Avrupa Bilim ve Teknoloji Dergisi
Özel Sayı 29, S. 397-414, Aralık 2021

Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
Special Issue 29, pp. 397-414, December 2021

Research Article

Veri Madenciliği ve Makine Öğrenimi Yaklaşımlarının Karşılaştırılması: Tekstil Sektöründe bir Uygulama

Filiz Ersöz[1],Yasemin Çınar[2]

[1*] Karabük Üniversitesi, Mühendislik Fakültesi, Endüstri Mühendisliği Bölümü, Karabük, Türkiye (ORCID: 0000-0002-4964-8487), fersoz@karabuk.edu.tr
[2] Karabük Üniversitesi, Mühendislik Fakültesi, Endüstri Mühendisliği Bölümü, Karabük, Türkiye, (ORCID: 0000-0000-0000-0000), cinarryasemen@gmail.com

(International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT) 2021 – 21-23 October 2021)

(**DOI:** 10.31590/ejosat.1035124)

ATIF/REFERENCE: Ersöz, F. & Çınar, Y. (2021). Veri Madenciliği ve Makine Öğrenimi Yaklaşımlarının Karşılaştırılması: Tekstil Sektöründe bir Uygulama. *Avrupa Bilim ve Teknoloji Dergisi*, (29), 397-414.

Öz

Her gün gelişmekte ve büyümekte olan teknoloji, modern dünyanın vazgeçilmez bir unsuru olmuştur. Teknolojinin hızla gelişmesiyle bilgisayar kullanımı artan dünyamızda daha fazla veri depolanmaya başlanmıştır. Oluşan bu büyük veriler tek başlarına bir anlam ifade etmemektedir. Ancak veri ve analitik alanda yetkinliklerin artırılması ile belirli örüntülere dayalı çıkarımlardan anlamlılık boyutu kazanırlar. Örüntülerin belirlenebilmesini sağlayan, yapılacak araştırmaya ve veri tipine uygun veri madenciliği ve makine öğrenimi teknikleri bulunmaktadır. Bu teknikleri ile veriler arasındaki kural, kalıp ve ilişkiler bulunur. Veri madenciliği ve makine öğrenimi teknikleri birçok farklı sektörde farklı amaçlarla kullanılabilmektedir. Bu çalışmada veri madenciliği ve makine öğrenimi arasındaki benzerlik ve farklılıklar ortaya konmaya çalışılmış ve bu disiplinlerin; veri bilimi, istatistik ve diğer disiplinler ile ortak ve ayrıştığı noktalar tespit edilmeye çalışılmıştır. Ayrıca çalışmada pantolon üreten bir tekstil firmasının verileri kullanılarak, R Studio, Python ve Knime makine öğrenimi programları yardımıyla, çoklu doğrusal regresyon, yapay sinir ağları ve karar ağaçları teknikleri uygulanmış, tahmini model sonuçlar bulunmuş ve model performansları karşılaştırılmıştır. Çalışmanın sonucunda tahminleme başarısında en iyi algoritmanın yapay sinir ağları ve en iyi makine öğrenimi programının RStudio programı olduğu sonucuna varılmıştır.

Anahtar Kelimeler: Tekstil Sektörü, Makine Öğrenimi, Çoklu Doğrusal Regresyon, Yapay Sinir Ağları, Karar Ağaçları, Python, RStudio, Knime

Comparison of Data Mining and Machine Learning Approaches: An Application in Textile Industry

Abstract

Technology, which is developing and growing every day, has become an indispensable whole of the modern world. With the rapid development of technology, more data has begun to be stored in our world, where the use of computers is increasing. These big data do not mean anything on their own. However, they gain a meaningful dimension from inferences based on certain patterns by increasing their competencies in data and analytics. There are data mining and machine learning techniques suitable for the research and data type to be made, enabling the determination of patterns. With these techniques, there are rules, namely algorithms, between the data. Data mining and machine learning techniques can be used for different purposes in many different sectors. In this study, the similarities and differences between data mining and machine learning have been tried to be revealed and these disciplines; It has been tried to determine the common and divergent points with data science, statistics and other disciplines. In addition, using the data of a textile company producing trousers, multiple linear regression, artificial neural networks and decision trees techniques were applied with the help of R Studio, Python and Knime machine learning programs, and estimated model results were found and model performances were compared. As a result of the study, it was concluded that the best algorithm in predicting success is artificial neural networks and the best machine learning program is RStudio.
Keywords: Textile Sector, Machine learning, Multiple Linear Regression, Neural Networks, Decision Trees, Python, RStudio, Knime

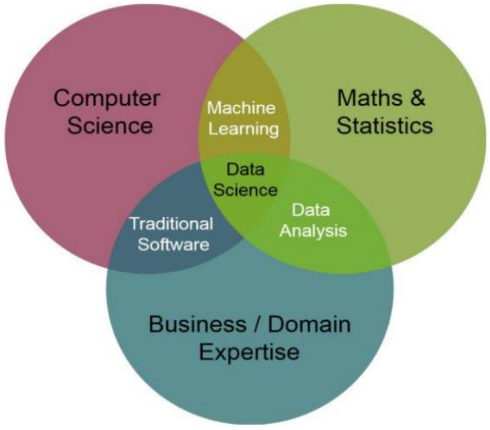
*Sorumlu Yazar: fersoz@karabuk.edu.tr

1. Giriş

Teknoloji, son yıllarda modern dünyanın vazgeçilmez bir parçası haline gelmiştir. Teknolojinin hızlı gelişmesiyle veri toplanması ve verilerin depolanması daha kolay bir hal almıştır. Teknoloji ve bilgisayar sistemlerinin kullanımının bu derece hızlı artmasıyla Dünya' da daha fazla veri depolanmaya başlamıştır. IDC (International Data Corporation), Dünya çapında verilerin yılda % 61'lik orana sahip büyüme hızıyla artabileceğini ve bu yıl 33 ZetaByte' tan 2025 yılına kadar 175 ZetaByte' a çıkacağını öngörmüştür. Günümüzde, 5 milyardan fazla kullanıcı her gün verilerle etkileşim halindedir, 2025 yılına kadar bu rakam 6 milyara veya dünya nüfusunun % 75'ine ulaşacağı ve ağa bağlı durumda olan her insanın 18 saniyede bir en az bir adet etkileşime sahip olacağı öngörülmektedir (Seagate Technology, 2020).

Günümüzde kurumlar, düşük maliyetler ile bilişim ve veri depolama sistemlerine sahip olabilmektedirler. Bu sayede bilgiye kolay erişebilmekte ve internet ağları üzerinden kolay bir şekilde yayabilmektedirler. Büyük miktarda veriyle çalışan çoğu sektör, makine öğrenimi teknolojisini ve veri madenciliğinin değerini kabul etmiştir. Sahip olunan verilerden anlamlı ve değerli ilişkileri ortaya çıkarabilmek ve bazende gerçek zamanlı olarak tahminler yaparak, sektörde daha verimli çalışabilir veya rakiplerine göre avantajlar elde edilebilmektedir. Ayrıca kurumların sahip olduğu büyük veri ve farklı veri çeşitleri, daha ucuz ve daha güçlü olan hesaplamalı işleme sahip veri depolama isteği de makine öğrenimini popüler kılmıştır.

Üretim ve hizmet yönetim işletmelerinde oluşan bu büyük veriler tek başlarına bir anlam ifade etmemektedir. Ancak veriler işlendiği zaman, belirli örüntülere dayalı çıkarımlardan anlamlılık boyutu kazanırlar. Veriye dayalı yöntemlerle (Veri madenciliği, büyük veri ve analitikleri, iş analitikleri vb.) ilgili süreç modelleri bu isimlerle tanımlanmasa da on yıllardır kullanılmaktadır (Mariscal ve diğerleri, 2010 ; Martinez-Plumed ve diğerleri, 2020). Günümüzde bu kavramlar artık daha net anlaşılmaktadır ve çoğunlukla veri bilimi adı altında tanımlanmaktadır. Veri bilimi; istatistik, bilimsel yöntemler, yapay zekâ (AI) ve veri analiği gibi birden çok disiplini birleştirmektedir ve veri bilimini uygulayanlara ise veri bilimcisi denmektedir. Veri bilimi ve ilişkili olduğu disiplinler Şekil 1.1'de verilmiştir. Veri bilimi; makine öğrenimi süreçleri, karmaşık araçlar ve algoritmalar, matematik, istatistik ve diğer benzer alanları kullanarak ham verilerden anlamlı ilişkiler çıkarılmasıyla müşteri davranışları ve eğilimleri vb. ile ilgilenen bir disiplin olarak da tanımlanabilir. Günümüzde bir veri bilimcisi; veri madenciliği, veri analizi, bilgisayar programlama, istatistik, makine öğrenimi, veri görselleştirme ve büyük veri analitiği gibi uzmanlık alanlarına sahip olmalıdır (Kdnuggets, 2020).



Şekil 1.1 Veri bilimi ve ilişkili disiplinler

Büyük veriler arasında ilişkilerin ve örüntülerin belirlenebilmesini sağlayan ve yapılacak araştırmaya, veri tipine uygun veri madenciliği ve makine öğrenimi teknikleri bulunmaktadır. Veri madenciliği veri tabanlarında bilgi keşfi olup (Fayyad, Piatetsky-Shapiro and Smyth 1996), büyük veri yığınlarından anlamlı ilişkileri ve kuralları ortaya çıkaran bir disiplindir ve makine öğrenimi tarafından geliştirilen algoritmalardan yararlanır (Fayyad, Piatetsky-Shapiro and Smyth 1996; Ersöz, 2019; Kulin vd., 2021).

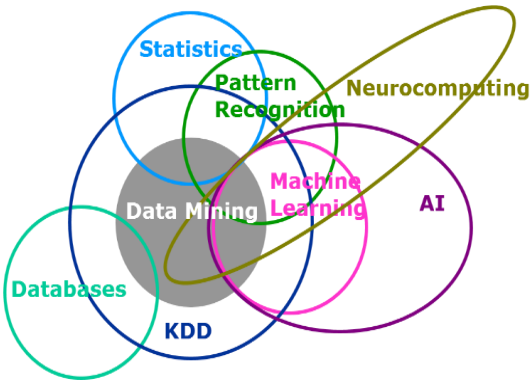
Tahmine dayalı analitik yaklaşımlar, bilinmeyen olaylar hakkında ve geleceğe yönelik tahmin yapmak için veri madenciliği, modelleme ve makine öğreniminde çeşitli istatistiksel teknikler gerekir (Eckerson, 2007; Ncye, 2007). Verilerden öngörüsel ve tahmini bilgiler elde etmek için algoritmalar (kurallar) geliştirmede çoğunlukla; istatistik, matematik ve bilgisayar teknolojileri kullanılmaktadır. İstatistik bilimi, tüm veri madenciliği ve makine öğrenimi algoritmalarının temeli olarak tanımlanabilir. Veri madenciliği ve istatistik biliminin temel amacı veri analizi yapmaktır, ancak bu iki disiplin farklılıklar içermektedir. Veri madenciliği sayısal ve sayısal olmayan büyük verilerden yararlanarak tahmini bir model oluşturmayı ve optimize etmeyi amaçlarken keşfedici bir süreç izler, hipotezlere gerek duymaz ve veri madenciliği için büyük veri gereklidir. Ancak istatistik bilimi büyük sayısal veri dışında genellikle küçük sayısal verilerden de çıkarımlar yaparak, doğrulayıcı bir süreç izler ve hipotezlere ihtiyaç duyar (Ersöz, 2019). İstatistik ve veri madenciliği biliminde ilgilenilen özellikler olarak tanımlanan değişken kavramında, hedef veya sonuç değişkeni "bağımlı değişken" olarak tanımlanırken, makine öğreniminde değişken "özellik" ve hedef veya sonuç değişkeni "etiket" olarak tanımlanır. Bununla birlikte istatistik ve veri madenciliği biliminde verilere yönelik "dönüşüm", makine öğreniminde "özellik oluşturma" olarak adlandırılır (SAS, 2021). Kısaca istatistik bilimi rakamları okur ve olasılıklı modellerle, özellikle verileri kullanan bu modellere ilişkin tanımlama ve çıkarımlar ilgilidir. Veri madenciliği ise bu veriler arasındaki ilişkileri ve kalıpları açıklar. Makine öğrenimi ise modellerle tahminde bulunur ve çoğunlukla makine öğrenimi yöntemi resmi bir olasılık modeli olarak formüle edilebilir. Bu nedenle makine öğrenimi bu anlamda istatistik ve veri madenciliği ile çok benzerdir ve yapay zekâ ile davranış ve nedenleri ortaya koyar.

Makine öğrenimi ve yapay zekâ kavramı birbirlerinin yerine kullanıldığı görülmesine rağmen, bu disiplinlerde farklı anlamlar taşımaktadır. Makine öğrenimi uygulamalarının tümü yapay

zekâyı kapsarken, yapay zekâ uygulamalarının tümü makine öğrenimi değildir.

Makine öğrenimi, istatistik, veri madenciliği ve analitik tahminler iç içe kavramlardır. Makine öğrenimi kavramı bazı çalışmalarda istatistiksel makine öğrenimi olarak adlandırılmaktadır (Patel ve ark., 2008; Sotirios, 2018; Tanzeel ve ark. 2019) ve modern yazılımın geliştirilmesinde bir araçtır. Bununla birlikte makine öğrenimi programlanmamış sonuçları bile ortaya çıkarabilen bir tür yapay zekâ ve insan müdahalesi olmadan sonuçları tahmin etmede daha doğru olmasını sağlayan bir yapay zekâ türü olarak da tanımlanabilir.

Veri madenciliği ve makine öğreniminin, veri tabanlarında bilgi keşfi (Knowledge Discovery in Databases-KDD) olarak ortak bir noktası vardır. KDD terimi, 1989'daki ilk Piatetsky-Shapiro tarafından ortaya çıkarılmış ve yapay zekâ ve makine öğrenimi ile popüler hale gelmiştir (Fayyad, Piatetsky-Shapiro and Smyth 1996). Ayrıca makine öğrenimi var olan verilerden yapay zekâ içeren bilgisayar veya cihazlara ilişkin çıkarımlar yapmaktadır. Veri madenciliği Şekil 1.2'de görüldüğü gibi istatistik, örüntü tanıma, veri tabanlarında bilgi keşfi (KDD), nöral hesaplama (nörobilgisayar), yapay zekâ, veri tabanı ve makine öğrenimi ile ilişki içindedir. Ancak makine öğrenimi; veri madenciliği, yapay zekâ, nörobilgisayar ve örüntü tanıma gibi disiplinlerin birleşimidir (Mitchell Guthrie, 2014).



Şekil 1.2 Makine öğrenimi ve ilişkili disiplinler

Genel olarak her büyüklükteki işletmede makine öğrenimi tekniklerinin kullanımı; işletmenin maliyetlerinin düşürülmesi, müşteri içgörüleri ve istihbaratının oluşturulması ve müşteri deneyimini iyileştirmesine yöneliktir (Algorithmia, 2020). Makine öğrenimi, yapay zekâ ve veri bilimi teknolojileri; sağlık, işletme, endüstri, güvenlik gibi birçok çalışma alanı üzerinde önemli bir etkiye sahiptir ve amacı; hesaplamalı ve istatistiksel yöntemler kullanarak çeşitli veri türlerinden otomatik olarak bilgi çıkarmaktır. Örneğin günlük yaşamda herhangi bir websitesinin bilgisayarınıza gelen çevrimiçi öneri teklifleri, müşteri ilişkileri yönetimi, dolandırıcılık tespiti, uber gibi ulaşım şirketleri tahmini varış süresini hesaplamak için veri madenciliği ve makine öğrenimi teknikleri kullanmaktadır. Devlette özellikle sensör verilerinin analiz edilmesi, müşteri tahmin analizleri ve risk belirlemede finansal hizmetler, giyilebilir cihazlar ve sensörler ile bir hastanın gerçek zamanlı hastalığını belirlenmesi, müşteri satın alma geçmişine bakarak müşterinin beğeneceği ürünlerin önerilmesi, yeni enerji kaynaklarının bulunması, rafineri sensör arızasının tespit edilmesi ve verimli ve uygun maliyetli petrol dağıtımı, verimli ve kârlı toplu taşıma rotalarının tespiti uygulamaları vb. makine öğrenimine örnek verilebilir (SAS, 2021).

2. Makine Öğrenimi ve Diğer Disiplinler ile Gelişimi

Makine öğrenmesi (ML), insanların öğrenme şekillerini taklit etmek için veri ve algoritmaların kullanımını sağlayan ve doğruluğunu kademeli olarak artıran bir yapay zekâ (AI) ve bilgisayar bilimi dalıdır (IBM SPSS, 2021). Günümüzde yapay zekânın iş üretkenliğini % 40'a kadar artırabildiği (Accentura, 2021) ve yapay zekâ girişimlerinin ise son yirmi yılda 14 kat büyüdüğü tahmin edilmektedir (Forbes, 2018). Günümüzde yapay zekâ ve makine öğrenimi, işletmeler tarafından benimsenmiş ve büyük gelişmeler kaydetmiştir. Bununla birlikte yazılım geliştiricileri tarafından yazılım araçlarının sayısı da paralel bir şekilde artmıştır.

Makine öğrenimi yöntemleri birçok farklı sektörde farklı amaçlarla kullanılabilmektedir. Tüm sektörlerde makine öğrenimi algoritmaları, üretimden müşteri memnuniyetine kadar hemen hemen her süreçte kullanılabilmesi mümkündür. Örneğin, üretimdeki hataların nedenleri, günlük üretilen ürünün en çok hangi parametrelere bağlı olduğu, müşteriler için önemli olan kriterler ve istekler gibi birçok problem, şirketlerin sağladıkları veriler kullanılarak işlenip, makine öğrenimi algoritmalarıyla çözülebilir. Öğrenme kelimesi, bilgi edinme süreci olarak tanımlanabilir. İnsanlar doğası gereği öğrenme sürecine doğdukları andan itibaren başlarlar. Bilgisayarlar tıpkı insanlar gibi öğrenir ve bunu algoritmalar aracılığıyla yaparlar (Portugal vd., 2018).

Makine öğrenimi, bilgisayar dili olmadan verilerden öğrenen bir yapay zekâ biçimi olarak tanımlandığı kadar (Malaca vd., 2019), büyük ve karmaşık verileri doğru tahmin edebilme yeteneğine sahip bir algoritma olarak da tanımlandığı çalışmalar vardır. Makine öğrenimi, sıradan görevlerin otomatikleştirilmesi ve akıllı tahminler sunmaya kadar her sektörde kullanılmaktadır. Ancak bunları yaparken bir cihaz gereklidir. Akıllı ev asistanları da bir makine öğrenimi olarak tanımlandığı gibi, makine öğreniminin temel olarak en önemli görevi; tüm sektörlerde üretim ve yönetimin sahip olduğu verilerin çoğunlukla istatistik ve veri madenciliği teknikleri kullanarak anlamlı ve değerli bilgilerin keşfedilmesini sağlamaktır. Tahmin analizleri, görüntü işleme, konuşma tanıma ve konuşulan kelimelerin metne çevrilmesi, tıbbi teşhisler ve özellikle finans ve ticaret sektörlerinde oldukça yaygın kullanıma sahiptir. Sosyal platformada ise facebook, makine öğrenimi algoritmalarını kullanarak davranışsal bilgileri toplayan en iyi platformdur.

Makine öğrenimini ve veri madenciliği tekniklerinin tarihçesine baktığımızda kökeninin, istatistik ve matematik bilimine dayandığı görülmektedir. Olasılık ve istatistik biliminin kökenleri 1650-1700 yıllarda şans oyunlarının matematiksel olarak ele alınmasında ve ölüm verilerinin sistematik olarak incelenmesi ile başlamıştır. 1700'lü yıllarda Bayes teoremi ortaya çıkmış, olasılık ve istatistiğin en önemli konularından regresyon yöntemi Legendre tarafından 1805'te (Legendre, 1805) ve 1809'da Gauss tarafından (Angrist&Pischke, 2008) yayınlanan en küçük kareler yöntemi ile başlamıştır. İkinci Dünya Savaşı ilk modern bilgisayarlar olan "Z3", John Von Neumann tarafından geliştirilen ve Turing tarafından kurulan "ENIAC" ve "Colossus" gibi devasa kod kırma makineleri kullanılmıştır. 1940 ve 1950'li yıllarda bilim adamları yapay bir beyin yaratma olasılığını tartışmaya başlamış ve 1956 yılında yapay zekâ araştırmaları olarak bir disiplin olarak doğmuştur (McCorduck, 2004).

1952 yılında Arthur Samuel tarafından da "Makine öğrenimi" terimi ortaya atılmıştır. IBM'den Arthur Samuel 1950'lerde dama oynamak için bir bilgisayar programı tasarlamıştır. Tasarım, dama tahtasındaki parçaların pozisyonlarını kullanan bir puanlama işlevi içermekte ve her iki tarafın kazanma şansını ölçmeye çalışmıştır. Program bir sonraki hamlesini bir minimaks stratejisi kullanmayı seçmiş ve bu strateji sonunda minimaks algoritması ortaya çıkmıştır. Ayrıca programının daha iyi olmasını sağlayan bir dizi mekanizma tasarlamıştır. Arthur Samuel "Ezberci öğrenme" adını verdiği bu programda daha önce oyunda görülen tüm pozisyonları kaydetmiş ve hatırlatmıştır. Uzman dama oyuncusu Robert Nealey ise 1962 yılında bu dama oyununu IBM 7094 bilgisayarında oynamış ve bilgisayara karşı kaybetmiştir (Dataversity, 2021).

1957 yılında Cornell Havacılık Laboratuvarı'nda çalışan Frank Rosenblatt, Donald Hebb'in beyin hücresi etkileşimi modelini, Arthur Samuel'in makine öğrenimi programı ile birleştirmiş ve algılayıcıyı (Perceptron) yaratmıştır. Algılayıcı başlangıçta bir program değil, görüntü tanıma için bir makine olarak (IBM 704) kurulmuştur. Daha sonra bu yazılım ve algoritmalar diğer makinelere aktarılabilir ve kullanılabilir hale getirilmiştir. Algılayıcılar önemli bir keşif olsa da sinir ağı/makine öğrenimi araştırması, 1990 yıllardan sonra görsel deseni (yüzleri) tanımada başarılı olmuştur.

1960'larda, çok katmanlı yapıların keşfi ve kullanımı, sinir ağı araştırmalarında yeni bir çığır açmıştır. 1967'de, temel örüntü tanımanın temeli olan en yakın komşu algoritması tasarlanarak, araç rotaları haritalamak için kullanılmış ve seyahat eden satış elemanının en verimli rotayı bulma sorununa çözüm bulmuştur. Çoklu katmanların kullanılması ileri beslemeli sinir ağlarını ortaya çıkarmıştır. 1970'lerde geliştirilen ileri ve geri beslemeli sinir ağları, günümüzde derin öğrenme olarak tanımlanan verileri eğitmek için kullanılmaktadır.

1965 yılında Lawrence J. Fogel tarafından evrimsel programlama uygulamaları ve gerçek dünya problemlerini çözmek için özel olarak evrimsel hesaplamayı uygulayan ilk veri bilimi (Decision Science) şirketi olmuştur (KDnuggets, 2021).

1970 yılı öncesinde makine öğrenimi, yapay zekâ için bir eğitim programı olarak kullanılırken, 1970-1980 yılları arasında yapay zekâ araştırmaları algoritmaların dışında mantıksal bilgiye dayalı yaklaşımları kullanınca, makine öğrenimi ve yapay zekâ kavramları ayrışmıştır. 1990'ların başında veri madenciliği Veritabanlarında Bilgi Keşfi (KDD- Knowledge Discovery in Databases) olarak tanımlanmış ve günümüzde "Veri Biliminde Bilgi Keşfi" olarak tanımlanan ve KDD'nin bir alt süreci olarak kabul edilmiştir. Bununla birlikte Fayyad ve arkadaşları tarafından önemsiz ve potansiyel verilerden, yararlı ve anlaşılabilir kalıpların çıkarılması olarak da veri madenciliği ve süreci tanımlanmıştır (Fayyad, Piatetsky-Shapiro and Smyth 1996). İlk veri madenciliği teriminin Lovell'in çalışmasında kullanıldığı görülmüştür (Lovell, 1983). Bir kodlama süreci olarak başlayan veri madenciliği, günümüzde kodlama becerisine sahip ve verileri temizleme, analiz etme ve değerlendirme konusunda "veri madencisi" olarak tanımlanan uzmanları ortaya çıkarmıştır.

1990'lı yıllarda olasılık ve istatistiksel yöntemlere odaklanınca, makine öğrenimi sinir ağlarına odaklanmaya ve ilişkileri anlamlandırmaya başlamıştır. Bununla birlikte dijital verilerin sürekli artması ve endüstri problemlerini pratik sorunları çözme becerisi de artmıştır. İnternetin gelişmesi ile bilgileri paylaşma yeteneği de artınca bu gelişme oldukça hızlı olmuştur. Yeni teknolojilerle birleştirilmiş makine öğrenimi algoritmaları, ölçeklenebilirliği desteklediği ve verimliliği artırdığı ve iş analitiği ile birlikte çalıştığında çeşitli kurumsal karmaşıklıkları çözebildiği görülmüştür. Makine öğrenimi modelleri, sürekli öğrenmeye uyarlanabilir hale gelmiş ve çalıştıkları süre boyunca doğruluk oranları artmaya devam etmiştir.

Bilgisayarların ortaya çıkışı ile veri madenciliği disiplini yoğun bir manuel kodlama süreci olarak başlamış, günlük olarak üretilen veri miktarının artması ve bilgisayar teknolojilerinin gelişimi ile bugün veri madenciliği ve makine öğrenimi araçları kullanılmaya başlamıştır. 2021 yılı en iyi veri madenciliği araçlarının; MonkeyLearn, RapidMiner, Oracle Data mining, IBM SPSS Modeler, Weka, Knime, H2O ve Orange yazılım programları olduğu görülürken (MonkeyLearn, 2021), 2021 yılı uzman görüşlerine göre en popüler makine öğrenimi araçlarının ise; Keras, Knime, Weka, Shogun ve Rapid Miner yazılım programları olduğu görülmüştür.

2.1. Makine Öğrenimi Araçları

Makine öğrenmesi yapmak için kullanılması gereken programlar ve araçlar vardır. Bu araçlar sayesinde makine öğrenmesi algoritmalarını kullanarak verileri analiz edilebilmekte ve tahmini sonuçlar elde edebilmektedir. Makine öğrenimi yapılırken programa veri setini öğretmek için kullanılacak olan büyük veri, eğitim verisi ve test verisi olarak iki parçaya bölünür. Eğitim verisi modelin eğitildiği veri setidir. Test verisi ise eğitim veri setinde oluşturulan modeli incelemektedir (Ryu vd., 2018).

Makine öğrenimine Dünyada yaşayan her insanın katkısı vardır (Gürsakal, 2018). İnternet dünyasında yapılan her hareket birer birer kayıt altına alınmaktadır. ML algoritmalarının görevi yüzlerce insanın oluşturduğu büyük veriyi ayrıştırıp, temizleyerek uygun bir model ile anlamlandırmaktır. Günümüzde makine öğrenmesi için çok fazla araç olduğundan dolayı, en yaygın olarak kullanılan ve bilinen araçlar bu bölümde verilmiştir.

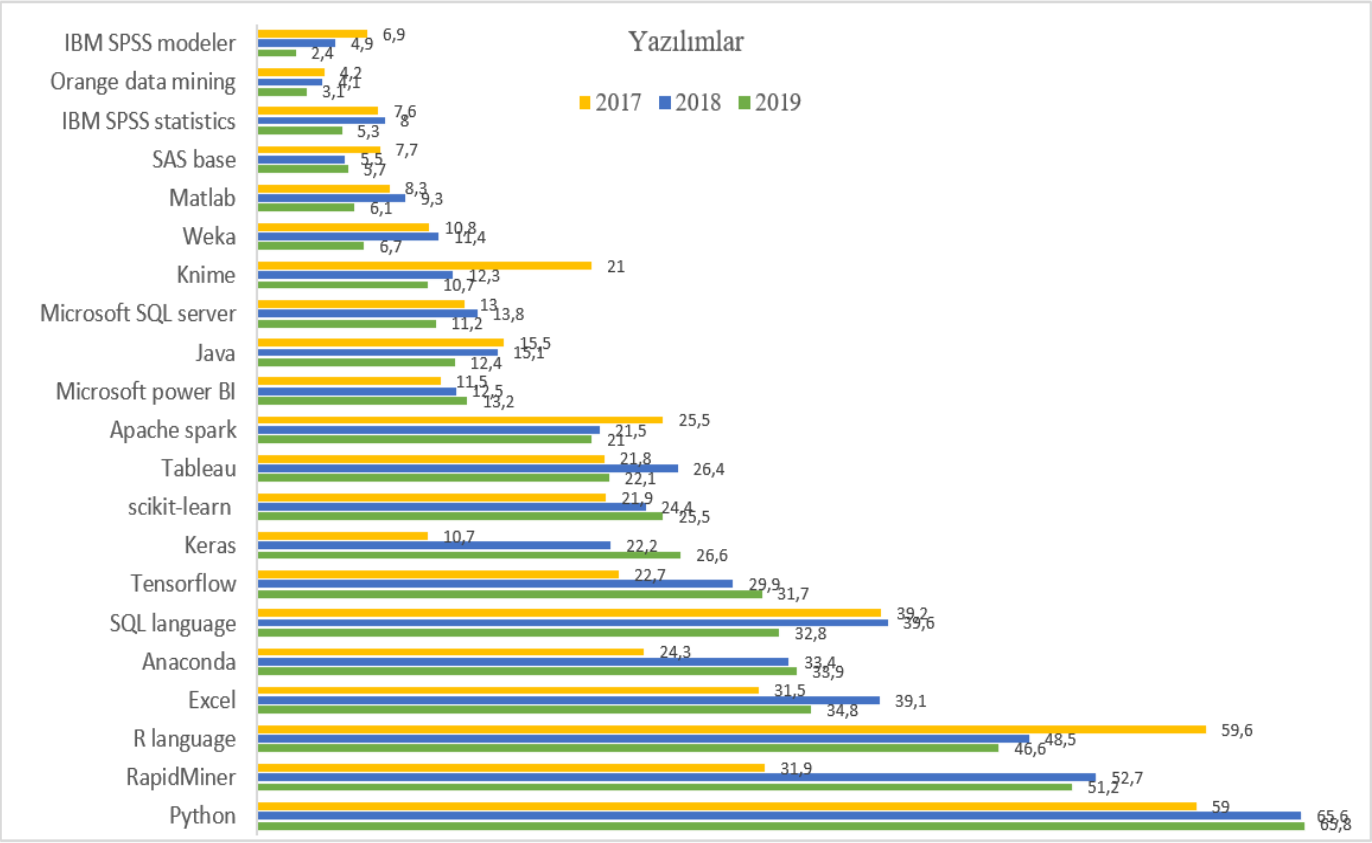
Python: Python yorumlanmış, genel amaçlı, üst düzey bir programdır. Python programının tasarımının temel felsefesi, girinti kullanımı ile kod okunabilirliğidir. Python, dil yapısı ve nesne yönelimli yaklaşımı ile programcıların küçük ve büyük ölçekli projeler için mantıklı, anlaşılır kod yazmasını kolaylaştırmayı amaçlamaktadır. Python programının; öğrenme kolaylığı, daha hızlı geliştirme ve işleme, güçlü paketler, topluluk desteği (herhangi bir sorun yaşadığınızda çözüm bulabileceğiniz bir alan) ve veri görselleştirme gibi birçok avantajı vardır (May, 2019).

R: R programlama dili istatistiksel hesaplamaları ve grafikleri destekleyen ücretsiz bir programlama dilidir. R programı istatistik ve veri madenciliği üzerinde çalışan kişiler arasında istatistiksel yazılım, veri analizini geliştirmek için kullanılır. İçinde bulunan paketler ile R, kullanılan istatistiksel prosedürlerin IBM SPSS gibi programlara göre daha derin bir şekilde anlaşılmasını teşvik etmektedir (Fox & Andersen, 2005). TIOBE (Programlama topluluğu endeksi, programlama dillerinin popülerliğinin bir göstergesidir) tarafından açıklanan popülerlik göstergesinde Mart 2020'de on birinci sırada yer alan R programı, Mart 2021 'de on üçüncü sırada yer almıştır (TIOBE, 2021).

Knime: Knime programı herhangi bir kod yazılımı olmadan kullanılabilen açık kaynak, çapraz platform veri analizi, raporlama ve entegrasyon sağlayan bir araçtır. Görselleştirme, modelleme ve veri analizi için temel veri ön işleme

fonksiyonlarını kullanıcılar kolay bir şekilde kullanabilmektir. Program kullanıcıların görsel olarak veri hatlarını oluşturmalarına olanak sağlar, uygulanmak istenen tüm analizlerin adımlarından sonra modelleri, sonuçları ve etkileşimli görünümlerin incelenmesini sağlamaktadır. Program, iş akışı ve iş analistlerinin alan bilgilerini uygulayarak tahmine dayalı analitik çözümlerini kolayca oluşturmaları için tasarlanmıştır (Birkhold vd., 2019). Knime program; ilaç araştırmalarında, iş zekâsı, müşteri ilişkileri yönetimi (Customer Relationship Management — CRM) ve finansal uygulamalarda kullanılmaktadır (Tiwari & Sekhar, 2007).

Şekil 2.2'de yer alan grafikte 2017, 2018 ve 2019 yıllarında oy veren kullanıcıların son 12 ayda gerçek bir projede kullanılan makine öğrenimi araçları verilmiştir.



Şekil 2.2 En yaygın kullanılan makine öğrenimi araçları (Piatestsky, 2019)

Şekil 2.2'de yer alan grafik incelendiğinde; 2017 yılında R programının kullanımı %59,9 kullanım oranına sahipken, Python programının kullanım oranı %59 'dur. Fakat zaman içinde Python kullanım oranı artığı ve R programı kullanım oranının düştüğü görülmüştür.

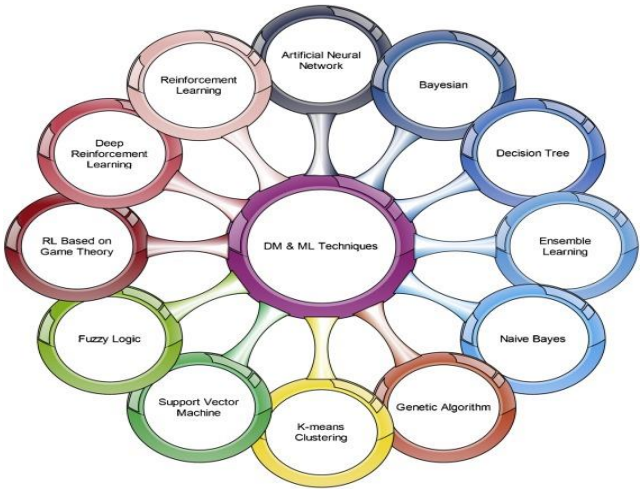
Dünya'da en çok kullanılan programları dilleri yıllara göre kullanım oranlarıyla beraber Tablo 2.1' de yer verilmiştir. Yapılan anketler yazılımla ilgilenen tüm insanları kapsamaktadır. Bilgiler incelendiğinde SQL dilinin yıllar içinde popülerliğinde azalma olmadığı görülmüş ve giderek arttığı gözlemlenmektedir. Python dili 2016'dan bu yana kolay ve anlaşılır kullanımı ile diğer programlama dillerine göre daha büyük artışlarla kullanım oranının arttığı gözlemlenmektedir. Python dilinin kullanım oranında 2017 yılından 2018 yılına geçerken %7'lik büyük bir artış olmuştur. R dilinin verilere bakıldığında 2018 yılında kullanım oranında %1,7 artış olduğu görülmektedir, fakat 2019 yılında programı kullanan kişilerin %0,3 azaldığı görülmüştür.

Tablo 2.1 Yıllara göre Dünya'da en çok kullanılan programlama dilleri

Programlama Dilleri	Kullanım Oranı (%)				
	2020	2019	2018	2017	2016
JavaScript	67,7	67,8	69,8	61,9	55,4
HTML/CSS	63,1	63,5	68,5	-	-
SQL	54,7	54,4	57,0	50,8	49,1
Python	44,1	41,7	38,8	31,7	24,9
Java	40,2	41,1	45,3	39,3	36,3
Bash/Shell/PowerShell	33,1	36,6	39,8	-	-
C#	31,4	31,0	34,4	33,8	30,9
PHP	26,2	26,4	30,7	27,9	25,9
TypeScript	25,4	21,2	17,4	9,4	-
C++	23,9	23,5	25,4	22,1	19,4
C	21,8	20,6	23,0	18,9	15,5
Go	8,8	8,2	7,1	4,2	-
Kotlin	7,8	6,4	4,5	-	-
Ruby	7,1	8,4	10,1	9,0	8,9
Assembly	6,2	6,7	7,4	4,9	-
VBA	6,1	5,5	4,9	-	-
Swift	5,9	6,6	8,1	6,4	-
R	5,7	5,8	6,1	4,4	-
Rust	5,1	3,2	-	-	-
Objective-C	4,1	4,8	7,0	6,4	6,5
Dart	4,0	1,9	-	-	-
Scala	3,6	3,8	4,4	3,5	-

Kaynak: (Szepesv´ari, 2009; Fazakis vd., 2016; Brownlee, 2020; Buffet vd., 2020)

Makine öğrenimi karar verme modellerini iyileştirmek için veri madenciliği ve hesaplamalı zekâ algoritmalarından yararlanmaktadır ve veri madenciliği ve makine öğreniminin iş kullanımlarına yönelik birçok uygulamalar içermektedir. Akademik çalışmalar incelendiğinde; tahmine yönelik çalışmalarda kullanılan algoritma ve tekniklerin, veri madenciliği veya makine öğrenimi başlığında verildiği görülmektedir. Ancak aralarında ortak noktalar olmasına rağmen, bu disiplinler farklı kavramlar olarak anılmaya başlanmıştır. Şekil 2.3'de araştırmacılar tarafından kullanılan en yaygın kullanılan veri madenciliği ve makine öğrenimi teknikleri görülmektedir (Shafiq ve ark., 2020).



Şekil 2.3 En yaygın kullanılan veri madenciliği ve makine öğrenimi teknikleri

2021 yılına ait istatistiklere en çok popüler veri madenciliği teknikleri incelendiğinde bunların öncelik sırasının; Appriori algoritması, Beklenti maksimizasyonu (Expectation-Maximization veya EM), Sayfa sıralaması algoritması (PageRank Algorithm), C4.5 Algoritması, Naive Bayes Algoritması, CART (Classification and Regression Trees), K-Means Algoritması, SVM (Support Vector Machines), Adaboost Algorithm (Adaptive Boosting'in kısaltması olan AdaBoost) ve KNN Algoritması olduğu görülmüştür (Analytics insight, 2021). En popüler makine öğrenimi algoritmaları incelendiğinde ise bunların öncelik sırasının; Doğrusal regresyon, Lojistik regresyon, KNN Algoritması, Naïve Bayes, Support Vector Machines, Random Forest, Adaboost, Gradian boost, XGBoost, LightGBM ve CatBoost algoritmaları olduğu görülmüştür (KDnuggets, 2021). Veri madenciliği ve makine öğrenimi algoritmalarının iş uygulamalarında sınıflayıcı temelli algoritmaların daha yaygın bir şekilde uygulandıkları görülmüştür.

3. Veri Madenciliği ve Makine Öğrenimi Arasındaki Benzerlikler ve Farklılıklar

Veri madenciliği, istatistik ve makine öğrenimi disiplinleri; sağlık, işletme, endüstri, güvenlik vb. gibi birçok kuruluşların daha iyi kararlar almasına yardımcı olan ve işletmenin büyümesini pozitif yönde etkileyen ve veriden öğrenen disiplinlerdir. Günümüzde ses ve yüz tanıma, otonom araçlarda arama analizi, veri madenciliği ve sektörel uygulamalar gibi alanlarda makine öğrenmesi kullanılmaktadır. Makine öğrenimi ve veri madenciliği kavramlarının her ikisi de birçok ortak kavramı kapsadığından dolayı bazen aralarındaki farkı görmek zordur. Veri madenciliği veri yığınları arasındaki anlamlı ve değerli ilişkileri ortaya koyarken, makine öğrenimi veri yığınları arasındaki ilişkiyi temsil eden bilgilerden yararlanarak modellere ilişkin tahmini sonuçlar bulur. Bu modeller, sonuca ulaşmada makinenin yapacağı işlemlerdir (softwaretesting, 2021). Veri madenciliği ve makine öğrenimi arasında bazı temel farklılıklar vardır (Educba, 2021; Knowlab, 2021; Kulin vd., 2021; Mitchell vd., 1990; Softwaretestinghelp, 2021; Javapoint, 2021). Söz konusu ortak noktalar ve farklılıklar Tablo 3.1'de verilmiştir.

Tablo 3.1. Veri madenciliği ve makine öğrenimi arasındaki ortak noktalar ve farklılıklar

Veri Madenciliği ve Makine Öğrenimi
Veri madenciliği ve makine öğreniminin her ikisi de büyük veriden öğrenir.
Veri madenciliği ve makine öğreniminin her ikisi de analitik süreçler olup, veri biliminin (Data science) temel bir parçasıdır.
Veri madenciliği ve makine öğreniminin her ikisi de işletmelerin veri kümelerini faydalı bilgilere dönüştürmek için kullanılır. İşletmelerin daha iyi iş kararlarına yol açabilecek eğilimleri analiz etmelerine ve anlamalarına yardımcı olur.
Veri madenciliği ve makine öğrenimi teknikleri araştırma konusuna göre tanımlanır, ancak kullanımlarında bazı teknikler aynıdır. Amaçları da genel olarak aynı olup, verileri anlamak, örüntülerin tanımaya yardımcı olmak ve kullanıcılar tarafından anlamlı modeller oluşturmaktır.
Makine öğrenimi algoritmalarını geliştirmek ve davranışını gelecekteki girdilere göre değiştirmek için veri madenciliği tekniklerini kullanır. Bununla birlikte makine öğrenimi, gelecekteki sonuçları tahmin edebilmesi için bazı verilerin arkasında neler olduğuna dair modeller oluşturmak için veri madenciliği tekniklerini ve diğer öğrenme algoritmalarını kullanır. Çoğunlukla matematiksel temellidir, ancak bu daha çok programlamaya yöneliktir.
Veri madenciliği ve makine öğreniminin ilk defa kullanımları örüntü tanımadan (Pattern recognition) ortaya çıkmıştır.
Makine öğrenimi algoritmaları için veri madenciliği algoritmasının "çıktısı" genellikle "girdi" olarak kullanılır.
Makine öğrenimi otomatikleştirilmiş bir süreç olduğundan, makine öğreniminin ürettiği sonuç, veri madenciliği ile karşılaştırıldığında daha kesin olacaktır.
Veri madenciliğinin kullanımında doğru algoritmaları seçebilen, parametreleri ayarlayabilen ve belirli bir problem için modelleri eğitebilen bir uzman gerektirir ve bu uzman makine öğrenimi araçlarıdır.

Makine öğrenimi ve veri madenciliği birbirlerinden ilham alan ve ortak noktaları olmasına rağmen, bazı farklılıkları olan kavramlardır. Söz konusu farklılıklar Tablo 3.2'de verilmiştir.

Tablo 3.2 Veri madenciliği ve makine öğrenimi arasındaki farklılıklar

Veri Madenciliği	Makine Öğrenimi
Veri madenciliği ham ve büyük veri yığınlarından anlamlı ilişki ve kurallar ortaya çıkarmaktır. Veri madenciliği istatistik, makine öğrenimi ve veritabanı sistemleri üzerine inşa edilmiştir.	Makine öğrenimi, makinelerin mevcut verilerden öğrendiği ve kendi kendine öğrenip geliştirdiği bir konsept üzerinde çalışır. Büyük verilerin yanısıra geçmiş deneyimlerden gelen algoritmayı ifade eder. Algoritmalar matematik ve programlama dilleri üzerine inşa edilmiştir.
Veri madenciliği verilerden kural çıkarmakla ile ilgilidir. Veri madenciliği teknikleri, hedef (Bağımlı) veri kümesini tanımlama veya makine öğrenimi algoritmalarını kullanarak sonuçları tahmin edebilmeye yönelik olarak iki ana amaca sahiptir.	Makine öğrenimi bir bilgisayara açıkça programlanmadan görünmeyen verileri tahmin etmek için verilerden nasıl öğrenileceğini öğretmekle ilgilidir ve performansı iyileştirmeye odaklıdır. Geçmiş verilerden analizler yaparak bir durumun modellenmesini ve bu tahmini model sayesinde yeni bir veri geldiğinde onu etiketlemeyi amaçlar.
Veri madenciliği birçok makine öğrenmesi tekniklerini kullanır, fakat çoğunlukla mantıksal olarak farklı hedefleri vardır.	Makine öğrenmesi de denetimsiz öğrenme ya da öğrenici doğruluğunu geliştirmek için ön işleme adımı gibi veri madenciliği tekniklerini kullanır.
Veri Madenciliği, sonuçları tahmin etmek ve yararlı bilgiler elde etmek için CRISP-DM teknolojisini kullanır ve bilgi keşfi için veri tabanı, veri madenciliği motoru ve örüntü değerlendirmesini kullanır.	Makine öğrenimi, gelecekteki sonuçları tahmin edebilmek ve karar verebilmek için belirli bilgilerin arkasında neler olduğuna dair tahmini model oluşturmada veri madenciliği tekniklerini, grafiksel modelleri, doğal dil işlemeyi, sinir ağlarını ve otomatik algoritmaları kullanır.
Veri madenciliği, makine öğrenimi algoritmaları dışında diğer birçok tekniği de kullanır. Veri madenciliği bir araç olarak bir makine öğrenimi algoritması kullanabilir, ancak veri madenciliği ham verilerden bir şeyler çıkarmak için başka bir araç olarak istatistikleri de kullanır.	Makine öğrenimi algoritmaları, veri madenciliği sürecinde kullanılabilir. Kullanılan sınıflayıcı teknikler çoğunlukla aynıdır ve hem veri madenciliğinde hemde makine öğreniminde kullanılabilir.

Veri madenciliği anlamlı ve değerli bilgiyi ortaya çıkarmada ve sonucu tahmin etmek için makine öğrenimi araçlarından yararlanır.	Makine öğrenimi kümeleme, sınıflandırma ve tahmin gibi veri madenciliği görevlerinde kullanılan hesaplama yöntemlerini kullanır.
Veri madenciliği veri bilimi ve iş analitiğinin bir alt kümesidir.	Makine öğrenimi veri bilimi ve yapay zekânın bir alt kümesidir.
Veri madenciliği, kullanıcıya yararlı ve anlamlı bilgileri çıkarmak için verileri derinlemesine inceler. Ayrıca veri madenciliği, makine öğrenimi için bir girdi kaynağı görevi görür.	Makine öğrenimi, makineleri eğitimli veri kümesiyle yinelemeli olarak besleyerek, makineleri mükemmele yakın hale getirmek için karmaşık algoritmaları geliştirme yöntemidir. Bir diğer ifade ile makine öğrenimi makineyi okur.
Veri madenciliği verilerden bilgi keşfi yaparken insan faktörüne ihtiyaç duyar. Veri madenciliği, otomatik olarak gelmeyen ve insan eliyle tanımlanan veriler ile tahmini model sonuçlarını üretir. Veri madenciliği sürecinde akıllı özellikler insan tarafından tanıtılarak akıllı hale getirilebilir. Ancak değişken değiştiğinde model değişir ve sonsuza kadar kullanılamaz.	Makine öğrenimi algoritmaları sürekli olarak çalışarak sistemin performansını otomatik olarak iyileştirir ve herhangi bir hatanın ne zaman ortaya çıkabileceğini de analiz edebilir. Bazı yeni veriler olduğunda veya değişiklik olduğunda, makine değişiklikleri yeniden programlamaya veya insan müdahalesine gerek kalmadan dahil edebilir. Makine öğrenmesinde algoritma tanımlanarak otomatik olarak öğrenir ve tasarlandıktan sonra, daha iyi olmak için bir insana ihtiyaç duymazlar. Bir kez uygulandıktan sonra sonsuza kadar kullanabilir.
Veri madenciliğinde model performansını artırmak için verinin büyüklüğü ve verinin temizlenerek modele hazır hale getirilmesi çok önemlidir. Verilerdeki aykırı uç değerler modelin doğruluk oranını azaltır veya model hatalarını artırır. Ayrıca veri madenciliği analizinde ilgili değişkenlerin modele alınması, veriye ve probleme uygun doğru model kurulması sonuçların güvenirliliğini artırır.	Makine öğreniminde model performansını artırmak için kendi kendine öğrenme algoritmalarını kullanır ve makine öğrenimi sonuç odaklıdır. Sunmuş olduğunuz veri ve parametreleri simüle ederek anlamlı tespitler yapan ve kendi kendini eğiten sistemlerdir. Ayrıca makine öğrenimi algoritmalarının parametreleri büyük ölçüde öğrenme sürecinin sonucunu etkiler. Modelin doğruluğunu artırmak için her parametre için optimum değeri bulmak ve bu parametreleri ayarlamak için bunların model üzerindeki bireysel etkilerini iyi anlamanız gerekir.
Veri madenciliği kendi kendine öğrenme yeteneğine sahip değildir. Önceden tanımlanmış yönergeleri takip eder. Veri analizinde aşamaları takip etmeniz gerekir ve belirli bir soruna cevap verir.	Makine öğrenimi algoritmaları kendi kendini tanımlar ve duruma göre kurallarını değiştirebilir ve belirli bir sorunun çözümünü bulabilir ve bu şekilde çözebilir.
Veri madenciliği genellikle gerçek zamanlı kullanıcılar ve yazılım çözümleri sağlayıcıları tarafından veritabanlarında bilgi keşfi (KDD) olarak tanımlanır ve mevcut bir veri kümesini veri ambarı gibi kullanır. Veri tabanlarında bilgi keşfi sürecinde ilişkiler, kurallar ve kalıplar bilinmemektedir. Söz konusu ilişki, kural ve kalıpları bulmada veri kümelerinden yararlanır.	Makine öğreniminde, makineye verilerden öğrenmesi ve anlaması için bazı değişkenler ve kurallar verilir. Makine öğrenimi, bilgisayara verileri nasıl anlamlandıracağını ve ardından yeni veri kümeleri hakkında tahminler yapmayı öğreten bir eğitim veri kümesi üzerinde eğitilir. Bir diğer ifade ile makine öğrenimi bilgisayarların programlanmadan hareket etmesini sağlama bilimi olarak da tanımlanabilir.
Veri madenciliği, mevcut verilerden kuralları elde etmek için kullanılır.	Makine öğrenimi bilgisayara, kuralların nasıl öğrenileceğini ve kavranacağını öğretir.
Veri madenciliğinde doğruluk oranı çok yüksek olmasa da, büyük verinin yanı sıra az sayıda veriyi işleyerek ve modele hazırlayarak değerli bilgi keşfedilebilir.	Makine öğrenimi algoritması, mevcut algoritmaların sınırlı olması nedeniyle, verilerin standart biçimde beslenmesine ihtiyaç duyar ve doğru sonuçlar için büyük miktarda veriye ihtiyaç duyar.

4. Makine Öğrenimi ve Veri Madenciliği Aşamaları

Veri madenciliği disiplini CRISM-DM metodolojisini izler. Bunlar; işi anlama, veriyi anlama, veriyi hazırlama, modelleme, değerlendirme ve sonuçların kullanmasıdır (Fayyad ve vd., 1996; Chapman P., Clinton J., 2000). Yapılan çalışmanın niteliği ya da amacı ne olursa olsun bu aşamalardan geçmek zorundadır. Mevcut sorunu herhangi bir zamanda başarılı bir şekilde çözüme ulaştırmak için bu adımlar büyük önem taşımaktadır (Chollet, 2017; Ersöz, 2019).

Makine öğrenimi (ML) adımlarının çoğu, Fayyad ve arkadaşlarının tanımladığı (1996) veri madenciliği aşamalarına benzemektedir. CRISP-DM'yi temel almakta, ancak kapsamının biraz daha geniş olduğu görülmektedir. Genel olarak makine öğreniminde yedi aşama vardır. Bunlar; veri toplama, veri hazırlama, model seçimi, model eğitimi, değerlendirme ve yorumlama, parametre ayarlama ve tahmin yapmadır [Guo, 2017; Chollet, 2017; Mayo, 2018). Endüstri genelinde iş akışında makine öğrenimi, veri bilimi ve veri madenciliği ile küçük farklılıklar olmasına rağmen, bu farklılık çoğunlukla geri besleme döngüleri gibi farklılıkları kapsamaktadır (Mayo, 2018). Tablo 4.1'de veri madenciliği ve makine öğrenimi iş akışlarının genel bir yapısı karşılaştırmalı olarak verilmiştir.

Tablo 4.1. Veri madenciliği ve makine öğrenimi aşamaları

Veri Madenciliği Aşamaları (CRISP-DM)	**Makine Öğrenimi Aşamaları (CRISP-ML)**	**Makine Öğrenimi Aşamaları**		
		Studer ve vd. (2021)	**Amershi ve vd.** (2019)	**Mayo** (2018) (Guo, 2017; Chollet, 2017)
İşi anlama	İşi anlama	İşi ve verileri anlama	Model ihtiyaçların ortaya konulması	Verilerin toplanması ve veri kümesinin birleştirilmesi
Verileri anlama ✓Verileri toplama ✓Verileri özetleme ✓Verilerin görselleştirilmesi	Verileri anlama ve toplanması			
Verilerin modele hazırlanması ✓ Veri temizleme ✓ Veri dönüşümü, ✓ Veri normalleştirme, ✓ Aykırı ve uç değer temizleme ✓ Boyut indirgeme vb.	Verilerin hazırlanması ✓ Veri temizleme, aykırı ve uç değer temizleme ✓ Veri dönüşümü ✓ Veri filtreleme ✓ Veri normalleştirme ✓ Eğitim ve değerlendirme setlerinin bölünmesi vb.	Verilerin hazırlanması	Verilerin toplanması	Verilerin hazırlanması ✓ Veri temizleme ✓ Veri dönüşümü, veri normalleştirme, aykırı ve uç değer temizleme, ✓ Eğitim ve değerlendirme setlerinin bölünmesi
			Verilerin temizlenmesi	
			Verilerin etiketlenmesi	
Modelleme (Veri madenciliği doğru model ve tekniklerin seçilmesi) ✓ Sınıflayıcı ✓ Kümeleyici ✓ Birliktelik kuralları	Özellik veya modelin seçilmesi	Modelleme	Özellik seçimi	Modelin seçilmesi
	Modelin eğitilmesi		Modelin eğitilmesi	Modelin eğitilmesi
Modelin değerlendirilmesi (Performans ölçümü) ✓ Doğru sınıflandırma başarısı (Accuracy) ✓ Kappa istatistiği (Duyarlılık analizi) ✓ Sensitivity (Hassaslık) ✓ Specificity (Belirginlik) ✓ Ortalama mutlak hata (MAE) ✓ Eğri altında kalan alan (ROC) ✓ Göreli mutlak hata (RAE) vb.	Modelin değerlendirmesi	Modelin değerlendirilmesi	Modelin değerlendirilmesi	Modelin değerlendirilmesi
Veri madenciliği sonuçlarının sunulması ve kullanılması	Model optimizasyonu (Hiperparametre ayarlama)			Modelin düzenli hale getirilmesi ve parametre ayarlama (Hiperparametre)
	Tahmin yapma	Model sonuçlarının sunulması ve kullanılması	Model sonuçlarının sunulması ve kullanılması	Tahmin yapma
	Modelin izlenmesi ve kestirimci bakım	Modelin izlenmesi ve kestirimci bakım	Modelin izlenmesi	

Veri madenciliği ve makine öğrenimi metodolojileri incelendiğinde, her ikisinde de CRISP-DM metodolojisinin ilk adımlarının benzediği görülmektedir. Ayrıca iş problemlerinin çözümlerinde kullanılan makine öğrenimi teknikleri ve veri madenciliği sınıflayıcı tekniklerinin çoğunlukla benzediği görülmektedir. Veri madenciliği "istatistiksel analiz ve modellemeler" ile "makine öğrenimi" tekniklerinin kullanılması ile devreye girmiştir. Bundan dolayı veri madenciliği, verinin içindeki bilginin ortaya çıkarılması için gelişmiş teknolojiler ve iş deneyimi birlikte kullanılmalıdır (Ersöz, 2019). Veri madenciliği ve makine öğreniminde temel kavramlar aşağıda açıklanmıştır.

İşin veya projenin tanımlanması: Problemlerin irdelenmesi aşamasında iş deneyimi ve uzmanlık önemlidir. Bu ilk adımda projenin amaç ve gerekliliklerinin anlaşılması ve iş perspektifinin ortaya çıkması gereklidir. Bu bilginin veri madenciliği problem tanımı olarak netleştirilmesi ve hedeflere ulaşma amaçlı planların oluşturulması gereklidir. Problemin tanımlanması aşaması, araştırmanın ve veri madenciliğinin amacını, mevcut durumun değerlendirilmesiyle planlama sürecinin belirlenmesini kapsar.

Bu aşamada ihtiyaçlar net bir şekilde tanımlanmalıdır. Bu basamakta amaçlar gerçekleştirilirken dikkat edilecek olan performans ölçütlerinin neler olduğuna ve son olarak bu süreç sonunda ortaya çıkan sonucun hangi durumlar için kullanılacağına karar verilir (Sumathi ve Sivanandam, 2006).

Verilerin toplanması: Bir modelin doğruluğu ve iyi sonuç vermesi kaliteli verilerin kullanılması sonucu elde edilir. Verilerin miktarı fazla oldukça model doğru sonuç elde eder. Veri toplama adımı makine öğrenimi sürecinin temelidir. Veriler araştırılmak istenen probleme uygun bir şekilde toplanmaktadır. Veri toplarken yanlış özelliklerin seçilmesi veya veri seti için sınırlı girdi türlerine odaklanma gibi hatalar, modeli geçersiz kılmaktadır (Yufeng, 2017).

Verilerin hazırlanması: Verilerin hazırlanma aşamasında, verilere makine öğrenme algoritmasının uygulanabilmesi için uygun hale getirilir. Uygun hale gelebilmesi için mevcut verilerin birtakım aşamalardan geçmesi gerekmektedir. Bunlar veri temizleme, verilerin bütünleştirilmesi, verilerin dönüştürülmesi ve verilerin indirgenmesidir. Bu işlemlerden hangisinin kullanılacağı verinin ihtiyacına göre belirlenir (Sherarer, 2000). Temizlenen veriler model uygulanabilmesi için genelde %80'ne %20 olmak üzere eğitim ve test verisi olarak ayrıştırılır. Veri setinin ayrılmasının nedeni modeli eğitim verisinde eğitilip, test verisinde test edilmesidir (Kuhlman, 2009).

Modelin seçilmesi: Veri bilimciler tarafından geliştirilmiş, farklı amaçlar için kullanılabilecek çeşitli mevcut modeller bulunmaktadır. Bu modeller farklı hedefler düşünülerek tasarlanmıştır. Bu aşamada istenilen hedefe uygun model seçimi söz konusudur (Chollet, 2017). Bu çalışmada veri seti üzerinde daha doğru sonuçlar elde edebileceği düşünülen makine öğrenimi ve veri madenciliği tekniklerinden seçilmiştir. Bunlar; çoklu doğrusal regresyon, yapay sinir ağları ve karar ağaçları tekniklerdir.

Modelin eğitilmesi: Makine öğrenimi sürecinin merkezinde modelin eğitimi yer alır. Algoritmanın eğitim verileriyle beslenmesini gerektiren bu aşamada öğrenmenin büyük bir kısmı yapılmaktadır. Veri setinin eğitim için ayrılan kısmı kullanılarak oluşturulan model eğitilir (Kubat vd., 1996). Çalışmada kullanılmak üzere seçilen modeller RStudio, Python ve Knime programlarında eğitilmiştir.

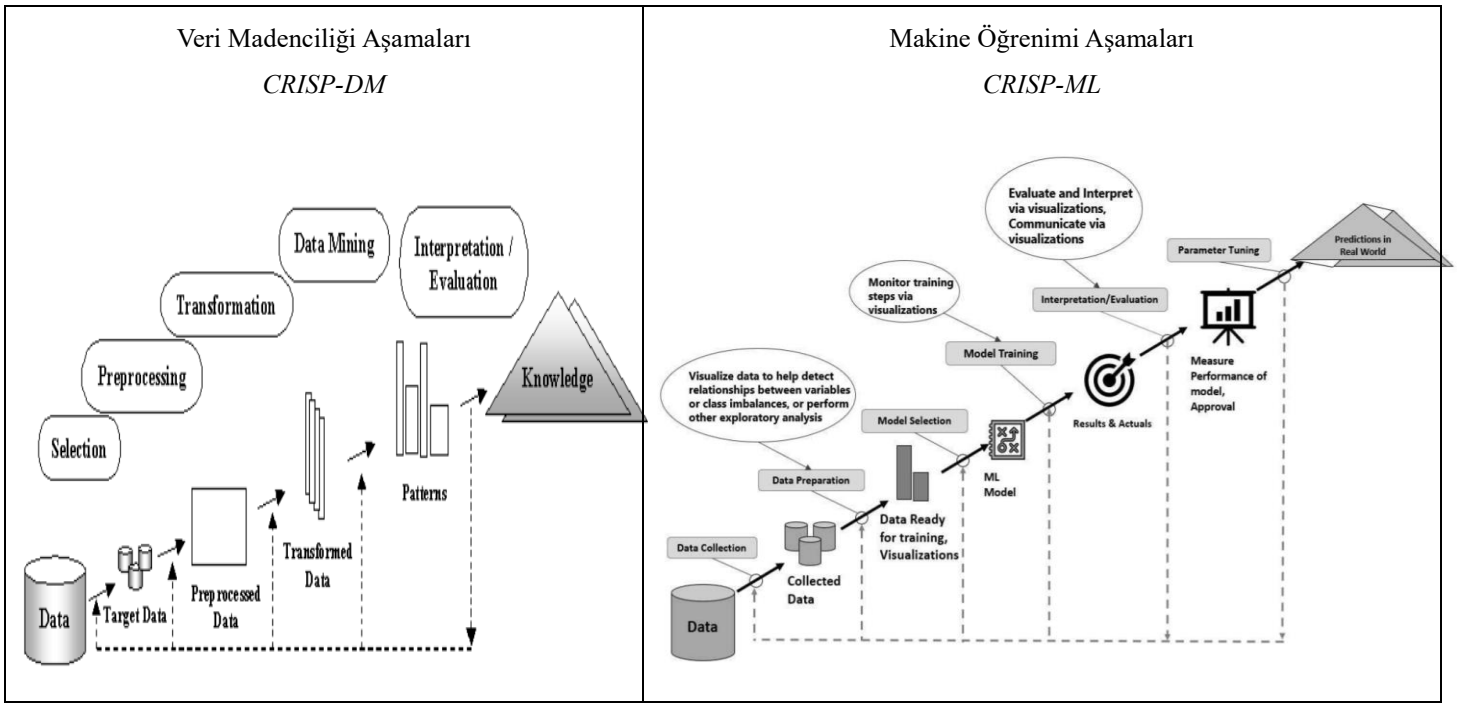
Modelin değerlendirilmesi: Bu aşamada eğitilen model değerlendirilmektedir. Bu nedenle değerlendirme için oluşturulan veri setinin bir kısmı modelin yeterliliğini kontrol etmek için kullanılır. Bu durum modeli eğitimin bir parçası olmayan durumlarda nasıl sonuç vereceğini test eder ve test sonucu modelin performansını belirlemektedir (Chollet, 2017). Yapılan çalışmada modellerin performans değerleri belirlenirken hataları ölçmek için hata kareler ortalamasının karekökü (Root Mean Squared Error – RMSE) kullanılmıştır. RMSE ölçeğe bağlı olduğundan, veri kümeleri arasında değil, belirli bir veri kümesi için farklı modellerin tahmin hatalarını karşılaştırmak üzere kullanılan bir doğruluk ölçüsüdür (Hyndman & Koehler, 2006). Modelin açıklama gücünü gösteren belirtme (determinasyon) katsayısı olan (R^2) kullanılmıştır. Regresyon modeli sonucunda bulunan belirtme katsayısı, kullanılan bağımsız değişkenlerin (X_1, X_2, X_n) bağımlı değişkeni (Y) açıklama oranıdır. Belirtme katsayısı ne kadar yüksek ise modeli açıklama gücünün o kadar yüksek olduğu söylenebilir (Ersöz & Ersöz, 2019). Model değerlendirme aşaması iş hedeflerinize ulaşmak için doğru yolda olduğunuzdan emin olmanıza olanak sağlar ve bir proje dağıtım aşamasına hazır olmadan önce önceki adımlara geri dönebilme imkânı sağlar.

Parametre ayarlama: Model performansını arttırmak için gelişmiş parametre performans ayarlarının yapıldığı adımdır (Yufeng, 2017). Hiperparemetre optimizasyonu değeri öğrenme sürecini kontrol etmek için kullanılan bir parametre değeridir. Aynı türden makine öğrenimi modeli, farklı veri modellerini genelleştirmek için farklı kısıtlamalar, ağırlıklar veya öğrenme oranları gerektirebilir. Hiperparametre optimizasyonu, belirli bağımsız veriler üzerinde önceden tanımlanmış bir kayıp fonksiyonunu en aza indiren optimal bir model sağlayan bir hiperparametre bulur. Oluşturulan modeller arasında en güvenilir ve en yüksek doğruluk derecesine sahip olanının saptanması gerekmektedir. Ayrıca model değerlendirme sürecinde, başarıyla tahmin edilen algoritmanın kendi içinde genelleştirip genelleştiremeyeceğini değerlendirmek gerekir. Bu değerlendirme yöntemlerinden birisi de k-kat çapraz geçerlemedir (k-fold cross validation). Veri seti k-kat çapraz geçerlemede k eşit parçaya ayrılır. Ayrılan k parçadan her defasında bir tanesi test, k-1 tanesi ise eğitim için kullanılması sağlanır. Sonuç olarak, k tane hata oranı oluşur ve bütün tahmin hatalarını hesaplamak adına hataların ortalaması alınmaktadır (Bergstra vd., 2012).

Tahmin yapma: Modelin uygulama yapmaya hazır olduğu aşamadır. Model kurarken ayrılan test verileri kullanılarak model tahmin edilir ve gerçek dünyada nasıl performans sağladığı incelenir (Mayo, 2018). Makine öğrenimi tahmin yapma aşamasında kurulan model artık insan unsurundan arınarak, makinenin kendi başına tahminlerde bulunma aşamasıdır.

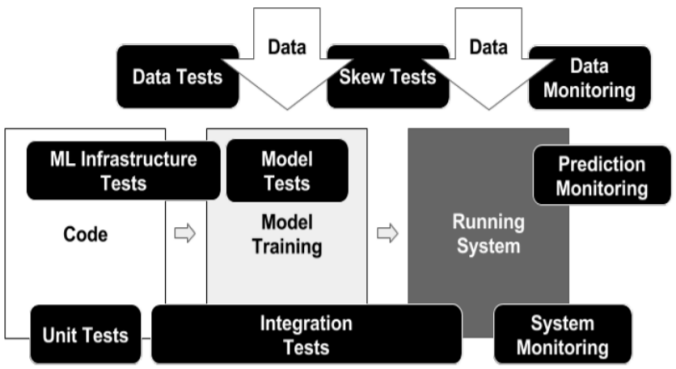
Model izleme ve bakım: Makine öğrenimi modeli gerçek dünyadaki herhangi bir sürecin istatistiksel bir gösterimidir ve veri kullanılarak süreç modellenir. Değişen bir ortamda modelin bozulma riskine karşı modelin izlenmesi ve kestirici bakımı önemlidir.

Şekil 4.1'de veri madenciliği ve makine öğrenimi sürecinin temel aşamalarının karşılaştırılması görsel olarak verilmiştir.



Şekil 4.1 Veri madenciliği ve makine öğrenimi süreçleri (Fayyad vd. 1996; Eisler& Meyer, 2020)

Makine öğrenimi (ML) sistem davranışı büyük ölçüde verilere ve modellere bağlı olarak değişir. Bunun için makine öğreniminde eğitim verilerin kod gibi test edilmesi gerektiği ve eğitimli bir ML modeli hata ayıklanabilirlik, geri alma ve izleme gibi uygulamalara ihtiyaç duyar (Breck ve vd., 2017). Şekil 4.2'de ML sistem tabanlı modelin test edilme ve izleme aşaması verilmiştir.



Şekil 4.2 Makine öğreniminde sistem tabanlı test etme ve izleme (Breck ve vd., 2017)

Makine öğreniminde model değişen bir ortama uyarlanabilir olmalıdır, aksi takdirde modelin performansının düşük olur ve modelinin kalıcı olarak izlenmesi ve bakımının sağlanabilmesi de zamanla bozulur.

5. Literatür İncelemesi

Bu çalışmada veri madenciliği ve makine öğrenimi teknikleri kullanılarak üretim miktarının tahmin edilmesi amaçlanmıştır. Literatür taramasında veri madenciliği ve makine öğrenimi tekniklerine ilişkin son yıllarda tekstil ve hazır giyim sektöründe yapılmış çalışmalar incelenmiştir. Makine öğreniminin tekstil sektörü üzerindeki etkileri inceleyen referans çalışmalar araştırılarak, literatüre yapılabilecek katkılar bulunmaya çalışılmıştır.

Hsu ve arkadaşları (2009) tarafından yapılan çalışmanın amacı, hazır giyim endüstrisinde standart boyutlu grafikler için endüstriyel standartları geliştirmek adına antropometrik veriler kullanılarak kalıplar ve kurallar oluşturmak oluşturmaktır. Tayvan'nın en büyük giyim şirketlerinden birinde 986 kadın bedeni ölçüleri ve 52 antropometrik değişken ile toplamda 51,272 adet antropometrik veri elde edilmiştir. Denekler ayrıştırılarak 956 adeti ileri analizler için kullanılmıştır. Vücut tiplerini belirleme, kümelere ayırma işlemleri Ward's minimum varyans yöntemi ve K-Means algoritması kullanılarak yapılmıştır. Faktör analizi sonucunda bel ölçüsü ve yükseklik faktörü konfeksiyon imalatında çok önemli değişkenler olduğu tespit edilmiştir. Ward'un minimum varyans yöntemi kullanılarak ilk kümeleme gerçekleştirilmiş, son kümeler K-Means algoritmasıyla gerçekleştirilmiştir. Çalışma sonucu K-Means algoritmasıyla en iyi beş küme bel ölçüsüne göre bulunmuştur. Ayrıca vücut tipi A, B, C, D ve Y yükseklikle ilgili antropometrik değişkenler için önemli farklar elde edilememiştir (Hsu, 2009).

Selvanayaki ve arkadaşları (2010) tarafından yapılan çalışmada, tekstil üretimi için önemli olan pamuğun özelliklerine odaklanarak kalite tahmini yapılmıştır. Çalışmada kullanılan 12 farklı özelliğe sahip veri seti, özel bir iplik fabrikasından toplanmıştır. Pamuğun kalitesine karar veren baskın özellikler arasında açıklık uzunluğu (mm), homojenlik oranı %, mukavemet (g/tex), mikronarie, tiftik, çepel, görünmez kayıp, olgunluk katsayısı bulunmaktadır. Weka programı kullanılarak yapılmış olup çalışmada; Multilayer perceptron, Naive bayes, J48 decision tree ve K-nearest neighbor algoritmaları kullanılmıştır. Modellerin performansı ise k-katlı çapraz kullanılarak ölçülmüştür. Araştırma sonucunda J48 decision tree algoritmasının, diğer algoritma tahminlerine göre daha iyi performans sağladığı görülmüştür (Selvanayaki vd., 2010).

Özbek ve Akalın (2011) çalışmalarında, Ocak 1995 - Aralık 2008 arasındaki 168 veri ile Türkiye'nin Almanya'ya olan denim pantolon ihracatının tahminlemesini yapmışlardır. Tahminleme için YSA modellerinden Çok Katmanlı Algılayıcı (MLP) ve Elman Tekrarlayan Sinir Ağları (ERNN) modelleri kullanılmıştır.

Modelde girdi olarak; pamuk fiyatı, su fiyatı, elektrik fiyatı, hazır giyim sektöründeki kredi kullanımı, denim pantolon ithalatı, reel efektif döviz kuru, Almanya'nın denim pantolon ithalatı ve Türkiye'ye kota uygulaması, Almanya'da kişi başına düşen gelir ve nüfusu, Almanya'daki işsizlik ve enflasyon verileri, Tl'nin ABD Doları karşısındaki değeri, ihracat kredileri, asgari ücret ve denim pantolon markaları, verileri kullanılmıştır. Kullanılan her iki modelin de tahmin açısından başarılı sonuçlar verdiği ve kot pantolon ihracatının tahmininde kullanılabileceği sonucuna varılmıştır. Elman Network'ün MLP Network'ten daha iyi tahmin performansına sahip olduğu belirlenmiştir. Çalışmada kurulan model ile ithalatın öngörülmesi ve gelecekteki ihracatlar için önemli çıktılar elde edilebileceği sonucuna varılmıştır (Özbek&Akalın, 2011).

Mozafary ve Payvandy (2014) tarafından yapılan çalışmada, tekstil endüstrisinde veri madenciliği teknikleri kullanılmıştır. Çalışmada kullanılan veriler kamgarn iplik fabrikasının kalite kontrol laboratuvarında bir yıl boyunca yapılan 70 değişken dahil olmak üzere 2241 deney çalışmasından oluşmaktadır. İncelenen değişkenler arasında; lif özellikleri, üretim süreci parametreleri ve iplik kalitesi parametreleri bulunmaktadır. K-means ile oluşturulan her küme için ANN algoritması ile iplik kalitesi tahmin etmiştir. Araştırma sonucunda K-means ve ANN tekniklerine ilişkin model doğruluk oranlarının, yapay sinir ağından daha doğru olduğu tespit edilmiştir (Mozafary & Payvandy, 2014).

Guler ve arkadaşları tarafından (2017) yapılan çalışmada, tekstil sektöründe yer alan bir firmada birden fazla veri madenciliği tekniği kullanılarak pantolon üretim miktarlarının, hangi faktörler tarafından etkilendiğini bulmayı amaçlamışlardır. Çalışmada kullanılan veri setinin değişkenleri; çalışan sayısı, çalışma saati, fazla mesai, toplam çalışma saati, günlük partide üretilen ürün sayısı ve kişi başına üretim olarak alınmıştır. Analiz tekniklerinden ise C&RT, çoklu doğrusal regresyon analizi, yapay sinir ağları ve Chaid teknikleri kullanılmıştır. Sınıflayıcı model performansları karşılaştırıldığında; ortalama mutlak hatanın en düşük olduğu ve doğrusal korelasyon dikkate alındığında en iyi tahmin sonucunu veren C&RT algoritması olduğu görülmüştür. C&RT algoritmasının sonuçlarına göre pantolon üretim miktarını etkileyen en önemli değişkenin kişi başına üretim olduğu görülmüş ve bunu sırasıyla üretim tarihi ve toplam çalışma saati izlemiştir (Guler vd., 2017).

Lin ve arkadaşları (2018) tarafından yapılmış olan çalışmada, tekstil endüstrisinde makine öğrenimi için anahtar operasyon parametresi ile kusur arasındaki ilişkiye odaklanıp, bir tavsiye sistemi (OPRS) tasarlamaktır. Veriler, Li Peng tekstil fabrikasının ERP sisteminden alınmıştır. Çalışılan veri seti 240963 satır uzunluğuna sahiptir ve değişkenler; özgü süreci, boyutlandırma süreci, ışınlanma süreci, dokuma işleminden oluşmaktadır. Çalışma, iplik özelliklerine göre işlem parametre tahmini için regresyon modeli ve üretim kalitesi işlem parametre tahmini için sınıflandırma modeli olmak üzere iki farklı model kullanılarak birleştirilmiştir. Regresyon modelleri olarak; Linear regression, Lasso regression, Ridge regression ve Elasticnet regression kullanılmıştır. Sınıflandırma teknikleri olarak; Decision Tree, Random Forest, Adaboost, Gradient Boosting, XGBoost kullanılmıştır. Araştırma sonucunda; iplik özelliklerine göre işlem parametresi tahmini için Lasso regression ve üretim kalitesi için XGBoost algoritmaları kullanılmış olup, on kat çapraz doğrulama testine dayanan sonuçlar ile modelin kalite seviyesi tahmininde % 90,8 doğruluk elde etmişlerdir. En iyi regresyon modelinin ortalama kare hatasını (MSE) %0.01' e düşürebileceğini göstermişlerdir (Lin vd., 2018).

Taur ve arkadaşları (2019) tarafından tekstil sektöründe kurutma işleminin tahmin ve analiz edilmesi üzerine yapılan çalışmanın amacı, kumaşların nem içeriği oranını tahmin etmek için birkaç makine öğrenimi modeli oluşturmaktır. Kumaşların kurutma prosesi senaryosu, kumaşı sekiz kurutma kutusundan geçiren bir bandı içerir ve kurutma sonrası kumaşın hedeflenen ölçüsü nem içeriği oranıdır. Kolaylık olması için sekiz kurutma kutusunun tümü, her çalışma için aynı sıcaklık derecesine sahiptir. Veri kümesi 117 adet veriden oluşmakta ve tür, genişlik, ağırlık, yoğunluk sıcaklığı, hız özelliklerine sahiptir. Çalışma iki aşamada gerçekleştirilmiştir; birinci aşama, veri özelliklerini incelemek ve uygun algoritma seçimini kolaylaştırmak adına temel modeller üzerinde bir parametre incelemesidir. Seçilen modeller Decision Forest Regression, Neural Networks ve K Nearest Neighbor regression modelleridir. İkinci aşamada, daha karmaşık complex ensemble öğrenme algoritmlarından olan Adaboost, Stacking Regressor, Voting Regressor metotlarının karşılaştırma sonuçları inclenmiştir. Modellerin ortak bir temelde nasıl performans gösterdiği RMSE ile belirlenmiş olup, Stacking Regressor'in en iyi model olduğu sonucuna ulaşmışlardır. (Taur vd., 2019).

Seçkin ve arkadaşları (2019) tarafından yapılan bu çalışmada, bir üretim sürecinin nasıl simüle edileceğine ve makine öğrenimi ile zaman serisi verilerinden regresyonun nasıl yapılacağına dair bir yöntem sunulmuştur. Çalışmanın oluşturulmasında K-Nearest Neighbors, Adaboost, Decision Tree, Random Forest ve Support Vector Regression algoritmaları kullanılmıştır. Modelin performansının ölçülmesinde k- çapraz doğrulama ve performans metrikleri (MAE, MSE, R^2) kullanılmıştır. Çalışma sonucunda, Support Vector Regression algoritması her parametre için en iyi tahmin sonucunu vermiştir (Seçkin vd., 2019).

Odabaş (2019) tarafından yapılan çalışmanın amacı, gömlek üretimi yapan bir tekstil firmasında üretim adetlerini etkileyen kumaş maliyeti, pamuk fiyatı, döviz kuru ve operasyonel maliyeti değişkenlerinin arasındaki ilişkileri çoklu doğrusal regresyon ile belirlemektir. Çalışmada kullanılan veriler bir tekstil firmasının 2011 Ocak – 2018 Aralık aylarında toplanan verilerden oluşmaktadır. Toplanan veriler Eviews ve Weka programlarında, talep tahmini analizi yöntemlerinden çoklu regresyon analizi ve zaman serileri analizi yöntemlerinden ağırlıklı hareketli ortalama yöntemi kullanılarak uygulanmıştır. Yapılan çalışma sonucunda Eviews programında çoklu doğrusal regresyon modeline göre en önemli değişkenin döviz kuru olduğu tespit edilmiştir (Odabaş, 2019).

Demir ve Dincer (2020) tarafından yapılan çalışmada, veri madenciliği ve makine öğrenmesi teknikleriyle bir tekstil firmasının ürettiği tekstil ürünlerinin kusurlu olup olmadığını incelemişlerdir. Çalışmada üretim hattında 250 farklı değişken ve 72959 satır veri bulunmaktadır. Çalışma Python programı kullanılarak yapılmış olup, lojistik regresyon ve KNN algoritmaları uygulanarak, modelin uygulanabilirliği ve başarı oranları değerlendirilmiştir. Modelin sonuçlarına göre, lojistik regresyon ve K-en yakın komşu algoritmalarının %90'ın üzerinde başarı oranı verdiği görülmüştür (Dincer & Demir, 2020).

Tozak (2021) tarafından yapılan çalışmada, tekstil sektöründe yer alan bir işletmenin satış verilerinin analizi yapılması ve tahmin edilmesi amaçlanmıştır. Çalışmada kullanılan veriler 2016 Ocak- 2019 Temmuz ayları arasında elde edilen 7281 adet gömlek satış verileridir. Veri setinin değişkenleri; siparişin ait olduğu müşteri, üretimin yapıldığı fabrika, sipariş grubunun departmanı, model numarası, üretimin yapıldığı ülke, sezonu ve siparişin

üretim termini, siparişin yılı ve ayı ve siparişin birim fiyatı ile aylık sevk adetidir. Uygulamada karar ağaçları algoritmaları kullanılmış olup, IBM SPSS modeler, Weka, RStudio ve Knime makine öğrenimi araçları kullanılarak programların model performansları ölçülmüştür. Model performanslarında; korelasyon katsayısı, MAE, RMSE, bağıl mutlak hata sonuçları karşılaştırmış ve çıkan sonucu göre modeli en iyi analiz eden programın Knime olduğu görülmüştür (Tozak, 2021).

6. Tekstil Sektöründe Bir Uygulama

Bir yapının temel süreçleri etrafında karar vermeyi oluşturan tekstil ürünlerine yönelik gelecekte oluşabilecek olan talebi tahmin etmek için doğru bir tahmin modeli oluşturulması çok önemlidir (Lorente-Leyva vd., 2021). Karmaşık verilerin analizinde ise veri madenciliğinin uygulama gücü birçok çalışmada kanıtlanmıştır (Taranto, 2021). Tekstil üretiminde veri madenciliği, makine öğrenimi ve yapay zekâ kavramları yeni değildir. Endüstriyel ve ticari alanlarda yaygın olarak kullanılan veri madenciliği ve makine öğrenimi araçları, üretim sorunlarının çözülmesi ve endüstriyel verilerden kuralları ve kalıpların çıkarılmasına kadar birçok uygulamada yararlanılmıştır.

Tekstil imalatında basit bir işlem veya standart bir ürün üretilmesinde bile büyük bir veri üretilir ve depolanır. Tekstil işlem parametreleri, lif ve iplik özellikleri, mukavemeti, kumaş performansı, hata tespiti vb. gibi birçok uygulama veri madenciliği ve makine öğrenimi teknikleri ile ortaya konabilmektedir. Ayrıca bu araçlarla birlikte yapay zekânın kullanımı ile kamera tabanlı bir denetim sistemi kurarak, ürünlerin görüntülerini gerçek zamanlı olarak yakalayabilir ve mevcut kumaş desen verileriyle karşılaştırılabilir.

Bu çalışmada, tekstil sektöründe faaliyet gösteren bir işletmenin aylık üretim verileri kullanılarak makine öğrenim teknikleri uygulanmıştır. Elde edilen veriler üzerinden analiz yapılırken üretimi etkileyen faktörlerde incelenmiştir. Çalışmada kullanılan veri setine hitap eden birden fazla teknik ve yazılım programı olduğu için, üretim verileri üzerinde aynı amaç üzerinden farklı makine öğrenimi algoritmaları ve programları kullanılmış olup, en iyi performansı sağlayan en iyi model ve yazılım programı bulunmaya çalışılmıştır.

Çalışmada kullanılan veriler, tekstil sektöründe faaliyet gösteren ve denim üretimi yapan bir firmanın üretim verilerini kapsamaktadır. 2017 yılının ilk üç ayını (Ocak, Şubat, Mart) kapsayan veri seti; gün, çalışan sayısı, çalışma saati, fazla mesai saati, toplam çalışma saati, günlük üretim, kişi başına üretim olmak üzere 75 adet veriden oluşmaktadır. Verinin yüksek miktarda olması, model tahmin gücünü ve güvenirliliğini etkilemesine rağmen, bu çalışmada az sayıda veri ile model kurulmaya çalışılmış ve kurulan her modelde yüksek model güvenirliği elde edilmiştir.

Veri setindeki değişkenlere ilişkin bilgileri aşağıda açıklanmıştır.

- ✓ Gün: 2017 yılının ilk üç ayı; Ocak, Şubat ve Mart kapsamaktadır.
- ✓ Çalışan sayısı: Normal çalışma saatlerinde çalışan kişi sayısını gösterir.
- ✓ Çalışma saati: Normal iş günündeki adam saat miktarıdır.
- ✓ Fazla mesai saati: İş yerinde çalışan personelin günlük olağan çalışma saatlerinin üzerinde çalıştığı saatleri ifade etmektedir.
- ✓ Toplam çalışma saati: Toplam çalışma süresi normal çalışma saatleri ve fazla mesai olarak belirtilir.
- ✓ Günlük üretim: Bir gün içinde ürün bandında üretilen ürün sayısını gösterir.
- ✓ Kişi başına üretim: Gün içinde çalışan bir kişinin üretime katkı sağladığı ortalama ürünü gösterir.

Çalışmada kullanılan değişkenlerin önemliliği doğrusal regresyon modeli altyapısına göre işlenmiştir. Çoklu doğrusal regresyon modelinin veri üzerindeki geçerliliği ve ayrıca verilerin daha doğru regresyon sonuçları vermesi adına değişkenler arasında çoklu bağıntı olup olmadığı kontrol edilmelidir. Bu doğrultuda modelin anlamlılığı Anova tablosuyla, çoklu bağıntı problemi ise varyans enflasyon faktörü (VIF) sonucuna ve tolerans değerlerine bakılarak karar verilmiştir. Şekil 6.1'de IBM SPSS Statistics programı ile elde edilen Anova tablosu ve regresyon katsayı değerlerine yer verilmiştir.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	25971,617	3	8657,206	2386,761	,000[b]
	Residual	257,530	71	3,627		
	Total	26229,147	74			

a. Dependent Variable: Production_per_emplooyer

b. Predictors: (Constant), Daily_production, Over_time, Working_time

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	215,657	3,601		59,884	,000		
	Working_time	-,001	,000	-,631	-37,459	,000	,488	2,050
	Over_time	-,001	,000	-,533	-43,348	,000	,916	1,092
	Daily_production	,038	,001	1,087	66,651	,000	,520	1,924

a. Dependent Variable: Production_per_emplooyer

Şekil 6.1 Anova ve katsayılar tablosu

Şekil 6.1'de yer verilen ilk kısımda yer alan Anova tablosu sonuçları incelendiğinde, çoklu doğrusal regresyon modelinin anlamlı ve önemli olduğu sonucuna varılır ($p=0,00<0,05$). Katsayılar tablosuna göre ise VIF değerleri beşten büyük ve tolerans değerleri 0,2'nin üstünde yer aldığı için çoklu bağıntının olmadığı sonucuna varılır (Ersöz, 2019). Regresyon modelinin anlamlı olması, bağımsız değişkenler arasında yüksek derecede ilişki olmamasına bağlıdır. Gün, çalışan sayısı ve toplam çalışma saati değişkenleri çoklu bağıntı oluşturduğu ve model hatasını arttırdığı için veri setinden çıkarılmıştır.

Regresyon analizi bir tahmin (Öngörüsel) analizi olup, bağımlı değişkenin bağımsız değişkenler yardımıyla tahmin edilmesini sağlar (Ersöz &Ersöz, 2019). Regresyon analizi ile değişkenler arasında oluşan ilişkiden bilgi elde edilebilir ve eğer ilişki var ise

regresyon bu ilişkinin gücü hakkında bilgi verebilir (Lukman & Natalina, 2019).

Şekil 6.2'de RStudio programında çoklu doğrusal regresyon modeli ekran çıktısı yer almaktadır.

```
Call:
lm(formula = .outcome ~ ., data = dat)

Residuals:
      Min        1Q    Median        3Q       Max
-0.074337 -0.010619 -0.000593  0.010486  0.073887

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.454168   0.002537  179.05   <2e-16 ***
working_time     -0.125578   0.003720  -33.75   <2e-16 ***
over_time        -0.096007   0.002637  -36.40   <2e-16 ***
daily_production  0.211381   0.003642   58.03   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01997 on 58 degrees of freedom
Multiple R-squared:  0.9891,    Adjusted R-squared:  0.9886
F-statistic:  1760 on 3 and 58 DF,  p-value: < 2.2e-16
```

Şekil 6.2 RStudio çoklu doğrusal regresyon modeli sonucu

Şekil 6.2'de yer verilen RStudio regresyon modeli sonucuna göre; tahmini regresyon katsayıları ve önem düzeyleri verilmiştir. Elde edilen çoklu doğrusal regresyon modelinin sabit katsayısı 0.454168 'dir. Çalışma zamanı ve fazla mesai saati değişkenleri kişi başına üretimde negatif bir etki yaratmaktadır. Çoklu doğrusal regresyon modeli sonucunda en önemli değişkeninin "günlük üretim" olduğu sonucuna varılmıştır.

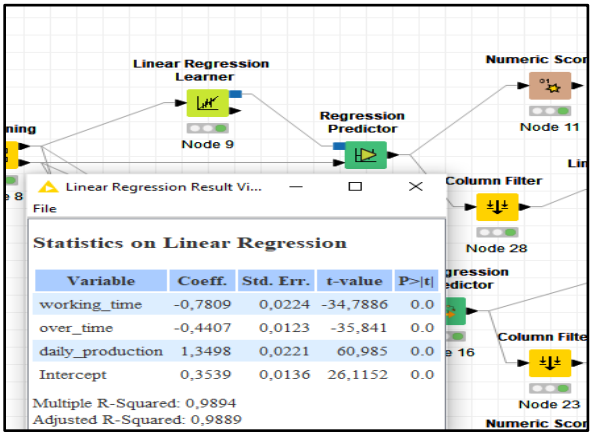
Şekil 6.3'de Python programından çoklu doğrusal regresyon modeli ekran çıktısı yer almaktadır.

```
model.intercept_ #sabit değer

array([2.13412689])

model.coef_ #bağımsız değişken katsayı

array([[-0.00079899, -0.00093322,  0.00038114]])
```

Şekil 6.3 Python çoklu regresyon modeli sonucu

Şekil 6.3'de yer alan çoklu doğrusal regresyon modeli incelediğinde, en önemli değişkenin "günlük üretim" olduğu sonucuna varılmıştır. Denklemi negatif yönde etkileyen değişkenler ise çalışma zamanı ve fazla mesai saati değişkenleridir.

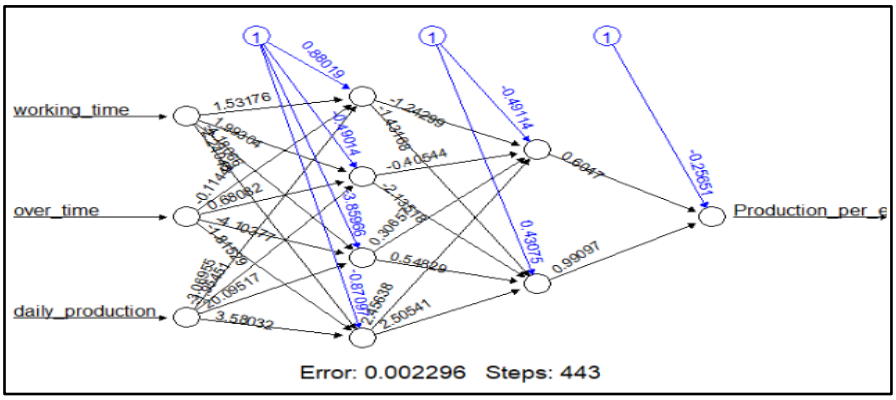
Şekil 6.4'de Knime programından çoklu doğrusal regresyon modeli ekran çıktısı yer almaktadır



Şekil 6.4 Knime çoklu regresyon modeli sonucu

Çoklu doğrusal regresyon modelleri Python, R ve Knime programlarında incelediğinde beklendiği gibi en önemli değişkenin "günlük üretim" olduğu görülmüştür.

Yapay sinir ağ yapısı gereği insan beyninin bilgi işleme sistemine benzemektedir. İnsan yapısında bulunan nöron hücrelerinin arasındaki sinaptik bağın dijital olarak modellenmesi yapay sinir ağlarıdır (Wang, 2003). Şekil 6.5'de RStudio YSA modeli sonucu yer almaktadır.



Şekil 6.5 RStudio YSA modeli sonucu

RStudio programı yapay sinir ağı modeli sonucu incelendiğinde modelin 443 adımda Şekil 6.5'de ekran resmi verilen şekle geldiği ve bu sinir ağını oluştururken 0.002296 hata yaptığı görülmektedir. Yapay sinir ağında bulunan hesaplamalar sonucunda kişi başına üretime (hedef değişken) gelen toplam mesaj topluluğunun sigmoid fonksiyonunun hesaplanması sonucu elde edilen değerin -0.25651 olduğu sonucuna varılmıştır.

Kurulan yapay sinir ağ modelinin elde edilen performans değerleri sonuçları; RMSE değeri 0.0085, R^2 değeri ise 0.997 'dir. Model hatasının oldukça düşük olması ve açıklama oranının yüksek olmasından dolayı modelin doğru ve güvenilir olduğu söylenebilir.

Şekil 6.6'da Python YSA modeli sonucu yer almaktadır.

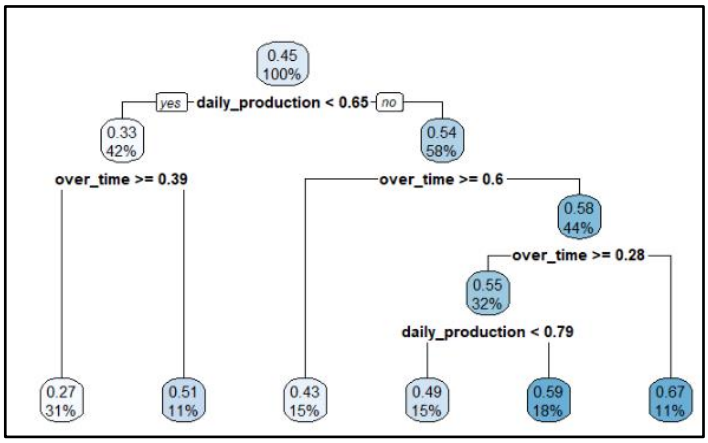
```
mlp_model = MLPRegressor(hidden_layer_sizes=(4,2),activation="relu" ,random_s

mlp_model

MLPRegressor(hidden_layer_sizes=(4, 2), max_iter=2000, random_state=99,
             solver='lbfgs')

•••

•••

y_pred = mlp_model.predict(X_test_scaled)
np.sqrt(mean_squared_error(y_test, y_pred))

0.08378773007212008

mlp_model.score(X_test_scaled, y_test)

0.7071912032599912
```

Şekil 6.6 Python YSA modeli sonucu

Şekil 6.6'da yer alan Python kodları incelendiğinde, oluşturulan modelin dört katmanlı ve iki gizli katmana sahip bir yapay sinir ağı tasarlandığı görülmektedir. Yapay sinir ağının hata oranını düşürmek ve daha doğru bir sonuç vermesini sağlamak için çözücü olarak "lbfgs" seçilmiştir. Bunun sebebi "lbfgs" çözücüsünün küçük veri setlerinde daha hızlı ve daha güçlü performans göstermesidir (Kuhlman, 2009). Modelin verdiği sonuçlar %70'lik bir doğruluk oranına sahipken, modelin hatası 0.08378'dir.

Knime programında oluşturulan yapay sinir ağı model sonuçlarına göre %97 model güvenirliliği ile 0.034 RMSE hata değerine sahiptir. Modelin hata değerinin oldukça düşük olduğu tespit edilmiştir.

Karar ağacı modeli basitlik, anlaşılabilirlik, parametresiz ve karma tip verileri işleyebilen özellikleri nedeniyle en başarılı öğrenme algoritmalarından biridir (Su & Zhang, 2006). Şekil 6.7'de RStudio programı kullanılarak elde edilen karar ağacı modelinin ekran görüntüsü yer almaktadır.



Şekil 6.7 RStudio karar ağacı modeli sonucu

Şekil 6.7'de yer verilen karar ağaçları modelinin görüntüsü incelendiğinde günlük üretim 0,65'ten küçük olduğunda tahmini kişi başına üretim %42 oranla 0,3, büyük olduğunda ise tahmini kişi başına üretim %58 oranla 0,5 olarak gerçekleşmektedir.

Fazla mesai 0,6 'ya eşit ve büyükse %15 oranla tahmini kişi başına üretim 0,4, fazla mesai 0.6'ya eşit ve büyük değilse %44 oranla tahmini kişi başına üretim 0,6 olarak elde edilir. Fazla mesai 0,28'e eşit ve büyükse %32 oranla tahmini kişi başına üretim 0.5, fazla mesai 0,28'e eşit ve büyük değilse %11 oranla tahmini kişi başına üretim 0,7 olarak elde edilir. Günlük üretim 0,79 'dan küçükse %15 oranla tahmini kişi başına üretim 0,5, günlük üretim 0,79'dan büyükse %18 oranla tahmini kişi başına üretim 0,6 olarak elde edilir.

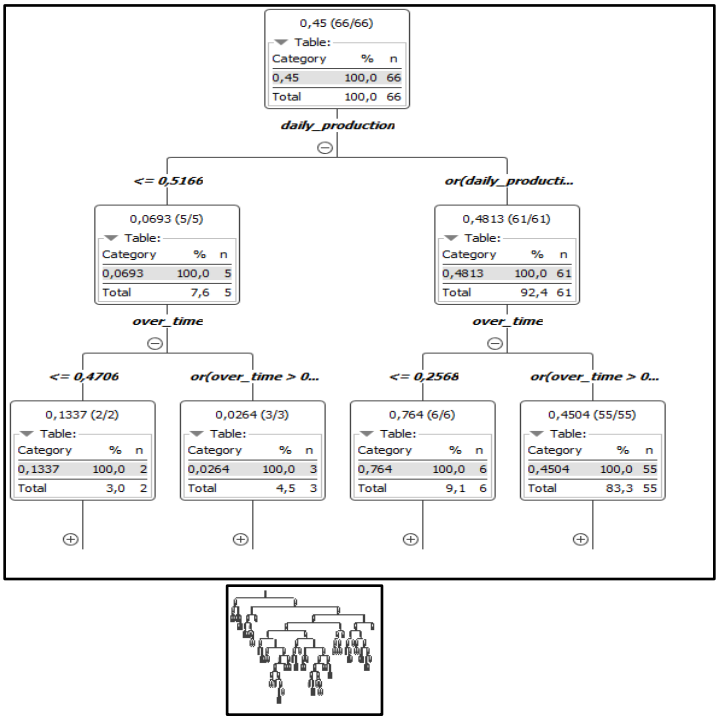
Kurulan karar ağacı modelinin elde edilen performans değerleri sonuçları; RMSE değeri 0,11835, R^2 değeri ise 0,59205 'tir. Bu sonuç çoklu doğrusal regresyon ve yapay sinir ağı modellerine göre karar ağacı modeli model performans sonuçlarına göre oldukça düşük tespit edilmiştir.

Şekil 6.8'de Python programı kullanılarak elde edilmiş karar ağacı modeli sonucu yer almaktadır.

```
from sklearn.metrics import mean_squared_error
np.sqrt(mean_squared_error(y_test, y_pred))

0.05452828012447613

dr_model.score(X_test, y_test)

0.8759870982093203
```

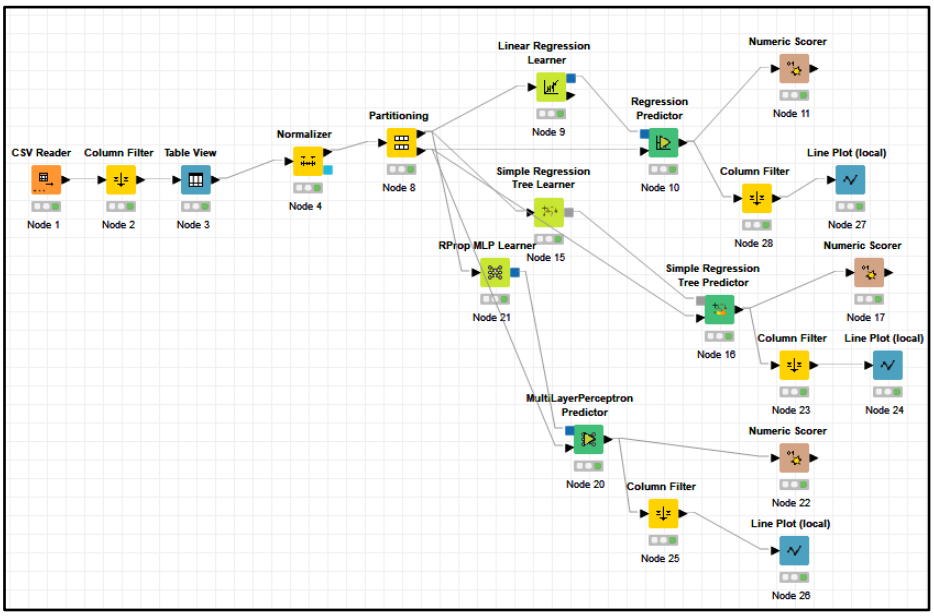
Şekil 6.8 Python karar ağacı modeli sonucu

Şekil 6.8'de Python programı kullanılarak oluşturulan karar ağacı modeli incelediğinde, kurulan modelin %87 oranında güvenilir olduğu ve model hatasının 0,0545 oranına sahip olduğu sonucuna varılmıştır. Modelin açıklama oranı kabul edilebilir bir düzeye sahiptir. Model 0,0545 hataya sahip olduğu görülmektedir ve hata oranı çok yüksek olmadığından modelin verdiği sonuçların doğruluk oranı yüksek olduğu sonucuna varılmıştır. Şekil 6.9'da Knime programı ile oluşturulan karar ağacı modelinin görüntüsü yer almaktadır.



Şekil 6.9 Knime karar ağacı modeli sonucu

Knime karar ağacı modeli görüntüsü incelendiğinde; günlük üretim 0,51'den küçük ve eşit olursa tahmini kişi başına üretim 0,07, küçük ve eşit olmazsa tahmini kişi başına üretim 0,5 olarak elde edilir. Fazla mesai 0,47'den küçük ve eşit olursa tahmini kişi başına üretim 0,1, küçük ve eşit olmazsa tahmini kişi başına üretim 0,03 olarak elde edilir. Fazla mesai 0,25'ten küçük ve eşit olursa tahmini kişi başına üretim 0,8, küçük ve eşit olmazsa tahmini kişi başına üretim 0,5 olarak elde edilir. Şekil 6.3.4'te Knime programında oluşturulan modellerin görüntüsü yer almaktadır.



Şekil 6.10 Knime modellerin ekran görüntüsü

6.1. Sınıflayıcı Tekniklerin Performans Karşılaştırması

Bu çalışmada tekstil denim üretimi verileri üzerinden makine öğrenimi algoritmaları ile RStudio, Python ve Knime programları kullanılarak doğrusal regresyon, yapay sinir ağı ve karar ağacı teknikleri uygulanmıştır. Tahmini model sonuçlarının karşılaştırılmasında R^2 ve RMSE değerlerinden yararlanılmıştır. Tablo 6.1.1'de modellerin performans değerlerinin sonuçlarının karşılaştırılması yer almaktadır**.**

Tablo 6.1.1 Modellerin performans değerlerinin karşılaştırılması

	Tahmini Model Performansları					
	Regresyon		**Yapay Sinir Ağı**		**Karar Ağacı**	
Programlar	**RMSE**	**R^2**	**RMSE**	**R^2**	**RMSE**	**R^2**
R	0.0193	0.9891	0.0085	0.9978	0.1210	0.5734
Python	0.0135	0.9923	0.0837	0.7072	0.0545	0.8760
Knime	0.0160	0.9940	0.0340	0.9740	0.0740	0.8720

Tekstil üretim verisine; RStudio, Python ve Knime farklı makine öğrenimi programları ve farklı teknikler uygulanmıştır. Çoklu doğrusal regresyon modeli sonuçları her üç makine öğrenimi program ile karşılaştırıldığında; Knime programında model belirtme katsayısının diğer programlara göre yüksek olduğu ve hata oranları incelendiğinde ise her üç program sonuçlarının yakın olduğu görülmüştür. Bu sonuçlara göre çoklu doğrusal regresyon modeli uygulamasında, Python ve Knime programlarının kullanımı tercih edilebilir.

Yapay sinir ağı modeli makine öğrenimi program sonuçları karşılaştırıldığında; R programı hata oranının, diğer programlara göre oldukça düşük olduğu görülmektedir. Ayrıca R programı belirtme katsayısının 0,99 ile oldukça yüksek olduğu tespit edilmiştir. Bu sonuçlara göre yapay sinir ağı modeli uygulamasında R programı tercih edilebilir.

RStudio, Python ve Knime farklı makine öğrenimi programlarında Karar ağacı modeli tahmini sonuçları incelendiğinde; Python ve Knime programlarının tahmini model hatalarının düşük ve model belirtme katsayılarının yüksek olduğu görülmüştür. Karar ağacı uygulamasında Python ve Knime makine öğrenimi programlarının kullanımı tercih edilebilir.

7. Sonuç veTartışma

Günümüzde firmalar için hızla artan rekabet ile pazarda güçlü kalarak ve sürekliliği sağlamak oldukça zor olmaya başlamıştır. Tekstil işletmeleri farklı yöntemler ile rakiplerine üstünlük sağlayıp, yeni müşteriler kazanarak sürekli gelişim ilkesini benimsemeleri gerekmektedir. Tekstil sektörü, Türkiye'nin önde gelen istihdam ve ihracat alanında rekabet oranı yüksek olan sektörlerden biridir. Yeni teknolojiler firmaların rekabet üstünlüğünü sağlamasında önemli roller üstlenmektedir. Makine öğrenimi ve veri madenciliği yaklaşımları yeni teknolojiler

arasında yer almaktadır. Bu sektörde başarıyı yakalayabilmek için önemli noktalardan biri de gelecek ile ilgili öngörülerde bulunabilmektir. Firmaların gelecek ile ilgili doğru ve güvenilir bilgiler elde edebilmesi, bulundukları pazarda büyük avantaj sağlayacaktır. Tekstil firmalarında söz konusu teknolojileri kullanarak; verilerin doğru şekilde analiz edilmesi, verilerden anlamlı yapıların ortaya konması ve geleceğe yönelik planlamalar ve stratejik kararlar alınabilir.

Bu çalışmada veri madenciliği ve makine öğrenmesi kavramları ve diğer disiplinler ile ilişkisi açıklanmış ve makine öğrenmesi ve diğer ilgili disiplinlerin gelişimi verilmiştir. Ayrıca genellikle kavram karmaşasına neden olan veri madenciliği ve makine öğrenmesi disiplinleri karşılaştırılmıştır. Uygulama bölümünde tekstil sektörünün denim üretiminde faaliyet gösteren bir işletmenin aylık üretim verileri kullanılarak makine öğrenimi ve veri madenciliği teknikleri uygulanmıştır. Elde edilen veriler üzerinden analiz yapılırken üretimi etkileyen faktörlerde incelenmiştir. Çalışmada kullanılan veri setine uygun birden fazla teknik ve program olduğu için üretim verileri üzerinde aynı amaç üzerinden farklı makine öğrenimi algoritmaları ve programları kullanılmış olup, en iyi performansı sağlayan en iyi model ve en iyi program bulunmaya çalışılmıştır. Çalışma sonucunda güvenilir ve daha doğru bir sonuç elde edebilmek adına RStudio, Python, Knime olmak üzere üç farklı program yardımı ile sınıflayıcı tekniklerden; çoklu doğrusal regresyon, yapay sinir ağları ve karar ağaçları algoritmaları kullanılarak makine öğrenimi yapılmıştır.

Araştırma sonucuna göre sınıflayıcı tekniklerin kullanımında, tahmini model performansları her programda farklı tespit edilmiştir. Çoklu doğrusal regresyon modeli sonuçları her üç program ile karşılaştırıldığında; model belirtme katsayısı yüksek ve hata oranı düşük olduğunun Python ve Knime programlarının tercih edilmesi önerilmektedir. Yapay sinir ağı modeli program sonuçları karşılaştırıldığında, model hata oranın düşük ve belirtme katsayısının yüksek olduğu R programının tercih edilmesi önerilmektedir. Karar ağacı modeli uygulamasında ise yine model hatası düşük ve belirtme katsayısı yüksek olan Python ve Knime programlarının tercih edilmesi önerilmektedir. Ayrıca tüm makine öğrenimi programları ve veri madenciliği teknikleri birlikte değerlendirildiğinde en iyi tahmini model sonuçlarını veren programın Python programı olduğu tespit edilmiştir. Bunu sırasıyla Knime ve R program izlemiştir.

Literatürde veri madenciliği ve makine öğrenimi teknikleri hemen hemen tüm sektörlerde uygulanmıştır. Tekstil sektöründe ise farklı yöntemler ve farklı makine öğrenimi programı çalışmaları mevcuttur. Ancak bu çalışmanın; tekstil sektöründe RStudio, Python ve Knime gibi farklı makine öğrenimi programlarında; çoklu doğrusal regresyon, yapay sinir ağı ve karar ağacı olmak üzere makine öğrenimi tekniklerinin kullanılıp, tahmini model performanslarının karşılaştırılması açısından literatüre katkı sağlayacağı değerlendirilmektedir. Ayrıca çalışmanın günümüz işletme problemlerinin çözümünde, iş geliştirmede ve müşterilerin ve olası sonuçları hakkında daha fazla bilgi edinmesine yardımcı olan; veri madenciliği, makine öğrenimi, veri bilimi ve yapay zekâ gibi disiplinlerin daha açık olarak tanımlanmasına yardımcı olacağı değerlendirilmektedir.

Günümüz işletmelerinin çoğu yazılımları sadece verileri toplamak ve bulgular ve raporlar sağlamak için kullanmaktadır. Ancak 21. yüzyılın yeni teknolojilerin ve daha akıllı ürünlerin ortaya çıkmasıyla oluşan büyük bir verilerin okunması ve anlamlandırılmasına yönelik; veri madenciliği, makine öğrenimi ve yapay zekâ uygulamalarının kullanılması işletmelere yüksek fayda sağlayacaktır. Gelecekte bu disiplinleri işletmenin temel yetenekleri haline getirmek, personelini eğitimlerle desteklemek, iş akışlarını buna göre entegre etmek ve en sonunda tüketicilere kişiselleştirilmiş hizmet sağlamak, işletmenin kârını artırmanın ve değer yaratmanın anahtarı olacaktır.

Kaynakça

Accentura (2021). *Artificial intelligince*. Erişim: 12 Eylül. 2021. https://www.accenture.com/in-en/insights/artificial-intelligence-summary-index.

AI, D. (n.d.). *Association Learning*. Deep AI. https://deepai.org/machine-learning-glossary-and-terms/association-learning

Algorithmia. (2020). *2020 State of Enterprise Machine Learning*. https://info.algorithmia.com/hubfs/2019/Whitepapers/The-State-of-Enterprise-ML-2020/Algorithmia_2020_State_of_Enterprise_ML.https://algorithmia.com/state-of-ml.

Analytics Insigth (2021). *Top Machine learning tools used by experts in 2021.* https://www.analyticsinsight.net/top-machine-learning-tools-used-by-experts-in-2021. Erişim 12 Ekim, 2021.

Angrist, J. D., & Pischke, J. S. (2008). *Mostly Harmless Econometrics:* An Empiricist's Companion. Princeton University Press.

Amershi, S., Begel, A., Bird, C., DeLine, R., Gall, H., Kamar, E., Nagappan, N., Nushi, B., & Zimmermann, T. (2019). *Software engineering for machine learning: a case study.* 2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP), pp. 291–300.

Analytic Insight (2021). *Top 10 data mining algorithms* 2021. Erişim: 21 Temmuz 2021. https://www.analyticsinsight.net/top-10-data-mining-algorithms-2021/

Bergstra, J., Ca, J. B., & Ca, Y. B. (2012). *Random search for hyper-parameter optimization Yoshua Bengio.* Journal of Machine Learning Research (Vol. 13). http://scikit-learn.sourceforge.net.

Birkhold, C., Tamagnini, P., & Schmid, S. (2019). *How to automate machine learning | KNIME*. https://www.knime.com/blog/how-to-automate-machine-learning.

Breck, E.; Cai, S.; Nielsen, E.; Salib, M.; Sculley, D. *The ML test score: A rubric for ML production readiness and technical debt*

reduction (2017). *In Proceedings of the 2017 IEEE International Conference on Big Data* (Big Data). Boston. MA, USA, 11–14 December. pp. 1123–1132.

Brownlee, J. (2020). *6 dimensionality reduction algorithms with python*. https://machinelearningmastery.com/dimensionality-reduction-algorithms-with-python/

Buffet, O., Pietquin, O., & Weng, P. (2020). *Reinforcement learning*. In arXiv (Vol. 3, issue 3, p. 1448). arXiv. https://doi.org/10.4249/scholarpedia.1448.

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

Chollet, F. (2017). *Deep learning with python*. Manning Publications.

Dataversity website (2021). A brief history of machine learning.

Erişim: 05 Eylül. 2021. https://www.dataversity.net/a-brief-history-of-machine-learning/

Educba website (2021). Data mining vs machine learning. Erişim: 21 Eylül 2021. https://www.educba.com/data-mining-vs-machine-learning/

Eisler, S., & Meyer, J. (2020). Visual analytics and human ınvolvement in machine learning. *ArXiv, abs/2005.06057*.

Ersöz, F. (2019). *SPSS ile istatiksel veri analizi*. Seçkin Yayıncılık.Ankara

Ersöz, F., & Ersöz, T. (2019). *Veri madenciliği teknikleri ve uygulamaları.* Seçkin Yayıncılık.Ankara

Fayyad, U.M., Piatetsky-Shapiro, G., and Smyth, P. (1996). *Knowledge discovery and data mining: towards a unifying framework.* In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96). AAAI Press, 82–88.

Fazakis, N., Karlos, S., Kotsiantis, S., & Sgarbas, K. (2016). *Self-Trained LMT for semisupervised learning*. Computational Intelligence and Neuroscience, 2016, 3057481. https://doi.org/10.1155/2016/3057481.

Forbes (2018). *5 Entrepreneurs on the rise in AI.* Erişim: 12 Eylül. https://www.forbes.com/sites/insights-intelai/2018/11/29/5-entrepreneurs-on-the-rise-in-ai/?sh=7c79e67cf99f

Fox, J., & Andersen, R. (2005). *Using the R statistical computing environment to teach social statistics Cources.* http://cran.r-project.org/.

Ersoz, F., Guler, E., Ersoz, T. (2017). *Knowledge discovery and data mining techniques in textile industry*. International Journal of Computer and Information Engineering. Vol. 11, No 7. 923-927.

Guo Yufeng. *The 7 steps of machine learning. 2017.* In: towardsdatascience.com

Gürsakal, N. (2018). *Makine Öğrenmesi*. Dora yayınları.

IBM Software (2021). *Machine learning.* Erişim: 28 Temmuz 2021. IBM Software Website: https://www.ibm.com/tr-tr/cloud/learn/machine-learning

Hyndman, R. J., & Koehler, A. B. (2006). *Another look at measures of forecast accuracy. International Journal of Forecasting*, *22*(4), 679–688. https://doi.org/10.1016/j.ijforecast.2006.03.001

Javapoint website (2021). Erişim: 16 Eylül 2021. https://www.javatpoint.com/data-mining-vs-machine-learning

KDnuggets website (2018). *The 7 steps of machine learning*. Erişim: 04 Eylül 2021. https://www.kdnuggets.com/2018/05/general-approaches-machine-learning-process.html

KDnuggets website (2020). *History of data mining*. Erişim: 20 Ekim 2021.https://www.kdnuggets.com/2016/06/rayli-history-data-mining.html

KDnuggets website (2021). *10 best data mining tools.* Erişim: 04 Eylül 2021. https://www.kdnuggets.com/2021/01/machine-learning-algorithms-2021.html.

Knowlab website (2021). Erişim: 10 Ekim 2021. from https://knowlab.in/machine-learning-vs-data-mining-whats-the-difference/

Kubat, M., Bratko, I., & Michalski, R. (1996). *A Review of Machine Learning Methods*.

Kuhlman, D. (2009). *A Python Book: Beginning Python, Advanced Python, and Python Exercises*. http://www.davekuhlman.org

Kulin, Merima & Kazaz, Tarik & De Poorter, Eli & Moerman, Ingrid. (2021). *A survey on machine learning-based performance ımprovement of wireless networks: PHY, MAC and Network Layer*. Electronics. 10. 318. 10.3390/electronics10030318.

Legendre, A.M. (1805). *Nouvelles méthodes pour la détermination des orbites des comètes*. Firmin Didot. Paris, 1805. "Sur la Méthode des moindres quarrés" appears as an appendix.

Lin, J.-Y., Lee, C.-Y., & Chang, R.-I. (2018). *Improve quality and efficiency of textile process using data-driven machine learning in industry 4.0. International Journal of Technology and Engineering Studies*, 4(2). https://doi.org/10.20469/ijtes.4.10004-2.

Lorente-Leyva, L. L., Alemany, M. M. E., Peluffo-Ordóñez, D. H., & Araujo, R. A. (2021). *Demand forecasting for textile products using statistical analysis and machine learning algorithms*. *Lecture Notes in Computer Science* (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 12672 LNAI, 181–194. https://doi.org/10.1007/978-3-030-73280-6_15

Lovell MC. *Data mining*. Rev Econ Stat 1983, 65:1– 11.

Lukman, I., & Natalina. (2019). *Association rules and regression linear model of the groundwater population by the evaluation of uranium*. *MATEC Web of Conferences*, *270*, 04017. https://doi.org/10.1051/matecconf/201927004017.

McCorduck, Pamela (2004), *Düşünen Makineler (2. baskı).* Natick, MA: AK Peters Ltd.. ISBN 978-1-56881-205-2, OCLC 52197627.

Malaca, P., Luis, ·, Rocha, F., Gomes, · D, Silva, J., Germano Veiga, ·, & Luis, B. (2019). *Online inspection system based on machine learning techniques: real case study of fabric textures classification for the automotive industry*. *J Intell Manuf*, *30*, 351–361. https://doi.org/10.1007/s10845-016-1254-6.

Mariscal, G., Marbán, Ó., & Fernández, C. (2010). *A survey of data mining and knowledge discovery process models and methodologies. The Knowledge Engineering Review*, 25, 137–166.

Martínez-Plumed F, Contreras-Ochando L, Ferri C, Orallo JH, Kull M, Lachiche N, Ramírez-Quintana MJ, Flach PA (2019) *CRISP-DM twenty years later: from data mining processes to data science trajectories*. IEEE Trans Knowl Data Eng 33(8):3048–3061.

May, S. (2019). *6 Reasons to learn data science with python | benefits of python data science training.* https://www.zeolearn.com/magazine/benefits-of-learning-data-science-with-python.

Mayo, M. (2018). *Frameworks for Approaching the Machine Learning Process-KDnuggets*. Erişim: 12 Eylül 2021. https://www.kdnuggets.com/2018/05/general-approaches-machine-learning-process.html

Mayo, Matthew. *The 7 Steps of Machine Learning*, In: KDnuggets.com,2018

Mitchell Guthrie, P. (2014). *Looking backwards, looking forwards: SAS, data mining, and machine learning*. https://blogs.sas.com/content/subconsciousmusings/2014/08/22/looking-backwards-looking-forwards-sas-data-mining-and-machine-learning/#prettyPhoto/0/)

Mitchell, T.; Buchanan, B.; DeJong, G.; Dietterich, T.; Rosenbloom, P.; Waibel, A. *Machine learning*. Annu. Rev. Comput. Sci. 1990, 4, 417–433.

Mozafary, V., & Payvandy, P. (2014). *Application of data mining technique in predicting worsted spun yarn quality. Journal of the Textile Institute*, *105*(1), 100–108.

https://doi.org/10.1080/00405000.2013.812552

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). *Definitions, methods, and applications in interpretable machine learning. Proceedings of the National Academy of Sciences of the United States of America*. *116*(44), 22071–22080. https://doi.org/10.1073/pnas.1900654116.

Özbek, A., Akalın, M. (2011). The prediction of Turkey's denim trousers export to Germany with ANN models. *Tekstil ve Konfeksiyon.* 21(4):313-322. İstanbul.

Patel, K., Fogarty, J., Landay, J., and Harrison, B. (2008). *Investigating statistical machine learning as a tool for software development. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '08). Association for Computing Machinery, New York, NY, USA, 667–676.

Piatestsky, G. (2019). *Python leads the 11 top Data Science, Machine Learning platforms: Trends and Analysis - KDnuggets*. https://www.kdnuggets.com/2019/05/poll-top-data-science-machine-learning-platforms.html/2

Portugal, I., Alencar, P., & Cowan, D. (2018). *The use of machine learning algorithms in recommender systems: A systematic review.* In *Expert Systems with Applications* (Vol. 97, pp. 205–227). Elsevier Ltd. https://doi.org/10.1016/j.eswa.2017.12.020

Ryu, S., Lee, H., Lee, D. K., & Park, K. (2018). *Use of a machine learning algorithm to predict individuals with suicide ideation in the general population*. *Psychiatry Investigation*, *15*(11), 1030–1036. https://doi.org/10.30773/pi.2018.08.27

SAS software (2021). Retrieved on september 14, 2021 from IBM Software Website: https://www.sas.com /en_us/insights/analytics/machine-learning.html

Seagate Technology. (2020). *SEAGATE*. Seagate. https://www.seagate.com/tr/tr/our-story/data-age-2025/

Selvanayaki, M., Vijaya, M. S., Jamuna, K. S., & Karpagavalli, S. (2010). *Supervised learning approach for predicting the quality of cotton using WEKA.* Communications in Computer and Information Science, *70*, 382–384. https://doi.org/10.1007/978-3-642-12214-9_61

Shafiq, Muhammad & Tian, Zhihong & Bashir, Ali & Jolfaei, Alireza. (2020). *Data mining and machine learning methods for sustainable smart cities traffic classification: A Survey.* Sustainable Cities and Society. 60. 10.1016/j.scs.2020.102177.

Sherarer, C. (2000). *The CRISP-DM model:the new blueprint for data mining. Journal of Data Warehousing,* 5(4), 1–15.

Softwaretestinghelp website (2021). Data mining vs machine learning vs artificial intelligence vs deep learning. Erişim: 04 Ağustos 2021. https://www.softwaretestinghelp.com/data-mining-vs-machine-learning-vs-ai/

Sotirios P. Chatzis, Vassilis Siakoulis, Anastasios Petropoulos, Evangelos Stavroulakis, Nikos Vlachogiannakis (2018). Forecasting stock market crisis events using deep and statistical machine learning techniques. *Expert Systems with Applications*. Vol. 112.353-371.

Studer, Stefan & Bui, Binh & Drescher, Christian & Hanuschkin, Alexander & Winkler, Ludwig & Peters, Steven & Müller, Klaus-Robert. (2021). *Towards CRISP-ML(Q): A machine learning process model with quality assurance methodology.* machine learning and knowledge extraction. 3. 392-413. 10.3390/make3020020.

Su, J., & Zhang, H. (2006). *A fast decision tree learning algorithm introduction and related work*. www.aaai.org

Sumathi, S., Sivanandam S.N., "*Data mining tasks, techniques and applications, studies in computational ıntelligence (SCI)*", Springer-Verlag, Berlin.189-216.

Szepesv´ari, C. (2009). *Algorithms for Reinforcement Learning*. Morgan & Claypool Publishers.

Taranto-Vera, G., P. Galindo-Villardón, J. Merchán-Sánchez-Jara, J. Salazar-Pozo, A. Moreno-Salazar and V. Salazar-Villalva, 2021. *Algorithms and software for data mining and machine learning: A critical comparative view from a systematic review of the literature*. J. Supercomputing. Vol. 2021.

Tanzeel U. Rehman, Md. Sultan Mahmud, Young K. Chang, Jian Jin, Jaemyung Shin (2019). *Current and future applications of statistical machine learning algorithms for agricultural machine vision systems*, Computers and Electronics in Agriculture. Volume 156. Pages 585-605,

IOBE. (2021). *TIOBE - The software quality company*. https://www.tiobe.com/tiobe-index/Tiwari, A., & Sekhar, A. K. T. (2007). Workflow based framework for life science informatics. In *Computational Biology and Chemistry* (Vol. 31, Issues 5–6, pp. 305–319). https://doi.org/10.1016/j.compbiolchem.2007.08.009

Wang, S.-C. (2003). *Artificial Neural Network. Interdisciplinary Computing in Java Programming.* 81–100. https://doi.org/10.1007/978-1-4615-0377-4_5

Wen, H., & Gu, Q. (2014). *The elements of supply chain management in new environmental era.* Lecture Notes in Electrical Engineering. 242 LNEE(VOL. 2), 867–880. https://doi.org/10.1007/978-3-642-40081-0_74

Yufeng, G. (2017). *The 7 Steps of Machine Learning* (pp. 1–13). https://livecodestream.dev/post/7-steps-of-machine-learning/