İnşaat Şirketi Müşterilerinin Gelecekteki Konut Satın Alma Davranışlarının Metin Madenciliği ve Makine Öğrenmesi ile Tahmin Modellerinin Oluşturulması

Araştırma Makalesi/Research Article

Haydar EKELİK[1]*, Şenol EMİR[2]

[1]İstanbul Üniversitesi (Ekonometri Bölümü, İstanbul Üniversitesi, İstanbul, Türkiye)
[2] İstanbul Üniversitesi (Ekonometri Bölümü, İstanbul Üniversitesi, İstanbul, Türkiye)

haydar.ekelik@istanbul.edu.tr, senol.emir@istanbul.edu.tr
(Geliş/Received:14.05.2024; Kabul/Accepted:19.10.2024)
DOI: 10.17671/gazibtd.1484123

***Özet*—** Bu çalışmada, inşaat sektöründe faaliyet gösteren bir işletmenin müşterileriyle yüz yüze veya telefonla yapılan görüşmelerinin kayıtlarına çeşitli metin madenciliği ve makine öğrenmesi teknikleri uygulanmıştır. Temel amaç, bu metin tabanlı doküman kümesinden (korpus), yeni görüşme yapılan herhangi bir müşterinin ileride şirketten konut satın alıp almayacağını doğru bir şekilde tahmin edebilecek bir model geliştirmektir. Bu amaçla metinsel verilere bir takım veri ön işleme aşamaları uygulandıktan sonra anahtar kelimeler ve vektör uzay modeli oluşturmuş ve metin tabanlı veri analize uygun formata dönüştürülmüştür. CART(Classification And Regression Tree), RF(Random Forest) ve XGBoost(eXtreme Gradient Boosting) makine öğrenmesi yöntemleri uygulanarak farklı tahmin modelleri oluşturulmuş ve daha sonra bu modeller farklı sınıflandırma ölçütlerine göre karşılaştırılmıştır. Sınıflandırma problemlerinde sınıflardaki gözlem sayıları arasında dengesizlikler olması durumunda yaygın sınıflandırma ölçütlerine göre modellerin karşılaştırılması yanlı sonuçlar verebilmektedir. Bu nedenle literatürde bu gibi durumlar için genel karşılaştırma ölçütlerine ek olarak yeni ölçütler geliştirilmiştir. Çalışmadaki uygulamada da sınıflar arası dengesizlik olduğundan bu ölçütlerden birisi olan PR (Precision- Recall) eğrileri kullanılmıştır. Analiz sonucunda, PR eğrileri dikkate alındığında, görüşme yapılan yeni müşterilerin ileride konut alıp almayacağını en iyi tahmin eden yöntemin Random Forest olduğu görülmüştür.

***Anahtar Kelimeler*—** metin madenciliği, pazarlama, random forest, CART, XGBoost

Predictive Modeling of Future Home Purchase Behavior in Construction Company Customers Using Text Mining and Machine Learning

***Abstract*—**In this study, various text mining and machine learning techniques were applied to the recordings of face-to-face or telephone interviews with customers of a company operating in the construction industry. The main objective is to develop a model from this set of text-based documents (corpus) that can accurately predict whether a new customer interviewed will purchase a house from the company in the future. For this purpose, a number of data preprocessing steps were applied to the textual data, then keywords and vector space model were created and the text-based data were converted into a format suitable for further analysis. Different prediction models were created by applying CART(Classification And Regression Tree), RF(Random Forest) and XGBoost(eXtreme Gradient Boosting) methods and then these models were compared according to different classification metrics. In classification problems, imbalances between classes make it difficult to compare models. For this reason, in literature new metrics have been developed in addition to the classical performance metrics. Since there is an imbalance between classes in the application in this study, PR (precision-recall) curves, one of the developed criteria, were used. As a result of the analysis, when the PR curves are taken into account, it is seen that Random Forest shows the best performance for predicting whether interviewed new customers will buy a house in the future.

***Keywords*—**text mining, marketing, random forest, CART, XGBoost

1. GİRİŞ (INTRODUCTION)

Günümüzde bilgi ve internet teknolojilerindeki gelişmeler ve bu teknolojilerin yaygınlaşmasıyla analizlerde kullanılabilecek veri kaynaklarının her geçen gün çeşitlendiğini gözlemlemekteyiz. İnternet kullanım oranlarında artış ile gerek internet sitelerinin gerekse sosyal medyanın kullanımının hızla arttığına tanık olmaktayız. Bütün bu gelişmeler bir yandan ulaşılabilen verinin hacimce büyümesini sağlarken bir yandan da verinin çeşitliliğini arttırarak araştırmacılara yapısal verinin (structured data) yanında yapısal olmayan veri (unstructured data) ve yarı yapısal veri (semi-structured data) türleri ile analiz yapma gereksinimi doğurmaktadır.

Yapısal olmayan veri, en kolay toplanabilecek veri türüdür [1]. Yapısal olmayan veri, analiz edilmesi için standart bir formatı olmayan veri olarak tanımlanır. Genellikle metinsel veriler, görseller, ses ve video dosyaları bu veri türüne örnek gösterilebilir [2]. Son yıllarda çeşitli sosyal ağlarda, web'de ve diğer bilgi işlem uygulamaları tabanlı uygulamalarda oluşturulan büyük miktarda metin verisi nedeniyle çeşitli metin verilerini etkin bir şekilde işleyebilecek yöntem ve algoritmalara ilgi giderek artmaktadır [1]. Yapısal olmayan verinin saklanması, sorgulanması ve analiz edilerek değer üretilmesi aşamalarında kullanılan teknikler geleneksel yöntemlerden farklıdır. Metin verisinden anlamlı bilgiler çıkartılarak değer üretilmesi aşamasında kullanılacak yöntemlerin başlıcalarından biri metin madenciliği (text mining) yöntemleridir.

Metin madenciliği, metin verisi üzerine klasik veri madenciliği yöntemlerinin uygulanması gibi basit tanımlardan "dünya hakkındaki gerçekleri ve yeni eğilimleri keşfetmek için büyük çevrimiçi metin verilerinin kullanılması" şeklinde sofistike tanımlara kadar geniş bir yaklaşım ve yöntem alanını kapsamaktadır [3]. Metin madenciliği genel olarak, veri madenciliği, dil bilimi, hesaplamalı istatistik ve bilgisayar bilimi ile etkileşimde olan disiplinler arası bir uygulama alanıdır. Metin madenciliği metin sınıflandırma (text classification), metin kümeleme (text clustering), belge özetleme (document summarization), ontoloji ve taksonomi yaratma (ontology and taxonomy), gizil korpus analizi (latent corpus analysis) gibi kendine özgü yöntemlerin yanı sıra ilişkili olduğu disiplinlerde kullanılan pek çok yöntem ve teknikten de yararlanmaktadır [4]. Metin madenciliği başka bir deyişle metin verisinden bilgi keşfi ilk kez Feldman ve Dagan (1995) tarafından metnin makine destekli analizi olarak tanımlanmıştır [5]. Metin madenciliği doğal dil işleme teknikleri ile birlikte bilgi elde etme, bilgi çıkarma süreçlerini kullanarak veri madenciliği, makine öğrenmesi, istatistiksel yöntemler ve algoritmalara bağlayan bir süreç olarak da görülebilir. Bu açıdan metin madenciliği, veri tabanından bilgi keşfi olarak da tanımlanan veri madenciliği ile benzer bir prosedürdür. Veri madenciliğinde veriye odaklanılırken metin madenciliğinde analizin temelini metinsel verileri içeren belgelerin kümesi (korpus) oluşturmaktadır [6].

Veri madenciliği ile metin madenciliği arasında kavramsal olarak kapsayıcı bir ilişki vardır. Literatürde metin madenciliğinin veri madenciliğinin bir alt dalı olarak da kabul edilebileceği yönünde yaygın bir görüş birliği vardır. Metin madenciliği de veri madenciliği gibi, ilgi çekici örüntülerin tanımlanması ve araştırılması yoluyla veri kaynaklarından faydalı bilgiler çıkarmayı amaçlamaktadır. Ancak veri madenciliğinden farklı olarak metin madenciliğindeki veri kaynakları yapılandırılmamış metin verileridir [7]. Veri madenciliğinde örüntülerin keşfi için yapılandırılmış veri tabanlarından kullanılırken metin madenciliğinde bilginin kaynağı çoğunlukla yapılandırılmamış doğal dil metinleridir [8]. Bu durumda iki kavram arasındaki temel ayrımın metin madenciliğindeki bilgi keşfi sürecinin doğal dil kalıplarının kullanılarak gerçekleştirilmesi olduğunu söylemek yanlış olmayacaktır. Metin madenciliğinin alt konuları ise metin sınıflandırma, metin kümeleme ve metni benzer metinlerle ilişkilendirme olarak genellenebilir [1].

Bu çalışmada inşaat sektöründe faaliyet gösteren ve konut satışı yapan bir firmanın müşterileriyle yaptığı görüşmelerin derlendiği metin tabanlı dosyalar metin madenciliği ve makine öğrenmesi yöntemleri kullanılarak analiz edilmiştir. Bu sayede bir müşteri ile görüşme yapıldıktan sonra ileride ilgili müşterinin konut satın alıp almayacağını tahmin eden bir model oluşturulmaya çalışılmıştır. Çalışmada metin madenciliği konusunda genel bir bakış açısı sunularak, müşteri ilişkileri yönetimi ve pazarlama konularında işletmelerin metin madenciliği ve makine öğrenmesi yöntemlerinin birlikte kullanılmalarından nasıl faydalanabilecekleri incelenmiştir.

Çalışmanın ikinci bölümünde literatür taraması, üçüncü bölümünde veri kümesi ve uygulanan yöntemler hakkında detaylı bilgi, dördüncü bölümünde analiz sonuçları yer almaktadır. Son bölümde ise sonuçların değerlendirilmiştir.

2. LİTERATÜR TARAMASI (LITERATURE REVİEW)

Metin madenciliği teknikleri ile makine öğrenmesi yöntemlerinin birlikte kullanıldığı birçok çalışma

bulunmaktadır. Farklı anlarda yapılan bu çalışmalardan birkaçı bu kısımda sunulmuştur. Ayrıca spesifik alarak inşaat yönetimi veya pazarlaması konularında metin madenciliğinin makine öğrenmesi ile birlikte kullanıldığı çalışmalar da bu literatür taramasının sonunda yer almaktadır.

Hosseini vd. (2023) tarafından yapılan çalışmada polis tarafından tutulan metin tabanlı kaza tutanakları incelenerek bir kazanın ters yönde sürüş sonucu oluşup oluşmadığını belirlemek için makine öğrenmesi yöntemlerinden faydalanılmıştır. Doğal dil işleme işlemleri için Bidirectional Encoder Representations from Transformers (BERT) modelleri kullanılmıştır. Daha sonra, Naïve Bayes, Support Vector Machine, Decision Tree, Random Forest, and Single Layer Perceptron makine öğrenmesi yöntemleri kullanılarak kazanın ters yönde sürüşten kaynaklanan bir kaza olup olmadığı sınıflandırılmıştır. Her bir sınıflandırma algoritmasının performansını değerlendirmek için çapraz doğrulama ve farklı performans ölçütleri kullanılmıştır. Oluşturulan modelin bu tür kazaların belirlenmesinde başarılı bir şekilde kullanılabileceği ve kaza tutanaklarından kaza sebebinin belirlenme süresini çok kısaltabileceği sonucuna varılmıştır [9].

Soleimani vd. (2021) yaptıkları çalışmada trenler ve arabalar arasındaki kaza sayısını azaltmak amacıyla gereksiz geçişlerin kapatılmasına ve böylece çarpışma risklerinin azaltılmasına yardımcı olacak bir Karayolu-Demiryolu Hemzemin Geçidi Konsolidasyon Modeli oluşturmuşlardır. Önceki çalışmalarından en iyi performansı gösteren XGboost makine öğrenimi algoritmasına ek olarak metin madenciliği teknikleri ve Jeo-uzamsal (Geo-spatial) analiz kullanılmıştır. Sonuçlar, önerilen model için %88'lik bir genel doğruluk göstermiştir. Çalışmada modelde yer alan değişkenlerin göreceli önemleri de rapor edilmiştir ve böylece modelin davranışının derinlemesine anlaşılması sağlanmıştır. Farklı eşik değerlere göre yapılan yorumlar soncunda incelenen bölgede mevcut karayolu-demiryolu hemzemin geçitlerinin %15'inin kapatılmasının gerektiği sonucuna varılmıştır[10].

Nilashi vd. (2021) yaptıkları çalışmada müşterilerin sosyal medyada restoranların hizmet kaliteleri hakkındaki görüşlerini inceleyerek müşteri segmentasyonunu sağlayan ve müşteri tercihlerini tahmin eden yeni bir yöntem geliştirmişlerdir. Bu amaçla çalışmada metin madenciliği yöntemleri (Latent Dirichlet Allocation), kümeleme analizi (Self Organizing Map) aşamalarından sonra CART (Classification and Regression Trees) yöntemi ile vejetaryen dostu restoranlarda müşterinin hizmet kalitesi hakkındaki görüşlerinden memnuniyet düzeyleri tahmin edilmiştir. CART yöntemi ile elde edilen modellerin restoranların kalite faktörlerine dayanarak müşterilerin tercihlerini iyi bir şekilde yansıtabildiğini göstermiştir. Toplanan büyük miktardaki çevrimiçi müşteri yorumlarının makine öğrenmesi yöntemleri ile birlikte kullanılmasının restoranlardaki hizmet kalitesinin iyileştirilebilmesi ve müşteri segmentasyonunun sağlanması için etkili bir yöntem olduğu belirtilmiştir [11].

Petropoulos ve Siakoulis (2021) Merkez Bankalarının yaptıkları açıklamaları inceleyerek açıklamalardaki en önemli sinyalleri filtrelemek ve gelecekteki finansal piyasa davranışını tahmin etmek için bir duyarlılık endeksi oluşturmuşlardır. Bu amaçla makine öğrenimi teknikleriyle (XGboost) birlikte doğal dil işleme tekniklerini kullanmışlardır. Daha önceden tanımlanmış veya açıklamaların yer aldığı korpustan elde edilen sözlüklerin XGBoost ile birlikte kullanılması sonucu elde edilen analizlerin sonucunda piyasalardaki çalkantıları başarılı bir şekilde öngerebilmeyi sağlayan bir duyarlılık endeksi oluşturmanın mümkün olduğu belirtilmiştir [12].

Chatterjee vd. (2021) yaptıkları çalışmalarda e-ticaret sektöründe sağlık ürünleri sunan şirketlerin müşteri memnuniyetini modellemeyi ve sektördeki alt gruplar arasındaki farklılıklarını analiz etmeyi hedeflemişlerdir. Bu amaçla, 2008 ve 2018 yılları arasında bir inceleme sitesinde yayınlanan 29 sağlık/sağlık ürünleri alt kategorisinden 619 e-ticaret firması hakkında 186.057 yorum üzerinde metin madenciliği, makine öğrenimi (Doğrusal Regresyon, XGboost, Random Forest, ve CART) ve ekonometri teknikleri kullanılmıştır. Böylece müşteri memnuniyetini belirleyen hizmetlerin temelde neler olduğu ve hangi duyguların ön plana çıktığı belirlenerek sağlık hizmetleri sunan e-ticaret platformlarının tasarımında ve sunumundan nelere dikkat edilmesi ortaya konmuştur [13].

Lin vd. (2022) çalışmalarında hisse fiyatlarının metin tabanlı veriler kullanılarak tahmin edilmesinde verilerin farklı finansal kaynaklardan alınmasının, farklı metin tabanlı öznitelik temsillerinin (TF-IDF, word embeddings gibi) kullanılmasının ya da farklı makine öğrenmesi yöntemlerinin çalıştırılmasının model performanslarını etkileyip etkilemediklerini analiz etmişlerdir. Bu amaçla Reuters, CNBC, The Motley Fool gibi farklı kaynaklardan alınan finansal veriler üzerinde TF–IDF, Word2vec, ELMo, BERT gibi farklı öznitelik temsillerini (feature representation) ve SVM, CNN, LSTM gibi farklı makine öğrenmesi yöntemlerini test etmişlerdir. Vardıkları sonuç en iyi tahmin

performanslarının CNN+Word2vec ve CNN+BERT kombinasyonları ile elde edildiğidir [14].

Allenbrand (2024) yaptığı çalışmada hastalıkların tedavisinde kişiselleştirilmiş ve optimize edilmiş bir tedavi programı oluşturabilmek için metin madenciliği ve makine öğrenmesinden faydalanmıştır. Hastaların aldıkları tedavinin faydaları ve yan etkileri hakkında beyan ettikleri görüşler yoluyla memnuniyeti tahmin edecek modeller oluşturulmuştur. Bu amaçla Naive Bayes, Support Vector Machines, Random Forests ve CART yöntemleri uygulanmıştır. Model performanslarını ölçmek amacıyla Matthews korelasyon katsayısı ve F1 ölçütü kullanılmıştır. En iyi sonuçları Random Forest yöntemi vermiştir. Önişleme adımında verideki gürültüyü azaltmak ve böylece tahmin modellerinin performansını arttırmak amacıyla kümeleme analizi de çalışmada yer almıştır [15].

Anagün vd. (2022) yaptıkları çalışmada finans kuruluşları için müşteri şikayetlerini otomatik olarak sınıflandıracak derin öğrenme tabanlı müşteri şikayet yönetim sistemi geliştirmişlerdir. Bu amaçla metin tabanlı veriler üzerinde modelin performansını arttırmak için yeni bir ön işleme tekniği tanımlanmıştır. Bu yöntemle %96 oranında bir başarı elde edilmiştir [16].

Işık vd. (2020) yaptıkları çalışmada spam ve spam olmayan epostaları sınıflandırmak için iki farklı öznitelik seçimi (feature selection) yöntemi uygulanmıştır. Bu amaçla eposta verileri analize uygun olacak şekilde yapılandırılmış ve üzerinde öznitelik seçimi yöntemi uygulamıştır. Tahmin için üç farklı öğrenme yöntemi kullanılmıştır. Elde edilen sonuçlar Türkçe gibi eklemeli bir dil içeren metinsel veriler üzerinde belirtilen öznitelik seçimi ve derin öğrenme yöntemlerinin birlikte kullanımının başarını oranını arttırdığı görülmüştür [17].

Yapılan çalışmalar incelendiğinde çalışmanın konusuna bağlı olarak (kazalar, hastalıklar, krizler, ağ saldırısı, sahtekarlık, alışveriş yapma, spam eposta vb) genellikle sınıflar arasında dengesizlikler bulunduğu görülmüştür. Makine öğrenmesi yöntemleri genellikle sınıf dağılımlarının benzer olduklarını varsayar. Yöntemlerin bazıları bu duruma karşı dirençliyken bazıları daha az dirençlidir.

Aşağıda inşaat yönetimi ve pazarlaması alanında yapılan metin analitiği çalışmalarının gözden geçirildiği kapsamlı çalışmalar yer almaktadır. Baek vd. (2021) derlemelerinde inşaat sektörüne odaklanan metin madenciliği çalışmalarının mevcut durumunu ve gelecekteki olası trendleri, karşılaşılan zorlukları, yeni araştırma alanlarını ve kullanılan farklı veri kaynaklarını gözden geçirmek için kapsamlı bir inceleme yapılmışlardır. Bu amaçla inşaat sektörü ile ilgili 103 akademik makale incelenmiştir. Alandaki analizlerde kullanılan metin veri kaynakları, farklı metin madenciliği yaklaşımları ve makine öğrenimi yöntemleri gözden geçirilmiştir. Yazarların vardıkları sonuç, bu alanda metin analitiği yöntem ve tekniklerinin anlamlı bilgi üretmek için yeterince gelişmiş olduğu ve metin madenciliği ve makine öğreniminin birlikte kullanımının ilerisi için daha yeni fırsatlar yaratabileceği yönündedir. Ayrıca, yakın gelecekte doğal dil işleme yöntemlerinin gelişmesi, kullanılan araçların yaygınlaşması ve dijital dönüşümün daha da hızlanması ile beraber emek yoğun metin tabanlı görevlerin yerini otomatik metin analizi araçlarının alacağı düşünülmektedir [18].

Yan vd. (2022) inşaat sektöründe metin madenciliği uygulamalarına ilişkin 2000-2021 yılları arasında yayınlanan 127 akademik dergi makalesi üzerinde hacimli bir araştırma gerçekleştirmiştir. Yayınlarda oluşan trendler, yayınların en çok yapıldığı ülkeler ve bölgeler, öne çıkan araştırmacılar ve anahtar kelimeler, hangi anahtar kelimelerin daha çok birlikte kullanıldığı gibi ayrıntıları belirlemek amacıyla VOSviewer yazılımından faydalanılmıştır. Çalışmada metin madenciliğinin öncelikli uygulama alanları detaylı olarak incelenmiştir. Ayrıca karşılaşılan temel zorluklar ve gelecekteki yönelimler (alan bilgisinin veya ontolojinin modellere entegrasyonu, yenilikçi teknolojilerin uygulanması ve sosyal medyadaki metinsel veri üzerinde duygu analizi yapılması) çalışmada yer almaktadır [19].

Shamshiri vd. 2024 tarafından inşaat sektöründe metin madenciliği ve doğal dil işleme yöntemlerinin derinlemesine incelenebilmesi için 205 akademik makale ele alınmıştır. Bu makaleler öncelikle uygulama alanlarına göre sınıflandırılarak incelenmiştir. Derleme çalışmasında metin madenciliği ve doğal dil işleme yöntemlerinin inşaat yönetiminin kısıt yönetimi, kapsam yönetimi, entegrasyon yönetimi, kaynak yönetimi, iletişim yönetimi, malzeme yönetimi, tedarik yönetimi, devreye alma ve başlatma ve proje kontrolü gibi alt alanlarındaki potansiyelleri hakkında yorumlar içermektedir. Ayrıca, çalışmada mevcut çalışmalardaki boşlukları tespit ederek gelecekteki inşaat sektöründeki uygulamaların daha az insana bağımlı ve daha az hataya eğilimli hale getirilmesi için öneriler bulunmaktadır [20].

Yapılan literatür araştırmasında metin madenciliğinin farklı amaçlarla birçok uygulama kullanıldığı görülmektedir. Fakat bildiğimiz kadarıyla bu

çalışmadaki gibi bir inşaat şirketinin potansiyel müşterileri ile yaptığı görüşmeler sonucu elde edilen metinsel veriler üzerinde yeni bir müşterinin ev alma niyetini farklı makine öğrenmesi yöntemleri ile tahmin eden bir çalışma bulunmamaktadır.

3. MATERYAL VE METODLAR (MATERIAL AND METHODS)

3.1 Veri Kümesi

Veriler, dijital mecralarda ve çeşitli gazetelerin internet sayfalarında reklam gösterimi yapan bir inşaat firmasından elde edilmiştir. Firma yaptığı reklamlar sayesinde internet kullanıcılarının kendi sitesine gelerek konut projeleri hakkında bilgi edinmelerini sağlamıştır. Daha detaylı bilgi almak isteyen kullanıcılar ise iletişim bilgilerini sitedeki bir form aracılığıyla firmaya göndermişlerdir. Kullanıcıların iletişim bilgilerini alan firma kullanıcılarla telefon görüşmesi veya yüz yüze görüşmeler yapıp onların görüşlerini yazılı olarak elektronik ortamda kayıt altına almıştır. Analiz için elde edilen metin verisi bu şekilde toplanmıştır. Yapılan görüşmelerin soncunda kullanıcıların konut satın alıp almadığı bilgisi de veri setinde yer almaktadır.

Oluşturulan bu veri seti üzerinde farklı makine öğrenmesi yöntemleri kullanılarak yeni bir kullanıcı ile yapılan görüşmenin sonucunun olumlu (konut alma) veya olumsuz (konut almama) olmak üzere iki seçenekten hangisi olacağını tahmin edecek modeller oluşturulmuştur. İkili sınıflandırma problemi şeklinde ele alınan uygulama yüksek performans gösterecek şekilde çözülmeye çalışılmıştır.

Yapılandırılmamış metin tabanlı veri kümesi R [21] dilinde metin madenciliği için en yaygın olarak kullanılan paket olan Text Mining (TM) [4] kullanılarak vektör uzayı modeli ile temsil edilen yapılandırılmış hale getirilmiştir. Bu paketin içerisinde tüm metin önişleme aşamaları için fonksiyonlar tanımlanmıştır.

Web sayfasına 1492 kullanıcı iletişim bilgilerini bırakmıştır. İletişim bilgileri doğru olan ve yüz yüze veya telefonla karşılıklı görüşme yapılan kişi sayısı ise 579 dur. Metin tabanlı veri seti bu kişilerle yapılan görüşmeler sonucu oluşturulmuştur. Dolayısıyla korpus her bir görüşmeye ait 579 metin dosyasından oluşmaktadır.

Her bir dosya üzerinde tüm karakterlerin küçük harfe dönüştürülmesi, noktalama işaretlerinin kaldırılması, cümlelerin kelimelere ayrılması (tokenization), bir anlam ifade etmeyen durma kelimelerinin (stop words) kaldırılması, kelimelerin köklerinin bulunması (stemming), aynı anlama gelen kelimelerin tek bir kelime altında birleştirilmesi gibi önişleme adımları yerine getirilmiştir. Veriler metin ön işleme aşamaları yapıldıktan sonra vektör uzay modeline dönüştürülmüştür. Vektör uzay modeliyle yapısal olmayan bir biçimden yapısal biçime dönüştürülen veriler analize hazır hale getirilmiştir. Veri önişleme aşaması birden fazla adımdan oluşması ve karşılaşılan çeşitli zorluklar nedeniyle uygulamanın en çok zaman harcanan aşaması olmuştur.

Örnek veri kümesi Tablo 4'de verilmiştir. Çıktı değerinin "1" olması görüşme sonucu satış yapıldığı, "0" satış işleminin yapılmadığını göstermektedir.

Tablo 1: Veri kümesi özet tablo

Görüşme No	**lokasyon**	**daire**	**metrekare**	**…**	**Sınıf Değişkeni**
1	1	1	1	…	1
2	0	1	0	…	0
3	0	1	0	…	0
4	0	0	1	…	1
5	0	1	1	…	0
⋮	⋮	⋮	⋮	⋮	⋮
579	0	0	0	…	0

Tablo 1'de gösterilen veri kümesi sınıf dağılımında "0" ile temsil edilen kategoriye ait 534, "1" ile temsil edilen kategoriye ait 45 gözlem bulunmaktadır. Bunun anlamı görüşmelerin %8 inin satın almayla sonuçlanırken, geri kalan %92 sinde herhangi bir satış işleminin olmadığını göstermektedir. Bu konut satışı gibi bir alanda beklenen bir sonuçtur. Fakat makine öğrenmesi açısından sınıfların dengesiz olması model performansını etkileyen faktörlerden birisidir.

Tablo 2: Anahtar kelime frekansları

Anahtar Kelimeler	**Frekanslar**
ofis	392
proje	332
daire	297
bilgilendirme	243
davet	208
fiyat	199
avantaj	159
lansman	126
satış	123
1+1	122

Tablo 2'de korpusta en çok yer alan ilk 10 kelime ve frekans sayıları verilmiştir.

Tablo 3: Veri kümesi sınıf dağılımı

Y-Sınıf değişkeni (Bağımlı değişken)		
0-sınıfı gözlem sayısı	1-sınıfı gözlem sayısı	Toplam Gözlem Sayısı
534 (92%)	45 (8%)	579 (100%)

Veri kümesi rastgele %80'i eğitim ve %20'i test verisi olarak ikiye ayrılmıştır. Eğitim kümesinde 465, test kümesinde ise 114 gözlem bulunmaktadır. Rastgele seçim yapıldığından eğitim ve test kümelerindeki sınıf dağılımları benzer olmuştur.

Sınıflandırma tabloları oluşturulurken kesim noktası satın alma gerçekleştirenler sınıfının yaygınlık değeri olan %8 olarak alınmıştır. Sınıflandırma analizi sonucu tahmin edilen değeri 0.08 değerinin üzerinde olan gözlemler "1" ile temsil edilen satın alma gerçekleştirmiş olanlar sınıfına, 0.08 değerine eşit ve altında olan gözlemler ise "0" ile temsil edilen satın alma gerçekleştirmeyenlerin olduğu sınıfa atanmışlardır.

3.2 Metin Madenciliği

Metin sınıflandırma probleminin amacı, metin verisini daha önceden belli olan kategorilere otomatik olarak atanmasıdır. Amaç, makine öğrenmesi algoritmalarını kullanarak, metinleri kategorilere (sınıflara) otomatik olarak atayacak sınıflandırıcıları oluşturmaktır. Veri madenciliğinde bu durum denetimli öğrenme olarak da bilinir. Dokümanların (metinlerin) otomatik olarak sınıflandırılabilmesi için vektörel olarak ifade edilmesi gerekir. Bunun için de "vektör uzay modeli" oluşturulmalıdır [1].

Metin sınıflandırma eğitim ve test aşaması olarak iki aşamadan oluşmaktadır. Eğitim aşamasına geçmeden önce belgeler vektör uzayı modeline dönüştürülür ve sınıflandırma algoritmaları kullanılarak model oluşturulur. Test aşamasında ise eğitim kümesinde olmayan veriler oluşturulan model yardımıyla sınıflandırılır ve model performansı değerlendirilir.

Metin verisinden bilgi çıkarımına giden süreçte genellikle takip edilen adımlar Şekil 1'de görülmektedir.

Şekil 1. Metin Sınıflandırma Süreci

Metin ön işleme, çoğu metin analizinde gerekli olan ve metin temsilini kolaylaştırmak için girdi belgelerini daha tutarlı hale getirmenin amaçlandığı aşamadır. Geleneksel metin ön işleme yöntemleri arasında durdurucu kelimelerden (stop-words) arındırma ve gövdeleme (stemming) yani kelime köklerini bulma işlemleri yapılır. Durdurucu kelimeleri kaldırma aşamasında sözcüklerin daha genel ve anlamsız olarak değerlendirildiği bir sözcük durdurma listesi kullanılarak (acaba, ama, ancak, için, madem vb.) kelimeler kaldırılır. Gövdelemede, kelimeler köklerine indirgemektedir. Örneğin, "izlemek", "izliyor", ve "izledi", ek almış kelimeleri "izlemek" olarak temsil edilir [1]. Ayrıca bu aşamada hatalı yazımları düzeltme, noktalama işaretleri ve gereksiz kelimeleri (edatlar, bağlaçlar) çıkarma işlemleri de yapılır. d_i tekil bir doküman olmak üzere, $D = \{d_1, d_2\ , \ \dots\ , d_n\}$ ile ifade edilen dokümanları bir arada bulunduran D veri seti "korpus" olarak adlandırılmaktadır. Korpus, üzerinde metin analizleri yapılacak veri setidir. Korpus nesnesi, metin madenciliğinde analizlerde kullanılan doküman terim matrisine (vektör uzay modeline) çevrilmeye hazır, tüm metinlerin bazı özellikleri ile birlikte tutulduğu matris benzeri bir yapıdır [22]. Ön işleme çalışmaları tamamlandığında çalışma matrisinin hazırlanmasına geçilir. Oluşan sözlüğe göre dokümanlar sayısal olarak ifade edilir. Vektör uzay modeli metinlerin sayısal hale (yapılandırılmış veri) getirilmesini sağlayan bir yöntemdir.

Bir metin kelime dizisi olarak ifade edilebilir. Bir eğitim setinin tüm kelimelerine sözlük veya özellik seti denir. Böylelikle, bir belge ikili vektörle gösterilebilir, eğer belge sözlükteki kelimeyi içeriyorsa 1 değerini, içermiyorsa 0 değerini alır. Bu, bir belgeyi $R^{|v|}$ alanına yerleştirmek olarak düşünülebilir. Burada, $|v|$ sözlükte yer alan kelime sayısını gösterir. Oluşturulan sözlüğe göre dokümanlar sayısal olarak ifade edilir Kelimelerin bu şekilde gösterilmesi vektör uzayı modeli olarak adlandırılır [23].

$$x_i(Kelime) = \begin{cases} 1, & i.kelime\ doküman\ içinde\ yer\ alıyorsa \\ 0, & i.kelime\ doküman\ içinde\ yer\ almıyorsa \end{cases}$$

Tablo 4: Vektör Uzay Modeli –Doküman Terim Matrisi

	Kelime 1	**Kelime 2**	**. . .**	**Kelime n**
Doküman 1	1	0	. . .	1
Doküman 2	0	1	. . .	0
Doküman 3	0	1	. . .	0
⋮ ⋱	⋮ ⋱	⋮ ⋱	. . .	⋮ ⋱
Doküman n	0	0		1

3.3 CART (Classification and Regression Tree) Yöntemi

CART (Sınflandırma ve Regresyon Ağaçları) her bir ara düğümün (internal node) bir değişkeni (öznitelik, feature) ve her yaprak düğümün (leaf node) bir sınıf etiketi taşıdığı akış şeması benzeri bir yapıya sahiptir. Ağaçta bulunan en üstteki düğüm, kök düğümdür (root node) ve bu düğüm de veri kümesindeki bir değişkeni temsil etmektedir [24]. Karar ağaçlarındaki kök düğüm ve ara düğümler entropi (entropy), gini endeksi (gini index), kazanç oranı (gain ratio), bilgi kazancı (information gain), twoing gibi bölünme kriteri (değişken seçim) algoritmaları ile oluşturulmaktadır. Bu kriterlerden hangisinin kullanılacağına göre karar ağacı algoritması farklılık göstermektedir.

CART algoritması karar ağaçlarının genelinde olduğu gibi yukarıdan aşağıya iteratif olarak böl ve yönet tarzında oluşan açgözlü (greedy) (geri izlemesiz) bir yaklaşım benimser. CART algoritmasına bölünme yöntemi olarak genellikle gini endeksi kullanılır. Karar ağaçları için çoğu algoritmada, eğitim kümesi iteratif olarak daha küçük alt kümelere bölünür [24]. Her bölünmede ağaç üzerinde aşağıya doğru iki yeni dal oluşur. Ağaç oluşturma sürecinin her aşamasında, en iyi bölünme, ileriye bakmak ve gelecekteki bazı adımlarda daha iyi bir ağaca yol açacak bir bölünme seçmek yerine, o aşamada yapılır [25].

3.4 Random Forest Yöntemi

Random Forest (Rastgele Orman), karar ağacı tabanlı sınıflandırıcılar topluluğudur, ormandaki her ağaç bağımsız olarak örneklenen bir rasgele vektörün değerlerine bağlı olarak oluşur. Bu yöntemi Breiman makalesinde şöyle açıklamaktadır. *k*.'ıncı ağaç için, geçmiş rastgele vektörlerden Q_1, ..., ,Q_{k-1} bağımsız ancak aynı dağılımla sahip Q_k rasgele vektörü oluşturulur; ve bir ağaç, *x*'in girdi vektörü olduğu eğitim veri seti ve Q_k kullanılarak $h(x,Q_k)$ sınıflandırıcısı oluşturulur. Rastgele bölünme seçiminde *Q, 1* ile *k* arasında bağımsız rastgele tam sayıdan oluşur. *Q*'nun yapısı ve boyutu, ağaç yapımındaki kullanımına bağlıdır. Rastgele orman yönteminde, torbalama ve rastgele değişken (özellik) seçimi birlikte kullanılır. Her yeni eğitim seti, orijinal eğitim setinden iadeli olarak (bootsrap yöntemiyle) çekilir. Ardından rastgele değişken (özellik) seçimi kullanılarak yeni eğitim setinde bir ağaç yetiştirilir. Yetiştirilen ağaçlarda budama yapılmaz [26]

Bu algoritmada her sınıflandırıcı eğitim kümesine eşit boyutta iadeli örnekleme yolu ile elde edilen veri kümeleri ile eğitilir. İadeli örnekleme kullanıldığı için eğitim kümesindeki gözlemler elde edilen veri kümelerinde birden fazlada görülebilir ya da hiç yer almayabilir. Yeni bir örneği sınıflandırmak için her sınıflandırıcı kendi sınıf tahminini oluşturur ve torbalanmış sınıflandırıcı (bagged classifier) en fazla tahmin edilen sınıfı sonuç olarak döndürür (oylama yöntemi - voting method) [27].

3.5 XGBoost (Extreme Gradient Boosting) Yöntemi

Gradyan arttırma karar ağacı (Gradient Boosting Decision Tree) sınıflandırma doğruluğu ve etkinliği nedeniyle yaygın olarak kullanılan bir makine öğrenme algoritmasıdır. GBDT sırayla eğitilmiş karar ağaçlarının bir topluluk modelidir [28]. GBDT zayıf modeller olarak genellikle regresyon ağaçlarını kullanır [29]. Her yinelemede, negatif gradyanları (hatalar olarak da adlandırılır) kullanarak karar ağaçları oluşturur [30].

XGBoost gradyan arttırma algoritmasının ek parametre ayarları ile gradyan artırmanın genelleştirilmiş hali olarak düşünülebilir. Tahmin performansı yüksek olmakla birlikte çok çekirdekli ve paralel (dağıtılmış) makine uygulamasına da sahiptir [31]. XGBoost algoritması yenilik olarak; seyrek veriler (sparse data) için yeni bir ağaç öğrenme algoritması sunar, ağırlıklı nicel çizim (weighted quantile sketch) ile ağaç öğreniminde örnek ağırlıklarını belirler ve paralel bilgi işleme ile öğrenmeyi daha hızlı hale getirir [32].

XGBoost, türevlenebilir kayıp fonksiyonuna ek olarak aşırı öğrenmeyi önlemek için düzenlileştirme (regularization) terimi kullanır. $\{(x_i, y_i)\}_{i=1}^{n}$ veri kümesi ve $l(y_i, \hat{y}_i)$ diferansiyallenebilir kayıp fonksiyon (loss function) olmak üzere;

$$L(\varphi) = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k)$$
$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda w^2 \quad (1)$$

olacak şekilde Denklem 1 elde edilir. $\Omega(f_k)$, f_k modelinin (karar ağacı) karmaşıklığı olarak adlandırılır ve bu terim aşırı öğrenmeyi engellemektedir. Bu terimin sıfır olması gradyan arttırma algoritmasına karşılık gelmektedir [32]. $\Omega(f_k)$ fonksiyonunda T, f_k ağacındaki yaprak sayısı w yaprak ağırlığıdır ve yapraklardaki tahmin edilen değerler yardımıyla hesaplanmaktadır. γT her ek ağaç yaprağı için sabit bir ceza puanı oluşturur ve λw^2 aşırı ağırlıkları cezalandırır. γ ve λ uygulayıcı tarafından belirlenen parametrelerdir [31].

3.6 Model Değerlendirme

Eğitilen modelin performansı bağımsız test verileri üzerindeki tahmin başarısına bağlıdır [33]. Model değerlendirme, gözlemlerin sınıf etiketini tahmin etmede ne kadar iyi olduğunu değerlendirmek için ölçümler sunmaktadır [24]. Sınıflandırma modellerini değerlendirme de genellikle hata matrisi (confusion matrix) kullanılır. İki sınıflı modellerin hata matrisi 2 satır ve 2 sütundan oluşmaktadır. Bir sınıflandırma modelinin iyi bir doğruluğa sahip olması için gözlemlerin çoğunun asal köşegen üzerinde yer alması ve asal köşegen dışında kalan yerlerinde sıfıra yakın olması istenir [24]. İki sınıflı bir sınıflandırma problemi için oluşturulan hata matrisi Tablo 5'de gösterilmektedir. Gözlemler pozitif ve negatif olarak etiketlendirilmiştir. *P* veri kümesinde yer alan pozitif gözlem sayısını, *N* ise negatif gözlem sayısını temsil etmekte olup $\boldsymbol{P'}$ ve $\boldsymbol{N'}$ ise model tarafından tahmin edilen pozitif ve negatif gözlem sayısını temsil etmektedir [24].

Tablo 5: Sınıflandırma tablosu- Hata Matrisi

		Gerçek Sınıf (Actual Class)		
		Hayır(0)	**Evet(1)**	**Toplam**
Tahmin Edilen Sınıf (Predicted Class)	**Hayır(0)**	*DN*	*YN*	N'
	Evet (1)	*YP*	*DP*	P'
	Toplam	*N*	*P*	*N+P*

Doğru Pozitif (DP): Sınıflandırıcı tarafından doğru şekilde tahmin edilmiş pozitif gözlemleri ifade etmektedir. *DP* doğru pozitif gözlemlerin sayısına karşılık gelmektedir.

Doğru Negatif [34]: Sınıflandırıcı tarafından doğru şekilde tahmin edilmiş negatif gözlemleri ifade etmektedir. *DN* doğru negatif gözlemlerin sayısına karşılık gelmektedir.

Yanlış Pozitif (YP): Gerçekte negatif olan ancak sınıflandırıcı tarafından pozitif olarak tahmin edilen gözlemleri ifade etmektedir. *YP* yanlış pozitif gözlemlerin sayısına karşılık gelmektedir.

Yanlış Negatif (YN): Gerçekte pozitif olan ancak sınıflandırıcı tarafından negatif olarak tahmin gözlemleri ifade etmektedir. *YN* yanlış negatif gözlemlerin sayısına karşılık gelmektedir.

Sınıflandırma modellerini birbiriyle karşılaştırabilmek için hata matrisinden birçok farklı ölçüt oluşturulmuştur. Tablo 6'da en yaygın kullanılan performans ölçütlerinden bazıları ve bunların hesaplanma yolları verilmiştir.

Tablo 6: Değerlendirme Ölçütleri

Değerlendirme Ölçütleri	**Formüller**
Doğruluk (Accuracy)	$\frac{DP + DN}{P + N}$
Yanlış Sınıflandırma Oranı (Misclassification rate)	$\frac{YN + YP}{P + N}$
Duyarlılık (Sensitivity) Hatırlama (Recall) Doğru Pozitif Oran (True Positive Rate)	$\frac{DP}{P}$
Özgüllük (Specificity) Doğru Negatif Oran (True Negative Rate)	$\frac{DN}{N}$
Yanlış Negatif Oran (False Negative Rate)	$\frac{YN}{P}$
Yanlış Pozitif Oran (False Positive Rate)	$\frac{YP}{N}$
Kesinlik (Precision)	$\frac{DP}{P'}$
Yaygınlık (Prevalence)	$\frac{P}{P + N}$

Sınıf dengesizliğinin olduğu problemlerde sınıflandırma modeli, çoğunluk sınıfının gözlemlerini doğru bir şekilde sınıflandırır. Ancak bu durumda azınlık sınıf gözlemleri de yanlış sınıflandırılabilir. Bu nedenle duyarlılık (sensivitiy) ve özgüllük (specificity) ölçülerini kullanmak daha tutarlı olmaktadır. Duyarlılık, doğru pozitif oran olarak da adlandırılır ve gerçekte pozitif olan gözlemlerin ne kadarının model tarafından doğru bir şekilde sınıflandırıldığının ölçüsünü vermektedir. Özgüllük ise doğru negatif oran olarak adlandırılır ve gerçekte negatif olan gözlemlerin model tarafından ne kadarının doğru bir şekilde sınıflandırıldığının ölçüsünü vermektedir [24]. Kesinlik (precision) ölçüsü ise model tarafından pozitif olarak etiketlenen gözlemlerin ne kadarının doğru bir şekilde sınıflandırıldığının oranını vermektedir [35].

Tablo 6'da hesaplanan tüm doğruluk ölçümleri eşik değer seçimine bağlı olarak hesaplanmaktadır. Eşik değer seçimi sınıflandırma hatalarında değişikliğe neden olabilmektedir. Bu durum dengesiz veri kümeleri için doğru bir yaklaşım değildir [36]. Dengesiz veri kümelerinde yüksek veya düşük düzeyde gözlemlenen yaygınlık (prevelence) bulunmaktadır. Yaygınlık, veri kümesinde her kategorinin ne kadar sıklıkta olduğunu veren ölçüdür [36]. Gözlenen (gerçek) ve tahmin

yaygınlığı olarak ikiye ayrılmaktadır. Birçok uygulamada, gözlenen ve tahmin edilen yaygınlığın benzerlik göstermesi önemlidir. Dolayısıyla hem tahmin edilen hem de gözlemlenen yaygınlık eşik değer için bir kriter olarak kullanılabilir [37]. Eşik değer seçimi ile birlikte sınıf tahminleri oluşur. Eşik değer seçimine bağlı olmayan yöntemlerden olan ROC (Receiver Operating Characteristics) ve kesinlik-doğru pozitif oran (Precision-Recall; PR) eğrileri yaklaşımı model performansını değerlendirmek için kullanılan daha objektif değerlendirme ölçüleridir [35].

İki sınıflı bir problemde, ROC eğrisi, modelin pozitif gözlemleri doğru bir şekilde sınıflandırdığı oran (duyarlılık) ile negatif gözlemleri yanlışlıkla pozitif olarak sınıflandırdığı oran arasındaki dengeyi görselleştirmemizi sağlar. ROC eğrisinin altındaki alan (Area Under Curve – AUC) ise modelin doğruluğunun bir ölçüsüdür [24]. ROC eğrisinin altındaki alan (Area Under Curve - AUC_{ROC}), genel model performansı için bir ölçü vermektedir. İyi modellerin AUC değeri 1'e yakınken, zayıf modellerin AUC değeri 0.5'e yakındır. Bu değerlere bakılarak model performansları değerlendirilir [36].

İkili sınıflama modellerinde sınıf dağılımında dengesizlik olması durumunda ROC eğrilerine alternatif olarak PR eğrileri önerilmiştir [35, 38]. ROC alanı ve PR alanı arasındaki önemli fark eğrilerin görselliğidir. PR eğrilerinde, ROC uzayında belirgin olmayan algoritmalar arasındaki farklılıklar ortaya çıkabilmektedir. PR alanında dikey eksende kesinlik (precision) ve yatay eksende duyarlılık (doğru pozitif oran, recall-sensitivity) bulunmaktadır [39]. ROC eğrisinde olduğu gibi PR eğrisinde de eğri altında kalan alan (Area Under Curve- AUC_{PR}) kullanılabilir ancak AUC_{PR} pozitif sınıfın yaygınlığına göre değişmektedir ve beklenen değeri veri kümesindeki pozitif sınıf oranına yakınıdır. AUC_{PR} değeri için alt sınır pozitif sınıfın yaygınlık değeridir. AUC_{PR} değeri yaygınlık değerinden ne kadar büyükse sınıflandırıcının o oranda iyi olduğu söylenir. [35]. AUC_{ROC} değerinde olduğu gibi iyi modellerin AUC_{PR} değerinin de 1'e yakın olması beklenir.

4. UYGULAMA (APPLICATION)

4.1 CART Sonuçlar

CART algoritmasında bölünme kriteri olarak gini endeksi kullanılmıştır. CART analizi rpart [40] paketi yardımıyla yapılmıştır. Terminal düğümlerde 2 veya fazla gözlem olduğu sürece bölünme işlemi devam ettirilmiştir. Bulunan bu değer denemeler sonucu test kümesindeki AUC değerini maksimum yapan değerdir ve bir parametre (minimum gözlem sayısı) olarak CART yönteminde kullanılmıştır. CART yöntemi kullanılarak elde edilen sınıflandırma tablosu eğitim ve test performansları sırasıyla Tablo 7 ve Tablo 8'de verilmiştir.

Tablo 7: CART (Eğitim kümesi performansı)

	Actual (Gerçek)	
Predicted (Tahmin)	**0**	**1**
0	363	5
1	70	27

Tablo 8: CART (Test kümesi performansı)

	Actual (Gerçek)	
Predicted (Tahmin)	**0**	**1**
0	88	8
1	13	5

Konut satışında önemli olan gerçekten satış yapılabilecek müşterilerin belirlenmesidir. Bu bağlamda CART yöntemi test kümesinde satış yapılan 13 müşteriden 5 tanesini doğru tahmin edebilmiştir. Bu değerler kesim noktasına bağlı olarak değişmektedir.

Tablo 9: CART (Farklı ölçütlere göre eğitim ve test kümesi performansları)

	Duyarlılık	**Özgüllük**	**Kesinlik**	**Hata**	**Doğruluk**
Eğitim	0.84	0.84	0.28	0.16	0.84
Test	0.38	0.87	0.27	0.18	0.82

Farklı sınıflandırma metriklerinin sonuçları eğitim ve test kümeleri için Tablo 9'da verilmiştir. CART yöntemi eğitim kümesindeki duyarlılığı %84 iken test kümesinde bu oran %38 olmuştur. Test doğruluk değeri ise 0.92 olarak bulunmuştur. Duyarlılık dışında diğer ölçütler eğitim ve test kümelerinde birbirine yakın değerler almıştır. Bu tablodaki hesaplanan tüm ölçütler, kesim noktasının alışveriş gerçekleştirenler sınıfının yaygınlık değeri olan %8 alınmasıyla elde edilmiştir. Kesim noktasının değişmesi durumunda hesaplanan değerler de değişiklik gösterecektir.

Tablo 10: CART (ROC-AUC ve PR-AUC değerleri)

	ROC-AUC	**PR-AUC**
Eğitim	0.91	0.71

Test	0.63	0.24

Tablo 10'da ise kesim noktasından bağımsız olarak hesaplanan ROC-AUC ve PR-AUC değerleri yer almaktadır. ROC-AUC ve PR-AUC değerleri PRROC [41, 42] paketi kullanılarak hesaplanmıştır. Bu değerlere bakıldığında eğitim ve test kümeleri arasında farklılıklar bulunmaktadır. Fakat bu tür dengesiz veri tiplerindeki performansı ölçmek için kullanılan PR-AUC değeri alışveriş gerçekleştirenler sınıfının yaygınlık değeri olan %8 değerinden daha büyük bulunmuştur. Bu durum da modelin sınıfları dengesiz dağılan bu veri kümesi üzerinde iyi bir performans gösterdiğinin bir ölçüsü olarak kabul edilir

4.2 Random Forest Sonuçlar

Random Forest yöntemi için önceden tanımlı parametrelerle model çalıştırılmış ve 25. ağaçta (ntree=25) minimum OOB (Out of Bag) hatasına (0.071) ulaşılmıştır. Ntree parametresi RF algoritmasında yer alan ağaç sayısını göstermektedir. Sonrasında 25 ağaçlı model için değişken sayısı (73) kadar döngü ile modeller oluşturulmuş ve rastgele değişken sayısının 8 olması durumunda model minimum OOB hatasına sahip olduğu görülmüştür. Rastgele orman analizi için R programında randomForest kütüphanesi kullanılmıştır [43].

Bu parametrelere göre Random Forest yönteminden elde edilen eğitim ve test performansları sırasıyla Tablo 11 ve Tablo 12'de hata matrisi biçiminde verilmiştir.

Tablo 11: Random Forest (Eğitim kümesi performansı)

	Actual (Gerçek)	
Predicted (Tahmin)	**0**	**1**
0	410	9
1	23	23

Tablo 12: Random Forest (Test kümesi performansı)

	Actual (Gerçek)	
Predicted (Tahmin)	**0**	**1**
0	92	10
1	9	3

Random Forest yöntemi test kümesinde satış yapılan 13 müşteriden sadece 3 müşteriyi doğru olarak tahmin edebilmiştir. Bu değerler kesim noktasına bağlı olarak değişmektedir.

Tablo 13: Random Forest (Farklı ölçütlere göre eğitim ve test kümesi performansları)

	Duyarlılık	**Özgüllük**	**Kesinlik**	**Hata**	**Doğruluk**
Eğitim	0.72	0.95	0.50	0.07	0.93
Test	0.23	0.91	0.25	0.17	0.83

Farklı sınıflandırma metriklerinin sonuçları eğitim ve test kümeleri için Tablo 13'de verilmiştir Random Forest yönteminin eğitim kümesindeki duyarlılığı %72 iken test kümesinde bu oran %23 olmuştur. CART yönteminden daha düşük bir duyarlılık değeri elde edilmiştir. Bu tablodaki hesaplanan tüm metrikler, kesim noktasının alışveriş gerçekleştirenler sınıfının yaygınlık değeri olan %8 alınmasıyla elde edilmiştir. Kesim noktasının değişmesi durumunda hesaplanan değerler de değişiklik gösterecektir.

Tablo 14: Random Forest (ROC-AUC ve PR-AUC değerleri

	ROC-AUC	**PR-AUC**
Eğitim	0.90	0.72
Test	0.66	0.35

Tablo 14'de ise kesim noktasından bağımsız olarak hesaplanan ROC-AUC ve PR-AUC değerleri yer almaktadır. PR-AUC değeri alışveriş gerçekleştirenler sınıfının yaygınlık değeri olan %8 değerinden daha büyük bulunmuştur. Bu durum da modelin sınıfları dengesiz dağılan bu veri kümesi üzerinde yine de iyi bir performans gösterdiğinin bir ölçüsü olarak kabul edilir. PR-AUC değeri CART yönteminden daha iyi sonuç vermiştir

4.3 XGBoost Sonuçlar

XGBoost algoritması için öncelikle modelde yer alan ağaç sayısı (iterasyon) belirlenmiştir. Ağaç sayısı belirlenirken minimum log-kayıp (log-loss) değeri dikkate alınmıştır. Bu değer eğitim ve test kümesi için ayrı ayrı belirlemiştir. Eğitim ve test kümlerindeki log-kayıp oranının iterasyon (yineleme sayısı, XGBoost algoritmasında ağaç sayısına da karşılık gelmektedir) sayısına göre test kümesi için log-kayıp oranı minimum değerini (0.32) 208.iterayonda almıştır. Öğrenme oranı (learning rate) ise 0.01 olarak alınmıştır. Öğrenme oranının küçük olması aşırı öğrenmeyi engellemektedir. Model önce 1000 iterasyon olacak şekilde çalıştırılmış ve test kümesi için minimum log-kayıp değerini 208. iterasyonda almıştır. XGBoost analizi için R dilinde hazırlanmış xgboost [44] kütüphanesi kullanılmıştır.

Yukarıda belirtilen parametrelere göre XGBoost algoritmasının çalıştırılmasıyla elde edilen eğitim ve test performansları sırasıyla Tablo 15 ve Tablo 16'de sunulmuştur.

Tablo 15: XGBoost (Eğitim kümesi performansı)

	Actual (Gerçek)	
Predicted (Tahmin)	**0**	**1**
0	222	1
1	211	31

Tablo 16: XGBoost (Test kümesi performansı)

	Actual (Gerçek)	
Predicted (Tahmin)	**0**	**1**
0	51	4
1	50	9

XGBoost yöntemi test kümesinde satış yapılan 13 müşteriden 9 müşteriyi doğru olarak tahmin edebilmiştir. Bu da CART ve Random Forest yöntemine göre duyarlılık yönünden daha iyi sonuçlar verdiğini göstermektedir.

Tablo 17: XGBoost (Farklı ölçütlere göre eğitim ve test kümesi performansları)

	Duyarlılık	**Özgüllük**	**Kesinlik**	**Hata**	**Doğruluk**
Eğitim	0.96	0.50	0.12	0.47	0.53
Test	0.69	0.50	0.15	0.47	0.53

Farklı sınıflandırma metriklerinin sonuçları eğitim ve test kümeleri için Tablo 17'de verilmiştir. RF yöntemi eğitim kümesindeki duyarlılığı %96 iken test kümesinde bu oran %69 olmuştur. CART ve RF yönteminden daha yüksek bir duyarlılık değeri elde edilmiştir. Bu tablodaki hesaplanan tüm metrikler, kesim noktasının alışveriş gerçekleştirenler sınıfının yaygınlık değeri olan %8 alınmasıyla elde edilmiştir. Kesim noktasının değişmesi durumunda hesaplanan değerler de değişiklik gösterecektir.

Tablo 18: XGBoost (ROC-AUC ve PR-AUC değerleri)

	ROC-AUC	**PR-AUC**
Eğitim	0.92	0.68
Test	0.67	0.31

Tablo 18'de ise kesim noktasından bağımsız olarak hesaplanan ROC-AUC ve PR-AUC değerleri yer almaktadır. PR-AUC değeri alışveriş gerçekleştirenler sınıfının yaygınlık değeri olan %8 değerinden daha büyük bulunmuştur. Bu durum da modelin sınıfları dengesiz dağılan bu veri kümesi üzerinde yine de iyi bir performans gösterdiğinin bir ölçüsü olarak kabul edilir. Test veri kümesi için PR-AUC değeri CART yönteminden yüksek ancak RF algoritmasından düşük bulunmuştur.

4.4 Genel Değerlendirme

Genel bir değerlendirme için eğitim veri kümesi sonuçları Tablo 18'de sunulmuştur.

Tablo 19: Yöntemlerin farklı ölçütlere göre eğitim kümesi performansları

Algoritma	**Duyarlılık**	**Özgüllük**	**Kesinlik**	**Hata**	**Doğruluk**
CART	0.84	0.84	0.28	0.16	0.84
RF	0.72	0.95	0.50	0.07	0.93
XGBoost	0.96	0.50	0.12	0.47	0.53

Tablo 19'a bakıldığında en yüksek doğruluk değeri Random Forest algoritmasında olurken duyarlılığı en yüksek algoritma ise XGBoost olmuştur. Bu tarz sınıf dengesizliğinin olduğu veri kümelerinde duyarlılık ve kesinlik değerlerine bakmak sonuçların daha anlaşılır yorumlanmasını sağlar. Bu tablodaki hesaplanan tüm ölçütler, kesim noktasının alışveriş gerçekleştirenler sınıfının yaygınlık değeri olan %8 alınmasıyla elde edilmiştir.

Kesim noktasının değişmesi durumunda eğitim veri kümesi için hesaplanan değerler de değişiklik gösterecektir. Bu sebeple kesim noktasından bağımsız olarak model performansını değerlendiren metriklerin dikkate alınması sonuçların daha objektif değerlendirilmesini sağlar. Eğitim veri kümesi için kesim noktasından bağımsız olarak hesaplanan ROC-AUC ve PR-AUC değerleri Tablo 19'da sunulmuştur..

Tablo 20: Yöntemlerin eğitim kümesindeki ROC-AUC ve PR-AUC değerleri

Algoritma	**ROC-AUC**	**PR-AUC**
CART	0.91	0.71
RF	0.90	0.72
XGBoost	0.92	0.68

Tablo 20'ye bakıldığında ROC-AUC ve PR-AUC değerlerinin birbirlerine yakın olmakla birlikte en yüksek ROC-AUC değeri XGBoost algoritmasında en yüksek PR-AUC değeri ise RF algoritmasında bulunmuştur.

Genel bir değerlendirme için test veri kümesi sonuçları Tablo 21'de sunulmuştur

Tablo 21: Yöntemlerin farklı ölçütlere göre test kümesi performansları

Algoritma	Duyarlılık	Özgüllük	Kesinlik	Hata	Doğruluk
CART	0.38	0.87	0.27	0.18	0.92
RF	0.23	0.91	0.25	0.17	0.83
XGBoost	0.69	0.50	0.15	0.47	0.53

Tablo 21'de test veri kümesi sonuçlarına bakıldığında en yüksek doğruluk değeri CART algoritmasında olurken duyarlılığı en yüksek algoritma ise XGBoost olmuştur. Bu tarz sınıf dengesizliğinin olduğu veri kümelerinde duyarlılık ve kesinlik değerlerine bakmak sonuçların daha anlaşılır yorumlanmasını sağlar. Bu tablodaki hesaplanan tüm metrikler, kesim noktasının alışveriş gerçekleştirenler sınıfının yaygınlık değeri olan %8 alınmasıyla elde edilmiştir. Kesim noktasının değişmesi durumunda hesaplanan değerler de değişiklik gösterecektir. Bu sebeple kesim noktasından bağımsız olarak model performansını değerlendiren metriklerin dikkate alınması sonuçların daha objektif değerlendirilmesini sağlar.

Test veri kümesi için kesim noktasından bağımsız olarak hesaplanan ROC-AUC ve PR-AUC değerleri Tablo 22'de sunulmuştur.

Tablo 22: Yöntemlerin test kümesindeki ROC-AUC ve PR-AUC değerleri

Algoritma	ROC-AUC	PR-AUC
CART	0.63	0.24
RF	0.66	0.35
XGBoost	0.67	0.31

Tablo 22'ye bakıldığında, ROC-AUC değerleri birbirine yakın bulurken, PR-AUC değerlerinde farklılar ortaya çıkmıştır. En yüksek PR-AUC değeri RF algoritmasında, ikinci en yüksek PR-AUC değeri de XGBoost algoritmasında bulunmuştur. Sınıf dengesizliğinin olduğu bu veri kümesinde PR-AUC değerine göre sonuçları değerlendirmek daha doğrudur. Bu sebeple veri kümesi için en iyi öğrenme Random Forest algoritmasında gerçekleşmiştir.

5. SONUÇLAR VE TARTIŞMA (RESULTS AND DISCUSSION)

Metin madenciliği, yapılandırılmamış veri türleri olan belgeler, sosyal medya gönderileri ve e-postalar gibi çok farklı formatta kaydedilmiş veri dosyalarından bilgi çıkarımı için kullanımı gittikçe yaygınlaşan bir yöntemdir. Metin madenciliği yöntemleri ile birlikte makine öğrenmesi yöntemleri kullanılarak büyük hacimli metinsel veriler içindeki kalıpların, eğilimlerin ve ilişkilerin ortaya çıkarılabilmesi karar verme süreçlerine yardımcı olarak işletmelerin veriye dayalı kararlar almasını kolaylaşmaktadır. En önde gelen uygulama alanlarından birisi de müşterilerin belirli bir ürün veya hizmet hakkındaki yorumlarını, görüşlerini ve sosyal medya gönderilerini analiz ederek işletmelerin müşteri duygularını, tercihlerini ve geri bildirimlerinin incelenmesidir. Büyük boyuttaki metinsel verilere göz atarak hemen anlaşılamayabilecek gizli kalıpları ve ilişkileri tanımlayarak bilgi keşfini kolaylaştırması metin madenciliğinin çok farklı alanlara uygulanabilmesini kolaylaştırmıştır. Ayrıca çok uzun metin dosyalarını analiz etmeyi, özetlemeyi sağlayan, belgede geçen temel konuları ortaya çıkaran metin madenciliği yöntemleri sayesinde zaman ve kaynak tasarrufu sağlanabilmektedir.

Metin madenciliği uygulamalarında yaygın olarak takip şu aşamalar bulunmaktadır. Problemin tanımlanması, metin tabanlı veri kullanılarak ulaşılmak istenen amacın belirlenmesidir. Örneğin müşteri görüşleri kullanılarak duygu analizinin yapılması, metin içindeki konuların belirlenmesi veya korpus içindeki belgelerin kümelenmesi vb.), veri toplama ve veri önişleme (analizde kullanılacak verilerin web, eposta, sosyal medya vb. gibi farklı ortamlardan ulaşılarak analizde kullanılacak korpus oluşturulduktan sonra verideki gürültüyü azaltabilmek için noktalama işaretlerinin ve özel karakterlerin kaldırılması, durma kelimelerinin kaldırılması, kelimelerinin eklerinin kaldırılarak köklerine ulaşılması vb.), öznitelik çıkarımı (metinsel verilerin tahminlerde kullanılacak makine öğrenmesi yöntemlerinin kullanımına hazır hale getirilebilmesi için nümerik hale getirilmesi gerekir. Bunun için Bag-of-Words, TF-IDF, Word Embedding yöntemlerinden birisi kullanılır), model seçimi ve eğitimi (Tahmin için kullanılacak makine öğrenmesi veya derin öğrenme algoritmalarının seçilmesidir. Bu seçimi veri büyüklüğü, verinin boyutu, oluşturulacak modelden beklenen açıklanabilirlik seviyesi gibi kriterler belirleyicidir. Analize hazır hale getirilmiş veri seti eğitim ve test kümesi şeklinde ayrılarak, eğitim kümesinde performansı en yüksek hale getirecek şekilde model parametreleri optimize edilir.), model değerlendirme

(Oluşturulan modellerin performanslarını belirlemek için örneğin sınıflandırma problemlerinde doğruluk, kesinlik, ROC-AUC gibi birçok farklı ölçüt kullanılır. Model performans ölçümünü daha güvenilir hale getirmek için çapraz geçerleme de kullanılabilir), modelin kullanımı (oluşturulan en iyi model daha önceden eğitimde kullanılmayan, yeni veriler üzerinde test edilir. Zaman içinde modelin performansını korumak veya geliştirmek için model parametrelerinde gerekli düzenlemeler yapılabilir), yorumlama ve görselleştirme (elde edilen sonuçlardan karar verme sürecinde nasıl faydalanabileceği yorumlanır, modelin davranışı hakkında daha detaylı bilgi alabilmek ve sonuçları görselleştirmek için hata matrisi, ROC eğrileri veya kelime bulutları gibi teknikler kullanılabilir)

Bu çalışmada konut üretimi ve satışı yapan bir inşaat şirketinin müşterileri ile yaptığı görüşmelerin kayıtlarının yer aldığı metin dosyaları incelenerek gelecekte bir müşterinin yapılan görüşme sonucunda konut alma kararı alıp almayacağını başarılı bir şekilde tahmin edecek bir model oluşturulmuştur. Veri kümesinde müşteriler ile telefonla veya yüz yüze yapılan görüşmelerin kayıtları ve daha sonrasında bu müşterilerin konut alıp almadıkları bilgisi yer almaktadır. Dolayısıyla problem makine öğrenmesi açısından bir ikili sınıflandırma problemi olarak ele alınmıştır. Beklenildiği gibi görüşme sonucu konut alma kararı veren müşteri sayısının, konut alma düşüncesinden vazgeçen veya erteleyen müşteriler göre çok daha düşük olmasından dolayı, sınıfların genelde dengeli olduğu ikili sınıflandırma problemlerinin aksine bu problem dengesiz bir ikili sınıflandırma problemidir. Bu nedenle model performansının biraz düşük olabileceği çalışmanın başında öngörülmüştür. Verileri eksik olan, analize uygun olmayan görüşme kayıtları çıkarıldığında 579 müşteri görüşmesi kaydından oluşan bir veri kümesine ulaşılmıştır. Bu veri kümesi yukarıda detayları verilen önişleme adımlarından sonra analize uygun hale getirilmiştir. Tüm görüşmeleri özetleyen vektör uzay modelini gösteren matris elde edilmiştir. Beklenildiği gibi bu seyrek (sparse) ve yüksek boyutlu (high dimensional) bir matristir. Bu iki problem makine öğrenmesi algoritmalarının performanslarını negatif yönde etkileyen faktörlerdir. Model oluşturma aşamasında Random Forest, Classification and Regression Tree (CART), ve XGBoost algoritmaları kullanılarak modellerin performansları karşılaştırılmıştır. Karşılaştırma için sınıflandırma problemlerinde yaygın olarak kullanılan kesinlik, doğruluk, özgüllük, ROC-AUC gibi ölçütlere ek olarak dengesiz sınıflandırma problemlerinin performansını analiz etmede daha anlamlı sonuçlar verdiği bilinen PR-AUC ölçütü de yer almıştır.

Duyarlılık ölçütüne göre analizde kullanılan CART, RF ve XGBoost yöntemleri test kümesi üzerinde sırasıyla 0.38, 0.23 ve 0.69 değerlerini vermişledir. Bu yönden XGBoost yönteminin gerçekte konut alan müşterilerin daha yüksek bir doğrulukla konut alacaklar sınıfında tahmin ettiği görülmektedir. Her iki sınıfın göz önüne alındığı tahminlerin genel olarak doğruluğuna bakıldığında ise elde edilen doğruluk değerleri sırasıyla 0.92, 0.83 ve 0.53 olarak elde edilmiştir. Bu yönden bakıldığında CART yönteminin önce çıktığı görülmektedir. Duyarlılık ve doğruluk gibi ölçütlerin elde edildiği hata matrislerindeki değerler çalışmada da ifade edildiği gibi tercih edilen eşik değerine göre değişmektedir. Çalışmada eşik değeri olarak tüm müşterilerin içinde görüşmeler sonucu konut alma kararı vermiş müşterilerin yüzdesi olan %8(0.08) değeri esas alınarak analizler yapılmıştır. Dolayısıyla farklı eşik değerleri tercih edilirse bu ölçütlerin değerleri de değişecektir. Eşik değerinden bağımsız olarak bir değerlendirme yapılabilmesini sağlayan PR-AUC değerlerine bakıldığında CART, RF, XGBoost yöntemleri sırasıyla 0.24,0.35 ve 0.31 bulunmuştur. Bu değerlerin hepsi seçilen eşik değeri olan 0.08 den büyük olduğu için dengesiz şekilde dağılan bu veri kümesi üzerinde yeterli performans gösterdikleri söylenebilir. PR-AUC değerine göre en iyi performansı Random Forest (RF) yöntemi göstermiştir.

Çalışma ileride daha fazla sayıda müşteri görüşmesinden elde edilen bir korpus üzerinde derin öğrenme yöntemleri (CNN veya RNN gibi) ve farklı öznitelik temsilleri ile (Word2Vect, BERT) gibi kullanılarak geliştirilebilir. Ayrıca verideki dengesizliğin sebep olduğu sorunları aşmak için literatürde yaygın olarak kullanılan SMOTE(Synthetic Minority Over-Sampling Technique) gibi yöntemlere yer verilerek sonuçlar incelenebilir.

KAYNAKLAR (REFERENCES)

[1] C. C. Aggarwal and C. Zhai, Eds. Mining Text Data (An Introduction to Text Mining. Springer, 2012.

[2] L. Duan and Y. Xiong, "Big data analytics and business analytics," Journal of Management Analytics, vol. 2, no. 1, pp. 1-21, 2015/01/02 2015, doi: 10.1080/23270012.2015.1020891.

[3] M. A. Hearst, "Untangling text data mining," presented at the Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, College Park, Maryland, 1999. [Online]. Available: https://doi.org/10.3115/1034678.1034679.

[4] I. Feinerer, K. Hornik, and D. Meyer, "Text Mining Infrastructure in R," Journal of Statistical Software, vol. 25, no. 5, pp. 1 - 54, 03/31 2008, doi: 10.18637/jss.v025.i05.

[5] R. Feldman and I. Dagan, "Knowledge Discovery in Textual Databases (KDT)," 06/28 1995.

[6] A. Hotho, A. Nürnberger, and G. Paass, "A Brief Survey of Text Mining," LDV Forum - GLDV Journal for Computational Linguistics and Language Technology, vol. 20, pp. 19-62, 07/01 2005, doi: 10.21248/jlcl.20.2005.68.

[7] R. Feldman, Ronen, Sanger, and James, The text mining handbook: Advanced approaches in analyzing unstructured data. 2007.

[8] D. Delen and M. Crossland, "Seeding the survey and analysis of research literature with text mining," Expert Systems with Applications, vol. 34, pp. 1707-1720, 04/01 2008, doi: 10.1016/j.eswa.2007.01.035.

[9] P. Hosseini, S. Khoshsirat, M. Jalayer, S. Das, and H. Zhou, "Application of text mining techniques to identify actual wrong-way driving (WWD) crashes in police reports," International Journal of Transportation Science and Technology, vol. 12, no. 4, pp. 1038-1051, 2023/12/01/ 2023, doi: https://doi.org/10.1016/j.ijtst.2022.12.002.

[10] S. Soleimani, M. Leitner, and J. Codjoe, "Applying machine learning, text mining, and spatial analysis techniques to develop a highway-railroad grade crossing consolidation model," Accident Analysis & Prevention, vol. 152, p. 105985, 2021/03/01/ 2021, doi: https://doi.org/10.1016/j.aap.2021.105985.

[11] M. Nilashi et al., "Big social data and customer decision making in vegetarian restaurants: A combined machine learning method," Journal of Retailing and Consumer Services, vol. 62, no. 102630, 2021.

[12] A. Petropoulos and V. Siakoulis, "Can central bank speeches predict financial market turbulence? Evidence from an adaptive NLP sentiment index analysis using XGBoost machine learning technique," Central Bank Review, vol. 21, no. 4, pp. 141-153, 2021/12/01/ 2021, doi: https://doi.org/10.1016/j.cbrev.2021.12.002.

[13] S. Chatterjee, D. Goyal, A. Prakash, and J. Sharma, "Exploring healthcare/health-product ecommerce satisfaction: A text mining and machine learning application," Journal of Business Research, vol. 131, pp. 815-825, 2021/07/01/ 2021, doi: https://doi.org/10.1016/j.jbusres.2020.10.043.

[14] W.-C. Lin, C.-F. Tsai, and H. Chen, "Factors affecting text mining based stock prediction: Text feature representations, machine learning models, and news platforms," Applied Soft Computing, vol. 130, p. 109673, 10/01 2022, doi: 10.1016/j.asoc.2022.109673.

[15] C. Allenbrand, "Supervised and unsupervised learning models for pharmaceutical drug rating and classification using consumer generated reviews," Healthcare Analytics, vol. 5, p. 100288, 2024/06/01/ 2024, doi: https://doi.org/10.1016/j.health.2023.100288.

[16] Y. Anagun, N. S. Bolel, S. Isik, and S. E. Ozkan, "DEEP LEARNING-BASED CUSTOMER COMPLAINT MANAGEMENT," Journal of Organizational Computing and Electronic Commerce, vol. 32, no. 3-4, pp. 217-231, 2022/10/02 2022, doi: 10.1080/10919392.2023.2210049.

[17] S. Isik, Z. Kurt, Y. Anagun, and K. Ozkan, "Spam E-mail Classification Recurrent Neural Networks for Spam E-mail Classification on an Agglutinative Language," International Journal of Intelligent Systems and Applications in Engineering, vol. 8, no. 4, pp. 221-227, 12/30 2020, doi: 10.18201/ijisae.2020466316.

[18] S. Baek, W. Jung, and S. H. Han, "A critical review of text-based research in construction: Data source, analysis method, and implications," Automation in Construction, vol. 132, p. 103915, 12/01 2021, doi: 10.1016/j.autcon.2021.103915.

[19] H. Yan, M. Ma, Y. Wu, H. Fan, and C. Dong, "Overview and analysis of the text mining applications in the construction industry," Heliyon, vol. 8, no. 12, p. e12088, 2022/12/01/ 2022, doi: https://doi.org/10.1016/j.heliyon.2022.e12088.

[20] A. Shamshiri, K. Ryu, and J. Y. Park, "Text mining and natural language processing in construction," Automation in Construction, vol. 158, p. 105200, 02/01 2024, doi: 10.1016/j.autcon.2023.105200.

[21] R: A Language and Environment for Statistical Computing. (2021). R Foundation for Statistical Computing, Vienna, Austria. [Online]. Available: https://www.R-project.org/

[22] E. Ikonomakis, S. Kotsiantis, and V. Tampakas, "Text Classification Using Machine Learning Techniques," WSEAS transactions on computers, vol. 4, pp. 966-974, 08/01 2005.

[23] E. Leopold and J. Kindermann, "Text Categorization with Support Vector Machines. How to Represent Texts in Input Space?," Machine Learning, vol. 46, no. 1, pp. 423-444, 2002/01/01 2002, doi: 10.1023/A:1012491419635.

[24] J. Han, M. Kamber, and J. Pei, Data mining : concepts and techniques, 3 ed. Morgan Kaufmann, 2012.

[25] G. James, D. Witten, T. Hastie, and R. Tibshirani, An introduction to statistical learning : with applications in R. New York : Springer, 2013.

[26] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.

[27] L. Rokach, Pattern Classification Using Ensemble Methods. Singapore: World Scientific Publishing, 2010.

[28] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," Annals of statistics, pp. 1189-1232, 2001.

[29] J. Son, I. Jung, K. Park, and B. Han, "Tracking-by-Segmentation with Online Gradient Boosting Decision Tree," in 2015 IEEE International Conference on Computer Vision (ICCV), 7-13 Dec. 2015 2015, pp. 3056-3064, doi: 10.1109/ICCV.2015.350.

[30] G. Ke et al., "LightGBM: a highly efficient gradient boosting decision tree," presented at the Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA, 2017.

[31] R. Mitchell and E. Frank, "Accelerating the XGBoost algorithm using GPU computing," PeerJ Comput. Sci., vol. 3, p. e127, 2017.

[32] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," presented at the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA, 2016.

[33] T. Hastie, R. Tisbshirani, and J. Friedman, The Elements of statistical learning : data mining, inference, and prediction. New York: Springer (in English), 2018.

[34] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, Applied Logistic Regression, 3 ed. John Wiley & Sons, 2013.

[35] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, Learning from Imbalanced Data Sets. Springer Nature Switzerland AG, 2018.

[36] E. A. Freeman and G. G. Moisen, "A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa," Ecological Modelling, vol. 217, no. 1, pp. 48-58, 2008, doi: https://doi.org/10.1016/j.ecolmodel.2008.05.015.

[37] J. S. Cramer, Logit Models from Economics and Other Fields. Cambridge University Press, 2003.

[38] T. Saito and M. Rehmsmeier, "The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets," PLoS ONE vol. 10, no. 3, p. e0118432, 2015.

[39] J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves," presented at the Proceedings of the 23rd international conference on Machine learning, Pittsburgh, Pennsylvania, USA, 2006. [Online]. Available: https://doi.org/10.1145/1143844.1143874.

[40] T. Therneau and B. Atkinson, "rpart: Recursive Partitioning and Regression Trees," 2019. [Online]. Available: https://CRAN.R-project.org/package=rpart.

[41] J. Grau, I. Grosse, and J. Keilwagen, "PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R," Bioinformatics, vol. 31, no. 15, pp. 2595-2597, 2015.

[42] J. Keilwagen, I. Grosse, and J. Grau, "Area under Precision-Recall Curves for Weighted and Unweighted Data," vol. 9, no. 3, 2014.

[43] A. Liaw and M. Wiener, "Classification and Regression by randomForest," R News, vol. 2, no. 3, pp. 18-22, 2002. [Online]. Available: https://CRAN.R-project.org/doc/Rnews/.

[44] T. Chen et al., "xgboost: Extreme Gradient Boosting," 2021. [Online]. Available: https://CRAN.R-project.org/package=xgboost.