Avrupa Bilim ve Teknoloji Dergisi
Sayı 52, S. 272-278, Kasım 2023
© Telif hakkı EJOSAT'a aittir
Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
No. 52, pp. 272-278, November 2023
Copyright © 2023 EJOSAT
Research Article

Türkçe Faturaların Sınıflandırılmasında Farklı Öznitelik Seçimi Yöntemleri ile Topluluk Öğrenme Algoritmalarının Etkilerinin İncelenmesi

İlker Yıldız [1], Ayberk Emin Kotan [2], Ayşe Berna Altınel Girgin [3*]

[1*] Ar-Ge Merkezi, NetBT Danışmanlık Hizmetleri A.Ş., İstanbul, Türkiye, (ORCID: 0000-0001-9167-2774), ilker.yildiz@net-bt.com.tr
[2] Ar-Ge Merkezi, NetBT Danışmanlık Hizmetleri A.Ş., İstanbul, Türkiye, (ORCID: 0000-0001-5085-2031), ayberk.kotan@net-bt.com.tr
[3*] Marmara Üniversitesi, Teknoloji Fakültesi, Bilgisayar Mühendisliği Bölümü, İstanbul, Türkiye, (ORCID: 0000-0001-5544-0925), berna.altinel@marmara.edu.tr

(İlk Geliş Tarihi 15 Ağustos 2023 ve Kabul Tarihi 30 Kasım 2023)

(**DOI:** 10.5281/zenodo.10439999)

ATIF/REFERENCE: Yıldız, İ., Kotan, A.B., Altınel Girgin, A.B. (2023). Türkçe Faturaların Sınıflandırılmasında Farklı Öznitelik Seçimi Yöntemleri ile Topluluk Öğrenme Algoritmalarının Etkilerinin İncelenmesi. *Avrupa Bilim ve Teknoloji Dergisi*, (52), 272-278.

Öz

Özellikle Covid-19 pandemisiyle birlikte insanların alışveriş tercihlerinin daha çok dijital ortamlara geçmesiyle birlikte şirketler ve tedarik zincirleri de ciddi manada genişledi. Bu genişleme beraberinde fatura sayılarında da artışı getiriyor. Kanunen fiziki faturaların dijitalleştirilmesi ve saklanması zorunlu hale geldi. Bu zorunlulukla beraber dijitalleşmiş faturaların otomatik olarak sınıflandırılması ve gerekli durumlarda istenilen bilgilerin otomatik olarak çıkartılması çok önemli bir ihtiyaç haline gelmiştir. Özellikle İngilizce dilindeki ve diğer dillerdeki faturaların otomatik yöntemlerle analiz edilmesi için farklı öğrenme algoritmaları içeren çeşitli çalışmalar yapılmıştır. Ancak Türkçe dilindeki faturaların otomatik olarak analiz edilmesi ve sınıflandırılması için yeterli miktarda çalışma ve kamuya açık veri kümesi olmadığı görülmektedir. Bu motivasyonla yola çıkarak bu çalışmada, farklı özellik seçimi yöntemlerinin Türkçe dilindeki faturaların topluluk öğrenme modelleri ile sınıflandırılması problemi üzerindeki performansını analiz etmeyi amaçladık. Deneylerimizi oluşturduğumuz 15k ve 50k boyutlarındaki 2 adet veri kümesi üzerinde gerçekleştirdik. Bu veri kümeleri üzerinde Bilgi Kazancı, Chi Kare, Kazanç Oranı, Geriye Beslemeli özellik seçimi yöntemlerinin K-En Yakın Komşu (KNN), Destek Vektör Makineleri (DVM), Naif Bayes (NB), Rassal Orman (RO), Adaboost topluluk öğrenme sınıflandırma algoritmalarının ve Serpme (Sprinkling) tekniğinin performans etkilerini gözlemledik. Deneysel sonuçlara göre en yüksek sınıflandırma başarısı Geriye Beslemeli özellik seçimi yöntemi ve Adaboost topluluk öğrenme algoritmasının birlikte kullanılması ile elde edilmiştir. Bildiğimiz kadarıyla bu çalışma Serpme (Sprinkling) tekniğinin topluluk öğrenme algoritmalarıyla beraber Türkçe faturaların sınıflandırılması probleminin çözümü üzerine ve bu kapsamda yapılmış ilk çalışma olma özelliğini taşımaktadır. Türkçe fatura analizi ile ilgili kaynakların yetersiz olmasından ötürü Türkçe fatura analizi üzerine yapılan çalışmalar da oldukça kısıtlı sayıdadır. Dolayısıyla, Türkçe fatura sınıflandırması alanında literatüre katkıda bulunabilmek için bu çalışmada kullanılan veri kümeleri ve geliştirilmiş algoritmalar diğer araştırmacıların erişimine açık hale getirilmiştir.

Anahtar Kelimeler: Finansal Analiz, Topluluk Öğrenme Algoritmaları, Özellik Seçimi Yöntemleri, Serpme tekniği, Makine öğrenmesi.

Analysis of the Effects of Different Feature Selection Methods and Ensemble Learning Algorithms in Classification of Turkish Invoices

Abstract

Companies and their supply chains have expanded significantly, especially with the Covid-19 pandemic, as people's shopping preferences shift to more digital environments. This expansion brings with it an increase in the number of invoices. By law, it has become mandatory to digitize and store physical invoices. With this necessity, automatic classification of digitalized invoices and automatic extraction of the requested information when necessary has become a very important need. Various studies involving different learning algorithms have been carried out, especially for the automatic analysis of invoices in English and other languages. However, there does not appear to be enough studies and publicly available datasets to automatically analyze and classify Turkish-language

[*] Sorumlu Yazar: berna.altinel@marmara.edu.tr

invoices. Based on this motivation, in this study, we aimed to analyze the performance of different feature selection methods on the problem of classification of Turkish language invoices with ensemble learning models. We performed 2 datasets of 15k and 50k sizes, in which we created our experiments. We observed the performance effects of Information Gain, Chi Square, Gain Ratio, Back-Feed feature selection methods on K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Naive Bayes (NB), Random Forest (RO), Adaboost ensemble learning classification algorithms and Sprinkling technique on these datasets. According to the experimental results, the highest classification success was obtained by using the Back-Feed feature selection method and Adaboost ensemble learning algorithm together. As far as we know, this study is the first study on the solution of the problem of classification of Turkish invoices using the Sprinkling technique with ensemble learning algorithms. Therefore, in order to contribute to the literature in the field of Turkish invoice classification, the datasets and improved algorithms used in this study have been made available to other researchers.

Keywords: Financial Analysis, Ensemble Learning Algorithms, Feature Selection Methods, Sprinkling technique, machine learning.

1. Giriş

Satın alma süreci, tedarikçi araştırılması sürecinden başlayarak faturalaşma ve sonrasında ödeme ile sonlanan ve sürekli tekrar eden bir süreçtir. Rekabetin gün geçtikçe arttığı günümüzde şirketlerin fatura sayılarına bağlı olarak muhasebeleşme, gider kontrolü ve maliyet dengeleme konularına odaklandığı gözlemlenmektedir. Bu konuların haricinde faturalar üzerinden elde edilen bilgilerin geçmişe ve geleceğe yönelik finansal analizlerinin önemi gün geçtikçe artmaktadır.

Fatura, satıcının müşteriye satışın ayrıntılarını sunmak ve ödeme yapıldığını göstermek için kullanılan bir ticari belgedir. Genellikle işletmeler arasındaki ticari işlemlerde kullanılmaktadır. Fatura üzerinde fatura numarası, tarih, satıcı bilgileri, alıcı bilgileri, ürün veya hizmet ayrıntıları, vergiler, ödeme koşulları, işletme logosu ve iletişim bilgileri gibi bilgiler yer almaktadır. Ayrıca fatura; muhasebeleştirme, vergi beyannamesi hazırlama, envanter yönetimi ve finansal raporlama gibi işlemlerde önemli bir rol oynamaktadır. Fatura birden fazla türde oluşturulabilmektedir. Fatura türleri genellikle ülkelere ve işletmelere göre değişiklik gösterebilmektedir. Türkiye'de e-Arşiv, e-Fatura, e-İrsaliye , e-Serbest Meslek Makbuzu gibi bir çok fatura türü bulunmaktadır. Bu fatura türlerinden bazıları elektronik ortamda oluşturulup elektronik ortamda teslim edilmektedir. Bazı faturalar ise elektronik ortamda oluşturulup müşteriye kağıt olarak ya da elektronik ortamda görüntü formatında teslimi gerçekleştirilmektedir. Teslim alınan faturaların gelir ve gider kaydı, vergi uyumu, mali analiz ve raporlama, denetim ve iç kontrol için tekrardan dijitalleştirilip muhasebeleştirilmesi gerekmektedir.

Bir faturanın muhasebeleştirilmesi sürecinde ilk olarak faturanın doğruluğu, tarihleri, tutarları ve diğer detayları insan gücü kullanılarak kontrol edilmektedir. Kabul edilen fatura muhasebe yazılımı veya muhasebe defterine kayıt altına alınır. Fatura kayıt edilirken faturanın türü, mal alımları veya hizmet alımları uygulanan muhasebeleştirilme adımları farklı olduğu için fatura içeriğinin belirlenmesi ve o alana uygun olarak kayıt edilmesi gerekmektedir. Bu nedenle yetkili çalışanlar tarafından faturanın türlerine ve mal hizmet bilgisine göre ayrıştırılarak kayıt altına alınmaktadır. Faturanın doğru alanlara kaydı yapıldıktan sonra KDV (katma değer vergisi) muhasebe işlemleri gerçekleştirilmektedir. Bu işlem gerçekleştirilirken varsa faturada bulunan farklı KDV değerleri ayrı ayrı incelenmesi ve sisteme işlenmesi gerekmektedir. Bu işlemlerin tamamında insan gücü kullanılması nedeniyle büyük ölçekli işletmelerde on binlerce faturanın muhasebeleştirilmesi sürecinde uzun çalışma sürelerine ve yoğun insan gücüne ihtiyaç duymaktadır.

Gelir İdaresi Başkanlığı muhasebeleştirme sürecinde yüksek insan gücünün ve hata oranlarının azaltılması için karekod geliştirmelerini zorunlu hale getireceğini duyurmuştur. Bu geliştirmeler incelendiğinde faturanın muhasebeleştirilmesi için yeterli bilgi bulunmadığı için gelecekte yüksek insan gücüne ihtiyacın devam edileceği görülmektedir. Bu nedenle fatura üzerinden mal/hizmet bilgilerin alınması ile otomatik muhasebeleştirmenin gerçekleştirilmesi ve bu alandaki insan gücünün minimuma indirilmesi daha da önem kazanmıştır. Ayrıca faturalar üzerinden elde edilen veriler kullanılarak yapay zeka destekli finansal analiz araçları ile işletmenin finansal, müşteri, ürün ve tedarik zinciri analizlerinin gerçekleştirilmesi de hem akademik hem de ticari platformlar nezdinde birer cazibe merkezi haline geldiği görülmektedir.

2022 itibariyle ülkemizde 600.000 aşkın e-fatura mükellefi bulunmaktadır. Ülkemiz ticaretinin %93,45'i elektronik faturalama üzerinden oluşmaktadır. Özellikle Avrupa'da pek çok ülke e-fatura sistemini devreye almıştır. İşletmeler arasındaki ticari işlemlerde standart belge formatlarının kullanılması şeffaflığı sağlanırken, kamu otoriterlerinin denetimini kolaylaştırmaktadır. Türkçe faturalar üzerine analiz ve sınıflandırma çalışmaları oldukça az sayıdadır. Ayrıca yine diğer araştımacıların erişimine açık hale getirilmiş veri kümesi ve geliştirme ortamı gibi kaynaklar da oldukça kısıtlıdır. Ülkemizde faturalama verilerine odaklanan benzer bir analiz ve otomatik muhasebeleşme ürünü yapılan araştırmalar sonucunda görülmemiştir. Bu çalışma ile Türkçe faturalar üzerine doğal dil işleme ve makine öğrenmesi yöntemleri ile sınıflandırma yapmak ve ayrıca farkli özellik seçimi yöntemleirnin bu sınıflandırma algoritmaları üzerindeki etikelerini gözlemlemeyi amaçladık.

Bu çalışmamızın literatüre katkıları şu şekildedir:

1)Bildiğimiz kadarıyla Türkçe faturaların sınıflandırılması amacıyla, hem doğal dil işleme hem de makine öğrenmesi tekniklerini topluluk öğrenmesi mimarisi içinde Serpme (Sprinkling) tekniğini de kullanan bu kapsamdaki ilk çalışmadır.

2) Veri kümelerini ve deney ortamını diğer araştırmacıların erişimine açık hale getirilerek bu anlamda literatüre katkı yapılması amaçlanmaktadır.

Bu çalışmanın geri kalan bölümleri şu şekildedir: Bölüm 2'de Türkçe dilindeki faturalar üzerine yapılan çalışmalardan farklı örnekler verilmektedir. Bu çalışmamız kapsamında kullanılan algoritmalar Bölüm 3'te detaylı anlatılmaktadır. Yine, bu çalışmamızda

kullanılan veri kümeleri, deneysel sonuçlar ve tartışmalar Bölüm 4'te listelenmektedir. Sonuç ve gelecek çalışmalar ise Bölüm 5'te sunulmaktadır.

2. Literatür Taraması

Özellikle İngilizce dilindeki ve diğer dillerdeki faturaların otomatik yöntemlerle analiz edilmesi için farklı öğrenme algoritmaları içeren çeşitli çalışmalar yapılmıştır[1],[2],[3],[4], [5],[6],[7],[8]. Ancak Türkçe dilindeki faturaların otomatik olarak analiz edilmesi ve sınıflandırılması için yeterli miktarda çalışma ve kamuya açık veri kümesi olmadığı görülmektedir.

Kılınç (2016) yaptığı çalışmada topluluk öğrenme modellerinin Türkçe metin sınıflandırma üzerindeki etkilerini incelemiştir. Dört temel sınıflandırıcı (NB, KNN, DVM, J48) ve üç topluluk öğrenme modeli (Bagging, Boosting, Rotation Forest) kullanarak TTC-3600 veri kümesinde deneysel değerlendirmelerde bulunmuştur. Veri kümesinde ekonomi, kültür, sanat, sağlık vb. alanlarından oluşan 3600 doküman içermektedir. Deney sonuçları kolektif öğrenme modellerinin Türkçe metin sınıflandırmada temel sınıflandırıcıların doğruluk sonucunu çoğunlukla artırdığını göstermektedir [5].

Wang vd. metin sınıflandırma doğruluğunu geliştirmek amacıyla metin uzunluğuna duyarlı uyarlanabilir bir Önyükleme Toplama (Bagging) topluluk öğrenme algoritması önermektedir. Önerdikleri algoritma farklı kategorilerin oranlarını korurken uzun ve kısa metin örneklerinin alt kümelerini oluşturmak için uyarlanabilir bir eşik grubuna dayalı rastgele örnekleme yöntemi kullanmaktadır. Temel sınıflandırıcı olmak üzere altı tipik derin öğrenme yöntemi kullanmışlardır: LSTM, TextCNN, RoBERT, LSTM_AT, DPCNN, FastText. Deneysel sonuçlar, önerilen algoritmanın F1, duyarlılık ve özgüllük açısından temel yöntemlerden daha iyi performans gösterdiğini doğrulamaktadır [6].

Arslan ve Uymaz (2022) faturaların dijital ortama aktarılmasını otomatikleştirmeyi amaçlamıştır. Bir banka sisteminden alınan dört farklı türde fatura görüntüleriyle orijinal bir fatura veri kümesi hazırlamıştır. Ayrıca orijinal veri kümesine Sıfır Doldurma, Parlaklık Artırımı uygulanarak iki veri kümesi daha elde edilmiştir. LeNet-5, VGG-19 ve MobileNetV2 adlı Evrişimli Sinir Ağları (CNN) mimarileri üç farklı veri kümesi ile eğitilerek fatura sınıflandırma sistemi geliştirmiştir. Veri büyütme yöntemi kullanılarak elde edilen veri kümesi ile eğitilen modelde %99,83 sınıflandırma başarısına ulaşılmıştır [7].

Tarawneh vd. (2019), elle yazılmış, baskı alınmış ve makbuz olmak üzere üç fatura türünü otomatik bir yaklaşımla sınıflandırmayı önermektedir. Önerilen yöntemde AlexNet derin evrişimli sinir ağı kullanılarak özelliklerin çıkarılması önerilmektedir. RO, KNN ve NB gibi çeşitli makine öğrenme algoritmaları kullanılarak elde edilen özellikler sınıflandırılmaktadır. %98,4 doğruluk oranı ile en iyi sınıflandırma sonucu KNN algoritması ile sağlanmıştır [8].

Literatürdeki bir başka çalışmada [12], geleneksel TF-IDF terim ağırlıklandırma tekniği ve Doc2Vec yöntemi ile sayısallaştırılan 4 farklı Türkçe veri kümesi üzerinde bireysel sınıflandırıcılar ve topluluk öğrenmesi algoritmalarının başarım değerleri kıyaslanmıştır. Geleneksel makine öğrenmesi yöntemleri olan Lojistik Regresyon (LR), KNN, NB, Karar Ağaçları (Decision Trees- DT), DVM, Çok Katmanlı Algılayıcılar (Multi-Layer Perceptrons- MLP) gibi bireysel sınıflandırıcılar ve RO, Adaboost (AB), Bagging (BG) gibi topluluk öğrenmesi metotları da kullanılmıştır. Topluluk öğrenmesi algoritmalarında karar verme aşamasında Çoğunluk oylaması (Majority Voting) tekniği kullanılmıştır. Çalışma içerisinde raporlanan deney sonuçlarına göre,IDF ve Doc2Vec metotları ile elde edilen vektörlerin birleşimine uygulanan sınıflandırma algoritmalarında diğerlerine kıyasla daha yüksek sınıflandrıma başarımı elde edildiği gözlenmiştir.

3. Yöntem

A. Sınıflandırma Algoritmaları

1) Destek Vektör Makineleri: DVM, regresyon ve sınıflandırma problemini çözmek için kullanılan bir makine öğrenmesi algoritmasıdır. DVM hiperdüzlem bulmaya çalışarak veri noktalarını farklı sınıflara ayırır. Bulunan bu hiperdüzlem sınıfların arasında bulunan en büyük mesafeyi maksimize etmeye çalışır. Model eğitimi sonucunda elde edilen hiperdüzlem ile yeni örneklerin hangi sınıflara ait olduğu tahmin edilmeye çalışır [9].

2) Rassal Orman Algoritması: RO, birçok karar ağacının bir araya gelerek oluşturulan bir makine öğrenmesi algoritmasıdır. Her bir ağaç rastgele özellikler ve örnekler kullanılarak eğitilmektedir. Sınıflandırma için her bir karar ağacı tahminleme işlemi gerçekleştirir ve en yüksek oy alan sınıf etiketi tahmin olarak seçilmektedir. Rassal orman algoritması aşırı uyum eğilimini azaltırken, daha iyi bir tahmin performansı sağlayabilen esnek bir makine öğrenmesi modelidir.

3) Naif Bayes Algoritması: NB, Bayes Teoremi'ne dayanan ve temel olarak istatistiksel bir sınıflandırma algoritmasıdır. Algoritma bir örneğin her bir sınıfa ait olma olasılığını hesaplar ve en yüksek olasılığa sahip sınıfı tahmin etmektedir. Ayrıca genel olarak birbirinden bağımsız öznitelikler, sınıf etiketlerine göre veri setindeki dağılımını da incelemektedir.

4) K-En Yakın Komşu Algoritması: KNN algoritması, veri kümesindeki örneklerin bir uzayda dağıldığını ve aralarındaki uzaklık mesafelerine göre verileri sınıflandırmaktadır. Sınıflandırma yaparken yeni örneği sınıflandırmak için çevresindeki k en yakın komşuyu incelemektedir. Bu komşular genellikle öklidyen mesafe veya benzeri bir metrik kullanarak hesaplanmaktadır. Hesaplama işleminden sonra k komşunun etkilerine bakarak test örneğinin sınıfını belirlemektedir.

5) Birleştirici Adaboost Algoritması: Birleştirici Adaboost algoritması zayıf öğrenicileri bir araya getirerek güçlü bir sınıflandırıcı oluşturmayı amaçlayan bir makine öğrenmesi algoritmasıdır. Adaboost algoritması zayıf öğrencileri bir araya getirilerek birlikte

çalışmasını sağlamaktadır. Bu çalışma sayesinde yüksek öğrenme seviyesine sahip öğrenciler, zayıf öğrenme seviyesinde öğrenme düzeyine sahip olan öğrencilerin zayıflığını dağıtılan ağırlık oranları ile telafi etmektedir. Bu sayede, daha yüksek bir sınıflandırma potansiyeline sahip bir model elde edilmesini amaçlayan bir algoritmadır.

6) Serpme (Sprinkling) Tekniği kullanılarak Türkçe Faturaların DVM ve NB Algoritmaları ile Sınıflandırılması:

Serpme tekniği, belgelerin sınıf etiketlerini, eğitim veri kümesine yeni ek terim olarak ekleyen bir tekniktir [11]. Bu teknik ile, eğitim aşamasında sınıf temelli ilişkileri güçlendirmek amaçlanmaktadır. Genel olarak serpme tekniği, *d* adet belge, *t* adet terim, *c* adet sınıf içeren *d x t* boyutlarındaki bir matrise *c* adet yeri öznitelik eklenmesi ile *d x (t+c)* boyutlarında yeni bir matris elde edilmesi şeklindedir.

B. Öznitelik Değerlendirme Yöntemleri

Öznitelik değerlendirme yöntemleri, özniteliklerin regresyon veya sınıflandırma problemlerinde önem sırasını belirlemek ve katkısını ölçmek için kullanılan yöntemlerdir. Bu araştırmada Kazanç Oranı *(Gain Ratio Attribute Eval)*, Ki-Kare Özellik Değerlendirici *(Chi-Squared Attribute Eval)*, Bilgi Kazancı *(Info-Gain Attribute Eval)*, Geriye Doğru Eleme *(Backward Elimination)* kullanılmıştır [10].

1) Kazanç Oranı: Kazanç oranı, bilgi kazancını sınıflandırma problemlerinde normalize etmek için kullanılmaktadır. Kazanç oranı, özniteliklerin bölünme yetenekleri ve veri setindeki sınıflar arasında homojenliği artırmadaki etkisini ölçmektedir. Bir özniteliğin bilgi kazancı yüksek olsa bile bölünme yeteneğinin düşük olması durumunda kazanç oranı Kazanç oranı yöntemi sınıflandırma algoritmalarında yüksek kazanç oranına sahip olan özniteliklerin sınıflandırmadaki etkisinin yüksek olduğunu belirtmektedir. Kazanç oranı formül hesabı Eşitlik (1)'de verilmiştir.

$$Kazanç\ Oranı = Bilgi\ Kazancı / H_{Özellik} \quad (1)$$

2) Ki-Kare Özellik Değerlendirici: Ki Kare, özellik değerlendirici bir veri kümesindeki özniteliklerin sınıflandırma ve regresyon yöntemlerinde önemini belirlemek için kullanılan istatistiksel bir yöntemdir. Bu yöntem özniteliklerin hedef değişkenle bağımsız olduğunu varsayılarak beklenen frekanslar ve gözlenen frekanslar arasındaki farkları incelemektedir. Yüksek frekans farklılıklarına sahip olan özniteliklerin önemli bir bilgi sağladığı varsayılmaktadır.

3) Bilgi Kazancı: Bilgi kazancı, özniteliklerin veri setindeki sınıflar arasındaki olasılıklarına ve dağılışını incelemektedir. Özniteliklerin veri kümesinde ne kadar iyi dağıldığına ve homojenliği ne kadar artırdığına bakarak bilgi kazandı parametresini hesaplamaktadır. Bilgi kazancı değeri yüksek olan öznitelikler sınıflandırmada daha fazla bilgi sağlandığı düşünülerek önemli öznitelikler olarak değerlendirilmektedir. Bilgi kazancının formül hesabı Eşitlik(2)'de verilmiştir.

$$Bilgi\ Kazancı = H_{Sınıf} - H_{Sınıf/Özellik} \quad (2)$$

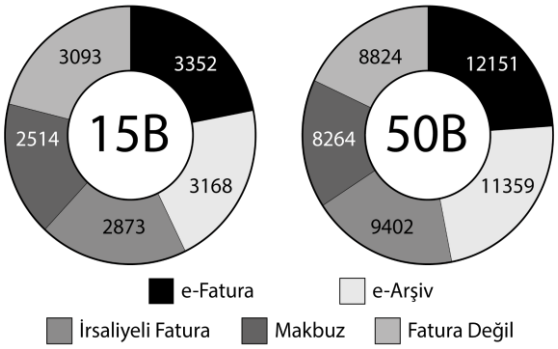
Geriye Doğru Eleme: Geriye doğru eleme, özellik seçiminde kullanılan bir yöntemdir. Belirlenen tüm özniteliklerden başlanarak istenmeyen öznitelikler adım adım veri kümesinden çıkartılarak eğitim gerçekleştirilmektedir. Eğitilen her modelin doğruluk oranı, performansı ve hata payı değerlendirilerek özniteliğin model üzerindeki etkisi gözlemlenmektedir. Bu işlem tüm öznitelikler veri kümesinden çıkarılana kadar devam etmektedir.

4. Veri Kümesi ve Deneysel Sonuçlar

4.1 Veri Kümesi

Birinci model eğitimi için iki farklı veri kümesi hazırlanmıştır. İki veri kümesinde de fatura türlerine göre e-fatura, e-arşiv, irsaliyeli fatura, makbuz ve fatura değil olarak beş farklı sınıf bulunmaktadır. İlk veri kümesi 15,000 adet fatura verisinden oluşmaktadır. Bu fatura verilerinden 3.352 adet e-fatura, 3.168 adet e-arşiv, 2.873 adet irsaliyeli fatura, 2.514 adet makbuz ve 3.093 adet fatura olmayan veri bulunmaktadır.

İkinci veri kümesi ise 50,000 adet fatura verisinden oluşmaktadır. Bu fatura verilerinden 12.152 adet e-fatura, 11.359 adet e-arşiv, 9.402 adet irsaliyeli fatura, 8.264 adet makbuz ve 8.824 adet fatura olmayan veri bulunmaktadır. Her iki veri kümesi için fatura örnekleri tamamen metinsel formattadır. Veri kümeleri için veri dağılımı Şekil 1'de gösterilmiştir.



Şekil 1. Veri Dağılımı (Figure 1. Data Distribution)

4.2 Veri Kümesi Ön İşleme ve Deney Ortamı

Veri kümesinde, fatura üzerinde bulunan verilerin anlamı daha net hale getirmek, gürültüyü azaltmak ve veri kümesini daha tutarlı hale getirmek amacıyla metin ön işleme yöntemleri kullanılmıştır. Bu metin ön işleme yöntemleri Zemberek kütüphanesindeki fonksiyonlar kullanılarak gerçekleştirilmiştir. Veri kümesi üzerinde boşluklu yapıların kaldırılması, kelime normalizasyon işlemleri, noktalama işaretlerinin kaldırılması, sayıların kaldırılması ve fatura türünün belirlenmesinde etki etmeyecek fakat kelime frekans sayısı çok olan kelimeler kaldırılmıştır. Elde edilen veriler üzerinde son olarak tokenize işlemi, morfolojik operasyonlar, kök bulma ve kök normalizasyon işlemleri uygulanmıştır.

Ön işleme yöntemleri uygulandıktan sonra veri kümesi üzerinde eğitim gerçekleştirmek amacıyla kelime frekans veri kümesi hazırlanmıştır. Elde edilen veri kümesinin %60'ı eğitim, %20'si test ve %20'si validasyon veri kümesi olarak bölünmüştür. Oluşturulan eğitim veri kümesi üzerinde özellik seçim yöntemleri uygulanarak hiper parametre optimizasyonu uygulanarak KNN, DVM, RO ve Birleşik Adaboost algoritmaları ile eğitim gerçekleştirilmiştir.

Bu çalışma HP Z4-G4 iş istasyonu üzerinde gerçekleştirilmiştir. İş istasyonunda Intel(R) Xeon(R) W-2155 CPU işlemcisi, NVIDIA Quadro RTX 6000 ekran kartı ,4 TB SSD depolama alanı, 32 GB DDR4 (2933Mhz) RAM bulunmaktadır. Çalışma ortamı olarak Ubuntu 20.04 kullanılmıştır. Programlama dili olarak Python, Zemberek, Sklearn kütüphaneleri kullanılmıştır. Hazırlanmış olan veri kümeleri ve deney ortamı [1] diğer araştırmacıların erişimine açık hale getirilmiştir.

4.3 Deneysel Sonuçlar

Veri ön işleme çalışmaları uygulanarak birbirinden bağımsız 15,000 ve 50,000 faturadan oluşan iki adet veri kümesi hazırlanmıştır. Bu veri kümeleri üzerinde kazanç oranı, ki-kare, bilgi kazancı ve geri doğru eleme özellik seçim yöntemleri kullanılarak. K-Yakın Komşu Algoritması, Rassal Orman, Naif Bayes, DVM ve Birleşik *(Ensemble)* Adaboost algoritmalarının eğitimleri gerçekleştirilmiştir.

Tablo 1. 15,000 Fatura ile Algoritmaların Sınıflandırma Başarımları - F1 Puan (%)

(Table 1. Classification Performance of Algorithms with 15,000 Invoices - F1 Score (%))

	Kazanç Oranı	**Ki-Kare**	**Bilgi Kazancı**	**Geriye Doğru Eleme**
KNN	94.52	93.97	92.92	93.47
DVM	94.79	94.58	93.74	92.76
NB	95.42	95.46	94.93	94.86
RO	94.98	95.01	94.82	95.91
Adaboost	94.62	95.32	94.97	**96.11**

15,000 adet faturadan oluşan veri kümesinde gerçekleştirilen eğitim sonuçları incelendiğinde en yüksek F1 Skor değeri Birleşik Adaboost ve Geriye Doğru Eleme özellik seçimi yöntemi kullanılan modelde %96.11 olarak elde edilmiştir. Bu değeri %95.91 ile Rassal Orman ve Geriye Doğru Eleme yöntemlerinin kullanıldığı model takip etmektedir. KNN ve DVM en iyi sonuçları özellik seçimlerinden Kazanç Oranı ile sağlamıştır. Naif Bayes en iyi sonucu özellik seçimi yöntemlerinden Ki-Kare kullanıldığında sağlamıştır. Rassal Orman ve Adaboost ise en iyi değeri Geriye Doğru Eleme yöntemi ile birlikte kullanıldığında sağlamıştır. Kullanılan diğer modellerin ve algoritmaların sonuçları Tablo 2' de verilmiştir.

[1] https://github.com/ayberk-kotan/TRInvoiceClassificationProject

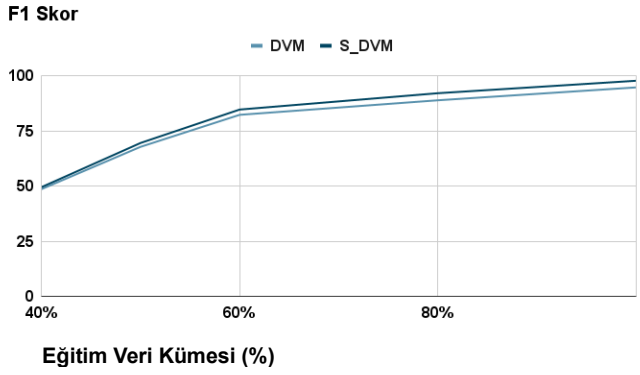
Tablo 2. 50,000 Fatura ile Algoritmaların Sınıflandırma Başarımları -F1 Puan (%)

(Table 2. Classification Performance of Algorithms with 50,000 Invoices - F1 Score (%))

	Kazanç Oranı	**Ki-Kare**	**Bilgi Kazancı**	**Geri Doğru Eleme**
KNN	90.91	89.98	90.51	92.58
DVM	91.45	91.04	92.72	93.81
NB	91.94	90.54	93.68	93.79
RO	90.78	91.67	93.91	93.68
Adaboost	92.94	92.04	92.17	**94.81**

50,000 adet faturadan oluşan veri kümesi üzerinde yapılan eğitim sonucunda en yüksek başarı oranı Geriye Doğru Eleme özellik seçimi yöntemi ve Birleşik Adaboost kullanılarak %94.81 ile elde edilmiştir. Birleşik Adaboost'un fatura sınıflandırma problemi üzerinde diğer bireysel sınıflandırıcılara göre daha yüksek sınıflandırma başarısı vermesi algoritmanın birden çok sınıflandırıcının güçlü yönlerini kullanarak her adımda kendini iyileştiren birleşik bir öğrenme modeli yapısına sahip olması ile açıklanabilir. Genel olarak özellik seçimlerinden Geriye Doğru Eleme yöntemi kullanıldığında modeller en iyi sonuçları vermektedir. Kullanılan diğer model ve yöntemlerin sonuçları Tablo II'de verilmiştir. İki veri kümesinde uygulanan yöntemler incelendiğinde en yüksek doğruluk oranının Birleşik Adaboost ve geriye doğru eleme yöntemi ile olduğu görülmüştür.

Serpme (Sprinkling) tekniği kullanılarak DVM deney sonuçları alındı ve geleneksel DVM deney sonuçları ile kıyaslandı. F1 Skoru açısından incelendiğinde Serpme tekniği kullanılan DVM modelinde geleneksel DVM modeline kıyasla 3,01 artış görülmektedir. Bu kıyaslamalar Şekil 2'de gösterilmektedir. Sınıflandırma başarımındaki bu artışın Serpme (Sprinkling) tekniğinin veri kümesinin vektörleşmesi aşamasında modele kattığı yeni bir etiketin, anlamsal boyutta sınıflandırma pozitif etkisi olmasından ötürü olabilir.



Şekil 2. Algoritmaların Sınıflandırma Başarımları - F1 Puan (%)

Figure 2. Classification Performance of Algorithms - F1 Score (%)

5. Sonuç ve Gelecek Çalışmalar

Özellikle pandemi süreciyle beraber dijital alışveriş platformları daha fazla kullanılmaya başlandı. Pandeminin bitmesine karşın birçok kişinin kazandığı bu alışkanlık konforunu terk etmemesi vesilesiyle eticaret siteleri çokça fatura üretmeye başladılar. İngilizce fatura analiz çalışmaları incelendiğinde birçok farklı spekturumda farklı içeriklerde faturaları otomatik olarak analiz edebilen çalışmalar olduğu görülmeketedir. Türkçe dilindeki faturaların otomatik olarak analiz edilmesi ve sınıflandırılması için yeterli miktarda çalışma ve kamuya açık veri kümesi olmadığı gözlemlenmiştir. Bu motivasyonla yola çıkarak bu çalışmada, farklı özellik seçimi yöntemlerinin Türkçe dilindeki faturaların topluluk öğrenme modelleri ile sınıflandırılması problemi üzerindeki performansını analiz ettik. Oluşturduğumuz 15k ve 50k boyutlarındaki 2 adet veri kümesi üzerinde gerçekleştirdiğimiz deney sonuçlarına göre en yüksek başarımın Geriye Beslemeli özellik seçimi yöntemi ve Adaboost topluluk öğrenme algoritmasının birlikte kullanılması ile elde edildiğini gözlemledik. Bildiğimiz kadarıyla bu çalışma Türkçe faturaları üzerine bu kapsamdaki ilk çalışma olma özelliğini taşımaktadır. Bu çalışmada oluşturulan veri kümesi ve geliştirilen algoritmalar, talep etmeleri halinde dünyanın her yerindeki tüm araştırmacılarla paylaşılabilir.

Oluşturduğumuz bu deney ortamını farklı öğrenme algoritmalarını entegre ederek daha da geliştirmeyi planlamaktayız. Ayrıca veri kümelerimizi büyütmek te yine gelecek çalışması olarak planladığımız maddeler arasındadır. Bununla birlikte, çizge tabanlı yapay sinir ağı modelellerini geliştirerek fatura analiz sistemimizde uygulamayı da düşünmekteyiz. Tüm bunlara ek olarak, elimizde bulunan görüntü formatındaki fatura verisinin oluşturduğumuz modele entegre edilmesi de yine gelecekte yapmayı düşündüğümüz maddeler arasındadır.

Kaynakça

[1] M. B. Wattar, "Analysis and Comparison of invoice data extraction methods," Doctoral dissertation, PhD thesis, University of Applied Sciences, 2021.Lee, K.-F., *Automatic Speech Recognition: The Development of the SPHINX SYSTEM,* Kluwer Academic Publishers, Boston, 1989.

[2] A. Khan, "Comparison of machine learning approaches for classification of invoices," Master's thesis, 2020.

[3] Ö. Arslan, "Evrişimsel sinir ağları ve metin benzerliği kullanılarak fatura görüntülerinde sınıflandırma," Master's thesis, Konya Teknik Üniversitesi, 2021.

[4] K. M. Yindumathi, S. S. Chaudhari and R. Aparna, "Analysis of Image Classification for Text Extraction from Bills and Invoices," 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kharagpur, India, 2020, pp. 1-6, doi: 10.1109/ICCCNT49239.2020.9225564.

[5] D. Kılınç, "The effect of ensemble learning models on Turkish text classification," Celal Bayar University Journal of Science, vol. 12, no. 2, 2016.

[6] Y. Wang, J. Liu, and L. Feng, "Text length considered adaptive bagging ensemble learning algorithm for text classification," Multimedia Tools and Applications, pp. 1-26, 2023.

[7] Ö. Arslan and S. A. Uymaz , "Classification of Invoice Images By Using Convolutional Neural Networks", Journal of Advanced Research in Natural and Applied Sciences, vol. 8, no. 1, pp. 8-25, Mar. 2022, doi:10.28979/jarnas.953634

[8] A. S. Tarawneh, A. B. Hassanat, D. Chetverikov, I. Lendak and C. Verma, "Invoice Classification Using Deep Features and Machine Learning Techniques", 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology

[9] (JEEIT), Amman, Jordan, 2019, pp. 855-859, doi: 10.1109/JEEIT.2019.8717504

[10] X. Hu and R. Zhang, "Text classification based on machine learning", 2022 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA), Dalian, China, 2022, pp. 911-916, doi: 10.1109/ICAICA54878.2022.9844556.

[11] H. Budak, "Özellik seçim yöntemleri ve yeni bir yaklaşım", Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi, cilt. 22, ss. 21-31, 2018.

[12] S. Chakraborti, R. Lothian, N. Wiratunga, S. Watt, Sprinkling: Supervised Latent Semantic Indexing. In European Conference on Information Retrieval 2006, 510-514. Springer Berlin Heidelberg.

[13] D. Kınık & A. Güran, "TF-IDF ve Doc2Vec Tabanlı Türkçe Metin Sınıflandırma Sisteminin Başarım Değerinin Ardışık Kelime Grubu Tespiti ile Arttırılması", Avrupa Bilim ve Teknoloji Dergisi, (21), 323-332, 2021.