Avrupa Bilim ve Teknoloji Dergisi
Özel Sayı 32, S. 227-234, Aralık 2021

Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
Special Issue 32, pp. 227-234, December 2021

Research Article

Veri Madenciliği Yöntemleri İle İşveren Sektörünün Sınıflandırılması

Elvan Kübra Doğan[1], Arafat Şentürk[2*]

[1]Düzce Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Düzce, Türkiye, (ORCID: 0000-0002-5530-9385), elvankubra95@gmail.com
[2]Düzce Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Düzce, Türkiye, (ORCID: 0000-0002-9005-3565), arafatsenturk@duzce.edu.tr

(International Conference on Design, Research and Development (RDCONF) 2021 – 15-18December 2021)

(**DOI:**10.31590/ejosat.1039844)

ATIF/REFERENCE: Doğan, E.K. & Şentürk A. (2021). Veri Madenciliği Yöntemleri İle İşveren Sektörünün Sınıflandırılması. *Avrupa Bilim ve Teknoloji Dergisi*, (32), 227-234.

Öz

Veri madenciliği algoritmalarının kullanımına hazır hale getirilmesi için, "Kaggle'da Veri Bilimi Anketi" isimli veri seti, veri madenciliği problem tanıma aşaması ile analiz edilmiştir. Analiz edilen veri seti ve karar verilen hedef doğrultusunda veri ön işleme aşaması da gerçekleştirilmiştir. Ön işleme aşaması çıktısı olarak elde edilen işlenmiş veri seti, veri madenciliği teknolojisi sınıflandırma yöntemine ait C4.5, Rastgele Orman ve K-En Yakın Komşu Algoritmaları ile modellenmiştir. Bu üç algoritmanın model başarı oranları hesaplanmıştır. Aralarındaki başarı oranı sapma değerleri analiz edilerek sapma değerlerine sebep olan durumlara değinilmiştir. Model başarı oranını etkileyen durumlara farklı bir açıdan daha bakılabilmesi için, bu üç algoritma ile yeni bir modelleme daha gerçekleştirilmiştir. İşlenmiş veri seti için karar verilen üç algoritma ve bu algoritmaların modelleme kriterleri ile gerçekleştirilen modelleme işlemi, orijinal veri seti kullanılarak da gerçekleştirilmiştir. İşlenmiş veri seti kullanılarak elde edilen modellerin başarı oranı hesaplamaları, orijinal veri seti için de hesaplanmış olup kıyaslamaları gerçekleştirilmiştir. Kullanılan veri seti, karar verilen yöntem, algoritma, algoritma kriter değerleri gibi model başarısını etkileyen etmenlerin kıyaslama işlemleri, gerçekleştirilen modelleme uygulamaları sayesinde somutlaştırılarak aktarılmıştır. Elde edilen bu kıyaslamalı örnekler referans alınarak, model başarı oranını etkileyen etmenler değerlendirilmiş olup, veri seti nitelik analizi ve Veri Madenciliği süreçleri hakkında sırasıyla çıkarımlar gerçekleştirilmiştir.

Anahtar Kelimeler: Veri Önişleme, C4.5, Rastgele Orman, KNN Algoritması, Karışıklık Matrisi, Model Başarı Oranı.

Classification of Employer Industry with Data Mining Methods

Abstract

In order to make data mining algorithms ready for use, the data set named "Data Science Questionnaire in Kaggle", was analyzed in the data mining problem recognition stage. In line with the analyzed data set and the decided target, the data preprocessing stage was also carried out. The processed data set obtained as the output of the pre-processing stage is modeled by C4.5, Random Forest and K-Nearest Neighbor classification algorithms of data mining technology. The model success rates of these three algorithms were calculated. The success rate deviation values between them were analyzed and the situations that caused the deviation values were mentioned. In order to look at the issues affecting the model success rate from a different perspective, new modeling was performed with these three algorithms. The modeling process, which was carried out with the three algorithms decided for the processed data set and the modeling criteria of these algorithms, was also carried out using the original data set. The success rate calculations of the models obtained using the processed data set were also calculated for the original data set and their comparisons were made. The comparison of the factors affecting the success of the model, such as the data set used, the method decided, algorithm, algorithm criterion values, were embodied and expressed thanks to the modeling applications carried out. By taking these comparative examples as a reference, the factors affecting the model success rate were evaluated, and inferences were made about the data set quality analysis and data mining processes, respectively.

Keywords: Data Preprocessing, C4.5, Random Forest, KNN Algorithm, Confusion Matrix, Model Success Rate.

* Sorumlu Yazar: elvankubra95@gmail.com

1. Giriş

Veri Madenciliği teknolojisi işleyiş ve alt yapı olarak bilişim sektörüne ait olsa da, bu teknolojinin hizmet verdiği sektörler günden güne artmaya devam etmektedir (Söyler, 2014). Veri Madenciliği kavramının, literatürde çalışma alanı araştırması gerçekleştiğinde Akademik, Bankacılık, Sigortacılık, Telekomünikasyon, Perakende, Pazarlama, İlaç, Sağlık, Endüstri, Mühendislik gibi birçok alanda,örnek uygulamalar ile karşılaşılmaktadır (Savaş et al., 2012). Hastalık teşhisi, kredi onay/red durumları, satış stratejileri vb. gibi örnekler literatürde sık rastlanılan Veri Madenciliği uygulamalarıdır(Yücebaş, 2018)(Sönmez, 2015).

Veri Madenciliği kavramı, hizmet sunarken birçok alana hitap eden bir teknoloji olduğu için, problemlere uygun modellemeler gerçekleştirebilen ve hedeflenen başarı oranı ile sonuçlara ulaşabilen bir alt yapıya sahip olması gerektiği düşünülmektedir. Farklı çalışma alanlarına hizmet edebiliyor olması, Veri Madenciliği teknolojisinin çeşitli problemler ile karşılaşabilmesine sebep olacağından dolayı Veri Madenciliği teknolojisi de yöntem olarak çeşit imkanı sunmaktadır (Özekes, 2003). Veri Madenciliği teknolojisi Sınıflandırma, Kümeleme ve Birliktelik Kuralları olmak üzere 3 ana başlık altında ele alınmaktadır (Coşlu, 2013). Yöntemler de içlerinde barındırdıkları algoritmalar ile başarılı modeller oluşturabilmek için çeşitlilik sağlamaktadır. Örneğin; Sınıflandırma yöntemine özgü Karar Ağaçları, Yapay Sinir Ağları, Genetik, Naive-Bayes ve Lojistik Regresyon vb., Kümeleme yöntemine özgü K-En Yakın Komşu, K-En Uzak Komşu ve K-Ortalamalar vb. mantığına dayalı algoritma çeşitleri bulunmaktadır (Sarıman, 2011).

Sağlık alanında hastalık teşhisi, mühendislik alanında model/algoritma başarı oranı iyileştirilmesi, bankacılık alanında kredi onay/red durumları, borsa alanında stratejik işlem belirlemeleri, eğitim sektöründe öğrenci başarısına katkı sağlayacak uygulamalar, ticarette satışa yönelik taktik/reklam çalışmaları, telekomünikasyon sektöründe ise müşteri kaybetmeme üzerinde kişi bazlı kampanya oluşumları vb. gibi birçok farklı alanda çalışmalar gerçekleştirilmiştir (Savaş et al., 2012). Bu çalışma kapsamında ise, bahsedilen veri madenciliği çalışma alanları, sınıf niteliğini belirten sektör bilgisi olarak kullanılmaktadır. Veri bilimcilerden alınan bazı teknik bilgiler doğrultusunda (yazılım dili, algoritma, çalışma aracı), veri bilimcilerin çalışmış oldukları sektörlerin sınıflandırılması, ele alınan problemin genel tanımıdır. Karar verilen algoritmalar doğrultusunda sınıflandırma işlemi gerçekleştirilmiş olup, algoritma kriterleri ve çıktıları analiz edilmiştir. Analiz edilirken aynı probleme farklı algoritma çeşitleri ve seçilen algoritmanın farklı kriter değerleri ile uygulama gerçekleştirilmesi ile elde edilen model başarı oranları referans alınarak, veri madenciliği süreci ile ilgili çıkarımlarda bulunulmuştur.

Veri Madenciliği süreci 4 temel adımda ele alınabilmektedir (Küçüksille, 2009)(Baykal, 2006.). İlk adım "Problem Tanıma ve Analiz" aşaması olup, girdi ve çıktı niteliklerinin verimli şekilde işlenip model oluşumuna olumlu yönde katkı sağlayabilmesi için problem ve hedeflenen çözüm ile ilgili fikir edinilen süreçtir. "Veri Ön İşleme" aşaması analiz edilen veri seti için model başarısının artırılmasına yönelik yapılan veri düzenleme işlemlerinin gerçekleştirildiği Veri Madenciliği aşamasıdır. "Veri Madenciliği Model Oluşumu" aşaması seçili yöntem, teknik ve algoritmaların kararlaştırılan veri seti değerlerinin kullanılması ile model oluşumunun gerçekleştirildiği aşamadır. Veri Madenciliği sürecinin dördüncü ve son adımı olan "Model Başarı Analizi" aşaması da, oluşturulan modellere ait doğruluk değerleri ve model analizlerinin gerçekleştirilmiş olduğu aşamadır.

Veri Madenciliği, büyük miktarda verinin, bahsedilen süreç aşamaları ile işlenerek anlam kazanması ve gerekli işlemlerden geçirilerek bilgiye dönüştürülmesi işlemi olarak tanımlanmaktadır (Şekeroğlu, 2010). Bahsedilen bu dönüşüm işlemi Veri Madenciliği teknolojilerinden Sınıflandırma, Kümeleme ve Birliktelik Kuralları ile gerçekleştirilmektedir.

Sınıflandırma yöntemi, ele alınan veri niteliklerinin incelenip, daha önceden belirlenmiş olan sınıf değerlerinin uygun olanına dâhil edilmesi mantığını benimsemektedir. Kümeleme yönteminde amaç, verilerin kendi aralarındaki ilişkilerine göre alt sınıflara ayrılmalarının sağlanmasıdır. Son olarak da Birliktelik Kuralları yöntemi ile büyük boyutlardaki veri tabanlarında yer alan, birbirleriyle bağlantılı verilerin ve aralarındaki bağlantıların belirlenmesi amaçlanmaktadır (Altun, 2017)

Çalışma alanı ve uygulama örnekleri genişliğinden bahsedilen Veri Madenciliği kavramına dair literatür taraması gerçekleştiğinde, en sık kullanılan yöntemin Sınıflandırma yöntemi olduğu tespit edilmiştir (Altun, 2017). Bu çalışma kapsamında ele alınan problemin de Sınıflandırma yöntemi ile çözüme ulaşabilmesi, örnek alınabilecek kaynak fazlalığı açısından avantaj sağlamaktadır.

Veri seti kriterlerinin ortak özelliklerine göre belirli sınıflara ayrılmasını amaçlayan sınıflandırma yöntemine dair çeşitli algoritmalar geliştirilmiştir (Çelik, 2009). Bu çalışmada, modellenecek probleme özgü yöntem seçimi yapılması ve yöntem içerisinde farklı işleyiş yapısına sahip algoritmalardan en uygununun seçilerek, en başarılı çözüm yolunun elde edilmesi hedeflenmektedir (Işık & Kapan Ulusoy, 2021). Sınıflandırma yöntemine ait başlıca teknikler şunlardır (Akpınar, 2000);

- Karar Ağaçları (Decision Trees)
- Yapay Sinir Ağları (Artificial Neural Netwoks)
- Genetik Algoritmalar (Genetic Algorithms)
- K-En Yakın Komşu (K-Nearest Neighbor)
- Bellek Temelli Nedenleme (Memory Based Reasoning)
- Naive-Bayes

Belirtilen sınıflandırma yöntemlerinden, seçili algoritmalar bu çalışma kapsamında ele alınan veri seti ile modellenerek başarı oranı kıyaslamaları gerçekleştirilmiştir. Bu sebeple Sınıflandırma yöntemi tekniklerini kullanan çalışmalar, literatür taraması sırasında incelenmiştir. Bu çalışmanın farklılığı, gerçekleştirilen literatür taraması ile ortaya çıkmıştır. Çalışma kapsamında da Sınıflandırma yöntemi ile modellemeler gerçekleştirilmiştir ancak farklılığı oluşturan durum, çalışmada kullanılan veri setidir. Literatürde, bu çalışma kapsamında kullanılan veri seti ile gerçekleştirilen Veri Madenciliği çalışmasına rastlanmamıştır.

2. Materyal ve Metot

Giriş bölümünde, değinilen Veri Madenciliği süreç aşamaları, bu çalışma kapsamında ele alınan "Kaggle'da Bulunan Veri Bilimi Anketi" isimli veri seti ile uygulamalı olarak ifade edilecektir.

2.1. Veri Seti

Veri Madenciliği çalışma alanının fazla olması durumundan, çalışma kapsamında sıkça bahsedilmektedir. Uygulama kısmında da çalışma alanı kavramını karşılayan nitelik sütunu ile Veri Madenciliğinin hizmet ettiği sektörlerin fazlalığı vurgulanmıştır. Bu çalışma kapsamında, ele alınan veri setinde veri bilimcilerden çalışma alanları hakkında alınmış olan cevapları içeren "İşveren Sektörü" nitelik sütunu bulunmaktadır.

"Kaggle'da Bulunan Veri Bilimi Anketi" isimli veri seti Numara (ID), Kullanılan Çalışma Aracı (Work Tools Select), Tercih Edilen Dil (Language Recommendation Select), İşveren Sektörü (Employer Industry) ve Katılımcılar Tarafından Kullanılan Algoritma (Work Algorithms Select) olmak üzere 5 sütun, 10153 satırdan oluşmaktadır (Abid Ali Awan, 2020).

Bahsedilen veri setine ait niteliklerin açıklamalarına kısaca değinilmek gerekirse; ID niteliği, veri setindeki satırları numaralandırmak, düzenli sıralamak açısından benzersiz değerlerden oluşan sütundur. Model girdisinde çıktı oluşumuna karar verme açısından katkı sağlamamaktadır. Geriye kalan nitelikler sütun isminden de anlaşılacağı üzere, veri bilimcilerin vermiş oldukları cevapları içermektedir. Kullanılan Çalışma Aracı niteliğinin %22'si, Tercih Edilen Dil niteliğinin %36'sı, Katılımcılar Tarafından Kullanılan Algoritma niteliğinin %28'i ve İşveren Sektörü niteliğinin %12'si"null" değere sahiptir. Veri Ön İşleme aşamasında "null" değeri içeren satırlar ile ilgili düzenleme işlemi gerçekleştirilmiştir.

2.2. Problem Tanıma

Ankete katılan veri bilimcilerin cevaplarından oluşan tercih edilen algoritma, yazılım dili, çalışma aracı değerlerinin yer aldığı niteliklerin model girdisi, hangi sektörde çalışma yapıldığı değerlerinin yer aldığı niteliğin de model çıktısı olduğu bir modelleme düzeneği oluşturulması hedeflenen durumdur.

Veri bilimciler bir nitelik değeri için birden fazla fikir belirtebildiğinden, veri setinde bulunan bir hücrede birden fazla veri değeri mevcut olabilir. Örneğin bir veri bilimci Katılımcılar Tarafından Kullanılan Algoritma Kriterine hem "CNNs (Convolutional Neural Network – Evrişimli Sinir Ağları)", hem "RapidMiner", hem de "SVMs (Support Vector Machine – Destek Vektör Makinesi)" tercih ettiğini belirtebileceğinden dolayı, aynı hücre içerisinde 3 farklı veri değeri bulunabilir.

Veri satırında "null" değere sahip hücre bulunmaması ve hücre içerisinde birden fazla veri değeri bulunmaması gibi durumlar veriler arasında tutarlılığı sağlayacağından model doğruluğunu ve verimliliğini olumlu yönden etkilemektedir. Veri Madenciliği süreçlerinden olan Veri Ön İşleme aşaması adımları sayesinde eksik verilerin temizlenmesi, gereksiz veri alanları tespiti, veri çeşidi birleşimi, veri artırımı ve aykırı değerlerin düzenlenmesi gibi işlemler gerçekleştirilerek sağlıklı veriler elde edilmesi hedeflenmektedir.

Veri ön işleme adımları gerçekleştirilmeden önce orijinal veri seti satırından bir örnek, Tablo 1'de yer almaktadır.

Tablo 1. Orijinal Veri Setinden Bir Satır

ID	Kullanılan Çalışma Aracı	Tercih Edilen Dil	Katılımcılar Tarafından Kullanılan Algoritma	İşveren Sektörü
11	C/C++, Jupyter notebooks, Python, TensorFlow	Python	CNNs, Neural Networks	Akademik

2.3. Veri Ön İşleme

Eksik Verilerin Temizlenmesi; herhangi bir hücresi "null" değere sahip veri satırının, veri setinden çıkartılması işlemini kapsamaktadır. Normalizasyon vb. yöntemler kullanılarak da "null" değerler anlamlı veriler haline dönüştürülebilir. Ancak bu çalışmada gerçekleştirildiği gibi, ele alınan veri setinde yeterli veri satırı olduğu düşünülüyor ise "null" değere sahip olan veri satırları veri setinden çıkartılabilmektedir.

Gereksiz Veri Alanları Tespiti; orijinal veri setinde bulunan Numara (ID) kriteri gibi, model mekanizmasının çıktı oluşturmasına katkısı olmayan veri değerlerinin veri setinden çıkartılması işlemidir. Ele alınan veri setindeki ID niteliği, satır düzeni oluşturulması için veri setinde bulunduğundan, girdi kriteri olarak tercih edilmeyerek veri setinden çıkarılmıştır.

Veri Çeşidi Birleşimi; kullanılan veri seti kapsamında sürüm farklılıklarından dolayı farklı isimlendirmelere sahip ancak alt yapı olarak aynı durumu işaret eden nitelik değerlerinin bulunduğu tespit edilmiştir. Veri birleşim işlemi sağlanarak veri setinde düzenleme gerçekleştirilmesinin veri seti verimliliği açısından uygun olduğu düşünülerek bu çalışma kapsamında veri birleşim işlemi de gerçekleştirilmiştir. Örneğin Kullanılan Çalışma Aracı niteliğinin sahip olduğu "KNIME (free version)" ve "KNIME (commercial version)" değerleri model girdisi kapsamında iki farklı veri çeşidi olarak değerlendirilmektedir. Ancak örnekte belirtildiği gibi sürüm farklılıkları vb. gibi durumlardan dolayı oluşabilecek gereksiz veri kalabalıklarının düzenlenmesi veri verimliliği açısından yarar sağlayacağından, bu veri çeşitleri "KNIME" şeklinde tek veri çeşidi altında ele alınabilmektedir.

Veri Artırımı; eksik verilerin temizlenmesi aşamasında da bahsedilen "null" veri değerlerinin normalizasyon vb. işlemler ile değer kazanarak veri setine katkı sağlaması, veri setinin fazlalaşmasını amaçlayan işlem adımıdır. Ele alınan veri setinde, tek hücre içerisinde birden fazla veri değerinin olduğu belirtilmişti. Bu hücre değerlerinin tek hücrede tek değer olacak şekilde düzenlenmesi ile veriler artış göstermiş olmaktadır. Bu şekilde ön işleme aşamalarından olan Veri Artırımı aşaması yapay değil gerçek veriler ile elde edilmiştir. Örneğin ele alınan satırda Tercih Edilen Dil niteliği için 3 farklı veri değeri bulunuyor ise, tek hücre tek veri mantığı devreye girerek, tek satırda belirtilen bu veri değerlerinin 3 satırda belirtilir şekilde düzenlenmesi sağlanmış olmaktadır.

Aykırı Değerlerin Düzenlenmesi; nitelik bazında ele alınan veri değerlerinin incelenip, incelenen veri grubundan aykırı duran değerlerin düzenleme işlemlerinin gerçekleştirilmiş olduğu adımdır. Bu çalışmada veri çeşitlerinin sayıları belirlenmiştir. Ortalama değerin altında kalan veri çeşitleri ortak özellik

barındıran diğer veri çeşitleri ile birleştirilip, model çıkarımında daha fazla tekrara sahip olarak sınıf değeri belirlenmesinde katkı sağlamış olmaktadır.

Aykırı Değerlerin Düzenlenmesi işlem adımının daha açıklayıcı olması açısından, kullanılan veri setindeki Katılımcılar Tarafından Kullanılan Algoritma niteliğinin aykırı değerlerinin tespit edilip düzenlenme işleminin gerçekleştirilmesi işlemi aşağıda belirtilmektedir.

Katılımcılar Tarafından Kullanılan Algoritma kriterindeki veri çeşitleri ve veri setindeki kullanım sıklıkları Tablo 2'de belirtilmiştir.

Tablo 2. Katılımcılar Tarafından Kullanılan Algoritma Kriterinin Çeşitleri ve Kullanıldıkları Satır Sayısı

Katılımcılar Tarafından Kullanılan Algoritma Kriter Verileri	Satır Sayısı
BayesianTechniques	621
DecisionTrees	715
RandomForest	700
Regression/LogisticRegression	920
CNNs	389
GradientBoostedMachines	254
Neural Networks	659
SVMs	555
HMMs	142
EnsembleMethods	465
EvolutionaryApproaches	205
Diğer	125
RNNs	265
GANs	63
MarkovLogicNetwoks	82

Veri setinde bulunan veri çeşitlerinin satır sayıları incelendiğinde, verilerin ortalama olarak 411 kez tekrarlandığı hesaplanmıştır. Ortalama altında kullanılan veri çeşitleri (CNNs, GradientBoostedMachines, HMMs, EvolutionaryApproaches, Diğer, RNNs, GANs ve MarkovLogicNetworks) aykırı değerlerin düzenlenmesi işlem adımında belirtildiği üzere ortak noktaları bulunan veri çeşitleri ile birleştirilmiştir. Bu sayede yeni oluşan veri çeşidi, sayısal olarak fazla olduğundan model çıktısının oluşumuna etkisi de fazla olmaktadır.

Orijinal veri setindeki hali, Tablo 1'de tek satırda karmaşık olarak belirtilen veri satırı, ön işleme aşamalarından sonra Tablo 3'teki 8 satırlık yalın ve net veri satırlarına dönüştürülmüştür. Tablo 1 ve Tablo 3'te somut olarak ifade edilen dönüşüm işlemi, tüm veri setine uygulanarak elde edilen işlenmiş veri seti ile model oluşum aşamasına hazır duruma getirilmiştir.

Tablo 3. İşlenmiş Veri Setinden Bir Satır

Kullanılan Çalışma Aracı	Tercih Edilen Dil	Katılımcılar Tarafından Kullanılan Algoritma	İşveren Sektörü
C/C++	Python	CNNs	Akademik
C/C++	Python	Neural Networks	Akademik
Jupyter	Python	CNNs	Akademik
Jupyter	Python	Neural Networks	Akademik
Python	Python	CNNs	Akademik
Python	Python	Neural Networks	Akademik
TensorFlow	Python	CNNs	Akademik
TensorFlow	Python	Neural Networks	Akademik

2.4. Veri Madenciliği Model Seçimi

Bilgiye ulaşmada ele alınan problemin analizi doğrultusunda, Veri Madenciliği yöntem seçimi gerçekleştirilmelidir. Ele alınan problem, veri bilimcilerin vermiş oldukları cevaplar esas alınıp, veri seti niteliklerinin incelenerek İşveren Sektörü niteliğinin tahmin edilmesini içermektedir. Veri Madenciliği yöntem tanımlamaları incelendiğinde, problem için hangi yöntemin kullanılmasının uygun olduğu ortaya çıkmış olmaktadır. Sınıflandırma tanımlamasındaki gibi, belirli nitelikler kullanılıp, belirli gruplama değerlerine göre dağılım gerçekleştirilmek hedeflenmektedir. Bu mantık, ele alınan problemin Sınıflandırma yöntemi ile modellenmesinin uygun olduğunu göstermektedir.

Veri seti niteliklerinin ortak özelliklerine göre belirli sınıflara ayrılmasını amaçlayan sınıflandırma yöntemine dair çeşitli algoritmalar geliştirilmiştir (Çelik, 2009).

Bu çalışma kapsamında, modellemenin gerçekleştirileceği tekniğe göre 3 farklı başlık altında ele alınan Entropi Tabanlı, Sınıflandırma-Regresyon Tabanlı ve Bellek Tabanlı algoritma çeşitlerinden birer algoritma seçilerek, "Kaggle'da Veri Bilimi Anketi" isimli veri setinde model denemeleri gerçekleştirilmiştir.

Entropi Tabanlı, Sınıflandırma-Regresyon Tabanlı ve Bellek Tabanlı algoritma çeşitlerinden sırası ile C4.5, Rastgele Orman ve K-En Yakın Komşu algoritmaları seçilerek orijinal ve işlenmiş veri seti üzerinde model denemeleri gerçekleştirilmiştir.

C4.5 Algoritmasına göre, veri setinde kullanılan niteliklerinin kontrolü sağlandıktan sonra, her niteliğin normalize edilmiş bilgi kazanım değerleri hesaplanmaktadır. En iyi bilgi kazanım değerini veren özellikler karar noktaları olarak belirlenip, karar düğümünün ardına alt liste oluşturularak alt karar ağacı inşa edilmesi ile modelleme gerçekleştirilmiş olmaktadır (Yıldırım, 2003).

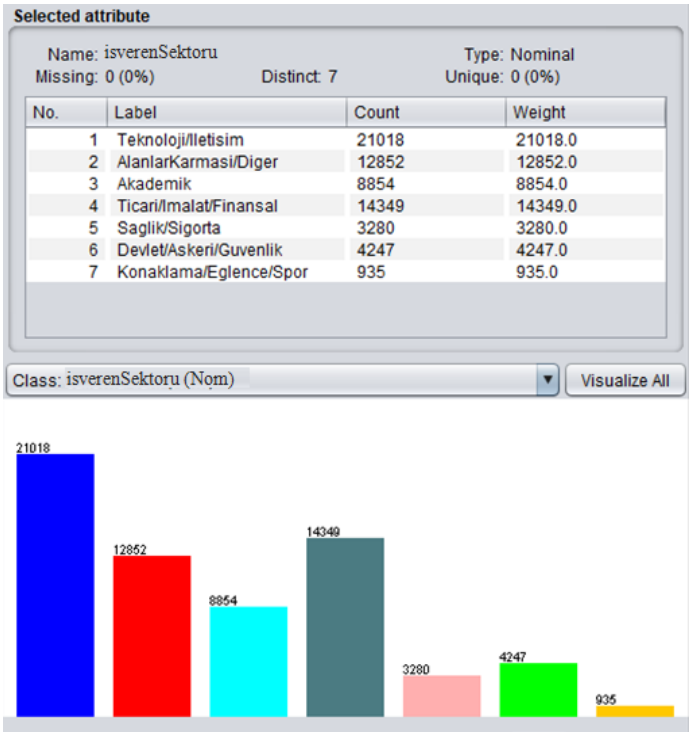
Rastgele Orman Algoritması, karar ağaçlarının birleşiminden oluşan bir algoritma olup, kullanılan karar ağaçları arasında doğruluğu ve bağımsızlığı en yüksek ağaçlar tercih edilmektedir (Breiman, 2001).

K-En Yakın Komşu (KNN) Algoritması, veri seti kullanılarak daha önceden oluşturulan çıkarımlar doğrultusunda, yeni gelen girdilere dair çıkarımların benzerliklerinin kıyaslanması mantığı ile uygun sınıf tahminini gerçekleştiren sınıflandırma algoritmasıdır (Mitchell, 1997).

2.5.Verilerin Benzetim Aracı Ortamına Aktarım ve Analizi

Ön işleme adımları gerçekleştirilmiş olan, işlenmiş veri seti 3 girdi, 1 çıktı (sınıf) niteliği olmak üzere 4 sütun 65535 satırdan oluşmaktadır. İşlenmiş veri seti, veri analizinin gerçekleştirileceği Weka ortamına uygun formata çevrilerek (.arff) aktarımı sağlanmıştır. Model çıktısı olarak elde edilmek istenilen sınıf değeri (İşveren Sektörü) baz alınarak işlenmiş veri setinin Weka'ya aktarılmış hali Şekil 1'de belirtildiği gibidir.

Şekil 1. Weka'da İşveren Sektörü Niteliğinin Analizi



3. Araştırma Sonuçları ve Tartışma

Veri Madenciliği teknolojisi yöntemlerinden Sınıflandırma yöntemi algoritmaları, çözüm üretilmek istenilen probleme uygulanarak, algoritmaların analiz ve kıyaslamaları gerçekleştirilmiştir.

3.1.Model Oluşumu ve Algoritma Analizleri

Sınıflandırma yöntemi, algoritmaları teknik olarak 3 farklı kategoride ele alındığından her bir kategoriden birer algoritma modeli seçilip uygulama gerçekleştirilmiştir. Entropi Tabanlı Algoritma olan C4.5, Sınıflandırma ve Regresyon tabanlı olan Rastgele Orman ve Bellek Tabanlı Sınıflandırma mantığına dayanan K-En Yakın Komşu Algoritmaları uygulama için tercih edilen algoritmalardır.

Weka ortamında "J48" olarak adlandırılan C4.5 algoritması, Weka ortamına yüklenmiş veriler doğrultusunda, varsayılan algoritma kriterleri kullanılarak modellenmesi ile oluşturulmuştur. Çapraz doğrulama seçeneği ve veri setinin 10 bölümleme ile modellemeye geçilmesi sonucu C4.5 Algoritmasının başarı oranı %34,3 şeklinde elde edilmiştir. Yüzde bölünmesi seçeneği %66 bölümleme ile modellemeye geçilmesi sonucu C4.5 Algoritmasının başarı oranı %34,5 şeklinde elde edilmiştir.

Weka ortamında Rastgele Orman algoritması ile modelleme seçimi gerçekleştirileceği belirtildikten sonra, varsayılan algoritma kriterleri ile modelleme gerçekleştirilmiştir. Çapraz doğrulama seçeneği ve veri setinin 10 bölümleme ile modellemeye geçilmesi sonucu Rastgele Orman Algoritmasının başarı oranı %34,5 şeklinde elde edilmiştir. Yüzde bölünmesi seçeneği %66 bölümleme ile modellemeye geçilmesi sonucu Rastgele Orman algoritmasının başarı oranı %34,5 şeklinde elde edilmiştir.

Weka ortamında "IBk" olarak adlandırılan KNN algoritması, yüklenmiş veriler doğrultusunda KNN algoritmasına özgü k parametresi hariç Weka'nın varsayılan algoritma kriterleri ile modelleme gerçekleştirilmiştir.

k değeri 1,11,13,17 ve 21 olarak belirlenerek Çapraz doğrulama seçeneği ile veri setinin 10 bölümleme ve Yüzde bölünmesi seçeneği ile %66 bölümleme ile elde edilen model başarı oranlarının çıktıları Tablo 4'te ifade edilmektedir.

Tablo 4.KNN Algoritması Başarı Oranları

Veri Seti Bölümlemesi / **k değeri**	**Çapraz doğrulama: Kıvrımlar 10**	**Yüzde bölünmesi: %66**
k=1	%33,463	%33,511
k=11	%33,578	%33,619
k=13	%33,582	%33,614
k=17	%33,533	%33,651
k=21	%33,507	%33,691

3.2.Model Başarı Oranı

Model başarılarının değerlendirilmesi, idealliğinin belirlenmesi için bazı ölçütler bulunmaktadır. Hesaplanan bu somut ölçütler ifade edilerek, bu değerler referans alınarak tartışma/yorumlama kısmı gerçekleştirilmiştir.

Bahsedilen başarı ölçütleri Doğruluk Oranı, Hata Oranı, Duyarlılık, Hassasiyet/Kesinlik, F-Ölçütü ve ROC Eğrisi olmak üzere Formül (1), Formül (2), Formül (3), Formül (4) ve Formül (5)'teki gibi somut hesaplamalar ile başarı oranlarını sunan kavramlardır. Bu kavramlar Tablo 5'te belirtilen model tahmininin doğru/yanlış durumuna göre hesaplanan karışıklık matris değerlerini kullanmakta olup devamında belirtilen şekilde hesaplanmaktadırlar (Esmer et al., 2020).

Tablo 5.Karışıklık Matris Yapısı

Tahmin Sınıfı / **Gerçek Sınıf**	**Pozitif**	**Negatif**
Pozitif	Doğru Pozitif (DP)(TP)	Yanlış Negatif (YN)(FN)
Negatif	Yanlış Pozitif (YP)(FP)	Doğru Negatif (DN)(TN)

$$Doğruluk\ Oranı = \frac{TP+TN}{TP+FN+FP+TN} \quad (1)$$

$$Hata\ Oranı = \frac{FP+FN}{TP+FN+FP+TN} \quad (2)$$

$$Duyarlılık = \frac{TP}{TP+FN} \quad (3)$$

$$Hassasiyet = \frac{TP}{TP+FP} \quad (4)$$

$$F\text{-}Ölçütü = \frac{2\ X\ (Duyarlılık) X\ (Hassasiyet)}{(Duyarlılık) + (Hassasyet)} \quad (5)$$

ROC Eğrisi; yatay eksen (x ekseni) FP, düşey eksen (y ekseni) TP oranlaması temsil edilerek oluşturulmaktadır.

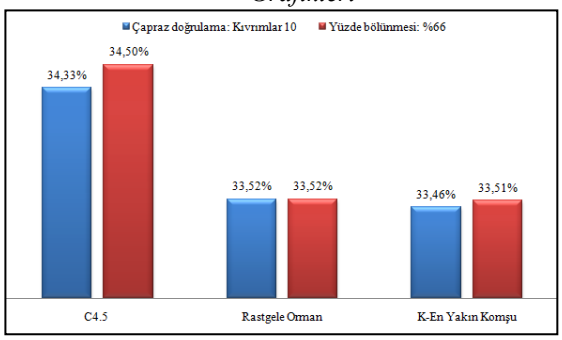
Ön işleme işlemi gerçekleştirilmiş ve modellemeye hazır hale getirilmiş olan veri seti, karar verilen 3 algoritma ile Weka ortamında modellenmiştir. Tablo 6'da işlenmiş veri setinin Weka seçeneklerinden olan 2 farklı veri seti bölümlemesinin kullanılması sonucu elde edilen başarı oranları ifade edilmektedir. Şekil 2'de de elde edilen başarı oranlarının grafiksel olarak ifadesi yer almaktadır.

Tablo 6.İşlenmiş Veri Setinin Algoritma Bazında Başarı Oranları

Veri Seti Bölümlemesi / Algoritma Çeşidi	**Çapraz doğrulama: Kıvrımlar 10**	**Yüzde bölünmesi: %66**
C4.5	%34,33	%34,50
Rastgele Orman	%33,52	%33,52
K-En Yakın Komşu	%33,46	%33,51

Aynı veri setinin farklı algoritmalar ile modellenmesi sonucu model başarı oranları analiz edildiğinde en yüksek başarı oranına C4.5 Algoritması ile oluşturulan modelin sahip olduğu gözlemlenmiştir.

Şekil 2. İşlenmiş Veri Setinin Algoritma Bazında Başarı Oranı Grafikleri

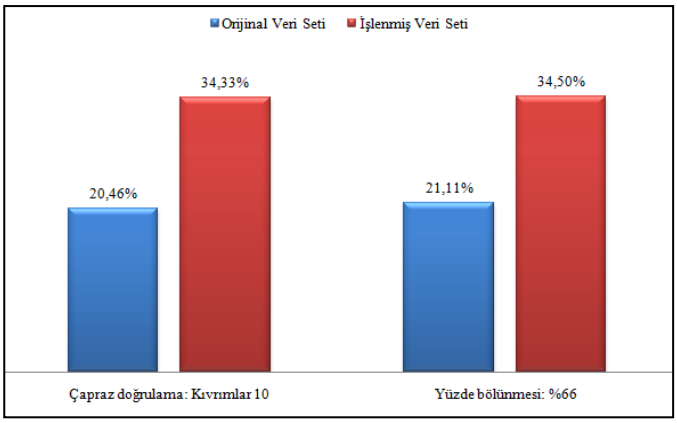


En yüksek başarı oranının elde edilmesini sağlayan C4.5 Algoritması, farklı veri seti için uygulandığı takdirde model başarı oranları arasındaki farkın gözlemlenebilmesi için de farklı bir senaryo oluşturulmuştur. Orijinal ve işlenmiş veri seti olmak üzere iki farklı veri setinin aynı algoritma ile (C4.5 Algoritması) modellenmesi sonucu elde edilen model başarı oranları Tablo 7'de, grafiksel olarak da Şekil 3'te ifade edilmektedir.

Tablo 7. Orijinal ve İşlenmiş Veri Setinin C4.5 Algoritması İle Modellenmesi Sonucu Başarı Oranları

Veri Seti Bölümlemesi / Veri Seti	**Çapraz doğrulama: Kıvrımlar 10**	**Yüzde bölünmesi: %66**
Orijinal Veri Seti	%20,46	%21,11
İşlenmiş Veri Seti	%34,33	%34,50

Şekil 3. Orijinal ve İşlenmiş Veri Setinin C4.5 Algoritması İle Modellenmesi Sonrasında Başarı Oranları



Ele alınan iki farklı durumdan çıkarımla, farklı algoritma çeşitlerinin aynı veri setini kullanarak oluşturdukları modellerin başarı oranları ve aynı algoritma çeşidinin farklı veri setlerini kullanarak oluşturdukları modellerin başarı oranları hesaplanmıştır.

Tablo 8. Model Başarı Oranları Kıyas Tablosu

Algoritma Çeşidi	Veri Seti	Veri Seti Bölümlemesi	Maksimum Başarı Oranı	Minimum Başarı Oranı	Başarı Oranı Farkı
Farklı *(C4.5, Rastgele, KNN)*	Aynı *(İşlenmiş Veri Seti)*	Aynı *(Çapraz doğrulama: Kıvrımlar 10)*	%34,33 (C4.5)	%33,46 (KNN)	%0,87
Farklı *(C4.5, Rastgele, KNN)*	Aynı *(İşlenmiş Veri Seti)*	Aynı *(Yüzde bölünmesi: %66)*	%34,50 (C4.5)	%33,51 (KNN)	%0,99
Aynı *(C4.5)*	Farklı *(Orijinal ve İşlenmiş Veri Seti)*	Aynı *(Çapraz doğrulama: Kıvrımlar 10)*	%34,33 (İşlenmiş Veri Seti)	%20,46 (Orijinal Veri Seti)	%13,87
Aynı *(C4.5)*	Farklı *(Orijinal ve İşlenmiş Veri Seti)*	Aynı *(Yüzde bölünmesi: %66)*	%34,50 (İşlenmiş Veri Seti)	%21,11 (Orijinal Veri Seti)	%13,39

Bu çalışma kapsamında gerçekleştirilen model denemelerinin başarı oranları, çalışmanın amacını vurgulamak üzere Tablo 8'de aktarılmak istenmiştir. Tablodan çıkarımla veriler, Veri madenciliği yöntemi, algoritmada kullanılan teknik ve veri seti bölümlemesi aynı olmuş olsa dahi farklı model başarıları elde etmek mümkün olmaktadır. Tablo 6 incelendiğinde de, işlenmiş veri seti ile sınıflandırma yöntemi, Rastgele Orman algoritması kullanılan ve tek fark veri seti bölümlemesi olan model sonucunun aynı başarı oranı yüzdesine sahip olduğu görülmektedir. Ancak aynı koşullar C4.5 ve K-En Yakın Komşu Algoritması için incelendiği takdirde algoritmalar kendi bazlarında farklı başarı oranlarına sahip olmaktadırlar. Oluşan bu farklar veri setinin büyüklüğü ile alakalı olarak da değişim gösterebilmektedir.

Tablo 8 incelendiği takdirde göze çarpan bir başka durum da "Başarı Oranı Farkı" sütunu değerleridir. Bu sütunun ilk iki satırı incelendiğinde yaklaşık %1'lik değer gözlenirken, son iki satırı incelendiğinde yaklaşık %13-14'lük değer gözlemlenmektedir. İlk iki satır ve son iki satırın benzer ve farklı nitelikleri incelendiğinde;

İlk iki satır:

→Farklı Algoritma Çeşidi

→ Aynı Veri Seti

→ Aynı Veri Seti Bölümlemesi

niteliklerine sahip olmakta olup, model başarı oranları farkı yaklaşık %1'dir.

Son iki satır:

→Aynı Algoritma Çeşidi

→ Farklı Veri Seti

→ Aynı Veri Seti Bölümlemesi

niteliklerine sahip olmakta olup, model başarı oranları farkı yaklaşık %13-14'tür.

4. Sonuç

Araştırma Sonuçları ve Tartışma bölümünde Tablo 4, Tablo 6 ve Tablo 7 olmak üzere model başarıları 3 farklı kıyaslama kategorisi ile ifade edilmiştir. Tablo 8'de de bu üç farklı durum göz önünde bulundurularak model başarı oranlarının minimum-maksimum olmasını etkileyen kriterlerin vurgulanması için, makalenin amacı doğrultusunda yapılan hesaplamalara yer verilmiştir.

Yapılan hesaplamalar sonucu elde edilen başarı oranları analiz edildiğinde, özellikte Tablo 8'de ifade edilen şekilde model başarı oranları incelendiğinde yaklaşık %1'lik başarı oranı farkına sahip ilk iki satır ve yaklaşık %13-14'lük başarı oranı farkına sahip son iki satır olmak üzere iki farklı durum göze çarpmaktadır.

İlk iki satır ilk durum, son iki satır ikinci durum olarak ifade edilip, iki durumun da incelemesi gerçekleştirilir ise;

İlk durumun sahip olduğu iki satır incelendiğinde model başarı oranı hesabını etkileyen kriterlerden olan algoritma çeşidi farklı, veri seti ve veri seti bölümlemesi kriterleri aynı değere sahiptir. İşlenmiş veri setinin, aynı veri seti bölümlemesi ile C4.5, Rastgele Orman ve KNN Algoritmaları kullanılarak oluşturulan modellerden en fazla model başarı oranına C4.5 Algoritmasıyla, en az model başarı oranına ise KNN Algoritması ile ulaştığı gözlemlenmektedir. Maksimum başarı oranına sahip C4.5 Algoritmasının model başarı oranı yaklaşık %34 iken, minimum başarı oranına sahip KNN Algoritmasının model başarı oranı da yaklaşık %33 olarak hesaplanmıştır. Aradaki fark yaklaşık olarak %1'dir.

İlk durum model oluşumunda kullanılan kriterlerden, algoritma çeşidi kriteri farklılık göstermektedir. Veri seti de, veri seti bölümlemesi de model oluşumunda aynı olduğundan dolayı, model başarı oranlarındaki oluşan %1'lik farkın seçilen algoritma çeşidinden kaynaklandığı ortaya çıkmıştır. C4.5 Algoritmasının bu veri seti üzerinde kullanılması, yaklaşık %1'lik oranla Rastgele Orman ve KNN algoritmalarına göre model oluşumunda daha verimli olmuştur.

İkinci durumun sahip olduğu iki satır incelendiğinde model başarı oranı hesabını etkileyen kriterlerden olan kullanılan veri seti kriteri farklı, kullanılan algoritma çeşidi ve veri seti bölümlemesi kriterleri aynı değere sahiptir. Orijinal ve İşlenmiş Veri Setlerinin C4.5 Algoritması ile aynı veri seti bölümleme değeri kullanılarak modellenmesi sonucu en fazla model başarı oranı İşlenmiş Veri Seti ile oluşturulan modelle, en az model başarı oranı Orijinal Veri Seti ile oluşturulan model ile elde

edildiği gözlemlenmiştir. İşlenmiş Veri Setinin kullanılmasıyla oluşturulan modelin başarı oranı yaklaşık %34 ile maksimum, Orijinal Veri Setinin kullanılmasıyla oluşturulan modelin başarı oranı yaklaşık %13-14 ile minimum olarak hesaplanmıştır.

İkinci durumda ele alınan satırlar incelendiğinde, model oluşumunda kullanılan Veri Seti kriteri farklılık göstermektedir. Farklı Veri Setleri, aynı algoritma çeşidi (C4.5 Algoritması) ve aynı veri seti bölümlemesi seçenekleri ile modellendiğinde, model başarı oranları arasında oluşan yaklaşık %13-14 oranındaki farkın, modellemedeki tek farklı kriter olan kullanılan Veri Setinden kaynaklandığı ortaya çıkmaktadır. C4.5, Rastgele Orman ve KNN algoritmaları ile İşlenmiş Veri Seti kullanılarak model oluşumu gerçekleştirildiğinde, Orijinal Veri Seti kullanılarak oluşturulan modellerden daha başarılı modeller elde edildiği görülmektedir.

İlk durumda farklı olarak kullanılan Algoritma Çeşidi kriteri model başarı oranları arasında %1'lik sapma gösterirken, ikinci durumda farklı olarak kullanılan Veri Seti kriteri model başarı oranları arasında %13-14'lük fark yaratmaktadır.

Bu iki durum arasındaki model başarı oranı değerlerinden de referans alınarak kullanılan Veri Seti kriterinin, kullanılan Algoritma Çeşidi kriterine göre model başarı oranına daha fazla etkisinin olduğu ifade edilebilmektedir. Aynı teknik alt yapıya sahip algoritmalar arasından seçilmesinin de etkisi ile, algoritma çeşidinin farklı olması, model başarısı sapmasında afakî bir değişim yaratmamıştır. Ancak aynı teknikli algoritmalar olsalar dahi, algoritmaların oluşum yapılarında farklı teknikler kullanılmasından dolayı aynı veri seti ile oluşturulan model başarı oranlarının da farklı olduğu gözlemlenmektedir. İkinci durumda farklılık olarak ele alınan Veri Seti kriterinin model başarı oranları arasındaki sapma değerine etkisinin daha etkili olduğu başarı oranı farkı ile ortaya çıkmıştır. Model oluşumunda kullanılan teknik, algoritma, veri seti bölümlemesi aynı olmuş olsa dahi farklı veri setleri (Orijinal ve İşlenmiş Veri Seti) kullanılarak elde edilen modellerin başarı oranlarında ilk duruma göre daha gözle görülür bir fark bulunmaktadır. Kıyaslanan veri setlerinin ön işleme aşaması öncesi Orijinal Veri Seti ve ön işleme aşaması sonrası oluşturulan İşlenmiş Veri Seti olması da Ön İşleme aşaması gerçekleştirmenin model başarısına katkısının olumlu yönde olduğunun ispatı olarak belirtilebilmektedir.

5. Teşekkür

Lisans dönemimden bu yana desteklerini esirgemeyip, vermiş olduğu fikirleri ile çalışmalarımda bana yol gösteren ve yakın ilgisini eksik etmeyen değerli hocam Dr. Öğr. Üyesi Arafat ŞENTÜRK'e en içten dileklerim ile teşekkür ederim.

Yaşamım boyunca elde etmiş olduğum tüm başarılarımın asıl emektarları olan sevgili annem Gül KOK ve babam Yasin KOK'a sonsuz teşekkür ederim.

Gerek akademik, gerekse özel hayatımda her kararımda beni sabırla destekleyen sevgili eşim Mehmet DOĞAN'a sonsuz teşekkür ederim.

Kaynakça

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

Altun, M. (2017). Veri Madenciliği ve Uygulama Alanları. *Akdeniz Üniversitesi, Eğitim Bilimleri Bölümü, EYTEPE ABD Doktora Programı, Doktora Seminer Raporu.*

Awan, A. A. (2020). *Data Science Survey on Kaggle*. Kaggle: Your Machine Learning and Data Science Community. https://www.kaggle.com/kingabzpro/datascience-survey-on-kaggle

Baykal, A. (2006). Veri Madenciliği Uygulama Alanları. *Dicle Üniversitesi Ziya Gökalp Eğitim Fakültesi Dergisi*.

Breiman, L. (2001). Random Forest. *Machine Learning, C:1, s:5-32*.

Coşlu, E. (2013). Veri Madenciliği. *Akdeniz Üniversitesi, 15. Akademik Bilişim Konferansı*.

Çelik, M. (2009). Veri Madenciliğinde Kullanılan Sınıflandırma Yöntemleri ve Bir Uygulama. *İstanbul Üniversitesi, Sosyal Bilimler Enstitüsü, Ekonometri Anabilim Dalı, Yüksek Lisans Tezi.*

Esmer, S., Uçar, M. K., Çil, İ., & Bozkurt, M. R. (2020). Parkinson Hastalığı Teşhisi İçin Makine Öğrenmesi Tabanlı Yeni Bir Yöntem. *Düzce Üniversitesi Bilim ve Teknoloji Dergisi, C:8, s:1877-1893*.

Işık, K., & Kapan Ulusoy, S. (2021). Metal Sektöründe Üretim Sürelerine Etki Eden Faktörlerin Veri Madenciliği Yöntemleriyle Tespit Edilmesi. *Gazi Üniversitesi, Journal of the Faculty, Engineering and Architecture, C:36, Sayı:4, s:1949-1962.*

Küçüksille, E. (2009). Veri Madenciliği Süreci Kullanılarak Portföy Performansının Değerlendirilmesi Ve İMKB Hisse Senetleri Piyasasında Bir Uygulama. *Süleyman Demirel Üniversitesi, Sosyal Bilimler Enstitüsü, İşletme Anabilim Dalı, Doktora Tezi.*

Mitchell, T. (1997). Machine Learning. *McGraw Hill, New York.*

Özekes, S. (2003). VERİ MADENCİLİĞİ MODELLERİ VE UYGULAMA ALANLARI. *İstanbul Ticaret Üniversitesi Dergisi, Sayı:3*.

Sarıman, G. (2011). Veri Madenciliğinde Kümeleme Teknikleri Üzerine Bir Çalışma: K-Means ve K-Medoids Kümeleme Algoritmalarının Karşılaştırılması. *Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi, C:15, Sayı:3, s:192-202*.

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

Sönmez, F. (2015). Kredi Skorunun Belirlenmesinde Yapay Sinir Ağları ve Karar Ağaçlarının Kullanımı: Bir Model Önerisi. *ABMYO Dergisi*.

Söyler, H. (2014). Veri Madenciliği ve Kanser Erken Teşhisinde Kullanımı. *İnönü Üniversitesi Sosyal Bilimler Enstitüsü Ekonometri Ana Bilim Dalı.*

Şekeroğlu, S. (2010). Hizmet Sektöründe Bir Veri Madenciliği Uygulaması. *İstanbul Teknik Üniversitesi, Fen Bilimleri Enstitüsü, Endüstri Mühendisliği, Yüksek Lisans Tezi.*

Yıldırım, S. (2003). Tümevarım Öğrenme Tekniklerinden C4.5'in İncelenmesi. *İstanbul Teknik Üniversitesi, Fen Bilimleri Enstitüsü, Savunma Teknolojileri, Yüksek Lisans Tezi*.

Yücebaş, S. C. (2018). Karmaşık Hastalıkların Teşhisinde Veri Madenciliği Yöntemlerinin Başarım Karşılaştırması. *ÇOMÜ Açık Erişim Sistemi*.