naive Bayes sınıflandırma algoritmaları ve k-means, bulanık c-ortalama kümeleme algoritmaları kullanılmıştır.

Tezin akışı şu şekilde planlanmıştır: Bölüm 2'de nanoteknoloji ve nanolifler ile ilgili bilgiler verilerek, nanofiber filtre standardı ve filtre test süreci anlatılmıştır. Bölüm 3'te veri, veri tabanı ve veri ambarı kavramlarından, veri madenciliği sürecinden ve tez kapsamında uygulanan veri madenciliği yöntemlerinden bahsedilerek model değerlendirme ölçütleri detaylandırılmıştır. Bölüm 4'te nanoteknoloji alanında yapılan veri madenciliği çalışmalarının literatür taraması belirli bir metodolojiye göre belirtilmiştir. Bölüm 5'te tez kapsamında kullanılan veri seti, normalizasyon işlemleri ve uygulanan veri madenciliği yöntemlerinin sonuçları gösterilmiştir. R dili ve R Studio kullanılarak analizi yapılan modellerin değerlendirilmeleri yapılmıştır. Araştırmaların bulgu ve yorumları Bölüm 6'da açıklanacaktır.

2. NANOTEKNOLOJİ

2.1 Nanoteknolojinin Tanımı ve Kullanım Alanları

Yunanca "cüce" anlamına gelen "nanos" sözcüğünden türetilen nano terimi, bir metrenin milyarda bir kısmını tanımlar. Bir nanometre düzeyinde yaklaşık olarak 10 atom sığabilmektedir (Miller ve diğ. 2004). ABD Ulusal Nanoteknoloji Girişimi'nin tanımına göre nanobilim ve nanoteknoloji; "Kimya, biyoloji, fizik, malzeme bilimi ve mühendislik alanları gibi birçok alanda uygulama alanı bulan, oldukça küçük maddelerin çalışma ve uygulama alanıdır." (Web-1)

Nanobilim, nano ölçekte ortaya çıkan yenilikleri kuantum kuramlarıyla anlamamıza yardımcı olan bilim dalıdır (Sevinç 2017). Nanoteknoloji ise nanometre ölçeğinin şeklini ve boyutunu kontrol ederek yapıların, cihazların ve sistemlerin tasarımı, üretimi ve uygulamasıdır (Web-2). Bir başka deyişle nanoteknoloji, nesnelerin yapıtaşı olan atomları istendiği şekilde düzenlenmesini, her alanda daha dayanıklı malzemelerin üretimini ve doğaya az zarar veren üretim yapılmasını sağlayan teknolojidir (Özdoğan ve diğ. 2006).

Nanoteknoloji ile ilgili ilk kavramları, yirminci yüzyılın önemli fizikçilerinden Richard Feynman 1959'da nanobilimin ve nanoteknolojinin başlangıcı olarak kabul edilen "Aşağıda bolca yer var" başlıklı konuşmasında ortaya koymuştur (Feynman, 1992). 1974'te Norio Taniguchi bir konferansta ilk kez nanoteknoloji terimini kullandıktan yedi yıl sonra IBM, icat ettikleri taramalı tünelleme mikroskobu ile firma logosunu Ksenon atomlarını dizerek nano ölçekte oluşturmuştur (Körözlü 2016). 1986'daki ilk atomik kuvvet mikroskobu icadı ile nano boyutta görüntüleme, ölçme ve malzeme işleme imkânı sağlanmıştır. Bundan sonra hızla ilerleyen bilimsel çalışmalar ile 2000'li yılların başından itibaren nanoteknoloji her alanda hayatımıza girmiştir (Körözlü 2016).

Nanoteknoloji, tek tek atomlardan veya moleküllerden mikron altı boyutlarına kadar değişen ölçeklerde fiziksel, kimyasal ve biyolojik sistemlerin üretimini ve uygulanmasını ayrıca elde edilen nanoyapıların daha büyük sistemlere entegrasyonunu kapsar (Bhushan 2017). Nanoteknolojinin bazı alt alanları kısaca şu şekildedir; "Nanoyapılar" boyutun nano ölçekli olduğu nesne ve yapılardır. "Nanoparçacık" nanoyapının en basit şekli olan boyutu 100 nanometrenin altındaki nano elemandır. "Nanotüp" biraz daha karmaşık nanoyapılar oluşturabilecek tek boyutlu nano elementtir (Ramakrishna ve diğ. 2005).

Nanoteknolojide "aşağıdan yukarı" ve "yukarıdan aşağı" olmak üzere iki farklı yaklaşım kullanılmaktadır. Aşağıdan Yukarı yaklaşımında, küçük bileşenler bir araya getirilerek daha karmaşık bileşenler haline getirilmeleri sağlanır. Bu yaklaşımdaki amaç aşamalı bir şekilde küçük bileşenlerin bir araya getirilmesi ve atomlar üzerinde değişiklik yaparak daha büyük yapıların elde edilmesidir (Schmid 2008). Yukarıdan aşağı yaklaşımında ise makro yapılardan nano yapılar elde edilmesi yöntemiyle büyük boyutlardaki materyallerin küçük boyutlara dönüştürülmesi amaçlanır (Miller ve diğ. 2004).

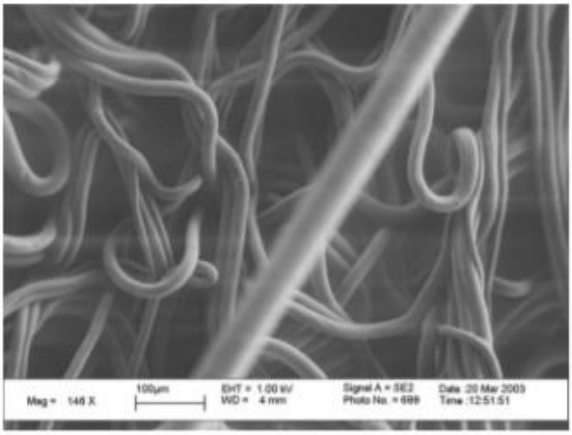
Nanoteknoloji; üretim, nanoelektronik, bilgi teknolojileri, elektrik-elektronik, biyomedikal, otomotiv, kimya, sağlık, enerji, biyoteknoloji ve ulusal güvenlik gibi farklı alanlarda büyük atılımlar vaat etmektedir (Gürmen ve Ebin 2008; Bhushan 2017). Bilgisayarlarda, moleküler biyolojide ve diğer birçok alanda gözlemlendiği gibi yakın gelecekte nanoteknolojinin ekonomi ve toplumda önemli bir rol oynayacağı tahmin edilmektedir (Amin ve Mohammadyani 2011). Nanometre ölçeklerindeki malzemelerin üstün fiziksel özellikleri kullanılarak teknoloji alanında devrim niteliğinde olacak yeni ürünler elde edilebilir. Malzemelerin özellikleri nano boyuttaki ölçeklerine yaklaştıkça yeni özellikleri ortaya çıkmaktadır.

Nanoteknoloji farklı kullanım alanına sahiptir ve her geçen gün buna yenileri eklenmektedir. *Malzeme ve imalat sektöründe,* malzemelerin atomik ve moleküler boyutlarda inşa edilmesi, geleneksel metotlarla elde edilen malzemelere göre daha dayanıklıdır. Bu malzemeler nispeten daha düşük hata seviyelerine ve dayanıklılığa sahip olduğu için mevcuttaki birçok endüstriyel süreç için devrim niteliğindedir (Özdoğan ve diğ., 2006). *Tekstil sektöründe* nanoteknoloji tekstil malzemelerinin mevcut işlevlerini ve performanslarını geliştirmiştir. Örneğin; spor kıyafetleri, dağ

kıyafetleri, askeri kıyafetler için yeni karma ürünler uygulanmaktadır. Bu ticari ürünlerin dışında tamamen yeni özelliklere ve fonksiyonlara sahip akıllı tekstil ürünlerinin de geliştirilmesini sağlamaktadır. Kendi kendini tamir eden ve temizleyen kumaşlar, zararlı ultraviyole ışınlardan koruyan giysiler, su tutmayan kumaşlar bunlara örnektir (Kaounides ve diğ. 2007). *Tıp ve sağlık sektöründe* bakteriler, virüsler ve DNA, onlarca nanometre boyutundadır; kırmızı kan hücreleri, nöronlar ve kılcal damarlar onlarca mikron boyutundadır. Tıbbi teşhis ve tedavi için kullanılan aletler nano düzeyde tasarlanan bilgisayar teknolojilerine dayanmaktadır. Kanser tedavisi, felçli hastalar için beyin-makine arayüzleri geliştirilmesi gibi çalışmalarda bu yüzden nanoteknolojiden yararlanılması gerekmektedir (Andrews, 2019). Nanoteknoloji sayesinde geliştirilen nanostent, kalp hastalarının iyileşme sürecine yardımcı olur ve kanın pıhtılaşmasını önler (Naschie 2006). Bunlara ek olarak, sadece hastalığın bulunduğu veya yayıldığı bölgelere ilaç veren makineler, insan vücudu içinde hareket edilmesine olanak sağlayan teşhis araçları nanoteknolojinin tıp ve sağlık sektöründeki potansiyel uygulama örneklerinden bazılarıdır (Özdoğan ve diğ. 2006). *Havacılık sektörü ve uzay* araçlarının imalatı yüksek maliyet gerektirmektedir. Araçların imalatında kullanılan malzemelerin ağırlığı maliyet artışını etkileyen faktördür. Nanoteknoloji bu malzemelerin ağırlığını azaltırken hem de maliyetini düşürecektir. Ayrıca nanotüplerin çekme direncinin çelikten daha yüksek olması sayesinde atmosfere kadar yükselecek yapıların inşa edilmesi potansiyel uygulama alanlarının içerisinde yer alır (Özdoğan ve diğ., 2006). *Enerji alanında* nanoenerji, dünyanın petrole dayanan enerji bağımlılığını azaltarak sürdürülebilir ve yenilenebilir enerji dönüşümünü hızlandıracaktır (Özer 2008). Nanoteknolojideki son gelişmeler, yeni nesil su temini sistemlerinin geliştirilmesi için önemli fırsatlar sunmaktadır. Yoğun taşıma ve merkezi kontrol gerektiren mevcut su arıtma, dağıtım ve deşarj uygulamaları yerine nanoteknolojinin sağladığı yüksek verimli, modüler ve çok işlevli süreçlerinin uygun fiyatlı su ve atık su arıtma çözümleri sağlaması öngörülmektedir (Qu vd, 2013). Nanoteknoloji askeri uygulamalar konusunda da kullanılmaktadır. Geliştirilmiş elektronik savaş kapasitesi, silah sistemi ve geliştirilmiş kamuflaj ve akıllı sistemler araştırmaların gerçekleştirildiği alanlardır (Özdoğan ve diğ., 2006).

2.2 Nanolifler ve Filtrasyon Uygulaması

Nanolifler, bir mikrondan daha küçük çapa sahip lifler olarak tanımlanır. Basit donanımlar ve az enerji ile üretilebilmeleri nanolifleri cazip hale getirmektedir (Kozanoğlu 2006). Nanolif, çapı açısından bir nano malzemedir ve birleşik nanolifler oluşturmak için nano parçacıklar ile doldurulursa nano yapılı malzeme olarak kabul edilebilir (Ramakrishna ve diğ. 2005). Yaklaşık 20 yıldır araştırmaları devam eden nanoliflerin çapı 50-300 nanometre arasında değişmektedir (Özdoğan ve diğ. 2006).



Şekil 2.1: Nanoliflerin Mikroskop Görüntüsü (Özdoğan ve diğ. 2006)

Nanolifleri üretmek için Kendiliğinden Tutunma (Self-assembly), Faz Ayırımı (Phase Seperation), Şablon Sentez (Template Synthesis), Eriyik Üfleme (Melt-blown), Elektro Çekme (Elektospinnig) gibi yöntemler kullanılmaktadır. Elektrospinning, bugünlerde geniş çapta sürekli nanolifler üretmek için en umut verici tekniktir ve lif çapı nanometreden mikrona ayarlanabilir (Fang ve diğ. 2008). Lifli nano malzemeler, yüksek gözeneklilik ve geniş yüzey alanları nedeniyle birçok uygulama için etkilidir. Elektro Çekme yöntemi (elektrospinning), polimerler, kompozitler, seramik gibi nanoliflerin üretilmesi için basit, çok yönlü bir tekniktir (Schaefer 2010). Elektrospinning tekniği üzerine ilk patent 1934'te yayınlanmasına rağmen, bu teknik son zamanlara kadar iyi bir şekilde kurulmamıştır (Fang ve diğ. 2008).

Elektrospinning yöntemi ile üretilen nanoliflerin kullanım alanları oldukça geniştir. Nanoliflerden oluşan yapıların birim ağırlıkta sağladığı yüksek alan özelliği, yumuşak tutumu, iyi mukavemet/birim ağırlık özelliği ve mikroorganizmalar ile ince parçacıklara engel oluşturması gibi başlıca sebeplerden dolayı nanolifler birçok endüstri alanında kullanılır (Çakmak 2011). Nanoliflerin kullanım alanları arasında tıbbi protezler, koruyucu giysiler, elektriksel ve optik sensörler, malzeme kuvvetlendirici kompozitler, bitki koruma örtüleri, yara örtücüler, dış cephe kaplama uygulamaları gibi alanlar bulunmaktadır (Kozanoğlu 2006).

Nanoliflerin uygulama alanlarından biri de filtrasyon uygulamalarıdır (Üstün, 2011). Filtreler, hem evlerde hem de endüstride, havadan veya sıvıdan maddelerin uzaklaştırılmasında yaygın olarak kullanılmaktadır. Çevreyi korumaya yönelik filtreler, kirletici maddeleri havadan veya sudan ayrıştırmak için kullanılır (Fang ve diğ. 2008). Endüstride, genelde filtre yapıları temiz hava sağlamak için kullanılır. Bu filtrelerin yaklaşık olarak 0,5 μm boyutundaki yağ parçacıklarını tutması gerekmektedir. Elektrospinning yöntemiyle mikrondan daha küçük çapa sahip olan lifler elde edildiği için bu parçacıkların uzaklaştırılması kolaydır (Graham ve diğ. 2002).

Nanolifler uzun filtre ömrü ve yüksek tutuş kapasitesi nedeniyle tercih edilmektedir (Grafe ve Graham 2003). Azalan lif çapı filtreyi daha dolgun hale getireceğinden dolayı akışa karşı koyan çarpma ataletleri ve engel olma isteği artarak parçacıklar daha sık yakalanacaktır (Kozanoğlu 2006). Ayrıca filtrasyon için kullanılan fibrilli materyallerin yüksek filtrasyon verimliliği ve düşük hava direnci avantajları vardır. Fiber inceliği ile yakından ilgili olan filtrasyon verimliliği filtre performansını belirleyen en önemli özelliktir (Çakmak 2011). Gözenekli yapısı ve yüksek yüzey alanı sayesinde nanoliflerden oluşturulmuş yüzeylerin filtrasyon amacıyla kullanılması da fayda sağlamaktadır. Bu yüzeyler 1 mikrondan küçük parçacıkların filtrelenmesini sağlar. Nanoliflerden oluşan kumaşların 100 nm'den daha küçük parçacıkları sıvıdan veya gazdan uzaklaştırması nanoliflerin filtre amaçlı kullanımını sağlamaktadır (Çakmak 2011).

Önemli endüstriyel uygulamaya sahip filtrelerin standardizasyonu üzerinde de durulmuştur. Avrupa'da hava filtreleri için birincil standart, Eurove 4/5 ve EN 779'dur. EN 779, Avrupa Standardizasyon Komitesi tarafından kontrol edilir.

Amerika Birleşik Devletleri'nde ise, havalandırma filtrasyonu için standartların geliştirilmesi ABD Çevre Koruma Ajansı ve ASHRAE tarafından ortaklaşa desteklenmektedir (Hutten 2007).

Hava filtreleri yakaladıkları parçacık büyüklüğüne göre kaba, ince ve hassas olmak üzere üç kategoriye ayrılmaktadır. Kaba ve ince parçacıkların yakalanmasında kullanılan filtreler EN 779 standardında (Genel Havalandırma İçin Partikül Hava Filtreleri) tanımlanmıştır (Dinçer ve diğ 2018).

Bu filtre standartları, filtreleri verimliliklerine göre sınıflandırır. Sınıf numarası yükseldikçe filtre verimliliği artar. Avrupa test standartlarında filtreler, yakalama ve toz lekesi verimliliklerine göre sınıflandırılır (Alan ve Tercan 2013). Bu sınıflandırma Şekil 2.2'de gösterilmektedir.

Filtre Grubu	Sınıf	Son Basınç (Pa)	Ortalama Yakalama Verimi (%)	0,4 µµm'daki Ortalama Verimi (%)	0,4 µm'daki Minimum Verimi (%)
Kaba	G1	250	50 ≤ Am < 65	-	-
	G2	250	65 ≤ Am < 80	-	-
	G3	250	80 ≤ Am < 90	-	-
	G4	250	90 ≤ Am	-	-
Orta	M5	450	-	40 ≤ Em < 60	-
	M6	450	-	60 ≤ Em < 80	-
Hassas	F7	450	-	80 ≤ Em < 90	35
	F8	450	-	90 ≤ Em < 95	55
	F9	450	-	95 ≤ Em	70

Şekil 2.2: EN 779:2012 Sınıflandırması (Web-4)

Filtre sınıfları aşağıdaki gibi belirtilmiştir (Web-4):

G Sınıfı Filtreler: 0,4 μm boyutundaki parçacıklara karşı ortalama verimliliği %40'dan küçük olan filtrelerdir. G sınıfı (G1 – G4) filtrelerin verimlilik değeri "< %40" olarak ifade edilir ve sınıflandırılması toz yükleme oranına bağlı ortalama yakalama değerine dayanmaktadır.

M Sınıfı Filtreler: 0,4 μm boyutundaki parçacıklara karşı ortalama verimliliği %40 ile %80 arasında olan filtrelerdir. Sınıflandırması 0,4 μm 'daki ortalama verimine göredir. Daha önce F5 ve F6 olarak ifade edilen filtre sınıfının diğer tüm özellikleri M5 ve M6 için geçerlidir.

F Sınıfı Filtreler: 0,4 μm boyutundaki parçacıklara karşı ortalama verimliliği %80 ve üzeri olan filtrelerdir. Sınıflandırması 0,4 μm 'daki ortalama verime ve test süresince gerçekleşen minimum verime göredir.

Nanofiber kaplı filtre malzemelerinin sınıflandırma işleminde TSI 8130 cihazı kullanılmaktadır. Sınıflandırma süreci aşağıda belirtilmiştir:

Operatör veya robot, filtre tutucunun alt yarısına bir filtre yerleştirir. Test, çift çalıştırma düğmelerine basılarak veya programlanabilir mantık denetleyicisi (PLC) aracılığıyla test cihazına bir "START" sinyali gönderilerek başlatılır. Hava basınçlı silindir, filtre tutucunun üst yarısını hızla düşürür ve aerosol filtreden geçirilir. İki ışık saçılımlı lazer fotometresi, aynı anda yukarı ve aşağı yönde aerosol konsantrasyon seviyelerini ölçer. Partikül penetrasyon değeri, bu iki okuma oranından belirlenir.

Bir yerine iki lazer fotometresi kullanılarak, ölçüm döngüsü süresi azaltılır ve ölçüm doğruluğu artırılır. Son derece hassas elektronik basınç transdüserleri, filtre direncini ve akış hızını belirler. Sıfır dengesi ve arka plan değerlerini belirlemek için her test arasında geliştirilmiş basınç ve fotometre değerleri alınır.

Test tamamlandığında, filtre tutucu otomatik olarak açılır. Tüm test verileri, yazıcı veya seri çıktı kullanılarak görüntülenebilir.

3. VERİ MADENCİLİĞİ

Veri madenciliği kavramı, veritabanları veri ambarları, Web, diğer bilgi kaynakları veya sisteme dinamik olarak akan veriden elde edilen büyük veri yığınları arasından ham veriyi işleyerek faydalı bilgiye dönüştürme süreci olarak tanımlanabilir (Han ve diğ. 2012). Veri miktarının artması ve teknolojinin gelişmesiyle bilgiye verilen önem artmıştır. Verilerin bilgiye dönüştürülmesi aşamasında kullanıcıların ya da iş analistlerinin aklına gelmeyen sorgular olabilmektedir. Bunun için de veri madenciliği ortaya çıkmıştır. Veriler arasında gizli kalmış bilgiler ve desenler veri madenciliği teknikleri ile ortaya çıkarılmaktadır.

Coşkun ve Baykal (2011)'a göre veri madenciliğindeki amaç, bilgi çıkarımı zor olan büyük verilerin analiz edilerek faydalı bilgiler ortaya çıkarılmasıdır. Buna ek olarak, ortaya çıkarılan bu bilgileri içeren model oluşturularak gelen yeni bir veri nesnesi hakkında yorum yapmak ve yeni veri hakkında tahminde bulunmayı sağlamakta amaçlanmaktadır. Baykasoğlu (2005)'na göre veri madenciliği istatistiksel bir yöntemler serisi olarak görülebilir. Amaç, modellerin kolaylıkla mantıksal kurallara veya görsel sunumlara çevrilebilmesidir. Şekil 3.3'de belirtildiği gibi veri madenciliği yapay zeka, makine öğrenmesi, istatistik, veri tabanı sistemleri ve veri görselleştirme gibi alanlarla yakın ilişkilidir.

Şekil 3.3: Veri Madenciliği Alanları (Baykasoğlu 2005)

Veri madenciliği konusunda çeşitli tanımlar yapılmıştır; Özkan (2008)'a göre veri madenciliği, büyük ölçekli veriler arasından "değerli olan" bilgiyi elde etme işidir. Terzi ve diğ. (2011) veri madenciliğini bir yöntem değil de süreç olduğunu belirterek; büyük veri yığınları içerisinden gelecekle ilgili tahminlerde bulunmamızı sağlayacak bağıntıların bilgisayar programıyla ortaya çıkarılması olarak tanımlanmıştır. Alpaydın (2000)'a göre veri madenciliği büyük miktardaki veri içerisinden gelecek ile ilgili tahminlerde bulunmamızı sağlayacak bağıntıların, kuralların bilgisayar programları kullanılarak aranmasıdır. Baykasoğlu (2005)'na göre veri madenciliği elde var olan bilgilerden üstü kapalı kalmış, net olmayan veya önceden bilinmeyen fakat potansiyel olarak kullanışlı olan bilginin ortaya çıkarılmasıdır. Han ve diğ. (2012) veri madenciliğini veri tabanları, veri ambarları, web veya diğer bilgi depolarından akan büyük miktardaki verilerden ilginç desenler ve bilgi keşfetme süreci olarak tanımlamıştır. Berry ve diğ. (1997) veri madenciliğini anlamlı desen ve kuralları keşfetmek için büyük miktardaki verilerin araştırılması ve analiz edilmesi olarak tanımlamıştır.

Yapılan birçok tanımda da belirtildiği gibi veri madenciliği, büyük veri yığınları arasından değerli olan bilginin elde edilmesidir. Bu sayede veriler arasındaki bağıntılar bulunabilir, ileriye dönük karar destek sistemlerinde kullanılabilir.

3.1 Veri, Veri tabanı ve Veri Ambarı

Teknolojinin ilerlemesiyle veri miktarında her geçen gün artış olmaktadır. Artan verilerin bilgiye dönüştürülmesiyle kurumlar için karar destek sistemleri oluşturulmaktadır. Veriler üzerinde yapılan analizlerde çeşitli istatiksel ve matematiksel yöntemler kullanılır. Ancak veri sayısının giderek artması sorunları da ortaya çıkaracaktır. Bu tür veriler üzerinde analizler yapabilmek için hem yeni veri tabanı kavramlarına hem de yeni çözümleme yöntemlerine gereksinim vardır. Veriyi yönetmek için "veri ambarı" , verileri analiz ederek yararlı bilgiye erişmek için "veri madenciliği" kavramları ortaya çıkmıştır (Özkan 2008).

Veri, günlük yaşam içerisinde doğal olarak veya bilgi teknolojilerindeki hızlı gelişmelerle elde edilebilmektedir. Ayrıca veri, işlenmemiş bilgi olarak da

tanımlanabilir. Veri ve bilgi kavramları veri tabanlarının, veri ambarlarının temelini oluşturur (Asilkan 2008).

Veri Tabanı

Veri tabanı, birbiriyle ilişkili olan verilerin bir arada tutulduğu, kullanım amacına göre düzenlenmiş olan veriler topluluğunun mantıksal ve fiziksel olarak tanımlarının yer aldığı bilgi deposudur. Başka bir deyişle veri tabanı, birbiriyle ilişkili olan verilerin tekrarlanmaksızın birden fazla amaçla kullanmaya imkân sağlayacak şekilde depolayan yazılımdır (Burma 2009).

Verinin sistematik olarak saklanmasını, güncellenmesini, bakımının yapılmasını gerektiren her uygulama veri tabanı oluşturmak zorundadır. Örnek olarak; marketlerde ürün stoklarının tutulduğu, bankalarda müşteri verilerinin, okullarda öğrencilere ait verilerin, hastanelerde hastalara, personellere ait verilerin saklandığı sistemlerin hepsi veri tabanı sistemine ihtiyaç duymaktadır (Kaya ve Tekin 2007). Veri tabanını mantıksal katman ve fiziksel katman olarak iki seviyeye ayırmak mümkündür. Mantıksal katman; tabloların yer aldığı düşünmesi ve kullanılması insanlar için daha kolay olan katmandır. Fiziksel katman ise disk üzerindeki bloklardan, segmentlerden oluşan bilgisayarın verileri nasıl tutulduğu gibi daha somut şeylerle ilgilenir (Şeker 2013).

Veri tabanı yönetim sistemi (VTYS), yeni bir veri tabanı oluşturmak veya mevcutta bulanan veri tabanını genişletmek, bakımını yapmak, yedeğini almak gibi işlemleri gerçekleştirebildiğimiz birden fazla programdan oluşan yazılım sistemidir. VTYS, kullanıcı ile veri tabanı arasında arabirim oluşturarak veri tabanına her türlü erişimin olmasına olanak sağlar (Burma 2009). VTYS programlarına Microsoft SQL Server, Oracle, MySQL, Microsoft Access, İnformix, Postroge SQL ve Sybase örnek verilebilir. VTYS yazılımlarının çoğunda SQL sorgulama dili kullanılır.

Veri modeline göre VTYS'yi aşağıdaki şekilde sınıflandırabiliriz;

Hiyerarşik Veri Tabanları: Hiyerarşik veri tabanı modelini ağaç yapısına benzetebiliriz. Öncelikle kök olarak bir kayıt ve bu kayıta bağlı dal kayıtlardan oluşur. Veri tabanları için kullanılan ilk model olarak belirtilmektedir. Bu veri tabanı kişisel bilgisayarlarda kullanılmayan sunucu ortamlarında çalışan yazılımlar

tarafından kullanılır (Burma 2009).İlk defa bu modeli IBM firmasına ait IMS kullanmıştır (Kaya ve Tekin 2007).

Ağ Veri Tabanları: Hiyerarşik veri tabanlarının yetersizliği sonucunda ağ veri tabanları geliştirilmiştir. Ağ veri tabanlarında, hiyerarşik veri tabanlarında bulunan ağaç yapısının daha gelişmiş hali olan graflarla verilerin saklanması sağlanır.

İlişkisel Veri Tabanları: E.F.Codd tarafından 1970'li yıllarda geliştirilmiştir. Veriler satır ve sütunlar halinde tablolarda saklanır. Tablolar arasında ilişkiler bulunmaktadır ve veri bütünlüğü bu şekilde sağlanır. Günümüzde kullanılan veri tabanı modelidir. Veriler iki boyutlu bir tablo olarak tutulur.

Nesneye Yönelik Veri Tabanları: Nesneye yönelik programlama ile yine nesneye dayalı bir dil kullanan veri tabanı olarak açıklanabilir. Üç boyutlu bir yapıya sahiptir.

VTYS'nin sağladığı avantajları aşağıdaki şekilde açıklayabiliriz(Burma 2009);

- VTYS programları standart sorgulama dilini kullanırlar.
- Veri tekrarının olmaması ve veri tutarlılığının sağlanmasından dolayı veri bütünlüğü söz konusudur.
- VTYS programı sayesinde verinin çoklu kullanıcı sistemlerde paylaşımı yapılır.
- Farklı veri tabanı programları arasında veri transferi işlemleri yapılabilir.
- Veri tabanı yöneticisi tarafından gruplar, roller oluşturularak yetkilerde değişiklikler yapılabilir. Bu sayede veriler üzerinde güvenlik ve gizlilik vardır.

Veri Ambarı

Veri tabanı sistemlerinin sayılan avantajlarına rağmen karar destek uygulamalarında gereksinimleri karşılamakta zorlanması bunun paralelinde verilerin farklı bir biçimde saklanması ve hızlı şekilde erişiminin sağlanması gereksinimlerinden dolayı "veri ambarı" kavramı ortaya çıkmıştır (Özkan 2008).

Veri ambarı teriminin yaratıcısı W.H.Inmon, "Building the Data Warehouse" kitabında veri ambarını şu şekilde tanımlamıştır; Veri ambarı, konu odaklı, bütünleşmiş, nispeten istikrarlı ve yönetimde karar vermeyi desteklemek için kullanılan tarihsel değişiklikleri yansıtan bir veri kümesidir (Shi ve Li 2010). Bir veri ambarı; metaveri, ayrıntılı veri, eski ayrıntılı veri, düşük düzeyde özetlenmiş veri ve yüksek düzeyde özetlenmiş veri olmak üzere beş seviyede sınıflandırılır (Inmon 2005).

Veri ambarı, farklı kaynaklardan toplanan verilerin veri temizleme, veri entegrasyonu, veri dönüştürme, veri yükleme ve veri yineleme işlemleriyle tek şema altında depolanmasıdır (Han ve diğ. 2012). Veri ambarı, günlük olarak kullanılan veri tabanlarının birleştirilmiş ve işlemeye uygun olan özetini saklamayı amaçlamaktadır (Alpaydın 2000). Bir veri ambarı genelde veri küpü denilen her boyut şemadaki bir nitelik veya bir dizi niteliğe karşı gelen ve her hücre sayım veya toplam gibi bazı ölçü değerlerini depolayan çok boyutlu veri yapısı ile modellenir. Veri küpü çok boyutlu veri görünümü sağlayarak özetlenen verilerin önceden hesaplanarak hızlı erişilmesini sağlar (Han ve diğ. 2012).

Veri ambarı konuya yönelik, bütünleşik, zaman değişkenli ve sadece okunabilen özelliklere sahiptir. Bu özelliklere sahip olması veri ambarını ilişkisel veri tabanı sistemlerinden, veri saklama sistemlerinden ayırt etmede yardımcı olmaktadır (Çakır 2012). Veri ambarının belirtilen özellikleri maddeler halinde detaylandırılmıştır.

Konuya yönelik: İşletmelerde günlük, aylık veya yıllık periyotlarla süreçler ve fonksiyonlar yer almaktadır. İşletimsel veri tabanları tüm bu konulara ağırlık verirken veri ambarı ise sadece karar destek sistemleri için kullanılacak olan bilgilere

odaklıdır. Karar destek sistemlerinde yararlı olmayacak veriler veri ambarının konusu değildir (Özkan 2008).

Örnek olarak bir işletmeye ait personel verileri, muhasebe verileri, stok bilgileri fonksiyonel veriler olurken, stok, müşteri, satıcı, çalışan, bordro gibi veriler veri ambarının konusudur.

Bütünleşik: İşletmelerde yer alan şirket içi ve dış kaynaklı verilerin değerli hale getirilebilmesi için bazı durumlarda birçok kaynaktan verileri birleştirmek gerekmektedir. Bu noktada da veri ambarı işin içine girer. Birçok farklı kaynaktan (ilişkisel veri tabanı, arşivler, dosyalar) toplanan veriler belirlenen tek bir format haline getirilerek bütünleştirilmesi sağlanır. Bütünleştirilen verilerden veri ambarı oluşmaktadır.

Örnek olarak A uygulamasında bir alan "Evet" , "Hayır" ifadelerini alırken B uygulamasında aynı alan "E" , "H" ifadelerini alabilir. Bu gibi durumlarda belirlenen standarda göre veriler üzerinde dönüştürmeler yapılarak veri bütünleştirilmesi yapılmış olur.

Zaman değişkenli: Veri ambarı bir bilginin geçmişteki değerleriyle beraber güncel değerlerini tutarken işlemsel veri tabanlarında ise güncel veriler tutulmaktadır. Veri ambarının bu özelliği sayesinde geçmişe dönük verilerde analizler yapılabilmektedir. Veri ambarında en az beş yıllık verilerin tutulması gerektiği yaygın olarak kabul edilen bir durumdur. İşlemsel verinin zaman boyutunun olmamasından dolayı verilerin güncelleştirme özellikleri bulunabilirken veri ambarında veri işlemsel sistemlerindeki verilerin belirli dönemlerindeki anlık görüntülerinden oluşmaktadır (Özkan 2008).

Sadece okunabilen: Sadece okunabilir olması özelliği, veri ambarında bulunan verilerin değiştirilememesi anlamına gelmektedir. Bu sayede veri ambarı oluştururken sadece veriye erişimin olması amaçlanmaktadır. İşlemsel veri tabanlarında ise bu durum geçerli değildir. İşlemsel veri tabanlarında güncelleme, silme gibi operasyonlar yapılabilmektedir.

3.2 Veri Madenciliği Uygulama Alanları

Bilim ve mühendislik, sağlık, bankacılık, finans ve borsa, eğitim, internet, pazarlama gibi birbirinden farklı birçok konu üzerinde veri madenciliği yöntemleri uygulanabilmektedir. Veri madenciliğinin uygulandığı birkaç alan açıklanmıştır;

- *Sağlık verileri:* Tıp ve sağlık alanındaki tarama testlerinden elde edilen veriler kullanılarak kanserler ile ilgili ön tanılar, kalp verileri analiz edilerek kalp krizi riskinin tespiti, acil servislerdeki hastaların semptomlarına göre risklerin ve önceliklerin belirlenmesi gibi geniş bir alanda uygulanabilir (Baykasoğlu 2005).

- *İş verileri:* Şirketlerde periyodik olarak devam eden iş süreçlerindeki veriler ile karar verme mekanizmaları kurgulanabilir. İnsan kaynakları departmanından elde edilen personel verileri analiz edilerek çalışanların performanslarını etkileyen nedenler ortaya çıkarılabilir.

- *Eğitim sektörü verileri:* Eğitim kurumlarının öğrenci veri tabanından elde edilen bilgiler analiz edilerek öğrencilerin başarı durumlarını etkileyen faktörler tespit edilebilir.

- *Doküman verileri:* Dokümanlarda yer alan anahtar sözcükler analiz edilerek dokümanlar arasındaki benzerlikler tespit edilebilir.

- *Pazarlama verileri:* Müşterilerin satın alma bağıntılarının belirlenmesinde, mevcut müşterilerin elde tutmak için sunulacak kampanyalarda, pazar sepeti analizinde, müşteri ilişkileri ve müşteri değerlendirme konularında kullanılabilir (Silahtaroğlu 2004).

- *Yüzey Analizi ve Coğrafi Bilgi Sistemleri:* Bölgelerin coğrafi özelliklerine göre sınıflandırılarak posta kutusu, ATM gibi hizmetlerin verilmesinde konum belirlemesi yapılabilir (Dinçer 2006).

Durmuşoğlu(2017) ‘nun 2006-2015 yılları arasını kapsayan veri madenciliği makaleleri üzerine yapmış olduğu analizde çalışma alanlarına göre dağılımları Şekil 3.4’de belirtildiği gibi verilmiştir.

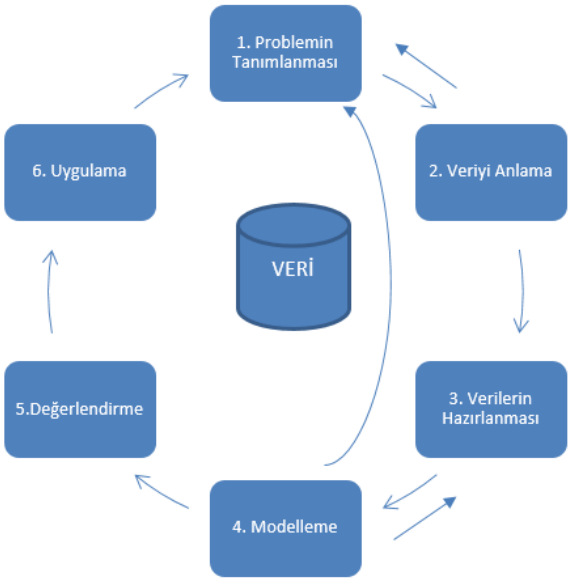
Sıra	Alan	Yayın Sayısı	Yüzde*
1	Bilgisayar Bilimleri	6707	48,18%
2	Mühendislik	3374	24,24%
3	Yöneylem Araştırması ve Yönetim Bilimleri	1270	9,12%
4	Matematik	935	6,72%
5	Biokimya ve moleküler biyoloji	813	5,84%
6	Matematiksel Hesaplamalı Biyoloji	605	4,35%
7	İş Ekonomisi	462	3,32%
8	Bilim Teknoloji ve Diğer Konular	454	3,26%
8	Kimya	429	3,08%
9	Bioteknoloji Uygulamalı Mikrobiyoloji	406	2,92%
10	Medikal Enformatik	342	2,46%
11	Otomasyon kontrol sistemleri	328	2,36%
12	Çevre Bilimleri Ekolojisi	315	2,26%
13	Farmakoloji	308	2,21%
14	Genetik	266	1,91%
15	Enformasyon ve kütüphane bilimleri	259	1,86%
16	Telekomünikasyon	232	1,67%
17	Tarım	186	1,34%
18	Fizik	214	1,54%
19	Sağlık hizmetleri ve bilimleri	186	1,34%
20	Jeoloji	178	1,28%

Şekil 3.4: Çalışma Alanlarına Göre Veri Madenciliği Makaleleri Analizi (Durmuşoğlu 2017)

3.3 Veri Madenciliği Süreçleri

Veri miktarı arttıkça büyük boyutlu verilerdeki eksik, hatalı veya tutarsız verilerin miktarında da artmalar olabilir. Bu durumdaki verilerden oluşabilecek analizler kalitesiz sonuçlar verebilir. Kural ve bağıntıların kaliteli olabilmesi, önemli bilgileri ortaya çıkarabilmek için verilerin güvenilir olması gerekmektedir. Bu sebeplerden dolayı başarılı bir veri madenciliği projesi için süreç modelleri geliştirilmiştir. SEMMA, CRISP-DM ve KDD en çok kullanılan yöntemlerdir.

SEMMA yöntemi SAS firması tarafından ortaya çıkarılmıştır. Sample, Explore, Modify, Model ve Assess kelimelerinin baş harflerinden oluşmaktadır. Örnekleme – Keşfetme – Dönüştürme – Model Oluşturma ve Değerlendirme adımları takip edilir. KDD (Knowledge and Data Discovery) yönteminde 5 işlem adımı gerçekleştirilmektedir. Öncelikle veriler arasından işlenecek olan veriler seçilir, seçilen veriler işlenecek hale dönüştürülür ve model oluşturularak değerlendirme süreci sonunda bilgi oluşturulur. Bu yöntem genelde araştırma amaçlı konularda kullanılır. CRISP-DM süreç modeli ürüne yönelik, endüstriye veya piyasaya yönelik durumlarda kullanılır. En çok tercih edilen modeldir. CRISP-DM süreç modeli adımları aşağıda daha ayrıntılı anlatılacaktır.



Şekil 3.5: CRISP_DM Süreç Modeli (Shearer 2000)

Problemin Tanımlanması

Veri madenciliği projelerinde ilk adım olarak problemin iyi tanımlanması ve anlaşılması gerekmektedir. Projenin amacı, gereksinimleri, hedeflenen başarının anlaşılarak problem tanımlanmalıdır. Problemin doğru algılanmadığı durumlarda süreç işlememektedir veya yanlış ilerlemektedir.

Veriyi Anlama

Veriyi anlama adımında yapılması gereken verinin toplanmasıdır. Birden fazla kaynaktan elde edilebilecek farklı formattaki bilgiler tek tablo haline getirilmelidir. Veriyi anlama adımının veriyi hazırlama adımından farkı veriyi anlamaya çalışmak ve veri seti üzerinde herhangi bir değişiklik yapılmamasıdır (Erkan 2006).

Verilerin Hazırlanması

Problem ve verilerin anlaşılmasından sonraki adım verilerin hazırlanmasıdır. Projelerde en çok zaman ve emek bu adımda harcamaktadır.

- **Veri Toplama**

Belirlenen problemin çözümüne yönelik verilerin toplanmasıdır. Hangi kaynaklardan veri alınacağı önemlidir çünkü az veri kaynağı veri madenciliği çalışmasında eksikliklere neden olacağı gibi, fazla veri kaynakları da veri kirliliğine yol açarak süreci uzatabilir (Terzi ve diğ. 2011).

- **Veriyi Değerlendirme**

Projenin amacına ve problemin çözümüne yönelik farklı kaynaklardan, veri tabanlarından toplanan veriler kontrol edilmelidir. Analiz edilecek veriler arasında tutarsızlıklar olabilir bu sebepten dolayı verilerin değerlendirilmesi gerekmektedir.

- **Veri Birleştirme ve Temizleme**

Veri tabanından elde edilen veriler her zaman istenilen şekilde olmayabilir. Konuya uygun olmayan veriler veya eksik olan veriler olduğu tespit edilebilir. Bu

tarz hatalı veya faydasız bilgiler gürültü olarak tanımlanır. Bunun için aşağıda verilen yöntemler kullanılabilir:

- ✓ Eksik değer içeren kayıtlar veri kümesinden tamamen silinebilir.
- ✓ Eksik olan verilerin yerine standart bir değer verilebilir.
- ✓ Eğer değişken sayısal veriler içeriyor ise sayısal verilerin ortalaması hesaplanarak eksik değerlerin yerine kullanılabilir.
- ✓ Değişken için uygun tahmin yöntemi uygulanarak eksik değer tahmin edilebilir ve eksik değerlerin yerine kullanılabilir (Han 2012).

Daha önce veri ambarının özelliklerinde bahsetmiş olduğumuz bütünleştirme veri madenciliği içinde geçerli olmaktadır. Birden fazla kaynaktan alınan veriler eğer farklı formatlarda ise bu verileri standart bir formata dönüştürülerek veri bütünleştirme işlemi yapılmış olur.

- **Veri İndirgeme**

Veri madenciliği işlemlerinde bazen verilerin fazlalığı işlemin uzun sürmesine neden olabilir. Bu durumlarda eğer analizden elde edilecek sonucun değişmeyeceğine inanılıyorsa değişkenlerin sayısı azaltılabilir. Veri indirgenme yöntemleri aşağıdaki gibidir (Han 2012):

- ✓ Veri birleştirme
- ✓ Boyut indirgeme
- ✓ Örnekleme
- ✓ Genelleme

- **Veri Dönüştürme**

Veri madenciliğinde kullanılacak özniteliklerin ortalama ve varyansları birbirinden farklı olabilir. Bu farklılıklardan dolayı bazı durumlarda değişkenlerin diğerleri üzerindeki baskısı daha az veya çok olabilir. Bu nedenle değişkenler üzerinde dönüşüm yöntemleri kullanılır (Özkan 2016).

Normalizasyon yöntemleri ve formülleri aşağıda belirtilmiştir (Cihan ve diğ. 2017)

Minimum Maksimum Normalizasyonu: Bu yöntemde veriler doğrusal olarak normalize edilmektedir. Veri genellikle 0 ve 1 arasındadır. Minimum bir verinin alabileceği en küçük değer, maksimum bir verinin alabileceği en büyük değerdir.

$$X^* = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{3.1}$$

Z-Score Normalizasyonu: Z-score normalleştirme genellikle -1,5 ve +1.5 arasında değişmektedir. Değişkenin herhangi bir y değeri, değişkenin standart sapması ve ortalamasına bağlı olarak değişen Z dönüşümü ile normalleştirilmektedir.

$$X^* = \frac{x_i - \mu}{\sigma_i} \tag{3.2}$$

Ondalık Ölçekleme: Veri genellikle -1 ve +1 arasındadır. Değişkenin değerinin ondalık kısmı hareket ettirilerek normalleştirme işlemi yapılır. Hareket edecek ondalık nokta sayısı değişkenin maksimum mutlak değerine bağlı olmaktadır.

$$X^* = \frac{x_i}{10^j} \tag{3.3}$$

Sigmoid Normalizasyonu: Veriler 0 ve 1 arasında veya -1 ve +1 arasındadır.

$$X^* = \frac{1}{1 + e^{-x_i}} \tag{3.4}$$

Modelleme

Veri madenciliği sürecinde algoritma uygulama adımına gelene kadar gerçekleştirilen işlemler başarılı bir şekilde tamamlandıktan sonra uygun veri madenciliği algoritması uygulanır. Algoritmalara ait detaylı bilgi tezin 4.bölümünde Veri Madenciliği Yöntemleri başlığı altında detaylı olarak açıklanacaktır.

Modelin Değerlendirilmesi

Aynı problem üzerinde birden fazla veri madenciliği yöntemi uygulanır. Konuya uygun algoritma uygulandıktan sonra ortaya çıkan sonuçlar değerlendirilir.

Modelin Uygulanması

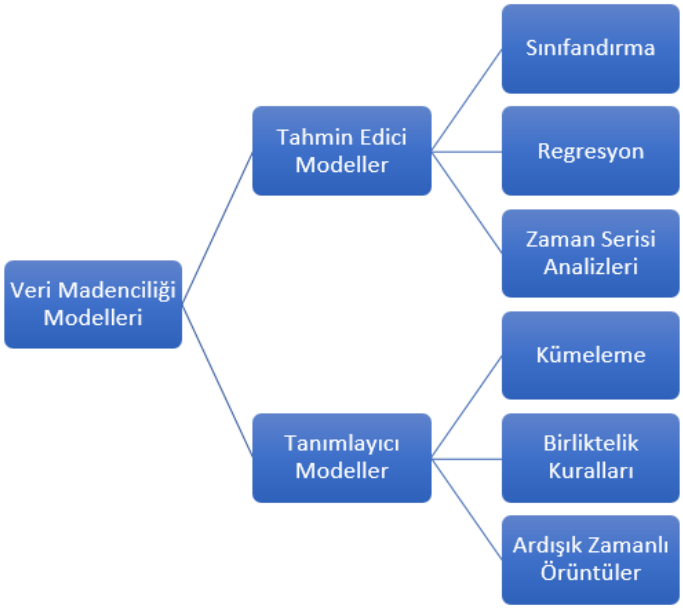
Değerlendirme sonuncunda ortaya çıkan sonuçlar uygulama adımına geçmiş olur. Projenin asıl başarılı olacağı nokta burasıdır. Örneğin bir işletme için tahmin analizi yaptıysanız uygulamanın işletmeye yardımcı olacağı adım burasıdır. Son kullanıcılara veya analistlere yazılım aracılığıyla yorumlanacak şekilde sunulabilir.

3.4 Veri Madenciliği ve Yapay Zekâ

Veri madenciliğinde veri setinin, uygulama alanlarının farklılığına göre farklı veri madenciliği yöntemleri kullanılmaktadır. Tek bir veri madenciliği tekniği kullanılabileceği gibi birden fazla veri madenciliği tekniği de beraber kullanılabilir.

Analizi yapılacak olan veri setinde çıktı değerleri biliniyorsa danışmanlı öğrenme (supervised learning) bilinmiyor ise danışmansız öğrenme (unsupervised learning) algoritmaları kullanılır. Danışmanlı öğrenme sınıflandırmayı temsil eder ve amaç sınıflandırmayı sağlayabilecek model oluşturulmasıdır. Girdi değerleri analiz edilerek çıktı değerlerinin tahmini yapılır. Bu öğrenme tekniğinde eğitim seti kullanılır. Danışmansız öğrenme ise kümelemeyi temsil eder ve birbirine yakın özellikleri benzerliklerine göre kümeleyerek veya ilişkileri temel alınarak birliktelik kurallarıyla model oluşturulmadan çözümlerin bulunmasıdır (Balaban ve Kartal 2015).

Sınıflandırma, regresyon ve zaman serileri analizleri tahmin edici model; kümeleme, birliktelik kuralları ve ardışık zamanlı örüntüler tanımlayıcı modeldir. Tahmin edici modeller sonuçları bilinen verileri kullanarak yeni veriler için tahmin yaparken tanımlayıcı modeller ise mevcut verideki benzerlikleri, birliktelikleri ortaya çıkarırlar.



Şekil 3.6: Veri Madenciliği Modelleri (Martin ve diğ. 2014)

3.4.1 Sınıflandırma ve Regresyon

Sınıflandırma, verilerde bulunan özellikleri kullanarak doğru bir model geliştirmek için girdi verilerini analiz etme ve daha sonra önceden tanımlanmış sınıflara yeni girdi verileri atamak için bu modeli kullanma işlemidir. Bir sınıflandırma algoritması, tahmin edicilerin değerleri ile hedefin değerleri arasındaki ilişkileri bulur. Sınıflandırma uygulamaları arasında belge kategorilendirme, teşhis tahmini, fiyat tahmini, risk değerlendirmesi ve duygu analizi yer almaktadır (Yıldırım ve diğ. 2018).

Sınıflandırma yönteminde verilerin bir kısmı eğitim bir kısmı test amacıyla kullanılmaktadır. Eğitim verileri analiz edilerek sınıflandırma kuralları oluşturulur. Test verilerine sınıflandırma kuralları uygulanarak doğruluğu test edilir. Doğruluk oranları yüksek ise yeni veriler için bu sınıflandırma kuralları kullanılarak tahminlerde bulunabilir.

Bankalar için hangi kredi başvuru sahiplerinin güvenli hangilerinin riskli olduğunu öğrenmek için veri analizi yapmalıdır. Elektronik satış yapan bir mağazanın pazarla müdürünün belirli bir profile sahip müşterinin yeni bir bilgisayarı

satın alıp almayacağını tahmin etmeye yardımcı olması için veri analizine ihtiyaç vardır. Bir tıp araştırmacısı bir hastanın alması gereken üç özel tedaviden hangisini alması gerektiğini karar vermek için göğüs kanseri verilerini analiz etmelidir. Bu örneklerde belirtildiği gibi kredi başvuruları için "güvenli" veya "riskli", pazarlama verileri için "evet" veya "hayır", sağlık verileri için "Tedavi A", "Tedavi B" veya "Tedavi C" gibi kategorilendirilerek sınıflandırma yapılmış olur. Eğer pazarlama müdürü bir müşterinin ne kadar harcama yapacağını tahmin etmek isterse burada sayısal bir model olacaktır. Regresyon analizi, sayısal tahmin için en sık kullanılan istatiksel bir metodolojidir (Han ve diğ. 2011).

Sınıflandırma ve regresyon modelleri aşağıdaki teknikleri içermektedir:

3.4.1.1 Karar Ağaçları Algoritması

Karar ağacı sınıflandırma algoritmalarında en çok kullanılan denetimli öğrenme algoritmasıdır. Karar ağaçları tümevarım yöntemiyle çalışmaktadır. Yukarıdan aşağıya doğru inen bir yapısı bulunmaktadır. Veri kümesi içerisinde kurallar oluşturularak parçalara ayrılır. İlk düğüme kök denir ve bir karar düğümünde bir veya birden fazla dallanma olabilir. Yaprak düğümü bir kararı veya etiketi belirtir. Bu şekilde kök düğümden yaparak düğümlere ilerlenerek kurallar oluşturulur.

Karar ağacı, "kök" adı verilen bir düğümle yönlendirilmiş bir ağaçtır. Giden kenarlara sahip bir düğüme dâhili veya test düğümü denir. Diğer tüm düğümlere "yaprak" denir. Bir karar ağacında, her iç düğüm, girdi özniteliklerinin değerlerini belirli bir ayırma işlevine göre iki veya daha fazla alt alana böler. Sayısal öznitelikler söz konusu olduğunda, koşul bir aralıktır. Her yaprak, en uygun hedef değeri temsil eden bir sınıfa atanır. Alternatif olarak yaprak, belirli bir değere sahip olan hedef özelliğinin olasılığını belirten bir olasılık vektörünü tutabilir. Örnekler, süreç boyunca yapılan testlerin sonucuna göre, ağacın kökünden bir yaprağa doğru seyrederek sınıflandırılır (Maimon ve Rokach 2010).

Karar ağacı algoritmalarının faydaları (Şeker 2013):

- Ön işleme aşaması diğer alternatiflere göre daha kısa sürede tamamlanır.
- Hem sayısal veriler hem de kategorik veriler üzerinde çalışma yapılabilir.
- Hızlı ve kolay şekilde veri işlenebilir bundan dolayı da düşük hesaplama karmaşıklığı bulunmaktadır.
- Algoritmanın her adımı görüntülenip yorumlanabilir.

ID3 Algoritması

ID3 algoritması kategorik niteliklerle çalışan ve entropiye dayalı karar ağacı algoritmasıdır. Entropi sistemin belirsizlik ölçütüdür. ID3 algoritması, özniteliklerin değerlerini test ederek nesnelerin sınıflandırmasını belirleyen bir karar ağacı oluşturma algoritmasıdır. Ağaç yukarıdan aşağıya doğru oluşmaya başlar. Ağacın her düğümünde bir özellik test edilir ve sonuçlar nesne setini bölümlemek için kullanılır. Bu işlem sınıflandırma ölçütlerine göre homojen olana kadar özyinelemeli olarak yapılır. Başka bir deyişle, aynı kategoriye ait nesneleri içerir. Daha sonra bir yaprak düğümü haline gelir. Her düğümde test edilecek özellik, bilgi kazancını maksimize etmek ve entropi en aza indirmek isteyen bilgi kuramsal ölçütlere göre seçilir (Joshi 1997).

ID3, birçok özniteliğin olduğu ve eğitim setinin birçok nesne içerdiği, ancak çok fazla hesaplama yapılmadan makul bir karar ağacının gerekli olduğu durumlar için tasarlanmıştır. Eğitim setindeki diğer tüm nesneler daha sonra ağaç kullanılarak sınıflandırılır. Ağaç tüm bu nesneler için doğru cevabı verirse, tüm eğitim seti için doğrudur ve işlem sona erer. Değilse, yanlış sınıflandırılmış nesnelerin bir seçimi pencereye eklenir ve işlem devam eder. Bu sayede, 50'ye kadar öznitelik olarak tanımlanan otuz bine kadar nesneden oluşan antrenman setleri için birkaç iterasyondan sonra doğru karar ağaçları bulunmuştur (Quinlan 1986).

Algoritma, bir düğümdeki tüm örneklerin bir sınıfa sahip oluncaya kadar her bir alt düğüm için yinelemeli olarak uygulanır. Karar ağacındaki yaprağa giden her yol bir sınıflandırma kuralını temsil eder. Böyle bir yukarıdan aşağı karar ağacı oluşturma algoritmasındaki kritik karar, bir düğümdeki bir özniteliğin seçimidir. ID3'ün öznitelik seçimi, karar ağacının karmaşıklığının, verilen öznitelik değerinin

taşıdığı bilgi miktarı ile ilişkili olduğu varsayımına dayanmaktadır (Kantardzic 2011). Karar ağaçlarında başlangıç düğümü seçimi önemlidir.

- Öncelikle bağımlı değişken entropisi hesaplanır.

$$H(T) = -\sum_{i=1}^{n} p_i \, log_2 \, p_i \tag{3.5}$$

- Daha sonra özniteliklerin niteliğe bağlı entropileri hesaplanır.

$$H(X,T) = \sum_{k=1}^{n} \frac{|X_k|}{|X|} \, H(X_k) \tag{3.6}$$

- Bağımlı değişkenin entropisinden özniteliğin entropisi çıkarılarak kazan ölçütü hesaplanır.

$$Kazanç\ (X,T) = H(T) - H(X,T) \tag{3.7}$$

- En büyük kazanca sahip olan karar düğümü seçilir.

C4.5 ve C5 Algoritması

C4.5 algoritması Quinlan tarafından 1993 yılında ID3 algoritmasının geliştirilmiş hali olarak oluşturulan bir karar ağacı algoritmasıdır.

C4.5 algoritmasının en önemli parçası, eğitim örnekleri kümesinden bir başlangıç karar ağacı oluşturma işlemidir. Algoritma karar ağacı biçiminde bir sınıflandırıcı oluşturur. Sınıfa işaret eden bir yaprak veya bir testin olası sonuçları için bir dal ve bir alt ağacı olan bir öznitelik değeri üzerinde gerçekleştirilecek bazı testleri belirten karar düğümü olacak şekilde iki tip düğüm içeren yapısı vardır (Kantardzic 2011).

CART (Classification And Regression Trees) Algoritması

Breiman ve diğ. (1984) tarafından geliştirilen sınıflandırma ve regresyon ağacı algoritmasıdır. Ağaçta yer alan her bir karar düğümünden sonra homojen olan dallara ayrılması ilkesine dayanır. Bu algoritma da entropiye dayanmaktadır. En iyi dallanma kriterlerini gerçekleştirirken Twoing ve Gini olarak iki yöntem kullanır (Karaibrahimoğlu 2014).

CART karar ağacı, hem hedef hem de öngörücü olarak sürekli ve nominal nitelikleri işleyebilen bir ikili yineleme bölümleme prosedürüdür. Veriler ham haliyle ele alınır. Ağaçlar, durdurma kuralı kullanılmadan maksimum bir boyuta genişletildikten sonra maliyet-karmaşıklık (cost-complexity) budaması ile köke geri verilir. Ağacın eğitim verisi üzerindeki genel performansına en az katkıda bulunan bölüm bir sonraki budama yapılacak bölümdür (We ve diğ. 2008).

Kumar ve Kurithika (2015) karar ağaçlarını karşılaştırılmasını aşağıdaki tablo üzerinde göstermişlerdir.

Tablo 3.1: Karar Ağaçlarının Karşılaştırılması (Kumar ve Kiruthika 2015)

Algoritmalar	ID3	C4.5	C5.0	CART
Data Tipi	Kategorik	Sürekli ve kategorik	Sürekli ve kategorik, tarih ve zaman	Sürekli ve nominal özellikli veri
Budama	Yok	Ön budama	Ön budama	Sonra budama
Hız	Düşük	ID3'e göre daha hızlı	Yüksek	Orta
Formul	Entropi ve bilgi kazancı kullanır	Split info ve gain radio kullanır	C4.5 ile aynı	Gini

3.4.1.2 k-En Yakın Komşu Algoritması

K-En Yakın Komşu; sınıflandırma yöntemlerinden denetimli öğrenme tekniği içerisindedir. Çoğunlukla yeni bir problemi çözerken daha önceden çözülen benzer problemleri inceleriz. k-En Yakın Komşu tekniği de bu şekilde çalışmaktadır. Benzer durum ve komşuları inceleyerek yeni bir durumun hangi sınıfta yer alacağına karar verir. Her sınıf için durum sayısını sayar ve yeni durumu komşularının ait olduğu aynı sınıfa atar. k-NN'yi uygulamak için öncelikle öznitelikler arasındaki mesafenin ölçüsü hesaplanır. Bu sayısal veriler için kolay olsa da, kategorik değişkenler özel işlem gerektirir (Edelstein 1999).

Komşuluk arası uzaklık Öklid ve Manhattan gibi uzaklık hesaplama yöntemleriyle bulunur. Bilinmeyen veriler k-en yakın komşuya en çok benzerlik

gösteren sınıf değerine atanır. k-en yakın komşu algoritmasında aşağıdaki adımları takip edilir (Harrington 2012):

- Belirlenen bir noktaya en yakın komşu sayısı olan k belirlenir.
- Belirlenen nokta ile diğer tüm noktalar arasındaki uzaklık hesaplanır.
- Bir önceki işlemde hesap edilen uzaklıklara göre kayıt sıralaması yapılarak bunlar arasındaki en küçük k seçilir.
- Seçilen kayıtlar bulunarak en fazla tekrar eden kategorinin seçimi yapılır.
- Seçilen kategori tahmin edilecek olan gözlemin kategorisi kabul edilir.

3.4.1.3 Rastgele Orman Algoritması

Birçok sınıflandırıcı üreten ve bir araya getiren "ensemble learning" olarak adlandırılan yöntemler bulunmaktadır. Bilinen bu iki yöntem boosting ve baggingdir. Boosting metodunda art arda gelen ağaçlar önceki tahminlerdeki yanlış tahmin edilen noktaya ekstra ağırlık verir. Ve sonunda ağırlıklı oylama alınır. Bagging metodunda, ardışık ağaçlar daha önceki ağaçlara bağlı değildir. Her bir ağaç bağımsız olarak veri kümesinin bir bootstrap örneği kullanılarak oluşturulur. Sonuç olarak, tahmin için basit bir çoğunluk oyu alınır. Breiman 2001 yılında bagging için ek bir rastgele katman olan Rastgele Orman önermiştir (Liaw ve Wiener 2002).). Bir Rastgele Orman, genellikle bagging yöntemi yoluyla eğitilen, eğitim setinin boyutuna ayarlanan maksimum örnek sayısı ile eğitilmiş bir Karar Ağacı topluluğudur. Rastgele Orman, Karar ağaçlarının toplanmasıyla oluşmaktadır.

Rastgele orman, ağaç yapılı sınıflandırıcılar $\{h(\boldsymbol{x}, \boldsymbol{k}), \boldsymbol{k} = 1, ...\}$ 'dan oluşan bir sınıflandırıcıdır. Burada $\{k\}$, birbirinden bağımsız olarak dağıtılmış rastgele vektörlerdir (Breiman 2001).

$h_1(x)$, $h_2(x)$, . . . , $h_K(x)$ sınıflandırıcıları ve rastgele Y vektöründen elde edilmiş rastgele eğitim seti ele alındığında X marjin fonksiyonun $I()$ gösterge işlevi olduğu şekilde tanımlar (Breiman 2001).

$$mg(X,Y) = av_k I(h_k(X) = Y) - maxav_k I(h_k(X) = j) \quad (3.8)$$

Marjin fonksiyonu *mg*, ne kadar büyük olursa, sınıflandırmaya olan güven daha fazla olur. Genelleme hatası Eşitlik 3.9'daki gibidir; (Breiman 2001).

$$PE^* = P_{X,Y}(mg(X,Y) < 0) \quad (3.9)$$

Karar ağaçları tahmin yapmak için eğitim setindeki özellikler ve etiketlere göre kurallar oluşturur. Rastgele Orman algoritması ise karar ağaçları oluşturmak için özellikler ve etiketleri rastgele seçer ve sonuçların ortalamasını alır (Web-3).

3.4.1.4 Yapay Sinir Ağları

Yapay sinir ağları insan beynindeki nöronlardan esinlenerek ortaya çıkmıştır. İlk olarak 1943 yılında Warren McCulloch ve Walter Pitts tarafından ortaya çıkarılmıştır. Yapay sinir ağları, nöronlardan oluşan giriş katmanı, gizli katman ve çıkış katmanından oluşmaktadır. Aşağıdaki şekilde nöronların birleştirildiği tipik bir sinir ağı gösterilmiştir. Burada her bir bağlantı ağırlık olarak adlandırılan nümerik bir sayı ile ilişkilidir (Wang 2003).

Sinir ağlarına örnek bir eğitim seti sunularak eğitilmesi sağlanabilir. Eğitim verilerinin istenen çıktıları bilinir böylelikle eğitimin amacı; bağlı nöronlar arasındaki ağırlıkları ayarlayarak yapay sinir ağı çıkışları ile istenen çıkışlar arasındaki hata oranını en aza indirgemektir. Sinir ağları mimarisi denenirken doğrulama seti olarak adlandırılan bağımsız bir veri seti uygulanabilir. Doğrulamadan sonra, yapay sinir ağının ne kadar güvenli olduğunu belirleyen yapay sinir ağının performans seviyesini belirlemek için test seti olarak adlandırılan başka bir bağımsız veri kümesi kullanılır. Bir sinir ağının eğitim setinde bulunmayan bir bilgiyi öğrenemeyeceği bilinmelidir. Bu nedenle eğitim setinin büyüklüğü sinir ağının eğitim setine gömülmüş özellikleri ezberlemesini sağlayacak kadar büyük olmalıdır (Wang 2003).

3.4.1.5 Naive Bayes Algoritması

Bayes sınıflandırıcıları Thomas Bayes teoremine dayanan istatiksel sınıflandırıcılardır. Bayes teorimi 18.yy'da olasılık üzerine çalışan İngiliz papaz Thomas Bayes adıyla anılır. Naive Bayes bir özniteliğin diğer öznitelikler üzerindeki etkisini ele alır ve nominal değerler üzerinde çalışmaktadır. Sayısal veriler varsa bunlar da nominal değerlere dönüştürülür.

Naive Bayes sınıflandırıcısının çalışma şekli aşağıda belirtildiği gibidir: (Han ve diğ. 2011)

- D ilişkili sınıf etiketleri olan değişkenler grubundan oluşan eğitim seti olsun. Her değişkenler grubu $X = (X_1, X_2, \ldots\ldots\ldots X_n)$ olarak belirtilen n boyutlu öznitelik vektörüyle temsil edilir. Sırasıyla $A_1, A_2, \ldots A_n$, n adet öznitelikten oluşan değişkenler grubu üzerinde yapılan n adet ölçümlerini belirtir.
- $C_1, C_2, \ldots\ldots, C_m$ sınıfları olduğunu varsayalım. X değişkenler grubunda sınıflandırıcı, X'in en yüksek sonsal olasılığa sahip olan sınıfa ait olduğunu öngörür ve bu durum Naive Bayes sınıflandırıcısının sadece aşağıdaki durumda X'in C_i sınıfına ait olduğunu tahmin eder;

$$P(C_i \setminus X) > P(C_j \setminus X) \quad 1 \leq j \leq m, j \neq i. \tag{3.10}$$

Böylelikle $P(C_i \setminus X)$ maksimize edilmiş olur. Maksimize edilmiş $P(C_i \setminus X)$ için C_i sınıfı maksimum sonsal hipotez olarak adlandırılır. Bayes teoremi formülü ise aşağıdaki gibidir;

$$P(B|A) = \frac{P(B|A) * P(A)}{P(B)} \tag{3.11}$$

$P(X)$ tüm sınıflar için sabit olduğundan sadece $P(X / C_i)\ P(C_i)$ maksimize edilmesi gerekir. Eğer sınıfların önceki olasılıkları bilinmiyorsa $P(C_1) = P(C_2) = \ldots\ldots = P(C_m)$ şeklinde sınıfların eşit olduğu varsayılır ve $P(C_i \setminus X)$ maksimize edilmiş olur. Aksi durumda $P(X/C_i)P(C_i)$ maksimize edilmiş olacaktır.

Birçok öznitelikli veri kümeleri verildiğinde $P(X/C_m)$ hesaplamak pahalı olacaktır. Bu hesaplamayı azaltmak için naive sınıf şartlı bağımsızlık varsayımı yapılır. Bu durum öznitelik değerlerinin koşullu olarak birbirinden bağımsız olmasını sağlayacaktır.

$$P(X|C_i) = \prod_{k=1}^{n} P(X_k|C_i) \quad (3.12)$$

$P(x_1 / C_i)$, $P(x_2 / C)$,……, $P(x_n / C)$ olasılıkları kolayca tahmin edilebilir. Burada X_k 'nin X için A_k öznitelik değerini ifade eder. Her bir öznitelik için kategorik mi yoksa sürekli değerli mi olduğuna bakılır.

Eğer A_k kategorik ise, $P(X_k / C_i)$, C_i sınıfının D içindeki A_k'nın X_k değerini aldığı D içindeki değişken grubu sayısının, D içindeki Ci sınıfının toplam değişken grubu sayısına bölümüdür.

Eğer A_k sürekli bir değişken ise μ ortalama ve σ, standart sapma ile bir Gaussian dağılımına sahip olduğu varsayılır.

$$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (3.13)$$

X tahmin etmek için her Ci sınıfı için $P(X / C_i)$ $P(C_i)$ değerlendirilir.

$$P(X|C_i)P(C_i) > P\left(X\middle|C_j\right)P \quad (3.14)$$

3.4.2 Kümeleme

Kümeleme genellikle önceden tanımlanmış öznitelikler üzerindeki veriler arasındaki benzerliğin belirlenmesiyle gerçekleştirilir. En benzer veriler kümelere ayrılmıştır. Kümeleme analizi, sınıflandırmada olduğu gibi gruplara ayrılır fakat burada sınıflandırmadan farkı şudur; sınıflandırmada gruplar daha önceden belirliyken kümelemede gruplar önceden belirlenmemiştir. Önceden belirlenmediği için de denetimsiz öğrenme vardır. Kümeleme alternatif olarak segmentasyon veya denetimsiz öğrenme olarak adlandırılabilir (Dunham 2006). Kümeleme analizinin sınıflandırmadan ayrılan bir başka özelliği de verilerin sadece sahip oldukları değerlere göre değil, diğer verilerle olan benzerliklerine göre kümelere ayrılmasıdır. Bundan dolayı da kümeleme sonuçları için dinamiktir denilebilir (Silahtaroğlu 2016).

Kümelemede temel prensip küme içi benzerliği maksimum, kümeler arasındaki benzerliği ise minimum yapmaktır. Kümeleme yönteminin kalitesi bu prensibi sağlaması ile ölçülür. Veri madenciliğinde birçok kümeleme yöntemleri bulunmaktadır. Hangi kümeleme yönteminin kullanılacağı verinin türüne ve uygulamanın amacına göre değişkenlik göstermektedir (Dinçer 2006).

Makine öğrenmesinde, kümeleme analizi genellikle denetimsiz öğrenmeyi ifade eder, çünkü bir nesnenin ait olduğu sınıflar önceden belirlenmez. Sınıflar, sınıf içi benzerliği yüksek ve sınıflar arası benzerliği düşük olan nesnelerin koleksiyonları olarak tanımlanır (Chen ve diğ. 1996).

Kümeleme, verilerin benzer nesnelerden oluşan gruplara bölünmesidir. Küme olarak adlandırılan her grup, kendi aralarında benzer ve diğer grupların nesnelerine benzemeyen nesnelerden oluşur. Verilerin daha az sayıda küme tarafından temsil edilmesi, ince ayrıntıları kaybeder ancak sadeleştirmeyi başarır. Çok sayıda veri nesnesini birkaç kümeye göre temsil etmektedir ve bu nedenle verileri kümeleriyle modellemektedir. Veri modellemesi kümelenmeyi; matematiğe, istatistiklere ve sayısal analizlere dayanan tarihsel bir perspektife koyar. Bir makine öğrenimi perspektifinden, kümeler gizli kalıplara karşılık gelir ve sonuçta ortaya çıkan sistem bir veri kavramını temsil eder. Bu nedenle, kümeleme gizli veri kavramının denetimsiz bir şekilde öğrenilmesidir (Berkhin 2006).

Silahtaroğlu(2016) kümeleme analizini Şekil 3.7'de sınıflandırmıştır.

Şekil 3.7: Kümeleme Analizi Sınıflandırma (Silahtaroğlu2016)

Hiyerarşik kümeleme ile kümelenmiş kümeler oluşturulur. Hiyerarşideki her seviye ayrı bir kümelere sahiptir. En düşük seviyede, her öğe kendi benzersiz kümelenmesindedir. En yüksek seviyede, tüm öğeler aynı kümeye aittir. Hiyerarşik kümeleme ile istenen sayıda küme girilmez (Dunham 2006).

Hiyerarşik yöntemler küme ağacı oluşturur. Aşağıdan yukarıya toplaşım kümeleme algoritmaları, yukarıdan aşağıya bölünür kümeleme algoritmaları olmak üzere iki gruba ayrılır. Toplaşım kümeleme algoritmaları başlangıçta veri tabanındaki her noktayı ayrı küme olarak belirler daha sonra bu kümeleri birleştirerek birbirinden ayrı kümeler oluşturur. Bölünür kümeleme algoritmalarında ise veri tabanındaki tüm noktaları tek bir küme olarak belirler daha sonra birbirinden ayrı kümeler oluşturur. Bölünür kümeleme algoritmalarında başlangıçta k adet kümeye bölüneceği bellidir (Silahtaroğlu 2016).

Daha büyük veri tabanlarında hedeflenen algoritmalar, veri tabanının örneklenmesi veya veri tabanının boyutundan bağımsız olarak belleğe sığacak şekilde sıkıştırılabilen veri yapıları kullanılarak bellek kısıtlamalarına uyum sağlayabilir. Kümeleme algoritmaları, örtüşen veya çıkmayan kümeler oluşturup oluşturmadıkları konusunda da farklılıklar gösterebilir. Yalnızca sonlandırılmamış kümeleri dikkate alsak bile, bir öğeyi birden çok kümeye yerleştirmek mümkündür. Sırasıyla, örtüşmeyen kümeler dışsal veya içsel olarak görülebilir (Dunham 2006).

Grid temelli algoritmalar; büyük veri tabanlarında kümeleme yapılabilmesi için yüksek miktarda bellek gerektirir. Bunun içinde numaralandırılmış çizgilerden oluşan hücresel yapılar yardımıyla kümeleme yapılır (Silahtaroğlu 2016).

Verilerin birbirine benzerlikleri, aralarındaki mesafelerin ölçülmesiyle değerlendirilir. x ve y noktalarının ne kadar uzaklıkta olduğunu belirten bir D(x,y) mesafe ölçüsü aşağıdaki gibi tanımlanabilir;

Bir dizi kümenin bir kümelenme olarak kabul edilebilecek kadar yakın olup olmadığını tartışmak için x ve y noktalarının ne kadar uzak olduğunu söyleyen bir mesafe ölçüsüne $d(\boldsymbol{x}, \boldsymbol{y})$ ihtiyacımız var. Bir mesafe ölçüsü d için olağan aksiyomlar (Ullman) :

1. d(***x***; ***x***) = 0 ise bir noktanın kendisine olan uzaklığı 0'dır.

2. d(***x***; ***y***) = d(***y***; ***x***) ise mesafe simetriktir.

3. d(***x***; ***y***) <= d(***x***; ***z***) + d(***z***; ***y***) ise üçgen eşitsizliği vardır.

Uzaklıklar aşağıdaki gibi hesaplanabilir:

Öklid Uzaklık Ölçüsü	$d(x,y) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2}$	(3.14)
Manhattan Uzaklık Ölçüsü	$d(x,y) = \sum_{k=1}^{n}(\lvert x_k - y_k \rvert)$	(3.15)
Minkowski Uzaklık Ölçüsü	$d(x,y) = [\sum_{k=1}^{n}\lvert x_k - y_k \rvert^m]^{1/m}$	(3.16)

3.4.2.1 k-means Algoritması

En çok kullanılan kümeleme algoritmalarından biri olan k-means algoritması, J.MacQueen tarafından 1967 yılında tanıtılmıştır. Veriler nitelik veya öz niteliklerine göre k adet kümeye ayırılır. Bu algoritmada verilerin en yakın veya benzer oldukları küme merkezleri etrafına yerleştirilir. Çalışmada Öklid bağıntısı temel alınarak işlemler gerçekleştirilir. K sayısı giriş parametresini belirtirken eğer küme sayısı belli değil ise k parametresi farklı değerler için tekrar tekrar uygulanarak en uygun parametre bulunur. Aşağıdaki özellikleri nedeniyle tercih edilmektedir:

- Küme sayısının parametrik olması analizi esnek hale getirmektedir.
- Algoritma hızlı çalışmakta ve kullanımı kolaydır.
- Sonuçları hem grafik olarak hem de metin ve sayısal olarak ifade edilebilmektedir (Dinçer 2006).

k-means algoritmasında başlangıçta *n* adet nesneden *k* adet küme sayısı seçilir ve kümenin merkezi belirlenmiş olur. Daha sonra ikinci olarak gelen nesneler en yakın olan küme merkezine göre kümeye dâhil olur. Mesafeyi ölçmek için birçok hesaplama yöntemi vardır. Her yeni gelen eleman ile kümenin ağırlıklı ortalaması değişir ve buna göre tekrar küme merkezi belirlenir. Kümeye dâhil edilecek başka eleman kalmayana kadar işlem bu şekilde devam etmektedir.

Başlangıçta küme merkezini belirlemek için aşağıdaki teknikler yer alır (Shimodaira 2015):

- Veriler küme merkezleri olarak rastgele seçilir.
- Veriler k adet kümeye rastgele atanır ve başlangıç merkezi olarak küme ortalamaları alınır.
- Küme merkezi en uç değerlerdeki verilerden seçilir.

k-means yönteminin adımları aşağıda belirtilmiştir (Altunkaynak 2017):

1. Küme sayısı k belirlenir.
2. Herhangi k adet gözlem küme olarak seçilir.
3. Geriye kalan n-k gözlem ile k gözlem arasındaki uzaklık hesaplanır.
4. n-k gözlem sırasıyla en yakın kümeye atanır. Her atama işleminden sonra kümeye ait merkez tekrar hesaplanır.
5. Kümeler boşaltılarak n adet gözlem yeniden küme merkezlerine göre uzaklıkları hesaplanarak atama işlemi baştan yapılır.
6. Kümeler arasındaki bu geçiş durana kadar adım 5 tekrarlanır.

3.4.2.2 Bulanık c-ortalama Algoritması

Bulanık c-ortalama kümeleme algoritması, 1973 yılında Dunn tarafından ortaya atılmıştır. 1981'de ise Bezdek tarafından geliştirilmiştir (Bezdek ve diğ. 1984). Bulanık kümeleme yöntemlerinin içinde uygulama kolaylığından dolayı en bilinen ve yaygın olarak kullanılanıdır.

Bulanık c-ortalama algoritmasında kümelere atanan verilere üyelik derecesi eklenmiştir. Bu nedenle veriler kümelerin her birine [0,1] aralığında değişen bir üyelik derecesiyle aittir. Bir verinin tüm kümelere olan üyelik derecelerinin toplamı 1 olması gerekmektedir. Bulanık kümedeki C-ortalamalar yöntemi kümeleme algoritmaları, uzaklık ölçütlerinin amaç fonksiyonlarını minimize ederek küme merkezlerini bulmayı hedeflemektedir (Özgür 2017).

Kümeleme işlemi yapılacak veri setinde her bir gözlem p boyutlu bir sütun vektörü içerisinde yer alan p sayıda özellikle gösterilir. Özellikler vektörü aşağıdaki eşitlikteki şekilde gösterilir (Martino ve Sessa 2009):

$$X_k = \{X_{1k}, \ldots\ldots, X_{pk}\}^T, X_k \in R^p \tag{3.17}$$

n gözlemli sonlu veri seti ve pxn boyutlu veri matrisi aşağıdaki şekilde gösterilir:

$$X = \{X \mid k = 1,2, \ldots., n\} \tag{3.18}$$

$$X = \begin{matrix} X_{11} & \ldots. & X_{1n} \\ \ldots & \ldots & \ldots \\ X_{p1} & \ldots. & X_{pm} \end{matrix} \quad (3.19)$$

Bulanık c-ortalamalar algoritması adımları şu şekildedir (Alpaslan ve diğ. 2011):

Adım 1: Başlangıç değeri belirleme: küme sayısı c, bulanıklık indeksi m, işlem bitirme kriteri Ɛ ve üyelik dereceleri matrisi U veya V küme prototipleri rastgele üretilir.

Adım 2: Eğer U küme prototipleri rastgele üretilirse bu değerler ile üyelik derecesi matrisi hesaplanır.

$$U_{tk} = \left[\sum_{f=1}^{c} \left(\frac{d_{ft}}{d_{fk}}\right)^{2/(m-19} \right]^{-1} \quad (3.20)$$

Adım 3: Adım2'ye göre U küme prototiplerin güncellenir.

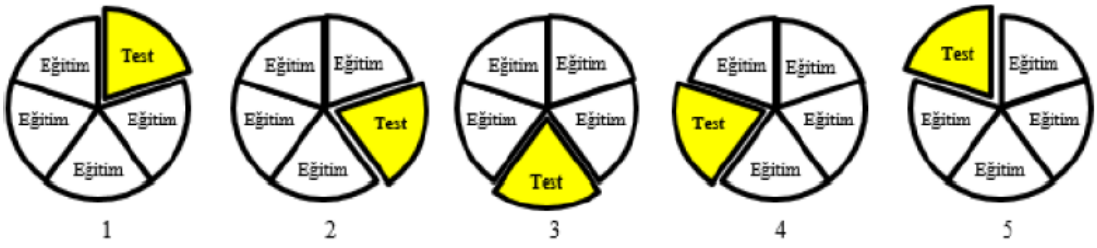
Adım 4: $\left\|U^t - U^{(t-1)}\right\| <$ □ ise iterasyon durdurulur değil ise Adım 2'ye geri dönülür.

3.5 Model Performans Değerlendirme ve Ölçütleri

Tez kapsamında kullanılan model performans değerlendirme ölçütleri bu başlık altında anlatılacaktır. Performans değerlendirme ve model seçim yöntemleri olarak hold-out ve k-kat çapraz geçerleme yöntemleri kullanılır.

Hold-out (Yüzdesel Bölme): Veri seti eğitim ve test veri seti olarak belirli bir oranda ayrımı yapılır. Eğitim veri seti model eğitmede kullanılırken test veri setiyle model doğrulaması yapılarak modelin performansı test edilir.

K-Kat Çapraz Geçerleme (k-fold cross validation): Veri seti k adet parçaya ayrılır. Parçalardan ilki test diğerleri eğitim için kullanılır. Çıkan sonucun ortalamasına bakılır. Örnek 5-kat çapraz geçerleme şeması Şekil 3.8'de gösterilmiştir.



Şekil 3.8: 5-kat Çapraz Geçerleme (Kartal 2015)

Danışmanlı öğrenmenin olduğu sınıflandırma algoritmalarında modelin performansını değerlendirmek için hedef niteliğinin sınıflarına ait gerçek ve tahmin edilen sayıların bir arada verildiği karışıklık matrisi (confusion matrix) kullanılır. Matrisin sütunları gerçek değerleri, satırları ise sınıflandırma sonucu elde edilen sonuçlara karşılık gelir. Örnek matris üzerinden aşağıda anlatılmıştır. (Cihan 2018).

Karışıklık Matrisi		Verinin Tahmin Edilen Sınıfı	
		Pozitif (Hasta)	Negatif (Sağlıklı)
Verinin Gerçek Sınıfı	Pozitif (Hasta)	True Positive (TP)	False Negative (FN)
	Negatif (Sağlıklı)	False Positive (FP)	True Negative (TN)

Şekil 3.9: İki Sınıf İçin Oluşturulmuş Karışıklık Matrisi (Cihan 2018)

TP: Gerçekte hasta olan ve sınıflandırma sonucunda da hasta olarak nitelendirilen örnek sayısı

FN: Gerçekte hasta olan ve sınıflandırma sonucunda sağlıklı olarak nitelendirilen örnek sayısı

FP: Gerçekte sağlıklı olan ve sınıflandırma sonucunda hasta olarak nitelendirilen örnek sayısı

TN: Gerçekte sağlıklı olan ve sınıflandırma sonucunda da sağlıklı olarak nitelendirilen örnek sayısı

Doğruluk Oranı: Doğru olarak sınıflandırılan örnek sayılarının toplam örnek sayılarına oranıdır (Balaban ve Kartal 2015).

$$Doğruluk = \frac{TP+TN}{TP+TN+FP+FN} \tag{3.21}$$

Duyarlılık (Sensitivity): Doğru sınıflandırılan pozitif örneklerin toplam pozitif örnek sayısına oranıdır (Balaban ve Kartal 2015).

$$Duyarlılık = \frac{TP}{TP+FN} \tag{3.22}$$

Seçicilik (Specificity): Doğru sınıflandırılan negatif örneklerin toplam negatif örnek sayısına oranıdır (Balaban ve Kartal 2015).

$$Seçicilik = \frac{TN}{TN+FP} \tag{3.23}$$

Kesinlik yada Pozitif Öngörü: Doğru sınıflandırılan pozitif örneklerin toplam pozitif doğru tahmin edilen örneklere oranıdır (Balaban ve Kartal 2015).

$$Kesinlik = \frac{TP}{TP+FP} \quad (3.24)$$

F-Ölçütü: Kesinlik ve duyarlılık ölçütlerinin her ikisini de ele alma imkanı sağlayan harmonik ortalamadır (Balaban ve Kartal 2015).

$$F - \text{Ölçütü} = 2 * \frac{Duyarlılık * Kesinlik}{Duyarlılık + Kesinlik} \quad (3.25)$$

Kappa Değeri: Sınıflandırıcı modelin doğru sınıflandırma başarımı hakkında bilgi vermektedir (Cihan 2018). Kappa değer 0 ila 1 arasında değişmektedir. Landis ve Koch (1997)'un belirtmiş olduğu gibi Kappa istatiği 0.00 altında ise zayıf, 0.00-0.20 aralığında ise önemsiz, 0.21-0.40 aralığında ise orta, 0.41-0.60 aralığında ise makul, 0.61-0.80 aralığında ise önemli, 0.81-1.00 aralığında ise neredeyse mümkün olarak tanımlanır.

Kümeleme algoritması performansının değerlendirmesinde iki tür doğrulama ölçütü bulunmaktadır. Bunlar dahili ölçütler ve dışsal ölçütlerdir. Dahili doğrulama ölçütleri verilere özgü olan bilgileri kullanır ve elde edilen kümelerin kalitesini ölçer. Dışsal ölçütlerde ise küme etiketleri ile verilen sınıf etiketleri karşılaştırılarak performans ölçümü yapılır. Purity, Rand İstatistikleri, Entropi, Jaccard Katsayısı, Fowlkes ve Mallows İndeksi, Minkowski skoru ve Goodman - Kruskals Katsayısı gibi birçok dış doğrulama vardır (Sripada ve Rao 2011).

Saflık(Purity) Ölçüsü: Kümedeki birimler gerçek durumdaki kümelere yayılışının ölçüsüdür. Bu değer [1 / *k*,1] aralığında değer almaktadır. Eğer C_1 kümesindeki birimler gerçek durumdaki kümelerden sadece birinde yer alıyor ise saflık değeri 1'dir. Gerçek kümedeki değerlere eşit sayıda dağıtılıyor ise saflık değeri 1 / *k* olur. Saflık değeri 1 e yaklaştıkça kümeleme kalitesi artmaktadır (Altunkaynak 2017).

i. küme için saflık ölçüsünün değeri aşağıdaki gibidir:

$$Purity_i = \frac{1}{n_j} \sum_{i=1}^{r} \max_{j=1} \{n_{ij}\} \quad (3.26)$$

C kümeleme yöntemi için saflık ölçüsü aşağıdaki gibi hesaplanır.

$$purity = \frac{1}{n} \sum_{i=1}^{r} n_i \,.\, (Purity_i\ ) = \frac{1}{n_j} \sum_{i=1}^{r} \max_{j=1}\{n_{ij}\} \tag{3.27}$$

Entropi Ölçüsü: Verideki belirsizliğe entropi denmektedir. Uygulanan kümeleme algoritmalarında her bir küme gerçek durumdaki kümelere denk gelir ise belirsizlik de sıfır olacaktır (Altunkaynak 2017).

C_i kümesine göre gerçek durumdaki entropi:

$$H(\,T \mid C_i\,) = -\sum_{j=1}^{k} \left(\frac{n_{ij}}{n_i}\right) \log_2 \left(\frac{n_{ij}}{n_i}\right) \tag{3.28}$$

C kümeleme yöntemine göre entropi aşağıdaki şekilde hesaplanır.

$$H\,(\,T|\,C\,) = \sum_{i=1}^{r} \left(\frac{n_i}{n}\right)\ H(\,T \mid C_i\,) = -\frac{1}{n} \sum_{i=1}^{r} \sum_{j=1}^{k} n_{ij}\ \log_2 \left(\frac{n_{ij}}{n_i}\right) \tag{3.29}$$

4. NANOTEKNOLOJİ ALANINDA VERİ MADENCİLİĞİ KULLANIMI

Literatürde nanoteknoloji alanında veri madenciliği ve yapay zekâ teknikleri kullanılarak yapılan bazı çalışmalara bu bölümde değinilmiştir. Çalışmalar nanolif üretim tekniğine göre, kullanılan veri madenciliği modeline ve yöntemine göre farklı başlıklar altında gruplandırılarak incelenmiştir.

4.1 Nanolif Üretim Tekniği

Tez kapsamında kullanılan veri seti elektrospinning yöntemi ile elde edilen nanoliflere ait veriler olduğu için literatür taramasında elektrospinning yöntemi üzerinde yoğunlaşılmıştır. Sarkar ve diğ. (2009) bir elektrospinleme işlemi tarafından oluşturulan nanoliflerin çapını tahmin etmek için sinir ağları kullanılmıştır. Sinir ağ modelini eğitmek ve test etmek için polietilen oksit sulu çözeltisi için deneysel veriler kullanılmıştır. Konsantrasyon, iletkenlik, akış hızı ve elektrik alan şiddeti sinir ağı modelinin giriş değişkenleridir ve k-kat çapraz doğrulama tekniği kullanılmıştır. Yapay sinir ağları nanoliflerin çapını tahmin etmede başarılı olmuştur.

Naghibzadeh ve Adabi (2014)'nin yaptığı çalışmada; yapay sinir ağı tekniği kullanılarak elektrospun jelatin nanolif çapının tahmini için etkili parametreler değerlendirilmiştir. Sıcaklık, uygulanan voltaj, polimer ve çözücü konsantrasyonları dâhil olmak üzere çeşitli elektrospinning süreçleri saf jelatin nanoliflerin üretilmesi için tasarlanmıştır. Elde edilen sonuçlar, SEM görüntüleri analiz edilerek üretilen nanoliflerin çapının 85 ila 750 nm arasında olduğunu göstermiştir. Verilerin hacmi nedeniyle veri ayarı için k-kat çapraz doğrulama kullanılmıştır. Dört giriş değişkenini, her katmanda sırasıyla 10, 18 ve 9 düğümlü 3 gizli katmanı ve bir çıkış katmanı içeren ağın test kümelerinde en iyi performansa sahip olduğu sonucuna varılmıştır. Elde edilen sonuçlar, seçilen yapay sinir ağ modelinin ilgili parametreleri değerlendirmek ve nanolif çapını tahmin etmek için kabul edilebilir bir performans sergilediğini göstermiştir.

Khatti ve diğ. (2017) yaptığı çalışmada; Tepki Yüzeyleri Metodolojisi ve yapay sinir ağı kullanılarak PCL-GT'nin elektrolif çekim yöntemi işlemini açıklayan iki tahmin modeli geliştirilmiştir. Her iki modelin özellikle de yapay sinir ağları modelinin, PCL-GT nanolif çapını doğru olarak tahmin edebildiği gösterilmiştir.

Sadan ve diğ. (2016) yapmış olduğu çalışmada yapay sinir ağları modeli, elektrospinleme parametrelerinin bir fonksiyonu olarak nano elyaf çapını tahmin etmek için kullanılmıştır. Bu çalışmada, YSA modeli aktivasyon fonksiyonu olarak sigmoid fonksiyonu kullanılarak geri yayılma algoritması ile eğitilmiştir. Yapay sinir ağı modeli, 19.5 ℃ sıcaklıkta, % 0.3 konsantrasyon, 0.3 mL / s akış hızında, 20 kV voltajda ve nozul toplayıcı mesafesi 6 cm olarak elde edilebilecek minimum 68 nm'lik bir elyaf çapını tahmin etmiştir. Bulgular, yapay sinir ağının etkili bir teknik olduğunu açıkça göstermiştir.

Nasouri ve diğ. (2013) elektrospining yöntemiyle üretilen poli nanoliflerin üretim hızını tahmin etmek için yapay sinir ağı ve tepki yüzey metodolojisi(RSM) yöntemleri kullanılmıştır. Regresyon katsayısı 0.988'dir ve bu yapay sinir ağı modelinin deneysel verilerle iyi uyuşma gösterdiğini gösterir. Elde edilen sonuçlar, YSA'nın performansının RSM'den daha iyi olduğunu göstermektedir.

4.2 Modelleme

4.2.1 Sınıflandırma ve Regresyon

Lee ve diğ. (200) tarafından nanomalzeme alanında yapılan çalışmada yapay sinir ağları kullanılarak patlayıcı gazların sınıflandırılması gerçekleştirilmiştir. PCA tekniği kullanılarak dokuz sensörden alınan sinyaller analiz edilmiş ve hata geri yayılımı öğrenme algoritması ile çok katmanlı yapay sinir ağları kullanılmıştır. Kullanılan yapay sinir ağı algoritması 9 giriş katmanı, 8 gizli katman ve 9 çıkış katmanından oluşmaktadır. Haj-Ali ve diğ. (2008) yapmış olduğu çalışmada yapay sinir ağları modellerini kullanarak doğrusal olmayan davranışlara sahip çeşitli malzemelerin nano sertlik testleri sırasındaki yanıtını simüle etmek için yeni bir yaklaşım sunmuştur. Liau ve Dai (2008) tarafından muntazam mikron altı

titandioksit kolloidlerinin hazırlanması için uygun değer işlem değişkenleri, yapay sinir ağları modellemesi ve süreç optimizasyonu algoritmaları kullanılarak tahmin edilmiştir. Bu çalışmada girdiler; [NH3], [H2O] ve reaksiyon sıcaklığıdır. Çıktılar ise titandioksit parçacık büyüklüğü ve parçacık büyüklüğü dağılımıdır. Girişler ve çıkışlar arasındaki ilişki, yapay sinir ağları yaklaşımı kullanılarak oluşturulmuştur.

Sarkar ve diğ. (2009) bir elektrospisnleme işlemi tarafından oluşturulan nanoliflerin çapını tahmin etmek için sinir ağları kullanılmıştır. Sinir ağ modelini eğitmek ve test etmek için polietilen oksit sulu çözeltisi için deneysel veriler kullanılmıştır. Konsantrasyon, iletkenlik, akış hızı ve elektrik alan şiddeti sinir ağı modelinin giriş değişkenleridir ve k-kat çapraz doğrulama tekniği kullanılmıştır. Yapay sinir ağları nanoliflerin çapını tahmin etmede başarılı olmuştur. Naghibzadeh ve Adabi(2014)'nin yaptığı çalışmada; yapay sinir ağı tekniği kullanılarak elektrospun jelatin nanolif çapının tahmini için etkili parametreler değerlendirilmiştir. Sıcaklık, uygulanan voltaj, polimer ve çözücü konsantrasyonları dâhil olmak üzere çeşitli elektrospinning süreçleri saf jelatin nanoliflerin üretilmesi için tasarlanmıştır. Elde edilen sonuçlar, SEM görüntüleri analiz edilerek üretilen nanoliflerin çapının 85 ila 750 nm arasında olduğunu göstermiştir. Verilerin hacmi nedeniyle veri ayarı için k-kat çapraz doğrulama kullanılmıştır. Dört giriş değişkenini, her katmanda sırasıyla 10, 18 ve 9 düğümlü 3 gizli katmanı ve bir çıkış katmanı içeren ağın test kümelerinde en iyi performansa sahip olduğu sonucuna varılmıştır. Elde edilen sonuçlar, seçilen yapay sinir ağ modelinin ilgili parametreleri değerlendirmek ve nanofiber çapını tahmin etmek için kabul edilebilir bir performans sergilediğini göstermiştir.

Shehadeh ve diğ. (2012) tarafından yapılan çalışmada; CdSe nanoyapı tipini tahmin etmek için istatistiksel ve veri madenciliği teknikleri kullanılmıştır. Tahmin için kullanılan yöntemler; multinominal lojistik regresyon, bir destek vektör makineleri ve rasgele orman algoritmasıdır. Sonuçlar ise duyarlılık ve seçicilik değerleriyle karşılaştırılmıştır. Rastgele orman modeli, destek vektör makineleri ve multinomial lojistik regresyon yaklaşımlarına göre nano yapı türünü tahmin etmede daha iyi performans göstermiştir. Khatti ve diğ. (2017) yaptığı çalışmada; Tepki Yüzeyleri Metodolojisi ve yapay sinir ağı kullanılarak PCL-GT'nin elektrolif çekim yöntemi işlemini açıklayan iki tahmin modeli geliştirilmiştir. Her iki modelin

özellikle de yapay sinir ağları modelinin, PCL-GT nanolif çapını doğru olarak tahmin edebildiği gösterilmiştir. Sözen ve diğ. (2018) nanokompozitlerin çekme testlerinde ortaya çıkan deformasyonu derin öğrenme ve yapay sinir ağları algoritmaları ile tahmin etmek için çalışma yapmışlardır. Farklı nanokomozitler üzerinden yapılan bu çalışmada, deformasyon oranını tahmin etmede her iki algoritmada tahminde yüksek doğruluğa sahiptir. Algoritmaların performanslarını değerlendirmek için korelasyon katsayısı ve yüzde hata kriterleri kullanılmıştır. Derin öğrenme yönetiminin daha yüksek performans sergilediği ortaya çıkarılmıştır.

Sadan ve diğ. (2016) yapmış olduğu çalışmada yapay sinir ağları modeli, elektrospinleme parametrelerinin bir fonksiyonu olarak nano elyaf çapını tahmin etmek için kullanılmıştır. Bu çalışmada, YSA modeli aktivasyon fonksiyonu olarak sigmoid fonksiyonu kullanılarak geri yayılma algoritması ile eğitilmiştir. Yapay sinir ağı modeli, 19.5 ℃ sıcaklıkta, % 0.3 konsantrasyon, 0.3 mL / s akış hızında, 20 kV voltajda ve nozul toplayıcı mesafesi 6 cm olarak elde edilebilecek minimum 68 nm'lik bir elyaf çapını tahmin etmiştir. Bulgular, yapay sinir ağının etkili bir teknik olduğunu açıkça göstermiştir. Salehi ve Razavi (2016) yapay sinir ağı ve ANFIS modellerini, atık su tuzlu suyunun nano filtrasyon işleminin ortalama permeat akı ve sodyum klorür reddini tahmin etmek için kullanılmıştır. ANFIS ve yapay sinir ağı modellerinin besleme konsantrasyonu, basınç ve sıcaklık olmak üzere üç girdisi vardır. Her iki model de toplam verilerin% 30'uyla eğitilmiştir. Sonuçlara göre yapay sinir ağı modelinin daha başarılı olduğu tespit edilmiştir.

Eren ve İleri (2007) DS 5 tipi nanofiltrasyon membran kullanan membran prosesinde tuz giderim verimini belirlemek için ileri beslemeli bir YSA modeli kullanılmıştır. Yapay Sinir Ağını(YSA) eğitmek ve test etmek için 238 adet deneysel veri literatürden toplanmıştır. Bu deneysel verilerin 178 tanesi yapay sinir ağının eğitim setinde ve 60 tanesi de test setinde kullanılmak üzere rasgele seçilerek iki kısma ayrılmıştır. YSA modeli beş girdi ve bir çıktıdan oluşmaktadır. Bu girdiler basınç, tuz konsantrasyonu, boya konsantrasyonu, yatay akis hızı ve pH iken çıktı ise tuz giderme verimidir. Test setinde YSA'nın performansını değerlendirmek için ortalama mutlaka yüzde hata (OMYH) ve ortalama karesel hata (OKH) metotları kullanılmıştır. Eğitim ve test sonuçlarının ortalama mutlak yüzde hata değerleri

sırasıyla 4,22 ve 3,84'dür. Bu değerlere göre deneysel sonuçlar ile YSA sonuçları arasında çok iyi bir uyum olduğu görülmüştür.

Nasouri ve diğ. (2013) elektrospining yöntemiyle üretilen poli nanoliflerin üretim hızını tahmin etmek için yapay sinir ağı ve tepki yüzey metodolojisi (RSM) yöntemleri kullanılmıştır. Regresyon katsayısı 0.988'dir ve bu yapay sinir ağı modelinin deneysel verilerle iyi uyuştuğunu ve YSA performansının RSM'den daha iyi olduğunu göstermektedir.

4.2.2 Kümeleme

AbuOmar ve diğ. (2013) yaptığı çalışmada, buharlı karbon nanolif (VGCNF) / vinil ester (VE) nanokompozitlerinin viskoelastik özellikleri hakkında bilgi edinmek için veri madenciliği ve bilgi keşif tekniklerini kullanmıştır. Bu çalışmada VGCNF / VE veri setini analiz etmek için denetimsiz veri madenciliği tekniği kullanılmıştır. Veri seti, beş girdi ve üç çıkıştan oluşan 240 adet veriden oluşmaktadır. Veri setine özdüzenleyici haritalar (self-organization maps-SOM) uygulanmıştır. Ek olarak temel bileşen analizi, nanokompozit verilerinin iki boyutlu gösterimini sağlamak için kullanılmıştır. Bu işlem, VGCNF / VE nanokompozitlerin fiziksel, mekanik özelliklerini karakterize etmek için bulanık c-ortalamalar (FCM) kümeleme algoritmasının uygulanmasını kolaylaştırmıştır. SOM'lar analiz edildikten sonra, viskoelastik madde tepkileri üzerinde en yüksek etkiye sahip olan sıcaklık; VGCNF / VE nanokompozitlerin baskın özelliği olarak belirlenmiştir. Son olarak, temel bileşen analizi tekniği kullanılarak başka bir veri analizi yapılmıştır. Ardından, ortaya çıkan yeni veri setine Gustafon-Kessel mesafe ölçüsü ile bulanık c-ortalamalar algoritması uygulanmıştır. FCM, numuneleri sıcaklık ve tan delta değerlerine göre kümelemiştir.

AbuOmar ve diğ. (2019) tarafından yapılan bir başka çalışmada daha önce analiz edilen VGCNF / VE nanokompozitinin; viskoelastik, sıkıştırma, gerginlik ve bükülme özellik verilerinden oluşan genişletilmiş veri setinde denetimsiz veri madenciliği ve bilgi keşif teknikleri uygulanmıştır. Veri madenciliği ve bilgi keşif teknikleri, buharlı karbon nanofiber (VGCNF) / vinil ester (VE) nanokompozitlerinin viskoelastik, eğilme, basma ve gerilme özellikleri hakkında yeni bilgiler elde etmek

için kullanılmıştır. Bu çalışmada self-organization maps(SOM) ve kümeleme tekniklerinin yer aldığı veri madenciliği ve bilgi keşif algoritmaları kullanılmıştır. Bu veri seti, üç farklı nanokompozit yapı, VGCNF / VE viskoelastik veriler, eğilme verileri, sıkıştırma ve gerilim verilerini temsil eden kaynaklar tarafından oluşturulmuştur. SOM'lar analiz edildikten sonra test sıcaklıkları ve tan deltaları, malzeme tepkileri üzerinde en yüksek etkiye sahip baskın özellikler olarak tanımlanmıştır. Bir başka veri analizi, temel bileşen analizi (PCA) kullanılarak yapılmıştır. Ardından, ortaya çıkan yeni veri setine Gustafson-Kessel (GK) mesafe ölçüsü ile bulanık C -ortalama algoritması uygulanmıştır.

4.3 Algoritma

4.3.1 Yapay Sinir Ağları

Lee ve diğ. (200) tarafından nanomalzeme alanında yapılan çalışmada yapay sinir ağları kullanılarak patlayıcı gazların sınıflandırılması gerçekleştirilmiştir. PCA tekniği kullanılarak dokuz sensörden alınan sinyaller analiz edilmiş ve hata geri yayılımı öğrenme algoritması ile çok katmanlı yapay sinir ağları kullanılmıştır. Kullanılan yapay sinir ağı algoritması 9 giriş katmanı, 8 gizli katman ve 9 çıkış katmanından oluşmaktadır.

Haj-Ali ve diğ. (2008) yapmış olduğu çalışmada yapay sinir ağları modellerini kullanarak doğrusal olmayan davranışlara sahip çeşitli malzemelerin nano sertlik testleri sırasındaki yanıtını simüle etmek için yeni bir yaklaşım sunmuştur.

Liau ve Dai (2008) tarafından muntazam mikron altı titandioksit kolloidlerinin hazırlanması için optimum işlem değişkenleri, yapay sinir ağları modellemesi ve süreç optimizasyonu algoritmaları kullanılarak tahmin edilmiştir. Bu çalışmada girdiler; [NH3], [H2O] ve reaksiyon sıcaklığıdır. Çıktılar ise titandioksit parçacık büyüklüğü ve parçacık büyüklüğü dağılımıdır. Girişler ve çıkışlar arasındaki ilişki, yapay sinir ağları yaklaşımı kullanılarak oluşturulmuştur.

Sarkar ve diğ. (2009) bir elektrospinleme işlemi tarafından oluşturulan nanolifin çapını tahmin etmek için sinir ağları kullanılmıştır. Sinir ağ modelini eğitmek ve test etmek için polietilen oksit sulu çözeltisi için deneysel veriler kullanılmıştır. Konsantrasyon, iletkenlik, akış hızı ve elektrik alan şiddeti sinir ağı modelinin giriş değişkenleridir ve k-kat çapraz doğrulama tekniği kullanılmıştır. Yapay sinir ağları nanoliflerin çapını tahmin etmede başarılı olmuştur.

Naghibzadeh ve Adabi (2014)'nin yaptığı çalışmada; yapay sinir ağı tekniği kullanılarak elektrospun jelatin nanolif çapının tahmini için etkili parametreler değerlendirilmiştir. Sıcaklık, uygulanan voltaj, polimer ve çözücü konsantrasyonları dâhil olmak üzere çeşitli elektrospinning süreçleri saf jelatin nanoliflerin üretilmesi için tasarlanmıştır. Elde edilen sonuçlar, SEM görüntüleri analiz edilerek üretilen nanoliflerin çapının 85 ila 750 nm arasında olduğunu göstermiştir. Verilerin hacmi nedeniyle veri ayarı için k-kat çapraz doğrulama kullanılmıştır. Dört giriş değişkenini, her katmanda sırasıyla 10, 18 ve 9 düğümlü 3 gizli katmanı ve bir çıkış katmanı içeren ağın test kümelerinde en iyi performansa sahip olduğu sonucuna varılmıştır. Elde edilen sonuçlar, seçilen yapay sinir ağ modelinin ilgili parametreleri değerlendirmek ve nanolif çapını tahmin etmek için kabul edilebilir bir performans sergilediğini göstermiştir.

Khatti ve diğ. (2017) yaptığı çalışmada; Tepki Yüzeyleri Metodolojisi ve yapay sinir ağı kullanılarak PCL-GT'nin elektrolif çekim yöntemi işlemini açıklayan iki tahmin modeli geliştirilmiştir. Her iki modelin özellikle de yapay sinir ağları modelinin, PCL-GT nanolif çapını doğru olarak tahmin edebildiği gösterilmiştir.

Sözen ve diğ. (2018) nanokompozitlerin çekme testlerinde ortaya çıkan deformasyonu derin öğrenme ve yapay sinir ağları algoritmaları ile tahmin etmek için çalışma yapmışlardır. Farklı nanokomozitler üzerinden yapılan bu çalışmada, deformasyon oranını tahmin etmede her iki algoritmada tahminde yüksek doğruluğa sahiptir. Algoritmaların performanslarını değerlendirmek için korelasyon katsayısı ve yüzde hata kriterleri kullanılmıştır. Derin öğrenme yönetiminin daha yüksek performans sergilediği ortaya çıkarılmıştır.

Sadan ve diğ. (2016) yapmış olduğu çalışmada yapay sinir ağları modeli, elektrospinleme parametrelerinin bir fonksiyonu olarak nano elyaf çapını tahmin

etmek için kullanılmıştır. Bu çalışmada, YSA modeli aktivasyon fonksiyonu olarak sigmoid fonksiyonu kullanılarak geri yayılma algoritması ile eğitilmiştir. Yapay sinir ağı modeli, 19.5 ℃ sıcaklıkta, % 0.3 konsantrasyon, 0.3 mL / s akış hızında, 20 kV voltajda ve nozul toplayıcı mesafesi 6 cm olarak elde edilebilecek minimum 68 nm'lik bir elyaf çapını tahmin etmiştir. Bulgular, yapay sinir ağının etkili bir teknik olduğunu açıkça göstermiştir.

Salehi ve Razavi (2016) yapay sinir ağı ve ANFIS modellerini, atık su tuzlu suyunun nano filtrasyon işleminin ortalama permeat akı ve sodyum klorür reddini tahmin etmek için kullanılmıştır. ANFIS ve yapay sinir ağı modellerinin besleme konsantrasyonu, basınç ve sıcaklık olmak üzere üç girdisi vardır. Her iki model de toplam verilerin% 30'uyla eğitilmiştir. Sonuçlara göre yapay sinir ağı modelinin daha başarılı olduğu tespit edilmiştir.

Eren ve İleri (2007) DS 5 tipi nanofiltrasyon membran kullanan membran prosesinde tuz giderim verimini belirlemek için ileri beslemeli bir YSA modeli kullanılmıştır. Yapay Sinir Ağını(YSA) eğitmek ve test etmek için 238 adet deneysel veri literatürden toplanmıştır. Bu deneysel verilerin 178 tanesi yapay sinir ağının eğitim setinde ve 60 tanesi de test setinde kullanılmak üzere rasgele seçilerek iki kısma ayrılmıştır. YSA modeli beş girdi ve bir çıktıdan oluşmaktadır. Bu girdiler basınç, tuz konsantrasyonu, boya konsantrasyonu, yatay akis hızı ve pH iken çıktı ise tuz giderme verimidir. Test setinde YSA'nın performansını değerlendirmek için ortalama mutlaka yüzde hata (OMYH) ve ortalama karesel hata (OKH) metotları kullanılmıştır. Eğitim ve test sonuçlarının ortalama mutlak yüzde hata değerleri sırasıyla 4,22 ve 3,84'dür. Bu değerlere göre deneysel sonuçlar ile YSA sonuçları arasında çok iyi bir uyum olduğu görülmüştür.

Nasouri ve diğ. (2013) elektrospining yöntemiyle üretilen poli nanofiberlerin üretim hızını tahmin etmek için yapay sinir ağı ve tepki yüzey metodolojisi(RSM) yöntemleri kullanılmıştır. Regresyon katsayısı 0.988'dir ve bu yapay sinir ağı modelinin deneysel verilerle iyi uyuşma gösterdiğini gösterir. Elde edilen sonuçlar, YSA'nın performansının RSM'den daha iyi olduğunu göstermektedir.

4.3.2 Bulanık c-ortalama

AbuOmar ve diğ. (2013) yaptığı çalışmada, buharlı karbon nanofiber (VGCNF) / vinil ester (VE) nanokompozitlerinin viskoelastik özellikleri hakkında bilgi edinmek için veri madenciliği ve bilgi keşif tekniklerini kullanmıştır. Bu çalışmada, VGCNF / VE veri setini analiz etmek için denetimsiz veri madenciliği tekniği kullanılmıştır. Veri seti, beş girdi ve üç çıkışı olan 240 adet veriden oluşmaktadır. Veri setine self-organization maps(SOM) uygulanmıştır. Ek olarak temel bileşen analizi, nanokompozit verilerinin iki boyutlu gösterimini sağlamak için kullanılmıştır. Bu işlem, VGCNF / VE nanokompozitlerin fiziksel, mekanik özelliklerini karakterize etmek için bulanık c-ortalamalar (FCM) kümeleme algoritmasının uygulanmasını kolaylaştırmıştır. SOM'lar analiz edildikten sonra, viskoelastik madde tepkileri üzerinde en yüksek etkiye sahip olan sıcaklık; VGCNF / VE nanokompozitlerin baskın özelliği olarak belirlenmiştir. Son olarak, temel bileşen analizi tekniği kullanılarak başka bir veri analizi yapılmıştır. Ardından, ortaya çıkan yeni veri setine Gustafon-Kessel mesafe ölçüsü ile bulanık c-ortalamalar algoritması uygulanmıştır. FCM, numuneleri sıcaklık ve tan delta değerlerine göre kümelemiştir.

AbuOmar ve diğ. (2019) tarafından yapılan bir başka çalışmada daha önce analiz edilen VGCNF / VE nanokompozitinin; viskoelastik, sıkıştırma, gerginlik ve bükülme özellik verilerinden oluşan genişletilmiş veri setinde denetimsiz veri madenciliği ve bilgi keşif teknikleri uygulanmıştır. Veri madenciliği ve bilgi keşif teknikleri, buharlı karbon nanofiber (VGCNF) / vinil ester (VE) nanokompozitlerinin viskoelastik, eğilme, basma ve gerilme özellikleri hakkında yeni bilgiler elde etmek için kullanılmıştır. Bu çalışmada self-organization maps(SOM) ve kümeleme tekniklerinin yer aldığı veri madenciliği ve bilgi keşif algoritmaları kullanılmıştır. Bu veri seti, üç farklı nanokompozit yapı, VGCNF / VE viskoelastik veriler, eğilme verileri, sıkıştırma ve gerilim verilerini temsil eden kaynaklar tarafından oluşturulmuştur. SOM'lar analiz edildikten sonra test sıcaklıkları ve tan deltaları, malzeme tepkileri üzerinde en yüksek etkiye sahip baskın özellikler olarak tanımlanmıştır. Bir başka veri analizi, temel bileşen analizi (PCA) kullanılarak yapılmıştır. Ardından, ortaya çıkan yeni veri setine Gustafon-Kessel (GK) mesafe ölçüsü ile bulanık c-ortalamalar algoritması uygulanmıştır.

5. YÖNTEM

Bir önceki bölümde nanoteknoloji alanında veri madenciliği konusunu kapsayan çalışmalar değerlendirilmiştir. Literatürde sıklıkla yapay sinir ağı yönteminin kullanıldığı ve tahminleme üzerine çalışmaların yapıldığı görülmüştür. Bu çalışmada farklı olarak nanofiber kaplı filtre malzemelerinden oluşan veri seti hem sınıflandırma hem de kümeleme algoritmaları kullanılarak model başarım ölçütleriyle değerlendirilmiştir.

Bu bölümde tezde kullanılan veri seti ve yöntemler açıklanmıştır. Nanofiber veri seti üzerinde sınıflandırma ve kümeleme algoritmalarıyla analiz işlemleri gerçekleştirilmiş olup model başarımları karşılaştırılmıştır.

Tez kapsamında kullanılan veri seti nanoteknoloji alanında faaliyet gösteren özel bir şirketten alınmıştır. Nanofiber kaplı filtre malzemelerinin belirli bir standarda göre sınıflandırılması için laboratuvar ortamında yapılan ölçümler sonucunda elde edilen verilere göre veri seti oluşturulmuştur. Belirli bir ebattaki filtreye hava akımı ve belirli boyutlarda parçacık gönderilerek filtre kalitesi ölçülebilmektedir.

Ham veri setindeki nitelikler Tablo 5.2'de gösterilmiştir.

Tablo 5.2 Ham Veri Setindeki Nitelikler

	Değişken	Veri Türü
1	Ürün Adı	Nominal
2	Lot No	Nümerik
3	Ebat	Nümerik
4	Hava Akımı	Nümerik
5	PartBuyuklugu	Nümerik
6	Resistance	Nümerik
7	Penetration	Nümerik
8	Filtre	Nominal

Nanofiber kaplı filtre numunesi üzerine hava akımı ile beraber parçacıklar gönderildiğinde resistans ve penetrasyon değerleri oluşmaktadır. Farklı boyutlarda ebatlar seçilerek ölçüm yapılmaktadır.

Resistans; hava akımına karşı gösterilen dirençtir. Filtre numunesine hava akımı uygulandığında numunenin hava giriş ve çıkışı arasında oluşan diferansiyel basınç olup birimi Pascaldır.

Penetrasyon; hava akım değerinde filtre numunesi üzerine gönderilen mikron büyüklüğündeki parçacıkların filtrenin diğer tarafına geçme yüzdesidir.

Ölçüm sonuçlarına göre filtre sınıflandırılması yapılmaktadır. F7, F8 ve F9 filtre sınıflarına göre en iyi kalitedeki filtre sınıfı F9'dur.

Tez kapsamında izlenen yöntem adımları aşağıda gösterilmiştir.

Şekil 5.10: Tez Süreç Adımları

Ham veri setinde yer alan Ürün Adı ve Lot Numarası nitelikleri bu çalışmada yer alan yöntemlere herhangi bir etkisi olmadığı için veri setinden kaldırılarak veri indirgeme işlemi yapılmıştır. İndirgeme işlemleri sonrasında veri seti 6 nitelikten ve 1128 satır veriden oluşmaktadır.

Laboratuvar ortamındaki ölçüm sonuçlarından alınan bu veri setinde eksik değer bulunmamaktadır. Literatürde belirtilen normalizasyon yöntemleri veri setine uygulanmıştır. k-en yakın komşu algoritması ile veri setine uygun olan normalizasyon yöntemi analiz edilmiştir.

Nanofiber kaplı filtre numunelerinin sınıflandırılmasında veri seti hold-out performansına göre %70 eğitim - %30 test, %75 eğitim - %25 test, %80 eğitim - %20 test veri seti olarak ayrılmıştır. K-kat çapraz geçerleme kullanılarak 2-kat, 5-kat ve 10-kat çapraz geçerleme uygulanmıştır. Sınıflandırma algoritmaları olarak C4.5 Karar Ağacı, Rastgele Orman, Naive Bayes ve yapay sinir ağları kullanılmış ve model başarı ölçütleri olan doğruluk, kesinlik, duyarlılık, F-ölçütü, Kappa istatistiği ile modeller karşılaştırılmıştır.

k-means ve bulanık c-ortalama kümeleme algoritmaları ile veri seti üzerinde kümeleme işlemleri yapılmıştır. Kümeleme başarısı entropi (entropy) ve saflık (purity) ölçütleri ile karşılaştırılmıştır.

Tüm bu normalizasyon işlemleri, sınıflandırma algoritmaları ve kümeleme algoritmaları R dilinde gerçekleştirilmiştir. Derleyici olarak RStudio kullanılmıştır.

R ücretsiz bir istatistiksel programlama dilidir. Bilimsel hesaplama, istatistik, veri madenciliği ve makine öğrenmesi konularında özellikle tercih edilmektedir. R dilinin açık kaynak kodlu olması, dokümantasyonunun gelişmiş olması ve geniş topluluk desteğinin olması diğer alternatif dillere göre daha çok ön plana çıkmasını sağlamaktadır (Kartal 2015). Yeni Zelandalı Ross Ohaka ve Robert Genteleman tarafından başlatılan R dili, bu iki yazılımcılarının isimlerinden gelmektedir (Balaban ve Kartal 2015). R programı temel program ve paketlerden oluşmaktadır. Bundan dolayı R programı indirildikten sonra içerisinde gelen standart paketler ile temel işlemler ve istatistiksel analizler yapılabilmektedir. Özel analizlerin veya grafik çizimlerinin gerekli olduğu durumlarda ek paketler indirilmesi gerekmektedir (Altunkaynak 2017).

Balaban ve Kartal (2015) tarafından yazılan “Veri Madenciliği ve Makine Öğrenmesi Temel Algoritmaları ve R Dili ile Uygulamaları” kitabı ile “R ile Veri

Madenciliği Uygulamaları" kitabı veri madenciliği ve R programlaması konusunda yol gösterici Türkçe kaynaklardır.

R dilinde veri setinin özet bilgisini göstermek için summary() fonksiyonu kullanılmaktadır. Veri indirgeme işlemi yapıldıktan sonraki veri setine ait özet bilgi Şekil 5.11'de yer almaktadır.

```
      Ebat          HavaAkim     PartBuyuklugu      Resistance     Penetration     Filtre
 Min.   :100.0   Min.   : 60.0   Min.   :0.2000   Min.   :102.6   Min.   : 2.597   F7:392
 1st Qu.:100.0   1st Qu.: 60.0   1st Qu.:0.2000   1st Qu.:124.5   1st Qu.:18.462   F8:416
 Median :100.0   Median : 60.0   Median :0.2000   Median :139.8   Median :27.537   F9:320
 Mean   :113.3   Mean   : 75.9   Mean   :0.2952   Mean   :220.4   Mean   :29.465
 3rd Qu.:100.0   3rd Qu.: 90.0   3rd Qu.:0.4000   3rd Qu.:285.5   3rd Qu.:33.147
 Max.   :200.0   Max.   :120.0   Max.   :0.6000   Max.   :589.1   Max.   :94.488
```

Şekil 5.11: Nanofiber Veri Seti Özeti

Veri setinde yer alan niteliklerin yapısını göstermek için str() fonksiyonu kullanılmaktadır. Veri indirgeme işlemi yapıldıktan sonraki veri setinin yapısı Şekil 5.12'de belirtilmiştir.

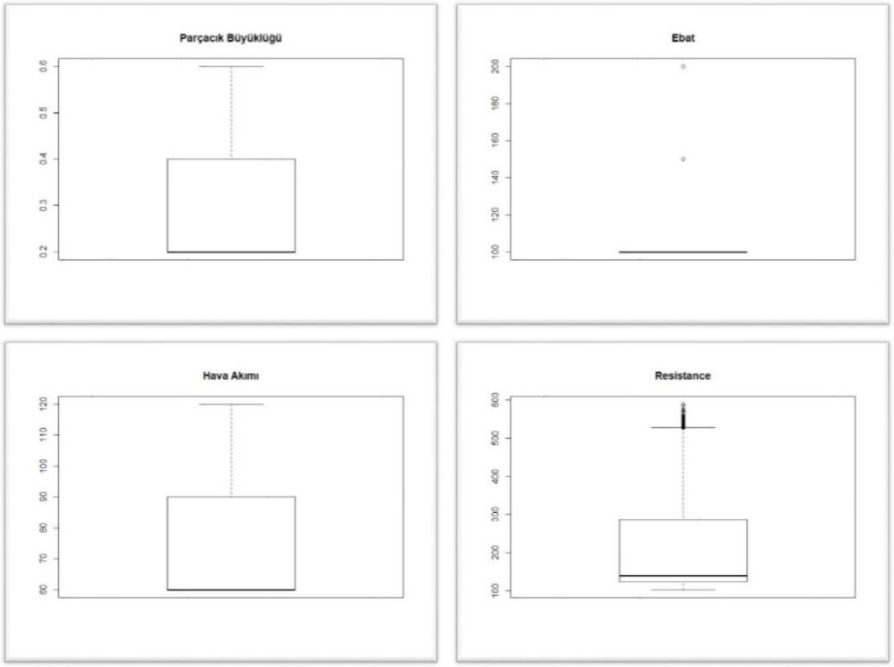
```
'data.frame':   1128 obs. of  6 variables:
 $ Ebat         : num  100 100 100 100 100 100 100 100 100 100 ...
 $ HavaAkim     : num  60 60 60 60 60 60 60 60 60 60 ...
 $ PartBuyuklugu: num  0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2 ...
 $ Resistance   : num  177 134 109 113 132 ...
 $ Penetration  : num  9.88 21.29 49.48 41.22 23.51 ...
 $ Filtre       : Factor w/ 3 levels "F7","F8","F9": 3 2 1 1 2 1 2 3 2 2 ...
```

Şekil 5.12: Nanofiber Veri Seti Yapısı

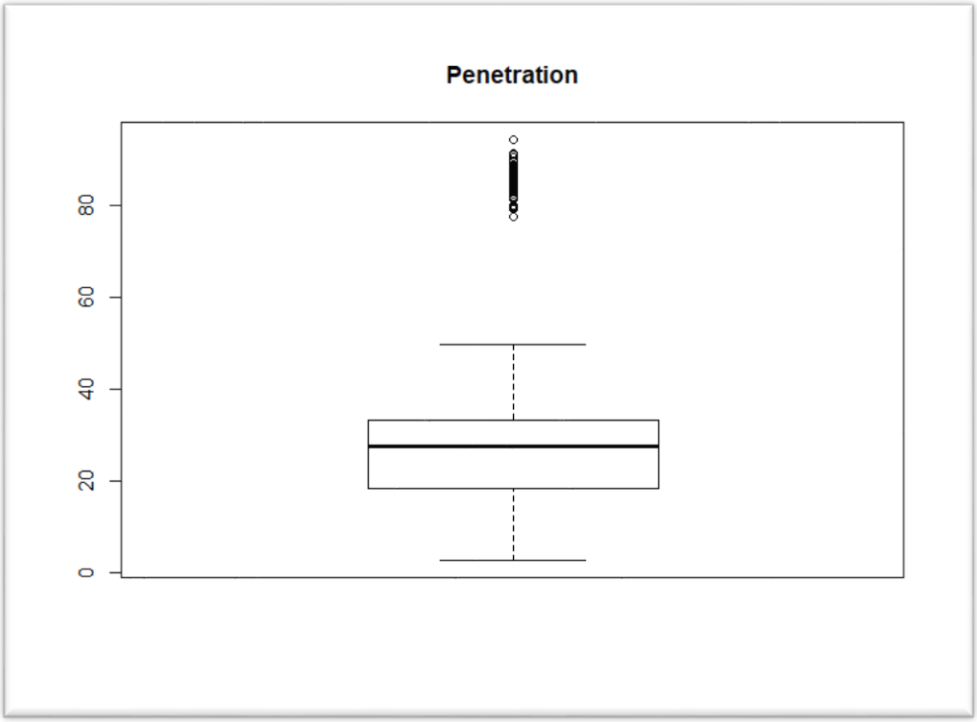
Laboratuvar ortamındaki ölçüm cihazının ebat olarak ölçebildiği numune büyüklüğü 100 cm^2 ve 200 cm^2 arasında değişmektedir. Cihazın hava akımı olarak ölçüm aralığı 60 lt/dk ile 120 lt/dk arasında değişebilmektedir. Cihazda yer alan partikül jeneratörünün ürettiği partiküllerin dağılımı 0,2 mikron ile 0,6 mikron arasındadır. Niteliklere ait minimum, maksimum, median değerleri summary() fonksiyonunda da gösterilmektedir. Veri setinde yer alan niteliklerin dağılımları Şekil 5.19 ve Şekil 5.20'daki Boxplot grafiklerinde gösterilmiştir.

Kutu (Boxplot) grafiğinde minimum değer, maksimum değer ve median değerine göre aşağı ve yukarı dağılım gösterilir. Kutu grafiği bize sayıların hangi aralıkta değiştiğini göstermektedir. Grafikte en üst ve en alttaki çizgiler sırası ile maksimum ve minimum değerleri, koyu renk çizgi ise median değerini verir.

Kutu grafiğinde dikdörtgen çerçevenin aşağı ve yukarı çizgileri ilgili niteliğin 3. ve 1. çeyrek değerlerine denk gelmektedir.

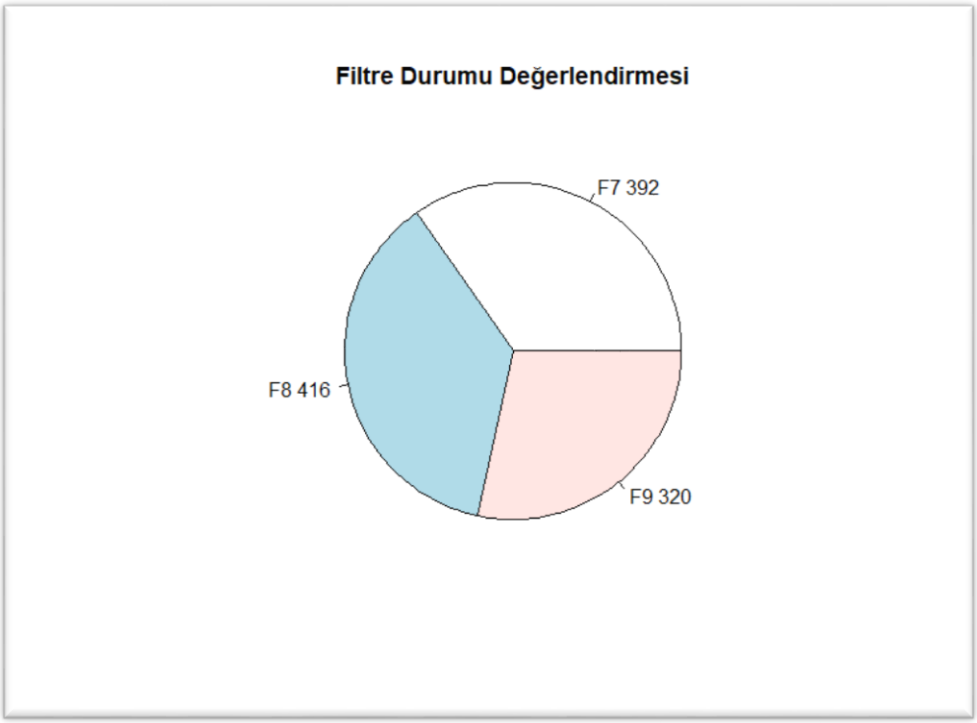


Şekil 5.13: Veri İndirgeme Sonrası Nitelik Kutu Grafiği



Şekil 5.14: Veri İndirgeme Sonrası Nitelik Kutu Grafiği

Nanofiber veri setindeki hedef değişkenin dağılımı Şekil 5.15'de dairesel grafik üzerinde gösterilmektedir. Grafikte de belirtildiği üzere F7 filtre sınıfında 392, F8 filtre sınıfında 416 ve F9 filtre sınıfında 320 adet değer vardır.



Şekil 5.15: Hedef Nitelik Dağılımı

5.1 Veri Ön İşleme Tekniklerinin Veri Setine Uygulanması

Veri setinde eksik veri bulunmamakla beraber ürün adı ve lot numarasında indirgeme işlemi yapılmıştır. Nümerik veriler arasındaki farklılığı azaltmak ve belirli bir aralıkta tanımlamak için normalizasyon adımı uygulanmıştır.

Literatürde yer alan min-max normalizasyonu, z-score normalizasyonu, ondalık ölçeklendirme ve sigmoid normalizasyonu veri setine uygulanmıştır.

Normalize edilen veri seti %70 eğitim ve %30 test olmak üzere ayrılmış ve model olarak k-en yakın komşu algoritması kullanılmıştır. Model başarı ölçütleri olarak doğruluk, kesinlik, duyarlılık, F-ölçütü, Kappa istatistiği karşılaştırılmıştır.

R dilinde min-max normalizasyonu için **mmnorm()** fonksiyonu, z-score normalizasyonu için **znorm()** fonksiyonu, ondalık ölçeklendirme için **decscale()** fonksiyonu, sigmoid normalizasyonu için ise **signorm()** fonksiyonu kullanılmıştır.

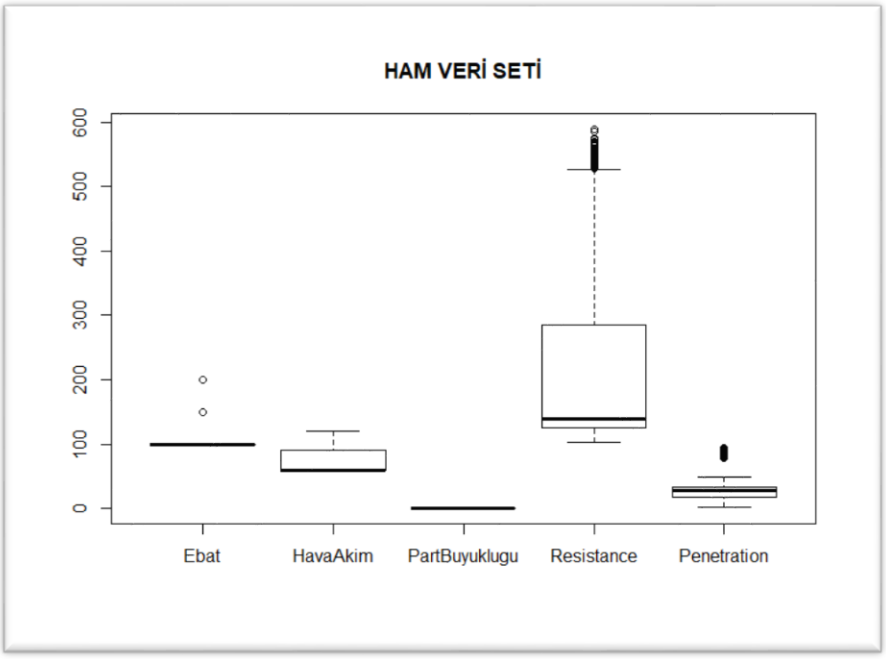
k-en yakın komşu algoritmasını uygulamak için R'da "caret" paketinde bulunan "knn" fonksiyonu kullanılmıştır.

Tablo 5.2: Normalizasyon Yöntemlerinin Karşılaştırılması

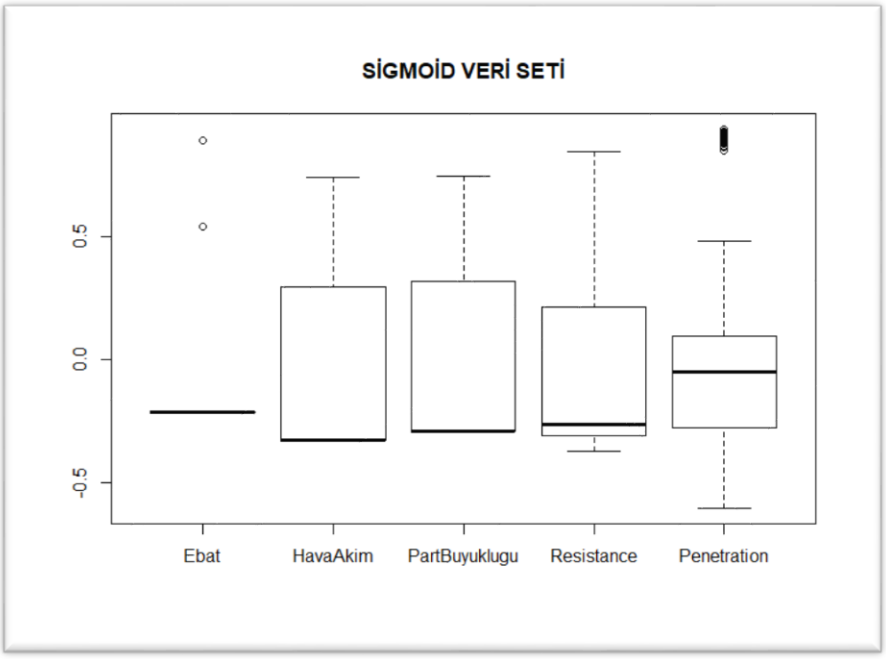
	Doğruluk	Kesinlik	Duyarlılık	F-Ölçütü	Kappa
Min-Max Normalizasyonu	0.8813	0.8988	0.8886	0.8737	0.8192
Z-Score Normalizasyonu	0.8872	0.9035	0.8741	0.8788	0.8282
Ondalık Ölçeklendirme	0.8991	0.9133	0.8854	0.8893	0.8464
Sigmoid Normalizasyonu	0.9139	0.9254	0.9001	0.9034	0.8690

Doğruluk değeri olarak min-max normalizasyon ve z-score normalizasyon yöntemleri yakın değerlere sahip iken sigmoid normalizasyon yöntemi en yüksek değeri vermektedir. Kesinlik ve duyarlılık ölçütlerini baz alan F-ölçütü değerlendirildiğinde yine sigmoid normalizasyonunun en yüksek değere sahip olduğu görülmektedir.

Ham veri setine ait kutu grafiği ve sigmoid normalizasyon işleminden sonraki veri setine ait kutu grafiği aşağıdaki gibidir. Normalizasyon öncesinde nitelikler arasındaki fark fazla iken sigmoid normalizasyon sonrasında nitelikler arasındaki farklılık azalmıştır.



Şekil 5.16: Orijinal Veri Seti Değer Aralıkları



Şekil 5.17: Normalize Edilmiş Veri Seti Değer Aralıkları

5.2 Sınıflandırma Yöntemleri ve Model Değerlendirme

C4.5 Karar Ağacı Algoritması ile Model Oluşturma

Tablo 5.3: C4.5 Karar Ağacı Algoritması Model Özeti

Algoritma	C4.5 Karar Ağacı
Performans Değerlendirme ve Model Seçimi	%70-%30 , %75-%25, %80-%20 Hold Out
	2-kat, 5-kat, 10-kat Çapraz Doğrulama
R Paket	Caret
R Fonksiyon	J48
Model Başarı Ölçütü	Doğruluk, Kesinlik, Duyarlılık, F-Ölçütü, Kappa Değeri

C4.5 karar ağacı algoritması oluşturulurken hold-out ve k-kat çapraz geçerleme performans değerlendirme yöntemi kullanılarak model başarısı ölçülmüştür. R'da "caret" paketindeki "J48" fonksiyonu kullanılmıştır. Model başarı ölçütleri olarak Doğruluk, Kesinlik, Duyarlılık, F-Ölçütü, Kappa istatistiği değerlendirilmiştir.

Rastgele Orman Algoritması ile Model Oluşturma

Tablo 5.4: Rastgele Orman Algoritması Model Özeti

Algoritma	Rastgele Orman
Performans Değerlendirme ve Model Seçimi	%70-%30 , %75-%25, %80-%20 Hold Out
	2-kat, 5-kat, 10-kat Çapraz Doğrulama
R Paket	Caret
R Fonksiyon	rf
Model Başarı Ölçütü	Doğruluk, Kesinlik, Duyarlılık, F-Ölçütü, Kappa Değeri

Rastgele Orman algoritması oluşturulurken hold-out ve k-kat çapraz geçerleme performans değerlendirme yöntemi kullanılarak model başarısı ölçülmüştür. R'da "caret" paketindeki "rf" fonksiyonu kullanılmıştır. Model başarı ölçütleri olarak Doğruluk, Kesinlik, Duyarlılık, F-Ölçütü, Kappa istatistiği değerlendirilmiştir.

Naive Bayes ile Model Oluşturma

Tablo 5.5: Naive Bayes Algoritması Model Özeti

Algoritma	Naive Bayes
Performans Değerlendirme ve Model Seçimi	%70-%30 , %75-%25, %80-%20 Hold Out
	2-kat, 5-kat, 10-kat Çapraz Doğrulama
R Paket	Caret
R Fonksiyon	nb
Model Başarı Ölçütü	Doğruluk, Kesinlik, Duyarlılık, F-Ölçütü, Kappa Değeri

Naive Bayes algoritması oluşturulurken hold-out ve k-kat çapraz geçerleme performans değerlendirme yöntemi kullanılarak model başarısı ölçülmüştür. R'da "caret" paketindeki "nb" fonksiyonu kullanılmıştır. Model başarı ölçütleri olarak Doğruluk, Kesinlik, Duyarlılık, F-Ölçütü, Kappa istatistiği değerlendirilmiştir.

Yapay Sinir Ağları ile Model Oluşturma

Tablo 5.6: Yapay Sinir Ağı Algoritması Model Özeti

Algoritma	Yapay Sinir Ağları
Performans Değerlendirme ve Model Seçimi	%70-%30 , %75-%25, %80-%20 Hold Out
	2-kat, 5-kat, 10-kat Çapraz Doğrulama
R Paket	Caret
R Fonksiyon	nnet
Model Başarı Ölçütü	Doğruluk, Kesinlik, Duyarlılık, F-Ölçütü, Kappa Değeri

Yapay sinir ağları oluşturulurken hold-out ve k-kat çapraz geçerleme performans değerlendirme yöntemi kullanılarak model başarısı ölçülmüştür. R'da "caret" paketindeki "nnet" fonksiyonu kullanılmıştır. Model başarı ölçütleri olarak Doğruluk, Kesinlik, Duyarlılık, F-Ölçütü, Kappa istatistiği değerlendirilmiştir.

Model Değerlendirme

Tablo 5.7: Hold-Out Model Performans Değerlendirme

	Hold-Out Yüzdesi	C4.5 Karar Ağacı	Rastgele Orman	Naive Bayes	Yapay Sinir Ağları
Doğruluk	*70%*	0.9139	0.9139	0.8754	0.9080
	75%	0.9078	0.9078	0.8652	0.8972
	80%	0.9289	0.9333	0.8533	0.9067
Kesinlik	*70%*	0.9283	0.9254	0.8966	0.9180
	75%	0.9241	0.9203	0.8893	0.9027
	80%	0.9397	0.9431	0.8910	0.9115
Duyarlılık	*70%*	0.8993	0.9001	0.8624	0.8947
	75%	0.8917	0.8926	0.8510	0.8830
	80%	0.9167	0.9219	0.8431	0.9018
F-Ölçütü	*70%*	0.9028	0.9034	0.8684	0.8953
	75%	0.8951	0.8958	0.8573	0.8851
	80%	0.9208	0.9260	0.8526	0.8979
Kappa Değeri	*70%*	0.8689	0.8690	0.8101	0.8601
	75%	0.8594	0.8595	0.7943	0.8436
	80%	0.8919	0.8986	0.7762	0.8583

Doğruluk değerleri göz önüne alındığında Rastgele Orman algoritmasının %80 eğitim - %20 doğrulama hold-out yönteminde 0.9333 oranıyla en başarılı sonucu verdiği görülmektedir. F-ölçütü, kesinlik (precision) ve duyarlılık (recall) değerlerinin ağırlıklandırılmış ortalamasını göstermektedir. Bundan dolayı sadece kesinlik ve duyarlılığa bakmak yerine F-ölçütüne bakmak daha doğru sonuç vermektedir. 0.9260 F-ölçütü oranıyla en yüksek başarıyı yine Rastgele Orman algoritması yöntemi göstermiştir. Kappa değerinin sınıflandırıcı modelin doğru sınıflandırma başarımı hakkında bilgi vermektedir. Kappa değerinin 0.40'dan büyük olmasının makul olduğu araştırmacılar tarafından bildirilmiştir.

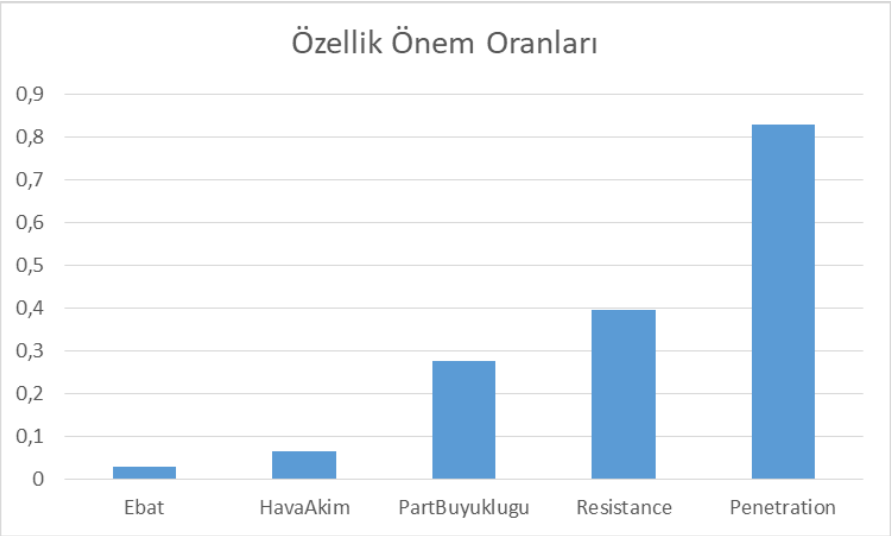
Tablo 5.8: k-kat Çapraz Geçerleme Model Performans Değerlendirme

	k-kat	C4.5 Karar Ağacı	Rastgele Orman	Naive Bayes	Yapay Sinir Ağları
Doğruluk	*2-kat*	0.9149	0.9096	0.7358	0.9096
	5-kat	0.9200	0.9022	0.7733	0.9022
	10-kat	0.9018	0.8929	0.7500	0.8839
Kesinlik	*2-kat*	0.9298	0.9148	0.7276	0.9232
	5-kat	0.9348	0.9071	0.7634	0.9215
	10-kat	0.9237	0.9074	0.7382	0.9128
Duyarlılık	*2-kat*	0.9000	0.8967	0.8350	0.8957
	5-kat	0.9062	0.8892	0.8627	0.8875
	10-kat	0.8854	0.8773	0.8647	0.8665
F-Ölçütü	*2-kat*	0.9040	0.8998	0.7353	0.8997
	5-kat	0.9109	0.8924	0.7757	0.8925
	10-kat	0.8895	0.8806	0.7522	0.8710
Kappa Değeri	*2-kat*	0.8703	0.8626	0.5948	0.8623
	5-kat	0.8782	0.8513	0.6225	0.8509
	10-kat	0.8502	0.8368	0.6165	0.8227

k-kat çapraz geçerleme performans değerlendirme yöntemine göre model başarısı ölçüldüğünde; doğruluk değeri en yüksek olan algoritma C54.5 karar ağacı algoritmasıdır. Sadece doğruluk model başarı ölçütünü değerlendirmek yerine diğer başarı ölçütlerine de bakılmalıdır. Hold out performans değerlendirme yönteminde de olduğu gibi F-Ölçütünü baz almak daha doğru sonuç verecektir. 0.9109 F-Ölçütü değeriyle 5-kat çapraz geçerleme C4.5 karar ağacı algoritması en yüksek değeri vermektedir.

Veri setinde yer alan her bir özellik bağımlı değişken hakkında tahminleyici bilgi taşımayabilir. Özellik seçimi, tüm özellik kümesi sütunlarından bağımlı değişkenle olan ilişkinin açıklanması, ilgisiz sütunların elenmesi ve açıklayıcı gücü yüksek sütun alt kümelerinin belirlenmesi işlemidir (Aktaş ve Kalıpsız 2015). Özellik seçiminde çeşitli teknikler kullanılmaktadır. Bunlardan biri de Bilgi Kazancı yöntemidir. Bilgi kazancına dayanan özellik seçme algoritması veri kümesinde bulunan ilgisiz, gereksiz, fazla veya bilgi kazancı düşük olan özellikleri atmayı amaçlamaktadır (Cihan 2018).

6 özellikten oluşan veri setine bilgi kazancına dayanan özellik seçim algoritması uygulanmıştır. Özellik seçimi için R'da "information.gain" metodu kullanılmıştır. Uygulama sonucunda özelliklerin önem dereceleri aşağıdaki şekilde gösterilmiştir.



Şekil 5.18: Bilgi Kazancı Yöntemine Göre Özelliklerin Önem Derecesi

En önemli özelliğin Penetrasyon olduğu görülmektedir. Penetrasyon özelliğiyle beraber Resistans özelliği de nanofiber kaplı filtre numunelerinin kalite standartlarına göre sınıflandırılmasında önemli olduğu belirlenmiştir.

5.3 Kümeleme Yöntemleri ve Model Değerlendirme

Kümeleme yöntemlerinden k-means ve bulanık c-ortalama algoritmaları ile veri seti analiz edilmiştir. Kümeleme sonucunda elde edilen küme etiketleri ile veri setine ait F7, F8, F9 etiketleri karşılaştırılarak kümeleme başarısı entropi ve saflık kriterlerine göre ölçülmüştür.

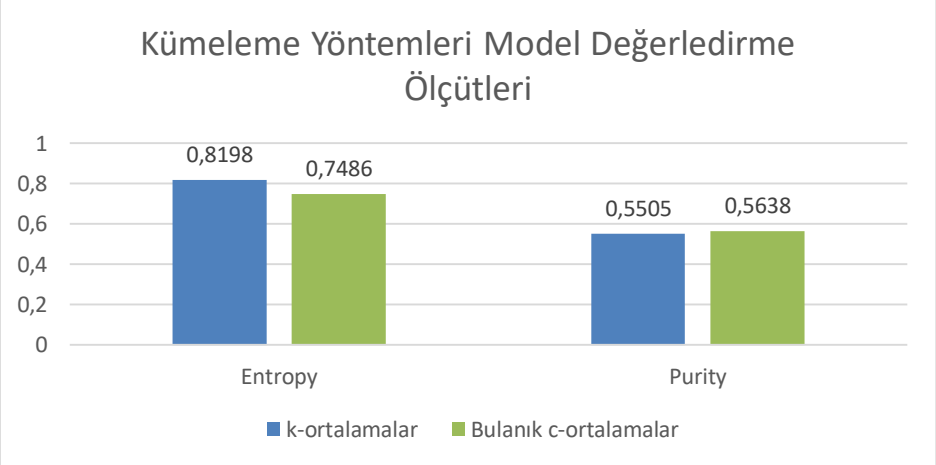
k-means algoritmasında "cluster" ve "clusterCrit" paketleriyle "kmeans" fonksiyonu kullanılmıştır. Bulanık c-ortalamalar algoritmasında ise "e1071" kütüphanesi ve "cmeans" fonksiyonu kullanılmıştır.

Kümelerin kalitesini belirlemede en sık kullanılan ölçütler entropi ve saflık ölçütleridir. Entropi düzensizliğin ölçütü olup değeri büyüdükçe kümelerin düzensiz

olduğunu gösterirken saflık ise entropi ile ters orantılıdır. Bir kümenin düzensizliği arttıkça saflığı azalmaktadır (Cihan 2018). Analiz sonuçlarında saflık değerinin yüksek entropi değerinin ise düşük olduğu metot daha başarılı olduğu kabul edilir. Bu durum baz alındığında saflık değerinin en yüksek olduğu bulanık c-ortalama algoritması daha başarılı olarak kabul edilmektedir.

Tablo 5.9: Kümeleme Yöntemleri Model Başarı Ölçütleri

	Entropi	Saflık
k-means	0.8198	0.5505
Bulanık c-ortalama	0.7486	0.5638



Şekil 5.19: Kümeleme Yöntemleri Model Başarı Ölçütleri Grafiği

6. SONUÇ VE ÖNERİLER

Bu tez çalışması kapsamında veri madenciliği, veri madenciliği süreci, veri madenciliği ve yapay zekâ teknikleri anlatılmıştır. Uygulama alanına konu olan nanoteknoloji ve nanofiber kavramları açıklanmıştır. Nanofiber kaplı filtre numunelerinin kalite standartları ölçümünden elde edilen laboratuvar sonuçları veri seti olarak kullanılmıştır. Literatürde ağırlıklı olarak nanolif çapının tahmini üzerinde durulmuştur. Yapılan literatür çalışmasında veri madenciliği ve yapay zekâ tekniklerinden yapay sinir ağları, regresyon analizi, bulanık kümeleme yöntemlerinin kullanıldığı görülmüştür. Bu çalışmada farklı olarak nanofiber kaplı filtre malzemelerinden oluşan veri seti hem sınıflandırma hem de kümeleme algoritmaları kullanılarak model başarım ölçütleri değerlendirilmiştir.

Nanofiber kaplı filtre malzemelerinin kalite standartlarına göre sınıflandırıldığı veri seti ön işleme aşamalarından geçirilmiştir. Verinin normalizasyonu aşamasında minimum-maksimum, ondalık ölçeklendirme, z-değeri ve sigmoid normalizasyon yöntemleri kullanılmıştır. Normalize edilmiş veri setine k-en yakın komşu algoritması uygulanarak doğruluk değeri en yüksek yöntemin sigmoid normalizasyon yöntemi olduğu tespit edilmiştir. Veri ön işleme adımlarından sonra veri seti hold-out ve k-kat çapraz geçerleme performans değerlendirme yöntemlerine göre modeller oluşturulmuştur. Bu modellere C4.5, Rastgele Orman, Naive Bayes ve Yapay Sinir Ağları sınıflandırma yöntemleri uygulanmıştır. Sınıflandırma algoritmalarının uygulanmasından sonra model başarım ölçütleri ile algoritmalar karşılaştırılmıştır.

Hold-out yöntemine göre oluşturulan modeller: Model 1: %70 eğitim - %30 test, Model 2: %75 eğitim - %25 test, Model 3: %80 eğitim - %20 test

k-kat çapraz geçerleme yöntemine göre oluşturulan modeller: Model 1: 2-kat çapraz geçerleme, Model 2: 5-kat çapraz geçerleme, Model 3: 10-kat çapraz geçerleme

Hold-out performans değerlendirme yöntemine göre sınıflandırma algoritmaları karşılaştırıldığında, Rastgele Orman yönteminin, C4.5 karar ağacı, Naive Bayes ve Yapay Sinir Ağlarına göre genel anlamda daha iyi sonuçlar verdiği gözlenmektedir. Rastgele Orman algoritmasıyla C4.5 Karar Ağacı algoritmasının yakın sonuçlar verdiği tespit edilmiştir. Toplamda başarılı sınıflandırma yüzdeleri açısından sıralama yapıldığında Rastgele Orman, C4.5, Yapay Sinir Ağları ve Naive Bayes şeklinde sıralama yapabiliriz. Modeller arasında kıyaslama yapıldığında Model 3'ün başarılı sınıflandırma yüzdesiyle en iyi tahmin modeli olduğu ortaya çıkmaktadır. k-kat çapraz geçerliliğe göre sınıflandırma algoritmaları karşılaştırıldığın 10-kat çaprazlama yöntemiyle C4.5 Karar Ağacı algoritmasının en iyi sonuca sahip olduğu görülmektedir.

Kümeleme algoritmalarında k-means ve bulanık c-ortalama karşılaştırıldığında k-means algoritmasının daha düşük saflık değerine ve yüksek entropi değerine sahip olduğu gözlenmiştir. Bulanık c-ortalama algoritmasında, kümelemenin daha fazla olduğunu gösterir. Bulanık c-ortalama kümelemenin, k-means algoritmasına kıyasla nanofiber veri kümesi için daha uygun olduğu anlamına gelir.

Çalışma kapsamında izlenen yöntem örnek alınarak farklı endüstri alanlarında benzer veri madenciliği uygulamaları yapılabilir. Farklı alanlarda uygulanmasıyla veri setinin model başarım ölçütüne olan etkisi analiz edilebilir. Bunun yanı sıra çalışma kapsamındaki tüm analizler R programlama dili ile gerçekleştirilmiştir. R programlama dilinden farklı veri madenciliği araçları ile çalışılarak algoritma ve zaman karmaşıklığı açısından performans karşılaştırması yapılabilir.

İşletmede üretilen nanofiber kaplı filtre numunelerinin fiziksel özelliklerinin, kalite standartlarına göre ölçüm değerlerinin bulunduğu teknik bilgi dosyaları hazırlanmaktadır. Hazırlanan bu teknik bilgi dosyaları müşterilere sunulur. Müşterilerden gelen özel isteklere göre ölçümler yapılarak teknik raporlar da hazırlanabilmektedir. Bu aşamada ölçüm yapmadan tez kapsamındaki modelden faydalanılarak filtre sınıfı belirlenebilir. Filtre sınıfının veri madenciliği yöntemiyle tahmin edilmesi işletme açısından zaman, kaynak, hammadde ve ekipman kullanımı konusunda fayda sağlamaktadır.

7. KAYNAKLAR

Abuomar, O., Nouranian, S., King, R., and Lacy Jr, T. E., "Application of Materials Informatics to Vapor-Grown Carbon Nanofiber/Vinyl Ester Nanocomposites Through Self-Organizing Maps and Clustering Techniques", *Computational Materials Science*, 158, 98-109, (2019).

Abuomar, O., Nouranian, S., King, R., Bouvard, J. L., Toghiani, H., Lacy, T. E., and Pittman Jr, C. U., "Data Mining and Knowledge Discovery in Materials Science and Engineering: A Polymer Nanocomposites Case Study.", *Advanced Engineering Informatics*, 27(4), 615-624, (2013).

Act, S. M., "Communıcatıon From The Commıssıon to The European Parlıament, The Councıl, The Economıc and Socıal Commıttee and The Commıttee of The Regions", (2012).

Aktaş, M. S., ve Kalıpsız, O., "Veri Madenciliğinde Özellik Seçim Tekniklerinin Bankacılık Verisine Uygulanması Üzerine Araştırma ve Karşılaştırmalı Uygulama" *9. Ulusal Yazılım Mühendisliği Sempozyumu (UYMS)*, 09-11, (2015).

Alan, G., ve Tercan, M., "Hava Filtrasyonu Amacıyla Kullanılan Tekstillerin Verimlilikleri ve Toz Tutma Kapasiteleri", *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi*, 19(4), 179-186, (2013).

Alpaslan, F., ve diğ., "Bulanık Kümelemede En Uygun Küme Sayısının Yapay Sinir Ağları ve Diskriminant Analizi ile Belirlenmesi" *Atatürk Üniversitesi İktisadi ve İdari Bilimler Dergisi*, 25, (2011).

Alpaydın, E., "Zeki Veri Madenciliği: Ham Veriden Altın Bilgiye Ulaşma Yöntemleri*", Bilişim 2000 Eğitim Semineri*, (2000).

Altunkaynak B., *Veri Madenciliği Yöntemleri ve R Uygulamaları Kavramlar - Modeller - Algoritmalar,* Seçkin Yayıncılık, (2017).

Amani, A., and Mohammadyani, D., "Artificial Neural Networks: Applications in Nanotechnology", *In Artificial Neural Networks-Application.* InTech, (2011).

Andrews R.J. "Nanotechnology: Managing Molecules for Modern Medicine", *In The Modern Hospital* (Latifi R. eds), Springer, Cham, (2019)

Asilkan, Ö., "Veri Madenciliği Kullanılarak İkinci El Otomobil Pazarında Fiyat Tahmini", Doktora Tezi, *Akdeniz Üniversitesi Sosyal Bilimler Enstitüsü*, Antalya, (2008).

Aydoğan, E. K., Gencer, C., ve Akbulut, S., "Churn Analysis and Customer Segmentation of A Cosmetics Brand Using Data Mining Techniques", *Journal of Engineering and Natural Sciences*, 26(1), (2008).

Balaban, M.E., Kartal E., *Veri madenciliği ve Makine Öğrenmesi Temel Algoritmaları ve R Dili Uygulamaları*, İstanbul: Çağlayan Kitapevi, (2015).

Baskaran, A., "UNESCO Science Report: Towards 2030", *Institutions and Economies*, 125-127, (2017).

Baykasoğlu, A., "Veri Madenciliği ve Çimento Sektöründe Bir Uygulama", *Akademik Bilişim 2005*, (2005).

Beasley, D., Bull, D. R., and Martin, R. R., "An Overview of Genetic Algorithms: Part 1, Fundamentals", *University Computing*, 15(2), 56-69, (1993).

Berkhin P., "A Survey of Clustering Data Mining Techniques", *In Grouping Multidimensional Data*, Berlin : Springer, (2006).

Berry, M. J., and Linoff, G., *Data Mining Techniques: For Marketing, Sales, and Customer Support*, John Wiley & Sons, Inc, (1997).

Bezdek, J. C., Ehrlich, R., and Full, W., "FCM: The Fuzzy C-Means Clustering Algorithm", *Computers & Geosciences*, 10(2-3), 191-203, (1984).

Bhushan, B.(ed), *Springer Handbook of Nanotechnology*, Springer, (2017).

Breiman, L., "Random Forests", *Machine Learning*, 45(1), 5-32, (2001).

Burma, Z.A., *Veri Tabanı Yönetim Sistemleri ve SQL/PL-SQL/T-SQL*, Ankara : Seçkin, (2005).

Chen, M. S., Han, J., and Yu, P. S., "Data Mining: An Overview From A Database Perspective", *IEEE Transactions on Knowledge and Data Engineering*, 8(6), 866-883, (1996).

Cihan, P., "Veri Madenciliği Yöntemleriyle Hayvan Hastalıklarında Teşhis, Prognoz ve Risk Faktörlerinin Belirlenmesi", Doktora Tezi, *Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü*, Bilgisayar Mühendisliği Anabilim Dalı, İstanbul, (2018).

Cihan, P., Kalıpsız, O., ve Gökçe, E., "Hayvan Hastalığı Teşhisinde Normalizasyon Tekniklerinin Yapay Sinir Ağı ve Özellik Seçim Performansına Etkisi", *Electronic Turkish Studies*, 12(11), (2017).

Coşkun, C., Baykal, A., "Veri Madenciliğinde Sınıflandırma Algoritmalarının Bir Örnek Üzerinde Karşılaştırılması", *Akademik Bilişim*, Malatya, 51-58, (2011).

Çakır, F., "Veri Madenciliği Yöntemleri Kullanılarak Küçük ve Orta Ölçekli İşletmelere İlişkin Bazı Verilerin Çözümlenmesi", Yüksek Lisans Tezi, *Hacettepe Üniversitesi Fen Bilimleri Enstitüsü*, İstatistik Anabilim Dalı, Ankara, (2012).

Çakmak S., "Elektroeğrilmiş Nanofiberlerin Uygulama Alanları", *Nanobülten Aylık Nanoteknoloji ve Nanotıp Bilim Dergisi*, (14), 12-20, (2011).

Eren, B., ve Ileri, R. "Yapay Sinir Ağlarını Kullanarak Nanofiltrasyon Membranları ile Tuz Giderim Veriminin Belirlenmesi", *Electronic Letters on Science&Engineering*, 3(2), 39-47, (2007).

Dikmen, H., Dikmen, H., Elbir, A., Ekşi, Z., ve Çelik, F., "Gezgin Satıcı Probleminin Karınca Kolonisi ve Genetik Algoritmalarla Eniyilemesi ve Karşılaştırılması", *Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, 18(1), 8-13, (2014).

Dinçer, E., "Veri Madeciliğinde K-Means Algoritması ve Tıp Alanında Uygulanması", Yüksek Lisans Tezi, *Kocaeli Üniversitesi Fen Bilimleri Enstitüsü*, Kocaeli, (2006).

Dinçer, K., Önal G., Selbes, M., ve Akdemir, A., "Nanofiber Tabakalı Hava Filtrelerinin Partikül Yakalama Performanslarının İncelenmesi", *Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, DOI-10, (2018).

Di Martino, F., and Sessa, S., "Implementation of the Extended Fuzzy C-Means Algorithm in Geographic Information Systems", *Journal of Uncertain Systems*, 3(4), 298-306, (2009).

Dunham, M. H., *Data Mining: Introductory and Advanced Topics*, India: Pearson Education, (2006).

Durmuşoğlu, A., " Veri Madenciliği Çalışmaları Üzerine Bir Analiz: Türkiye Adresli Yayınlar", *Elektronik Sosyal Bilimler Dergisi*, 16(62), 1111-1122, (2017).

Edelstein H.A., *Introduction to Data Mining and Knowledge Discovery*, Two Crows Corporation, (1999).

Erken Ş., "Veri Madenciliği Yöntemleri ve Optimizasyona Dayalı Modeller Üzerine Bir Araştırma ve Bir Uygulaması", Yüksek Lisans Tezi, *Dokuz Eylül Üniversitesi Sosyal Bilimler Enstitüsü*, İzmir, (2017).

Fang, J., Niu, H., Lin, T., and Wang, X., "Applications of Electrospun Nanofibers", *Chinese Science Bulletin*, 53(15), 2265-2286, (2008).

Feynman, R.P., “There's plenty of room at the bottom”, *Journal of Microelectromechanical Systems* , 1(1), 60-66, (1992).

Grafe, T. H., and Graham, K. M., “Nanofibers Webs From Electrospinning”, *Nonwovens in Filtration.-In Fifth International Conference*, (2003).

Graham, K., Ouyang, M., Raether, T., Grafe, T., McDonald, B., and Knauf, P., “Polymeric Nanofibers in Air Filtration Applications”, *In Fifteenth Annual Technical Conference & Expo of the American Filtration & Separations Society*, 9-12, (2002).

Gürmen, S., Ebin, B., “Nanopartiküller ve Üretim Yöntemleri-1”, Metalurji, 150, 31-38, (2008).

Gürsoy, U. T. Ş., *Uygulamalı Veri Madenciliği: Sektörel Analizler* , Pegem Akademi, (2011).

Haj-Ali, R., Kim, H. K., Koh, S. W., Saxena, A., and Tummala, R., “Nonlinear Constitutive Models From Nanoindentation Tests Using Artificial Neural Networks”, *International Journal of Plasticity*, 24(3), 371-396, (2008).

Han, J., Pei, J., and Kamber, M., *Data Mining: Concepts and Techniques*, Elsevier, (2012).

Harrington, P., *Machine Learning in Action*, Shelter Island, NY: Manning Publications Co, (2012).

Hatipoğlu, B., Aslan, Z., Zontul, M., ve Güneş, A., ”Dershane Eğitiminin Öğrencinin Üniversiteye Yerleşmesindeki Etkisi”, *İstanbul Aydın Üniversitesi Dergisi*, 12, 13-50, (2011).

Hutten, I. M., *Handbook of Nonwoven Filter Media*, Elsevier, (2007).

Inmon, W. H., *Building the Data Warehouse*, John Wiley & Sons, (2005).

Jones, D. E., Ghandehari, H., and Facelli, J. C., “A Review of the Applications of Data Mining and Machine Learning for the Prediction of Biomedical Properties of Nanoparticles”, *Computer Methods and Programs in Biomedicine*, 132, 93-103,(2016).

Joshi, K. P., “Analysis of Data Mining Algorithms”, University of Minnesota, (1997).

Kantardzic, M., *Data Mining: Concepts, Models, Methods, and Algorithms*, Canada: John Wiley & Sons, (2011).

Kaounides, L., Yu, H., and Harper, T., "Nanotechnology Innovation and Applications in Textiles Industry: Current Markets and Future Growth Trends", *Materials Technology*, 22(4), 209-237, (2007).

Karaibrahimoğlu, A., "Veri Madenciliğinden Birliktelik Kuralı ile Onkoloji Verilerinin Analiz Edilmesi: Meram Tıp Fakültesi Onkoloji Örneği", Doktora Tezi, *Selçuk Üniversitesi Fen Bilimleri Enstitüsü*, Konya, (2014).

Kartal, E., "Sınıflandırmaya Dayalı Makine Öğrenmesi Teknikleri ve Kardiyolojik Risk Değerlendirmesine İlişkin Bir Uygulama", Doktora Tezi, *İstanbul Üniversitesi Fen Bilimleri Enstitüsü*, İstanbul, (2015).

Kaya, Y., Tekin, R., *Veri Tabanı ve Uygulamaları*, İstanbul : Papatya, (2007).

Khatti, T., Naderi-Manesh, H., ve Kalantar, S. M., "Application of ANN and RSM Techniques For Modeling Electrospinning Process of Polycaprolacton", *Neural Computing and Applications*, 1-10, (2017).

Khatti, T., Naderi-Manesh, H., and Kalantar, S. M., "Prediction of Diameter in Blended Nanofibers of Polycaprolactone-Gelatin Using ANN and RSM." ,*Fibers and Polymer*s, 18(12), 2368-2378, (2017).

Kıyak, E., "CRISP-DM Yöntembilimi Kullanılarak Deniz Kuvvetleri Verisi Üzerinde Veri Madenciliği Sınıflandırma Tekniklerinin Karşılaştırılması", Yüksek Lisans Tezi, *Kocaeli Üniversitesi Fen Bilimleri Enstitüsü*, Kocaeli, (2006).

Kozanoğlu, G. S., "Elektrospinning Yöntemiyle Nanolif Üretim Teknolojisi", Yüksek Lisans Tezi, *İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü*, İstanbul, (2006).

Körözlü, Y. D. D. N., "Bilim ve Teknolojinin Geleceği Nanoteknoloji", *Ayrıntı Dergisi*, 4(39), (2016).

Kumar, S.V.K., and Kiruthika, P., "An Ovrview of Classification Algorithm in Data Mining", IJARCCE, vol.4, (2015).

Landis, J. R., and Koch, G. G., "The Measurement of Observer Agreement for Categorical Data", *Biometrics*, 159-174, (1977).

Lee, D. S., Jung, H. Y., Lim, J. W., Lee, M., Ban, S. W., Huh, J. S., and Lee, D. D., "Explosive Gas Recognition System Using Thick Film Sensor Array and Neural Network", *Sensors and Actuators B: Chemical*, 71(1-2), 90-98, (2000).

Liau, L. C. K., ve Dai, W. W., "Process Optimization of Preparing Spherical Titania Colloids with Uniform Distribution Using Artificial Neural Networks", *Colloids and Surfaces A: Physicochemical and Engineering Aspects*, 320(1-3), 68-73, (2008).

Liaw, A., and Wiener, M., "Classification and Regression by RndomForest", *R news*, 2(3), 18-22, (2002).

Mainmon, O., and Rokach, L., *Data Mining and Knowledge Discovery Handbook*, London: Springer US, (2010).

Martín, L., Baena, L., Garach, L., López, G., and de Oña, J., "Using Data Mining Techniques to Road Safety Improvement in Spanish Road", *Procedia-Social and Behavioral Sciences*, 160, 607-614, (2014).

Maulik, U., and Sanghamitra B., "Genetic Algorithm-Based Clustering Technique" ,Pattern Recognition, 33(9), 1455-1465, (2000).

Melanie, M., *An Introduction to Genetic Algorithms*, London: MIT Press, (1999).

Miller, J. C., Serrato, R., Represas-Cardenas, J. M., and Kundahl, G., *The handbook of nanotechnology: Business, policy, and intellectual property law*. John Wiley & Sons, (2004).

Naghibzadeh, M., and Adabi, M., "Evaluation of Effective Electrospinning Parameters Controlling Gelatin Nanofibers Diameter via Modelling Artificial Neural Networks", Fibers and Polymers, 15(4), 767-777, (2014).

Naschie, M.S.E., "Nanotechnology for the Developing World", Chaos Solitons&Fractals, 30(4), 769-773, (2006).

Nasouri, K., Shoushtari, A. M., and Khamforoush, M., "Comparison Between Artificial Neural Network and Response Surface Methodology in The Prediction of The Production Rate of Polyacrylonitrile Electrospun Nanofibers", *Fibers and Polymers*, 14(11), 1849-1856, (2013).

Özdoğan, E., Demir, A., ve Seventekin, N., "Nanoteknoloji Ve Tekstil Uygulamaları", *Tekstil ve Konfeksiyon*, 16(3), 159-168, (2006).

Özer, Y., "Nanobilim ve Nanoteknoloji: Ülke Güvenliği/Etkinliği Açısından Doğru Modelin Belirlenmesi.", Yüksek Lisans Tezi, *Kara Harp Okulu Savunma Bilimleri Enstitüsü* , Ankara, (2008).

Özgür, S., "Bulanık C-Ortalamalar Kümeleme Analizi ve Sağlık Alanında Uygulaması", Yüksek Lisans Tezi, *Ege Üniversitesi Sağlık Bilimleri Enstitüsü*, İzmir, (2017).

Özkan, Y., *Veri Madenciliği Yöntemleri*, İstanbul : Papatya Yayıncılık Eğitim, (2016).

Ramakrishna, S., Fujihara, K., Teo,W., Lim,T. and Ma, Z., (Eds.), *An Introduction to Electrospinning and Nanofibers*, Singapore: World Scientific Publishing Co. Pte. Ltd., (2005).

Sadan, M. K., Ahn, H. J., Chauhan, G. S., and Reddy, N. S., "Quantitative Estimation of Poly (Methyl Methacrylate) Nano-fiber Membrane Diameter by Artificial Neural Networks", *European Polymer Journal*, 74, 91-100, (2016).

Salehi, F., and Razavi, S. M., "Modeling of Waste Brine Nanofiltration Process Using Artificial Neural Network and Adaptive Neuro-Fuzzy Inference System", *Desalination and Water Treatment*, 57(31), 14369-14378, (2016).

Sarkar, K., Ghalia, M. B., Wu, Z., and Bose, S. C., "A Neural Network Model for the Numerical Prediction of the Diameter of Electrospun Polyethylene Oxide Nanofibers", *Journal of Materials Processing Technology*, 209(7), 3156-3165, (2009).

Schaefer, H. E., *Nanoscience: The Science of the Small in Physics, Engineering, Chemistry, Biology and Medicine*, Springer, (2011).

Schmid, G. (Ed.), *Nanotechnology: Volume 1: Principles and Fundamentals*, Wiley-VCH, (2008).

Sevinç, E., "Nanoteknoloji İnovasyon Sistemi: Türkiye Tekstil Sektörü Örneği", Doktora Tezi, *Marmara Üniversitesi Sosyal Bilimler Enstitüsü*, İstanbul, (2017).

Shearer, C.. "The CRISP-DM Model: The New Blueprint For Data Mining." *Journal of Data Warehousing*, 5(4), 13-22, (2000).

Shehadeh, M., Ebrahimi, N., and Ochigbo, A., "Predicting the Type of Nanostructure Using Data Mining Techniques and Multinomial Logistic Regression", *Procedia Computer Science*, 12, 392-397, (2012).

Shi, J. Y., and Li, L. L., "The Research of Data Mining in Telecom Data Warehouse", *2010 International Conference on In System Science, Engineering Design and Manufacturing Informatization (ICSEM) IEEE*, 2, 239-242, (2010).

Shimodaira, H., Clustering and Visualisation of Data [online], (12 Haziran 2018), http://www.inf.ed.ac.uk/teaching/courses/inf2b/learnnotes/inf2b-learn-note03-2up.pdf, (2015).

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

Silahtaroğlu, G., "Veri Madenciliğinde Kümeleme Analizi ve Öğretim Başarısının Değerlendirilmesine İlişkin Bir Uygulama", Doktora Tezi, *İstanbul Üniversitesi Sosyal Bilimler Enstitüsü*, İşletme Anabilim Dalı, İstanbul, (2004).

Sözen, E., Bardak, T., Aydem,r, D., ve Bardak, S., "Yapay Sinir Ağları ve Derin Öğrenme Algoritmaları Kullanarak Nanokompozitlerde Deformasyonun Tahmin Edilmesi", *Journal of Bartin Faculty of Forestry,* 20(2), 223-231, (2018).

Sripada, S. C., and Rao, M. S., "Comparison of Purity and Entropy of k-Means Clustering and Fuzzy c-Means Clustering", *Indian Journal of Computer Science and Engineering*, 2(3), 343-6, (2011).

Süpüren, G., Kanat, Z. E., Çay, A., Kırcı, T., Gülümser, T., ve Tarakçıoğlu, I., "Nano Lifler (Bölüm 2)", *Tekstil ve Konfeksiyon,* 17(2), 83-89, (2007).

Şeker, Ş. E., *İş Zekası ve Veri Madenciliği*, İstanbul : Cinius Yayınları, (2013).

Terzi, Ö., Küçüksille, E.U., Ergin, G. ve İlker, A., "Veri Madenciliği Süreci Kullanılarak Güneş Işınımı Tahmini", *SDU International Technologic Science,* 3(2), 29-37, (2011).

Ullman,J.D.,"Clustering[online]",(21Haziran2018),http://infolab.stanford.edu/~ullman/mining/cluster1.pdf

Üstün,A., "Hava Filtrasyonu için Nanolif Üretimi", Pamukkale Üniversitesi Fen Bilimleri Enstitüsü Tekstil Mühendisliği Anabilim Dalı Yüksek Lisans Tezi, (2011).

Yayık, A., "Yapay Sinir Ağları ile Kriptoloji Uygulamaları", Yüksek Lisans Tezi, *Mustafa Kemal Üniversitesi Fen Bilimleri Enstitüsü*, Enformatik Anabilim Dalı, Hatay, (2013).

Yildirim, P., Birant, D., ve Alpyildiz, T., "Data Mining and Machine Learning in Textile Industry", *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(1), (2018).

Yıldırım, P., ve Birant, D., "Bulut Bilişimde Veri Madenciliği Tekniklerinin Uygulanması: Bir Literatür Taraması", *Pamukkale University Journal of Engineering Sciences,* 24(2), 336-343, (2018).

Qu, X. Alvarez, P.J.J., Li, Q., "Applications of Nanotechnology in water and Wastewater Treatment", Water Research, 47, 3931-3946, (2013)

Quinlan, J. R., *Induction of Decision Trees*, Machine Learning, 1(1), 81-106.

Wang, S. C., *Interdisciplinary Computing in Java Programming*, Canada: Springer Science & Business Media,(2003).

Web-3 The Random Forest Algortihm, https://devhunteryz.wordpress.com/2018/09/20/rastgele-ormanrandom-forest-algoritmasi/, (Erişim Tarihi: 20.12.2018)

Web-4 EN779:2012 Sınıflandırma Standartları, http://www.ulpatek.com/tr/filtre-teknolojisi/en-779/, (Erişim Tarihi: 20.12.2018)

Web-5 Filter Tester Model 8130A, http://www.atselektronik.com.tr/media/1200/filter-tester 8130a_a4_5001771_web.pdf, (Erişim Tarihi: 18.12.2018)

Web-1 What is Nanotechnology?, http://www.nano.gov/nanotech-101/what/definition, (Erişim Tarihi: 15.01.2019)

Web-2 Opportunities and Risks of Nanotechnologies, http://www.oecd.org/science/nanosafety/44108334.pdf, (Erişim Tarihi: 28.03.2019)

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., Zhou, Z. H., "Top 10 Algorithms in Data Mining", Knowledge and Information Systems, 14(1), 1-37,(2008).

8. ÖZGEÇMİŞ

Adı Soyadı	: Aylin SABANCI
Doğum Yeri ve Tarihi	: Denizli 29.10.1992
Lisans Üniversite	: Süleyman Demirel Üniversitesi
Elektronik posta	: aylinsabanc@gmail.com