SETSCI conference proceedings
SETSCI Conference Proceedings
6 (1), 55-58, 2023

Cognitive Models and Artificial Intelligence Conference
26-28 October 2023, Ankara, Türkiye

https://doi.org/10.36287/setsci.6.1.020
2687-5527/ © 2023 The Authors. Published by SETSCI

BMYZ 2023

Fatura Yönetimi Uygulamasında Faturaların Makine Öğrenmesi ile Sınıflandırılması

Bilgehan AVCI[1*], Can AYDIN[2]

[1]Dokuz Eylül Üniversitesi, Yönetim Bilişim Sistemleri Bölümü, İzmir, Türkiye
[2]Dokuz Eylül Üniversitesi, Yönetim Bilişim Sistemleri Bölümü, DEÜ-KALMER, İzmir, Türkiye
[*](ba.bilgehanavci@gmail.com) Email of the corresponding author

Öz – Dijitalleşmenin yaygınlaşmasıyla birlikte işletmelerin muhasebe ve finans gibi yoğun ve rutin hesaplama ağırlıklı iş yükü olan departmanlarındaki iş ve işlemler dijital muhasebe uygulamaları, robotik süreç otomasyonları gibi farklı bilgi teknolojilerinden yararlanılarak gerçekleştirilmeye başlamıştır.

Bu çalışmada; aktif olarak kullanılmakta olan ve çok sayıda kullanıcısı olan bir muhasebe uygulamasında kullanıcılar tarafından gerçekleştirilmekte olan faturaların sınıflandırılması işlemini otomatik hale getirecek bir karar destek sistemi geliştirilmesi amaçlanmıştır. Mevcut durumda sisteme yüklenen faturalar ilk olarak sınıflandırmadan sorumlu kullanıcılar tarafından incelenerek hangi sınıfa ait olduğu belirlenmektedir. Daha sonra belirlenen sınıfın kullanıcılarına iletilmektedir. Bu sınıflandırma işleminin kullanıcılar tarafından yapılması hem iş gücü kaybına hem de özellikle iş yoğunluğunun yüksek olduğu dönemlerde gecikmelere yol açmaktadır. Bu doğrultuda uygulamanın veri tabanında kayıtlı yaklaşık 450 bin adet faturanın verileri düzenlenerek makine öğrenmesi teknikleri ile bir model eğitilmiştir.

Anahtar kelimeler: Makine Öğrenmesi, Sınıflandırma, Karar Destek Sistemi, Fatura Yönetimi.

Classification of Invoices with Machine Learning in Invoice Management Application

Abstract – With the widespread adoption of digitization, businesses have started to carry out tasks and processes in departments with heavy and routine calculation-based workloads, such as accounting and finance, by utilizing various information technologies such as digital accounting applications and robotic process automations.

In this study; it is aimed to develop a decision support system that will automate the classification process of invoices carried out by users in an accounting application that is actively used and has many users. In the current situation, the invoices uploaded to the system are first examined by the users responsible for classification and the class they belong to is determined. It is then forwarded to the users of the specified class. Performing this classification process by users causes both labor loss and delays, especially during periods of high workload. In this direction, a model was trained with machine learning techniques by editing the data of approximately 450 thousand invoices registered in the application's database.

Keywords: Machine Learning, Classification, Decision Support System, Invoice Management.

I. GİRİŞ

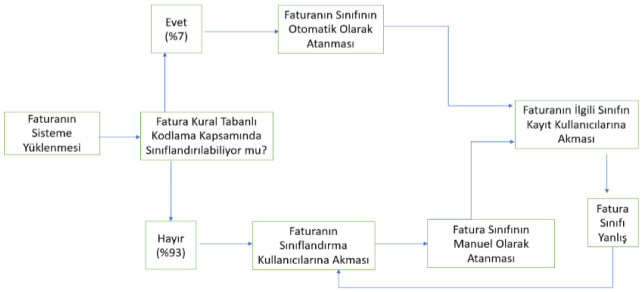
Günümüzde iş dünyası sistematik zorluklarla doludur. İş bölümleri kapsamlı ve birbiriyle bağlantılı birçok boyuttan oluşmaktadır. Pek çok araştırma karmaşıklaşan iş modellerinin hızla gelişen ve değişen iş dünyasında verimlilik ve başarının sürdürülmesinin önündeki büyük bir engel olduğunu göstermektedir. İş modelinde verimlilik tekrarlanabilir, sistematik ve düzenli süreçlerle sağlanabilir. Bu doğrultuda iş modelinin değişen şartlara göre güncellenmesi, ihtiyaca göre veya belli periyotlarla etkin bir şekilde denetimi ve iyileştirilmesi işletmeler açısından oldukça önemlidir (Sebetçi ve diğerleri, 2018).

İşletmeler bilgisayar ve internetle birlikte dijital muhasebe uygulamalarına geçmiş ve pek çok alanda olduğu gibi muhasebe departmanlarında da bilgi teknolojilerinin kullanılarak çeşitli yazılım uygulamalarından yararlanılması durumu giderek artmıştır (Tektüfekçi, 2012).

Şirketler çalışma şekillerine, büyüklüklerine ve evrak yoğunluklarına göre çok sayıda muhasebe personeli istihdam edebilmektedirler. Yine çalışma şekillerine bağlı olarak muhasebe departmanlarında iş bölümü ve görev paylaşımı yapılmaktadır.

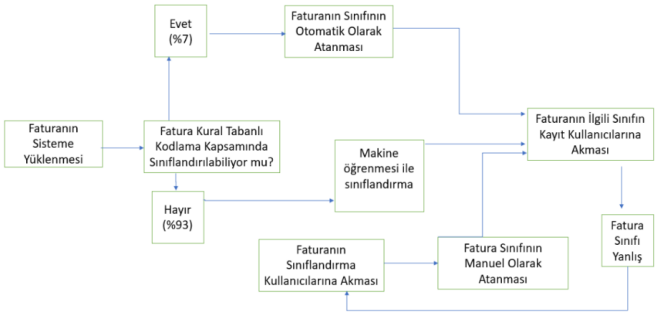
Bu çalışmada kapsamında ele alınan uygulama; kurumsal bir şirketin muhasebe departmanında aktif olarak kullanılmakta olan ve 100'ün üzerinde kullanıcısı olan bir fatura yönetimi uygulamasıdır. Uygulamadaki iş sürecinde sisteme yüklenen faturalar ilk olarak sınıflandırma işleminden sorumlu olan kullanıcıların görev ekranına düşmektedir. Sınıflandırma kullanıcıları tarafından faturanın hangi sınıfa ait olduğu manuel olarak belirlendikten sonra faturalar belirlenen sınıfın kullanıcılarının görevlerine iletilmektedir. Bu sınıflandırma işleminin manuel olarak gerçekleştirilmesi iş sürecini yavaşlatmakta ve departman verimliliğini olumsuz yönde etkilemektedir. Faturaların sınıflandırılma işlemi tamamlanana kadar kayıt aşamasına geçilememesi özellikle iş

yoğunluğunun yüksek olduğu dönemlerde kayıtlarda ve ödemelerde gecikmelere sebebiyet verebilmektedir. Ödemelerdeki gecikmeler tedarikçilerle ilişkileri olumsuz yönde etkilemekte ve vade farkı / gecikme faizi gibi ekstra maliyetlere de yol açabilmektedir.



Şekil 1. Uygulamanın geliştirme öncesi iş akış şeması.

Şekil 1'de yer alan şemada uygulamanın mevcut işleyişindeki iş akış aşamaları detaylı olarak yer almaktadır. İş akış şemasından da anlaşılacağı üzere sınıflandırma işlemi için kural tabanlı kodlama yöntemi uygulanmış olup bu yöntemin başarı oranı %7'dir. Bu yöntemde başarı oranının artırılması çok sayıda kullanıcı tarafından yeni kuralların tespit edilmesi ve tespit edilen yeni kuralların sisteme dahil edilmesi ile mümkündür. Bu çalışmada ise; uygulama veri tabanında yer alan sınıflandırılmış fatura verileri kullanılarak sınıflandırma işleminin makine öğrenmesi algoritmaları ile gerçekleştirilmesi amaçlanmıştır.



Şekil 2. Uygulama için önerilen geliştirme sonrası iş akış şeması.

Şekil 2'de yer alan şemada ise makine öğrenmesi modelinin, mevcut uygulama yaşam döngüsünün hangi aşamasında uygulamaya dahil edilebileceğinin gösterimi yer almaktadır. Mevcut sistemde kural tabanlı kodlama ile sınıflandırılamayan faturalar direkt olarak sınıflandırma kullanıcılarına iletilirken, şekil 2'de yer alan yeni yaşam döngüsünde bu aşamaya makine öğrenmesi modelinin dahil edilmesi önerilmektedir. Modelin hatalı sınıflandırdığı faturalar ise mevcut sistemde olduğu gibi; kayıt kullanıcıları tarafından, yeniden sınıflandırılmak üzere sınıflandırma kullanıcılarına iletilmeye devam edilecektir.

Faturaların makine öğrenmesi algoritmaları ile sınıflandırılması kapsamında, makine öğrenmesi yöntemi ile yapılan sınıflandırma çalışmaları araştırıldığında çok sayıda ve farklı konularda çalışmaya rastlanmıştır. 2021 yılında Yangın vd. tarafından yapılan çalışmada ABD'de 130 farklı hastanenin kayıtlarında yer alan 1999-2008 yılları arasında karşılaşılmış toplam 70 bin sağlık vakıası kayıtlarından elde ettikleri veri setini kullanarak hastaların diyabet rahatsızlıklarını sınıflandırmışlardır (Yangın vd. 2021). Bu çalışmada sınıflandırma işlemi için veri setini Karar Ağaçları, K-En Yakın Komşu, Lojistik Regresyon, Naive Bayes ve Rassal Orman algoritmaları olmak üzere 5 farklı sınıflandırma algoritması ile eğittiklerinde en iyi ve en doğru sınıflandırma başarısını Rassal Orman algoritması ile elde etmişlerdir. Bağımsız 22 değişkenin yer aldığı veri setinin kullanıldığı çalışmada bireylerin diyabet rahatsızlığı olma olasılıklarını %84,78 doğruluk oranı ile tahmin edilebileceği sonucuna ulaşmışlardır. Bardelli vd. 2020 yılında gerçekleştirdikleri çalışmalarında elektronik faturalarda yer alan bilgileri kullanarak muhasebecilerin işini basitleştirecek akıllı bir sistem geliştirmeyi amaçlamışlardır. Çalışmada, muhasebe işlemleri yapılırken özel kodlar halinde sınıflandırılan faturaların hesap ve vergi kodlarının tahmin edilebilmesini sağlayacak çok sınıflı bir sınıflandırma algoritması önerilmiştir (Bardelli vd. 2020). Farklı sınıflandırma algoritmalarının başarı oranlarının karşılaştırıldığı çalışmada en yüksek sınıflandırma başarı oranının Rassal Orman algoritması ile elde edildiği gözlemlenmiştir. Kazan vd. 2019'daki çalışmalarında bir e-ticaret sitesinden elde ettikleri, ürün bilgileri etiketlenerek oluşturulmuş veri setini eğiterek ürünleri 9 farklı sınıfa ayıracak şekilde tahminleme yapmayı amaçlamışlardır (Kazan vd. 2019). Çalışmada veri seti farklı makine öğrenmesi algoritmaları ile eğitilerek algoritmaların ürünlerin kategorilerini doğru tahminleme başarıları karşılaştırılmıştır. Eğitim aşamasında denedikleri makine öğrenmesi algoritmaları Karar Ağacı, Random Forest, Multinominal Naive Bayes (Multinominal NB): Lojistik Regresyon, Yapay Sinir Ağları (YSA) ve Destek Vektör Makineleri (DVM) algoritmalarıdır. Çalışma sonunda ürünler 6 kategoride sınıflandırılmış ve sonuçların başarı oranları birbirleriyle karşılaştırılmıştır. Birbirine yakın özellikleri taşıyan veri setlerinde sınıflandırmadaki başarı oranının diğer veri setlerinde elde edilen başarı oranlarına göre daha düşük çıktığı gözlemlenmiştir. Çalışmada kullanılan sınıflandırma algoritmalarının ortalama başarıları karşılaştırıldığında Destek Vektör Makineleri (DVM) ve Yapay Sinir Ağları (YSA) ((Multi-Layer Perceptron (MLP) (aktivasyon='logistic sigmoid')) algoritmalarının yaklaşık olarak %97 başarı oranları ile diğer yöntemlere göre daha yüksek sonuçlar verdiği görülmüştür.

Çalışmanın algoritma seçimi aşamasında literatür taramasın kapsamında incelenen çalışmalarda sıklıkla uygulanan ve yüksek doğruluk oranı elde edilen algoritmalar tercih edilmiştir.

II. YÖNTEM

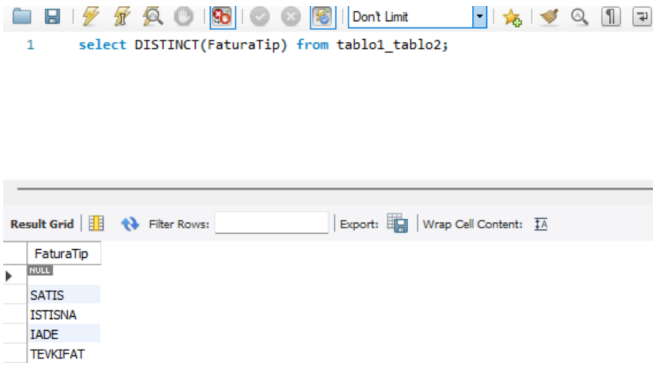
II.I Verilerin Toplanması

Bu aşamada uygulama verilerinin yer aldığı MySQL veri tabanı tabloları incelenmiş ve veri setine dahil edilebilecek olan fatura verileri ve fatura sınıf bilgisi verilerinin 3 farklı tabloda yer aldığı tespit edilmiştir. Ayrıca veri tespiti aşamasında uygulama kullanıcıları ile toplantılar gerçekleştirilerek uzman görüşüne de başvurulmuştur. Veri setine dahil edilecek veriler belirlenirken; ilerleyen süreçte geliştirilen modelin sisteme dahil edilmesi durumunda anlamlı bir şekilde modelden faydalanılabilmesi için verilerin herhangi bir işlem yapılmamış, faturaların uygulamaya yüklendikleri ilk aşamada veri tabanına otomatik olarak kayıtları gerçekleşen verilerden oluşmasına özen gösterilmiştir. Bu tablolardan birinde kayıtların birden fazla sayıda tutulabildiği anlaşılmış ve bu tablodaki fazla veriler temizlenmiştir. Sql sorguları ile uygulamada faturaların kaç farklı sınıfa ayrıldığı ve hangi

sınıftan kaç tane fatura bulunduğu bilgileri elde edilmiştir. Faturaların 7 farklı sınıfa ayrıldığı ve ulaşım ve reklam sınıflarına dahil faturaların toplam faturaların yaklaşık olarak %2'sini oluşturduğu tespit edilmiştir. Bu sınıflara ait verilerin veri setinden çıkarılmasına karar verilmiştir. Sonrasında tablolar veri bütünlüğü korunacak şekilde birleştirilmiş ve Excel formatında lokal bilgisayara aktarılmıştır.

II.II. Veri Ön İşleme Aşaması

Veri ön işleme aşamasında Excel formatına aktarılan veriler tekrar incelenerek veri temizleme işlemine devam edilmiştir. Sayısal olmayan verilerin kaç farklı tipte yer alabildikleri SQL sorguları ile tespit edilmiş ve bu değerler numerik hale getirilmiştir.



Şekil 3. Farklı fatura tipi değerlerinin tespit edildiği SQL sorgusu.

Şekil 2'de veri setine dahil edilmiş olan fatura tipi değerlerinin farklı olarak hangi isimlerle yer alabildiklerinin elde edildiği SQL sorgusu örneği yer almaktadır. Bu bilgiler elde edilirken, veri setine dahil edilen veri başlıkları için ayrı ayrı hangi veri tipinden kaç adet sonuç döndüğü de farklı SQL sorguları ile incelenmiş ve hatalı kayıtlardan veya yazım yanlışları gibi hatalardan kaynaklı olarak çok az sayıda yer alan değerler veri setinden çıkarılmıştır.

II.III. Algoritma Seçimi

Eğitim işlemleri makine öğrenmesi algoritmalarından Lojistik Regresyon, Naive Bayes, Rassal Orman, Destek Vektör Makineleri ve K- En Yakın Komşu Algoritması (K-NN) algoritmaları ile ayrı ayrı gerçekleştirilmiş ve başarı sonuçları kontrol edilerek karşılaştırılmıştır. Çalışmada veri seti %90 eğitim kümesi ve %10 test kümesi olacak şekilde belirlenmiş, verilerin %90'lık kısmı ile eğitim çalışmaları gerçekleştirilirken eğitim kümesinde kullanılmayan %10'luk kısmı ile de oluşturulan modellerin başarılarının test edilmesi amaçlanmıştır.

Lojistik Regresyon: Lojistik regresyon, ikili veya kategorik sonuçları tahmin etmek için kullanılan istatistiksel bir makine öğrenmesi algoritmasıdır. Bağımlı ve bağımsız değişkenler arasındaki ilişkiyi açıklamak için kullanılmaktadır. Lojistik regresyonda amaç, bağımsız değişkenlerin değerlerine dayalı olarak ortaya çıkan sonucun olasılığını tahmin etmektir.

Lojistik regresyonun sıralı lojistik regresyon, multinominal lojistik regresyon ve ikili lojistik regresyon olmak üzere üç farklı türü vardır. Bunlardan sıralı lojistik regresyon kategoriye göre sınıflandırmada belli periyotlarla sıralı tahminler veren çalışmalarda kullanılmaktayken multinominal lojistik regresyon belli bir sırası olmayan en az üç veya daha fazla kategorik tahminin olduğu durumlarda kullanılmaktadır. Lojistik regresyonun bir diğer türü olan ikili lojistik regresyon ise kategorik olarak sadece iki farklı sonucun olduğu durumlarda kullanılmaktadır (Aydın,2020).

Naive Bayes: Sınıflandırma işlemini, sınıflarda yer alan değişkenlerin farklı özelliklere sahip olduğu varsayımından hareketle yapan ve verideki farklı sınıfların yapılarını öğrenip yeni bir veri eklendiğinde bu yeni verinin hangi sınıfa ait olduğunu öğrenmiş olduğu sınıf yapılarına göre tespit eden basit ve etkili bir yöntemdir (Silahtaroğlu,2013). Verilerin eğitilmesi işleminin hızlı gerçekleşmesi ve tahminlemedeki başarı oranının yüksek olması Naive Bayes algoritmalarının avantajlı tarafları olarak kabul edilmekle birlikte, dezavantaj olarak ise sınıflandırma problemlerindeki karmaşıklığın yüksek olduğu durumlarda bu algoritmanın yetersiz kaldığı karşımıza çıkmaktadır (Uzun, 2007).

Rassal Orman: Denetimli makine öğrenmesi yöntemlerinden birisidir. Boyut sayısının çok yüksek olduğu verilerin eğitiminde kullanılır. Karar ağaçlarında sık rastlanan problemlerden birisi olan aşırı öğrenme probleminin aşılmasında çözüm olarak bu yöntem kullanılabilmektedir. Birbirinden bağımsız karar ağaçlarının eğitim ve test verilerindeki girdileri örneklemeye dayanmaktadır. Sınıflandırma çalışmalarında kullanıldığında bağımsız karar ağaçlarından birer tane sınıf oyu alarak oy çokluğuna göre sınıflandırma işlemini yapmaktadır. Regresyon problemlerinde ise tahmin ortalamalarına göre işlem yapmaktadır (Dangeti, 2017). Rassal Orman yönteminde veri kümesi de değişken kümesi de rastgele kullanılarak ağaçlar oluşturulur. Rastgele veri ve değişkenlerde oluşan bu rastgele ağaçlar bir araya geldiğinde rastgele bir ormanı tanımlar. Bu yöntemde sonuç iki farklı inanca dayanır. Bunlardan birincisi her bir ağacın verinin bir bölümü için doğru tahminde bulunmasıdır. İkincisi ise değişik yerlerde hatalarla karşılaşıldığıdır. Bu şekilde karar, karar ağaçları arasındaki oylama sonucuna göre verilmektedir (Gollapudi ve Laxmikanth 2016).

Destek Vektör Makineleri: Yaygın olarak sınıflandırma problemlerinde kullanılır. Gözetimli öğrenme yöntemlerinden birisidir. Bir düzlem üzerindeki farklı sınıflara ait olan noktaları, bu noktalara maksimum uzaklıktaki bir çizgi ile bölmeyi amaçlar. Karmaşık ama veri miktarının fazla büyük olmadığı çalışmalar için uygun yöntemlerden birisidir. Genellikle makine öğrenimi yöntemi olarak adlandırılan karar destek sistemleri, belirli durumlarda açıkça tanımlanamayan işlemler için, verilen girdilere karşılık gelen çıktının, mevcut bilgi ve deneyimlerin yardımıyla tahmin edildiği bir işlemdir (Alpaydın, 2004).

K- En Yakın Komşu Algoritması (K-NN): Verilerin sınıfını tespit etmek amacıyla regresyon ve sınıflandırma çalışmalarında kullanılan denetimli öğrenme yöntemlerden birisidir. Sınıflandırma işlemini, bir düzlem üzerindeki noktaların birbirine olan mesafelerini baz alarak yapmaktadır (Özkan, 2008). K-NN, bir öğrenme aşaması bulunmayan, tembel (lazy) bir öğrenme yöntemidir. Eğitim verilerini öğrenme yöntemiyle değil veri setini ezberleme yöntemiyle çalışır. Bir verinin sınıfı bu yöntemle tahmin edilmeye çalışıldığında, sınıfı tahmin edilmeye çalışılan verinin veri setinin tamamındaki komşularına bakarak tahmin sonucu vermektedir (Karadağ ve diğerleri, 2020).

II.IV. Veri Setinin Eğitilmesi

Verilerin eğitilmesi işlemi Python programlama dili ile gerçekleştirilmiştir. Bu işlemde Python'da yer alan numpy, pandas ve sklearn kütüphanelerinden faydalanılmıştır.

Çalışmada veri seti %90 eğitim kümesi ve %10 test kümesi olacak şekilde belirlenmiş, verilerin %90'lık kısmı ile eğitim çalışmaları gerçekleştirilirken eğitim kümesinde kullanılmayan %10'luk kısmı ile de oluşturulan modellerin başarılarının test edilmesi amaçlanmıştır.

III. SONUÇ

Makine öğrenmesi algoritmaları ile veri seti eğitildiğinde en yüksek başarı oranı %88 doğrulukla Karar Ağacı ve Rassal Orman algoritmaları ile elde edilmiştir. Bu algoritmalar yeni veriler girilerek test edildiğinde Karar Ağacı algoritmasında aşırı öğrenme sorunuyla karşılaşılırken, Rassal Orman algoritmasında bu soruna rastlanmadığı gözlemlenmiştir. Bu sonuç ulaşım ve reklam sınıflarına ait olanlar dışındaki faturaların makine öğrenmesi yöntemi ile %88 oranında doğru sınıflandırılabileceğini göstermiştir. Eğitim verisinde dahil edilmeyen faturaların, toplam fatura sayısının yaklaşık olarak %2'sini oluşturduğu bilinmektedir. Eğitim verisine ulaşım ve reklam sınıflarına ait faturaların verilerinin dahil edilmediği dolayısıyla bu sınıflara ait faturaların model tarafından doğru sınıflandırılamayacakları göz önünde bulundurulduğunda bütün fatura sınıfları için modelin kullanılması durumunda yaklaşık olarak %2 düzeyinde bir kayıpla %86 seviyesinde bir başarı oranının elde edilebileceği anlaşılmaktadır. Gelecek çalışmalarda derin öğrenme yöntemleri ile daha başarılı sonuçlar elde edilebilir.

REFERENCES

[1] SEBETCİ, Ö., GÜNAY, M. B., & SEBETCİ, E. (2018). İş Süreç Yönetimi (Bpm) ve İş Akış Yönetimi (Wfm) Kavramlarına Yaklaşım. AJIT-e: Academic Journal of Information Technology, 9(33), 115-126.

[2] Tektüfekçi, F. (2012). BİLGİ TEKNOLOJİLERİNİN MUHASEBE UYGULAMALARINA ENTEGRASYONU VE BÜTÜNLEŞİK SİSTEMLERLE OLAN ETKİLEŞİM. Organizasyon ve Yönetim Bilimleri Dergisi, 4(2), 51-59.

[3] BAŞER, B. Ö., YANGIN, M., & SARIDAŞ, E. S. (2021). Makine öğrenmesi teknikleriyle diyabet hastalığının sınıflandırılması. Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 25(1), 112-120.

[4] Bardelli, C., Rondinelli, A., Vecchio, R., & Figini, S. (2020). Automatic electronic invoice classification using machine learning models. Machine Learning and Knowledge Extraction, 2(4), 617-629.

[5] Kazan, S., & Karakoca, H. (2019). Makine öğrenmesi ile ürün kategorisi sınıflandırma. Sakarya University Journal of Computer and Information Sciences, 2(1), 18-27.

[6] Aydın, K.E. (2020). Web İçerik Sınıflandırması İçin Makine Öğrenmesi (Yayınlanmamış Yüksek Lisans Tezi). İstanbul Teknik Üniversitesi, İstanbul.

[7] Silahtaroğlu, G. (2013). Veri Madenciliği Kavram ve Algoritmaları (3. Basım). Papatya Bilim Yayıncılık.

[8] Uzun, E. (2007). İnternet tabanlı bilgi erişimi destekli bir otomatik öğrenme sistemi.

[9] Dangeti, P. (2017). Statistics for machine learning. Packt Publishing Ltd.

[10] Gollapudi, S. Ve Laxmikanth, V. (2016). Practical Machine Learning. Birmingham,UK: Packt Publishing.

[11] Alpaydın, E. (2004). Introduction To Machine Learning. United States Of America: MIT Press.

[12] Özkan, Y. (2008). Veri Madenciliği Yöntemler (3.Baskı). İstanbul: Papatya Yayıncılık.

[13] Karadağ, B., Bölükbaş, O. ve Ünal, M. A. (2020). Makine Öğrenmesi İle Bireysel Müşteriler İçin Finansman Ürün Önerilmesi. Academic Perspective Procedia, 3(1), 438-444.