T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

MAKİNE ÖĞRENMESİ YÖNTEMLERİYLE DEMOGRAFİK VERİLER ÜZERİNDEN YABANCI DİL SEVİYESİ TAHMİNİ

Gonca Gül DUMAN ORAL

YÜKSEK LİSANS TEZİ

Endüstri Mühendisliği Anabilim Dalı

Sistem Mühendisliği Programı

Danışman

Prof. Dr. Selin SONER KARA

Şubat, 2024

T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

MAKİNE ÖĞRENMESİ YÖNTEMLERİYLE DEMOGRAFİK VERİLER ÜZERİNDEN YABANCI DİL SEVİYESİ TAHMİNİ

Gonca Gül DUMAN ORAL tarafından hazırlanan tez çalışması 08.02.2024 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü Endüstri Mühendisliği Anabilim Dalı, Sistem Mühendisliği Programı **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Prof. Dr. Selin SONER KARA
Yıldız Teknik Üniversitesi
Danışman

Jüri Üyeleri

Prof. Dr. Selin SONER KARA, Danışman
Yıldız Teknik Üniversitesi _____

Prof. Dr. Selçuk ÇEBİ, Üye
Yıldız Teknik Üniversitesi _____

Prof. Dr. Ali Osman KUŞAKÇI, Üye
İbn Haldun Üniversitesi _____

Danışmanım Prof. Dr. Selin SONER KARA sorumluluğunda tarafımca hazırlanan "Makine Öğrenmesi Yöntemleriyle Demografik Verileri Üzerinden Yabancı Dil Seviyesi Tahmini" başlıklı çalışmada veri toplama ve veri kullanımında gerekli yasal izinleri aldığımı, diğer kaynaklardan aldığım bilgileri ana metin ve referanslarda eksiksiz gösterdiğimi, araştırma verilerine ve sonuçlarına ilişkin çarpıtma ve/veya sahtecilik yapmadığımı, çalışmam süresince bilimsel araştırma ve etik ilkelerine uygun davrandığımı beyan ederim. Beyanımın aksinin ispatı halinde her türlü yasal sonucu kabul ederim.

Gonca Gül DUMAN ORAL

İmza



Aileme

TEŞEKKÜR

Tez çalışmalarımda bana her zaman destek olan danışmanım Prof. Dr. Selin SONER KARA'ya ve her daim beni destekleyen aileme çok teşekkür ederim.

Gonca Gül DUMAN ORAL

İÇİNDEKİLER

TEŞEKKÜR iv
SİMGE LİSTESİ vi
KISALTMA LİSTESİ vii
ŞEKİL LİSTESİ viii
TABLO LİSTESİ ix
ÖZET x
ABSTRACT xii
1 GİRİŞ 1
1.1 Literatür Taraması 2
2 KULLANILAN YÖNTEMLER 8
2.1 K En Yakın Komşu 8
2.2 Karar Ağaçları 9
2.3 Destek Vektör Makineleri 11
2.4 Naïve Bayes 13
2.5 Lojistik Regresyon 14
2.6 Rastgele Orman 16
2.7 GBM 17
2.8 Topluluk Öğrenmesi 18
2.9 Çalışmada Kullanılan Anahtar Performans Göstergeleri 19
3 MATERYAL VE METOT 21
3.1 Veri Setinin İncelenmesi, Temizliği ve Hazırlığı 21
3.2 Makine Öğrenmesi Modelleri 27
3.2.1 Tekil Modeller 27
3.2.2 Topluluk Öğrenmesi 30
4 SONUÇLAR 33
4.1 Tekil Model Sonuçları 33
4.1.1 Aşırı Uyum Testi Sonuçları 35
4.2 Topluluk Öğrenmesi Modelleri Sonuçları 36
4.2.1 Topluluk Öğrenmesi Modelinde Aşırı Uyum Testi Sonuçları 41
KAYNAKÇA 42
TEZDEN ÜRETİLMİŞ YAYINLAR 50

SİMGE LİSTESİ

F_1	F_1 anahtar değerlendirme gösterge skoru

KISALTMA LİSTESİ

DVM	Destek Vektör Makineleri
GBM	Gradient Boosting Machine (Gradyan Arttırma Makineleri)
SMOTE	Synthetic Minority Over-Sampling Technique

ŞEKİL LİSTESİ

Şekil 2.1 K en yakın komşu örnek şeması **9**
Şekil 2.2 Karar ağaçları örnek gösterim **11**
Şekil 2.3 Destek vektör makineleri örnek gösterim **12**
Şekil 2.4 Destek vektör makineleri, ikinci derece polinom ile kümeleme **13**
Şekil 2.5 Sigmoid eğrisi gösterimi **15**
Şekil 2.6 Örnek karar ağacı gösterimi **17**
Şekil 2.7 Karmaşıklık matrisi örnek gösterim **19**
Şekil 3.1 Değişkenin önemi çalışması grafiği **24**
Şekil 3.2 Eğitim düzeyi ile yabancı dil bilgisi ilişkisi **25**
Şekil 3.3 Mezun olunan bölüm ile yabancı dil bilgisi ilişkisi **26**
Şekil 4.1 Test edilen parametreler ve en iyi sonuçları veren değerleri **33**

TABLO LİSTESİ

Tablo 1.1 Paylaşılan akademik başarı tahminlemesi çalışmalarında en başarılı sonucu veren modeller **5**

Tablo 1.2 Paylaşılan çalışmaların veri kaynakları **7**

Tablo 3.1 Yabancı dil bilgi düzeyleri sınıf açıklaması, örneklem sayıları ve yüzdelik dilimleri **22**

Tablo 3.2 Sentetik veri üretimi sonrası yabancı dil bilgi düzeyleri sınıf ve örneklem sayıları **23**

Tablo 3.3 En baskın ilk on değişken **24**

Tablo 3.4 Topluluk öğrenmesi kombinasyonları **30**

Tablo 3.5 Topluluk öğrenmesi metotları **31**

Tablo 4.1 Tekil makine öğrenmesi modellerinin performans değerleri **34**

Tablo 4.2 Rastgele orman karmaşıklık matrisi **34**

Tablo 4.3 Eğitim veri seti ile çalıştırılan rastgele orman modelinin anahtar performans gösterge sonuçları **35**

Tablo 4.4 Topluluk öğrenmesi F_1 skorları **37**

Tablo 4.5 Topluluk öğrenmesi doğruluk değerleri **38**

Tablo 4.6 Topluluk öğrenmesi kappa değerleri **39**

Tablo 4.7 Topluluk öğrenmesi kesinlik ve duyarlılık değerleri **39**

Tablo 4.8 Topluluk öğrenmesi 2. kombinasyon ile yığma metodunun karmaşıklık matrisi **40**

Tablo 4.9 En başarılı sonucu veren topluluk öğrenmesi modelinin eğitim verisi ile anahtar performans gösterge sonuçları **41**

ÖZET

Makine Öğrenmesi Yöntemleriyle Demografik Veriler Üzerinden Yabancı Dil Seviyesi Tahmini

Gonca Gül DUMAN ORAL

Endüstri Mühendisliği Anabilim Dalı

Sistem Mühendisliği Programı

Yüksek Lisans Tezi

Danışman: Prof. Dr. Selin SONER KARA

Bu çalışmada Türkiye İstatistik Kurumu'ndan alınan Yetişkin Eğitimi Araştırması Mikro Veri Seti kullanılmıştır. Çalışmanın amacı bireylerin demografik verilerinin, yabancı dil bilgisinin tahmininde kullanılabileceğini paylaşmaktır.

Çalışmaya literatür araştırması ile başlanmıştır. Ardından, verisetindeki tekil değişkenler ile hedef değişken olan yabancı dil seviyesi arasındaki ilişkiler incelenmiştir. Yer yer bağlantılar kurulabilmiş olsa da, çoğunlukla tekil incelemelerin yetersizliği öne çıkmıştır. Bu nedenle, bireylerin tüm bilgilerini sürece dahil edebilmek için makine öğrenmesi modellerinin kullanımına ihtiyaç doğmuştur.

Öncelikle, veri seti makine öğrenmesi modellerinde kullanılabilir hale getirilmiştir. Varolan boş hücreler verisetinden kaldırılmıştır. Veride tahminlenecek sınıflar arasında dengesizlik tespit edilmiş olup, literatürde de sıkça kullanılan sentetik veri üretme yöntemi izlenerek bu dengesizlik giderilmiştir. Ardından tekil makine öğrenmesi ve topluluk öğrenmesi yöntemleri kullanılarak, değişkenler ve hedef değişken arasındaki ilişki kurulmuştur. Modellerin sonuçları doğruluk, kesinlik, F_1

skoru gibi değerlendirme metrikleri ile karşılaştırılmış, ve en yüksek değeri veren modelin en başarılı olduğu kabul edilmiştir. Model başarı metrik değerleri oldukça yüksek olduğundan, çalışmanın bireylerin demografik verilerinin yabancı dil bilgisi ile ilişkili olduğu hipotezinin doğrulandığı söylenilebilir.

Doğrulanmış hipotezden yola çıkılarak demografik veriler incelendiğinde, çalışmanın eğitim planlayıcılar için iyileştirme noktalarının tespiti açısından önemli bir kaynak olacağına inanılmaktadır.

Anahtar Kelimeler: Makine Öğrenmesi, Topluluk Öğrenmesi, Demografik Veriler, Eğitim Seviyesi, Yabancı Dil Tahmini

ABSTRACT

Prediction of Foreign Language Proficiency via Machine Learning Methods Using Demographic Data

Gonca Gül DUMAN ORAL

Department of Industrial Engineering

Master of Science Thesis

Supervisor: Prof. Dr. Selin SONER KARA

This study utilized the Micro Data Set from the Turkish Statistical Institute's Adult Education Survey. The aim of the study is to demonstrate the potential use of individuals' demographic data in predicting their proficiency in a foreign language.

The study starts with a thorough literature review. Subsequently, the relationships between individual variables in the dataset and the target variable, which is the level of proficiency in a foreign language, were examined. Although connections were occasionally established, the insufficiency of singular examinations was evident. Consequently, there arose a necessity to incorporate all available information about individuals into the process, prompting the use of machine learning models.

Initially, the dataset was prepared to be utilized in machine learning models. Empty cells within the dataset were removed. Imbalance among the classes to be predicted was detected, and this imbalance was eliminated by employing a synthetic data generation method commonly used in literature.

Following this, machine learning and ensemble learning methods were employed to establish relationships between variables and the target variable. Model comparisons were made using evaluation metrics such as accuracy, precision, and

F_1 score. The model yielding the highest values in these performance metrics was deemed the most successful. Given the considerably high success metric values of the model, it can be asserted that the hypothesis stating a relationship between individuals' demographic data and proficiency in a foreign language has been confirmed.

Building upon the confirmed hypothesis, when demographic data is examined, it is believed that this study will serve as a significant resource for identifying improvement areas for educational planners.

Keywords: Machine Learning, Ensemble Learning, Demographic Data, Education Level, Foreign Language Prediction

1
GİRİŞ

Bir ülkedeki insanların eğitim seviyesi bir ülkenin kültürel, sosyoekonomik ve toplumsal gelişiminde oldukça önemli faktörlerdir. Eğitim seviyesine ek olarak yüksek yabancı dil bilgisi, ülkeler ve farklı milliyetten bireyler arasındaki iletişimi kolaylaştırır, istihdam şansları yaratır ve kültürler alışverişini arttırır. Yabancı dil seviyesi gibi yetkinlikler diplomatik ilişkilerde de büyük öneme sahiptir. Dolayısıyla, bir ülkenin vatandaşlarının eğitim düzeyi ve bununla birlikte yabancı dil bilgi seviyesinin ülke için ekonomik, stratejik ve toplumsal kalkınmaya yönelik bir değeri bulunmaktadır. Eğitim planlaması, noksan kalan alanların tespiti, iyileştirmeler için de bireylerin aralarında eğitim seviyeleri dahil olmak üzere bilgilerini incelemek, verilerini gözlemlemek ve çıkarımlar yapmak iyileştirme alanlarının tespitini ve fırsatlarını sağlar.

Yabancı dil bilgi seviyesi yüksek olan bireylerin genel itibarıyla verilerini incelemek korelasyonları ortaya çıkarmak açısından yeterli gibi gözükse de, temelde eksik bir yaklaşım olabilir. Zira, çok farklı sosyoekonomik değerlere sahip insanların yabancı dil bilgi seviyeleri yüksek olabilir. Örneğin, bir kişinin tamamladığı lisansüstü öğretimi, bu kişinin iyi bir yabancı dil bilgisine sahip olma ihtimalini arttırırken, başka bir lise mezunu da lisansüstü mezunu ile benzer dil seviyesine sahip olabilir. Dolayısıyla, yabancı dil eğitim planlaması gibi çalışmalarda yalnızca tekil parametreler incelenirse hatalı sonuçlar karşılaşılabilir ve çalışma yanlış eğitim planlamalarıyla sonuçlanabilir.

Bu hususta, çok sayıda bireyin farklı alanlardaki verilerinin detaylıca incelenmesi, verideki örneklemler arasındaki bağlantıların kurulması, tek bir parametre ile neden sonuç ilişkisi kurmaktan daha sağlıklı sonuçlar verecektir. Bu gibi büyük veri incelemeleri de makine öğrenimi metotları ile oldukça kolaylaşmaktadır. Bu çalışma, bireylerin farklı sosyokültürel ve demografik bilgilerini inceleyerek, yabancı dil seviyeleri arasında bir ilişki kurmayı ve bu parametreler arasında bir ilişki olduğu iddiasını ispatlamayı hedeflemektedir. Çalışmanın başlangıcında, yapay zekanın eğitim alanındaki kullanımları ile hem güncel hem de geçmiş akademik verileri, demografik bilgileri, ekonomik düzeylere ait bilgileri eğitim

seviyesi tahmini için kullanan, ya da eğitim seviyelerini kullanarak demografik yapı ve sosyoekonomik düzey tahminlemesi yapan makine öğrenmesi çalışmaları incelenmiştir.

Bu tezde, literatür taraması ile başlayan birinci bölüm, çalışmanın amacı ile sonlanmaktadır. İkinci bölümde hem literatürdeki hem de uygulamadaki farklı makine öğrenmesi modelleri ile ve bu modellerin başarısını ölçecek anahtar performans göstergeleri açıklanmıştır. Üçüncü bölüm Türkiye İstatistik Kurumundan alınan Yetişkin Eğitimi Araştırması Mikro Veri Seti'ndeki [1] verilerin hazırlığı ile başlayarak, tezde kullanılan makine öğrenmesi modellerini ve hiperparametre optimizasyonlarını paylaşmaktadır.

Veri setinde bireylere ait iki farklı yabancı dil bilgisi bulunuyor olsa da, ikinci yabancı dili bilenler halihazırda en az bir yabancı dil bildiğinden, ikinci yabancı dil bilgisi çalışmaya dahil edilmemiştir. Dördüncü bölümde tüm modellerin sonuçları paylaşılarak tartışılmıştır. Bir sonraki bölümde tezde faydalılan kaynaklar yer almaktadır. Son bölümde ise tezden üretilmiş yayın belirtilerek tez çalışması sonlandırılmıştır.

1.1 Literatür Taraması

Makine öğrenmesi ve yapay zekanın gün geçtikçe hayatın her dalında kendine yer bulduğu yadsınamaz bir gerçektir. Mühendislikten sosyal sanatlara, hukuk alanından tıpa kadar birbirlerinden bağımsız sektörlerde dahi katkı sunabilir hale gelmiştir. Eğitim de, tıpkı yukarıda sayılan sektörlerde olduğu gibi bir ülke için en kritik alanlardan biridir, ve yapay zeka (AI) etkisini bu alanda da göstermektedir.

Örneğin, makine öğrenmesi ile kişiye özelleştirilmiş eğitim setlerinden, sınav ve ödev notlandırmaya, ders seçim robotlarına [2], öğrenciler için kariyer patikası tavsiyelerine, öğrenim hızı tespiti ve iyileştirme önerilerine kadar birçok alanda AI kullanımı göze çarpmaktadır [3]. Öğrenici modelleri sayesinde beden dili takibi, ses tonu ve kullanımı gibi veriler ile, anlık olarak sınıfta eğitimin iyileştirme önerileri yapılması [4] gibi fütüristik çalışmalar da yine yapay zekanın eğitim alanında getirebileceği yeniliklerdendir.

Yapay zeka ve makine öğrenmesinin eğitime olan katkılarına bahse olan bir diğer konu ise, eğitim alanındaki büyük verinin gelecekteki akademik ve eğitim başarılarının tahminlenmesi yönündedir.

Tsiamaki ve ark. [5], bir grup öğrencinin, eğitim için oluşturulmuş bir web sitesinin ne kadar ziyaret edildiğini, sayfalarda geçirilen süreyi, hangi eğitim portfolyölerinin kullanılıp kullanılmadığını inceleyerek öğrencinin devam durumunu, ve eğer devamsızlık durumun söz konusu değil ise de öğrencinin final notunu tahmin eden bir makine öğrenmesi çalışmasını autoML (automated machine learning) algoritmasını kullanarak yapmıştır.

Hu ve ark. [6] ise yine çevrimiçi bir öğrenme arayüzünde öğrencilerin başarısızlığı önceden keşfetme üzerine bir çalışma sunmuş olup, 330 öğrenci için tahminlerini Sınıflandırma ve Regresyon Ağaçları (CART) metoduyla en yüksek doğruluk oranlarını yakalamıştır. Bununla birlikte, zamana bağlı birey değişkenlerinin (yaş, gelir, vs.) hedef değişken tahminlemesinde kritik olduğunu vurgulamaktadır.

Öğrencilerin başarısını tahmin etmek için birden fazla alandaki veriyi birleştiren çalışmalar da mevcuttur [7]. Trakunphutthirak ve Lee internet bazlı bir eğitim arayüzünü ve internet tarayıcısındaki çerezleri kullanarak başarısızlık ihtimali bulunan öğrencilerin tespitini yapmıştır. Ardından, bu bilgileri demografik veriler ile de birleştirerek eğitmenler için anlamlı hale getirmiştir ve sebep sonuç ilişkisi kurulmasını sağlamıştır.

Lakkaraju ve ark. [8], 200000 öğrenci ile yaptıkları çalışmada Destek Vektör Makineleri, Rastgel Orman, Lojistik Regresyon, Adaboost ve Karar Ağaçları metotlarını test ederek, okulu zamanında bitirmeme riski bulunan öğrencilerin tespitini en iyi Karar Ağaçları metoduyla tahminlediğini paylaşmıştır. Cinsiyet, yaş, etnik köken, oturulan şehir ve mahalle, geçmişteki ve güncel not ortalaması, sağlık durumu, ana dildeki başarı durumu gibi değişkenler kullanılmıştır.

Al-Shehri ve ark. [9], 395 öğrencinin devam ettiği okul, cinsiyeti, oturduğu mahalle, ailesinin büyüklüğü, ebeveynlerinin meslekleri gibi demografik verileri ile, K En Yakın Komşu ve Destek Vektör Makineleri modellerini kullanarak, Destek Vektör Makineleri modeli ile daha başarılı bir tahminleme yaptığını paylaşmıştır. Veri sayısının az olması burada dikkat çeken etmenlerdendir.

Benzer bir çalışmada ise Tomasevic ve ark. [10], öğrencilerin geçmiş akademik seviyelerindeki sınav sonuçları, ödev ve quiz sonuçları, akademik kaynakları ne kadar kullanım frekansı bilgileri kullanarak, dersi bırakma eğilimi gösteren öğrencilerin tahminlemesini yapmıştır. Farklı makine öğrenmesi modelleri kullanırken, en iyi sonuçların Yapay Sinir Ağları modeli ile elde ettiği paylaşılmıştır. Bununla birlikte çalışmada öğrencilerin başarısını demografik bilgilerin fazla etkilemediği, eğitim ile ilgili verilerin daha baskın olduğu sonucuna varılmıştır.

Benzer bir çalışmada Şahin [11] öğrenci gruplarına ait sosyoekonomik özellikler ve ders notları verilerini değerlendirerek yıl sonu not ortalamalarının makine öğrenmesi yöntemleri ile tahmin edilmesi üzerine bir çalışma sunmuştur. Yapılan çalışma neticesinde Rastgele Orman metodu en iyi sonucu vermiştir. Yine öğrencilerin demografik bilgilerinden ziyade geçmiş dönemdeki akademik başarılarının güncel eğitimlerindeki başarılarına etkisinin daha yüksek olduğu paylaşılmıştır.

Bir başka çalışmada Mohammadi ve ark. [12] bir üniversitedeki öğrencilerin yine hem akademik hem de demografik bilgilerini kullanarak, yıl sonu not ortalamalarını tahminleme çalışması yürütmüş olup, K En Yakın Komşu, Karar Ağaçları ve Naive Bayes yöntemlerini kullanmıştır. En başarılı sonucun K En Yakın Komşu ile elde edildiği sunulmuştur.

Literatürde paylaşılan bir kısım çalışmalara bakılarak, güncel eğitim başarılarını geçmiş akademik başarıların etkilediği, demografik bilgilerin eğitim alınan seviyedeki başarılara etkisinin ise düşük olduğundan söz edilebilse de, aksi yönde, bir başka deyişle sosyoekonomik düzeylerin de akademik başarıyı etkileyeceğini iddia eden çalışmalar [13] da mevcuttur. Öğrencilerin güncel not durumuna ek olarak, daha subjektif bilgiler olan öğrenim hızı, zaman yönetme becerileri, test çözme becerileri ve hızı, motivasyonları gibi parametreleri kullanarak gelecek akademik başarıyı tahminleyen bir çalışma da [14] literatürde görülmektedir.

Martins ve ark. [15], öğrencilerin demografik değerlerini ve sosyoekonomik bilgilerini 3 değişkenli çoklu sınıflandırma metodunu kullanarak öğrencilerin derse kaydolma, bırakma ve başarıyla bitirmesini tahminlemiştir. Kurulan modellerde en başarılı sonucun Yükseltme (Boosting) metotları ile geldiği gösterilmiştir.

Çalışmada veri setindeki dengesizlikten de bahsedilmiş olup, bu dengesizliğin sentetik veri üretme metoduyla çözüldüğü paylaşılmıştır. Gök [16], ortaokul öğrencilerinin eğitim başarılarını etkileyen demografik, ailesel, vb. gibi farklı kaynaklı pek çok değişkeni kullanıp çok sınıflı makine öğrenmesi modelleri oluşturmuştur. Yürütülen çalışmada öğrencilerin dönem sonu puanları en iyi Rastgele Orman metoduyla tahminlenmiştir. Bir başka çalışma ise, farklı bir bakış açısı ile gelir gibi ekonomik düzeylerin akademik başarıya etkisini değil, akademik başarıların gelire etkisini incelemiştir [17].

Özetle, yukarıda paylaşılan tüm yayınlara ek olarak, öğrencilerin akademik başarılarını yine geçmiş akademik başarıları ya da eğitim verilerini kullanarak önceden tahminlemeye çalışan hem tekil makine öğrenmesi hem de topluluk öğrenmesi kullanan çalışmalar [18,19,20,21,22,23], dersi bırakma riski olan öğrencilerin tespitinde kullanılan modeller [24,25], öğrencilerin ya da bireylerin demografik verilerini kullanarak akademik başarı tahmininde bulunan çalışmalar [26,27,28] literatürde bolca görülebilmektedir. Başka bir perspektiften konuya yaklaşan bir çalışmada [29] ise, öğrencilerin duygularına ve online öğrenme platformlarındaki birbirleri ile olan etkileşimlerine göre başarı tahminlemesi yürütmektedir. Yapay sinir ağları ve yinelemeli sinir ağları, rastgele orman, gradient boosting gibi metotlar testlere tabi tutulmuştur. Bireylerin ya da öğrencilerin duygu durumlarına ya da spor yapma, düzenli beslenme gibi davranışsal faktörlere göre akademik performans tahminlemesi de yine literatürde [30, 31, 32, 33] görülen yaklaşımlardandır.

Tablo 1.1 Paylaşılan akademik başarı tahminlemesi çalışmalarında en başarılı sonucu veren modeller

Referans Numarası	**Makine Öğrenmesi Modeli**
10, 19, 26, 30	Yapay Sinir Ağları
11, 16, 18	Rastgele Orman
8, 25, 31, 32	Karar Ağaçları
12, 33	K En Yakın Komşu

Tablo 1.1 paylaşılan akademik başarı tahminlemesi çalışmalarında en başarılı sonucu veren modeller (devamı)

9, 28	Destek Vektör Makineleri
17	Naive Bayes
22, 23	Topluluk Öğrenmesi
5, 6, 13, 14, 27, 31	Diğer (autoML, CART, Regresyon, Boosting, İstatistiksel: Chi-kare)

Tablo 1.1, eğitim başarısı tahminleyen çalışmalarda en başarılı tahmini yapan modelleri paylaşmaktadır. Buna bakılarak, Yapay Sinir Ağları ve Karar Ağacı'nın en sık ve en doğru sonucu verdiği gözlemlenmektedir. En sık ikinci yüksek sonuç verdiği rastlanılan model ise Rastgele Orman modeli olduğu gözlemlenmiştir. Bir başka deyişle, akademik başarı tahmini çalışmalarında, veri kaynağı ne olursa olsun ağaç bazlı modellerin yüksek doğruluk sunduğu görülmektedir.

Öte yandan, bir topluluğun demografik datalarının makine öğrenmesi metotları ile farklı alanlarda tahminleme çalışmalarına da literatürde rastlanmaktadır. Örneğin tıp alanında genetik ve kişiye özgü ailevi geçmiş ve güncel sağlık durumu gibi demografik veriler ile kanser riski tahmini [34], yaş, çalışma durumu gibi demografik veriler ile hastaların antidepresana karşı verdikleri olumlu ya da olumsuz cevabın tahminlemesi [35], ya da yine sosyodemografik veriler vasıtasıyla topluluk öğrenmesi modelleri kullanılarak kişilerin alzheimer riskini tahmin etmek [36] gibi çalışmalar gözlemlenmiştir.

Benzer şekilde farklı demografik gruplardan gelen bireylerin yaş, cinsiyet, medeni durum gibi verileri ile grupların mutluluk tahminlemesi [37], toplum sağlığı tahmini [38], demografik yapının belli bir bölgedeki bir sektörün ekonomik büyümeye katkısını [39], ve hatta iklim ve çevre şartlarını tahminlemede [40], bir ülkenin gelecek yıllardaki nüfusunu tahmin etmeye yönelik çalışmalarda [41] dahi kullanılmaktadır.

Tablo 1.2'de, literatür taramasında paylaşılan çalışmaların kullanmış olduğu veri kanyakları gösterilmektedir. Elbette ki, paylaşılan makaleler arasında veri

kaynaklarını birleştirerek kullananlar da mevcuttur. Kullanılan veri kaynakları arasında keskin bir ayrım yapmak çok mümkün olmamaktadır.

Tablo 1.2 Paylaşılan çalışmaların veri kaynakları

Geçmiş ve güncel akademik veriler
Demografik Veriler
Davranışsal (Duygusal ve Rutin) Verileri
Sosyoekonomik veriler

Özetlemek gerekirse, güncel literatürde sosyokültürel ve demografik veriler tıp sektöründen çevreye, sağlıktan eğitime kadar çok farklı disiplinlerde tahminleme çalışması yapmak için kullanılmaktadır. Eğitim alanında ise yalnızca demografik veriler kullanılarak değil, geçmiş akademik başarılar kullanılarak da makine öğrenmesi modelleri çalışılmıştır. Bu alanda yapılan çalışmalar incelenilerek makine öğrenmesi ve yapay zekanın eğitim sektörü içerisinde de bulunarak farklı noktalarda katkı verdiğini görmek mümkündür. Gerek eğitimin kaynak noktasında iyileştirilmesi, gerekse de durum tespitleri ve iyileştirme noktası arayışları için makine öğrenmesi sıkça kullanılmaktadır. Literatürde gözlenen tüm bu akademik başarı tahminleme çalışmaları da, bu tezin odak noktası olan yabancı dil tahminlemesi ile ortak bir pencerede yer almaktadır.

2
KULLANILAN YÖNTEMLER

Bu bölümde, çalışmada kullanılan makine öğrenmesi modelleri ve bu modelleri değerlendirmede faydalanılan anahtar performans göstergeleri açıklanmıştır.

2.1 K En Yakın Komşu

K En Yakın Komşu algoritması literarürde tahminleme yapmak için sıkça kullanılan bir algoritmadır. Basitliği ve az sayıda veride başarılı sonuçlar vermesi nedeniyle uygulamalarda sıkça tercih edilmektedir. K En Yakın Komşu algoritması gözetimli makine öğrenmesi yöntemlerindendir. Sınıflandırma problemlerinde yaygın olarak kullanılmaktadır.

Bir veri setine yeni bir veri eklenmesi durumunda, hangi sınıfa dahil edileceğini belirlenmesi için kullanılır. Algoritma önce veriyi inceler, k değeri belirlenir, ilgili değer verilen bir noktaya en yakın komşu sayısını ifade etmektedir. Daha sonra uzaklık hesabı yapılır. Uzaklık hesabı için Öklid mesafesi, Minkowski mesafesi, Manhattan mesafesi fonksiyonlarından biri kullanılabilir. Bu fonksiyonlar, sırası ile aşağıda paylaşılmıştır.

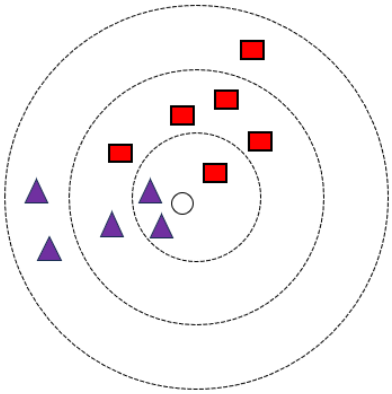
$$D(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{2.1}$$

$$D(x,y) = (\sum_{i=1}^{n}|x_i - y_i|^p)^{1/p} \tag{2.2}$$

$$D(x,y) = \left(\sum_{i=1}^{n}|x_i - y_i|\right) \tag{2.3}$$

Uzaklık belirlendikten sonra tahmin edilecek veriye en yakın komşu bulunur ve bu veri etiketlenir. Bir başka deyişle her bir adımda en yakındaki veriyi alarak sınıflama işlemi yapılmaktadır. Algoritmanın temel çalışma prensibi verileri en yakınındaki bulunan veriye göre gruplayarak, yeni tahmin edilecek verinin hangi grupta yer alacağını belirlemektir.

Örnek olarak Şekil 2.1'e bakıldığında KNN algoritmasıyla ortadaki yeni verinin üçgen veya dikdörtgen gruplarından hangisine dahil olacağına karar verilmesi gösterilmiştir. Buna göre, k=3 olması durumunda en yakın 3 veriden ikisi üçgen olduğundan dahil olcağı grup üçgendir. Eğer k=7 olursa, en yakın 7 veriden 4 tanesi dikdörtgen olduğundan dahil olacağı grup dikdörtgendir. Eğer k=11 olursa, en yakın 11 veriden 6 tanesi dikdörtgen olduğundan dahil olacağı grup dikdörtgendir.



Şekil 2.1 K en yakın komşu örnek şeması

2.2 Karar Ağaçları

Karar Ağaçları [43], sıkça kullanılan sınıflandırma algoritmalarındandır. Bir karar ağacı kök, karar düğümleri ve yapraklar olarak 3 bölümden oluşmaktadır. Bir karar ağacı kök ile başlar ve bu nokta verinin tümünü sembolize eder. Kökteki koşula göre gözlem verisi sınıflandırılır. Kök hücrelerinin altında karar düğümleri yer almaktadır. Her bir gözlem verisi karar düğümleri ile sınıflandırılır. Karar düğümü adeti çoğaldıkça modelin karmaşıklığı artmaktadır. Karar ağacının en alt kısmında ise yapraklar yer almaktadır. Yapraklar sınıflandırmaya ait nihai sonucu vermektedir.

Literatürde Karar Ağaçlarında kullanılan en yaygın algoritma uygulaması ID3 algoritmasıdır [44]. Bu algoritma, kök düğümden en alt dala doğru uzanan bir

yapıya sahiptir. Bu algoritma, sınıflandırma işini entropi adı verilen bir kavrama göre tamamlar. Entropi, mühendislik disiplinlerinde düzensizlik anlamına gelmektedir. Bu bilgi ışığında söylenebilir ki, ID3 algoritmasında entropi düşük olduğunda örneklemler benzer özelliklere sahip olacak ve tek bir sınıfa ait olacaktır.

$$Entropi(Sistem) = -\sum_{i=1}^{n} p_i log_2 p_i \quad (2.4)$$

Yukarıda paylaşılan entropi formulayonunda, n değeri hedef değişkenin kaç farklı sınıfa sahip olduğu gösterir. p_i, bir sınıf etiketinin doğru olma olasılığıdır. Veri setindeki bir örnekleminin entropisi, hedef değişkendeki tüm sınıflar için hesaplanır ve matematiksel olarak toplanır.

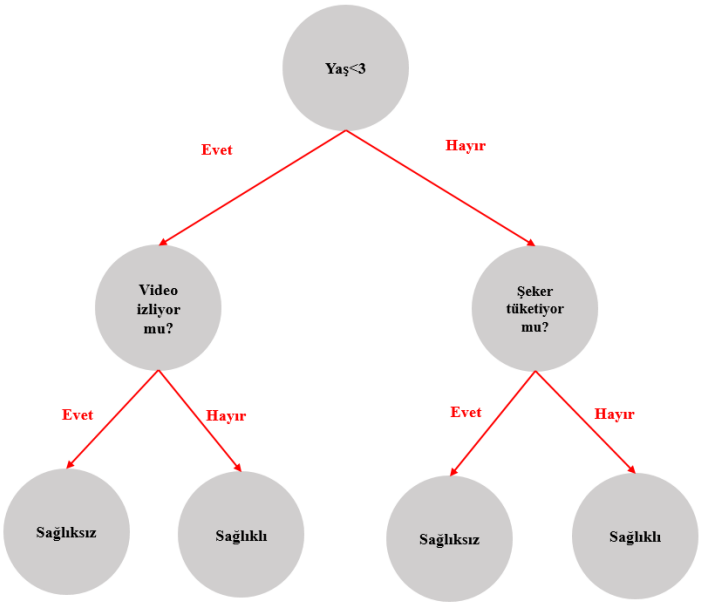
Entropinin düşük olması, örneklemlerin tek bir sınıf etiketinde toplanacağı anlamına geldiğinden, entropideki düşüş kazanç olarak yorumlanır. Entropiyi düşüren, nihai hedefi etkileyen her öznitelik için de ayrıca entropisi hesaplanır ve hedef değişkenlerin entropisinden çıkarılır. Kazanç formulü aşağıda paylaşılmıştır.

$$Kazanç(S, A) = Entropi(Sistem) - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{S} \times Entropi(S_v) \quad (2.5)$$

Burada $Kazanç(S, A)$, A adındaki bir özelliğin yapılan bölme işlemleri sonucundaki kazancı temsil eder. $Entropi(Sistem)$ tüm veri kümesinin entropisini, S örneklem sayısını, S_v ise A özelliğinin değeri v olan verikümesini temsil eder. $Entropi(S_v)$ ise S_v veri kümesinin düzensizliğini, yani entropisini ölçer.

Bu hesaplamalar, kazanç değeri en büyük çıkan değişken belirlenene kadar devam eder ve o değişken kök düğüm seçilir. Bir sonraki adımda entropi yeniden hesaplanır ve kazancı yine en büyük çıkan değişken karar düğümü seçilir. Bu süreç, yaprağa ulaşana kadar devam edecek, ve sonunda karar ağacı oluşturulacaktır. Yaprağa ulaşıldığında, entropi 0 olmalıdır. Bir başka deyişle, örneklemler homojenliğe kavuşmuş ve düzenli olmalıdır.

Aşağıdaki şekilde örnek bir karar ağacı yapısı yer almaktadır.



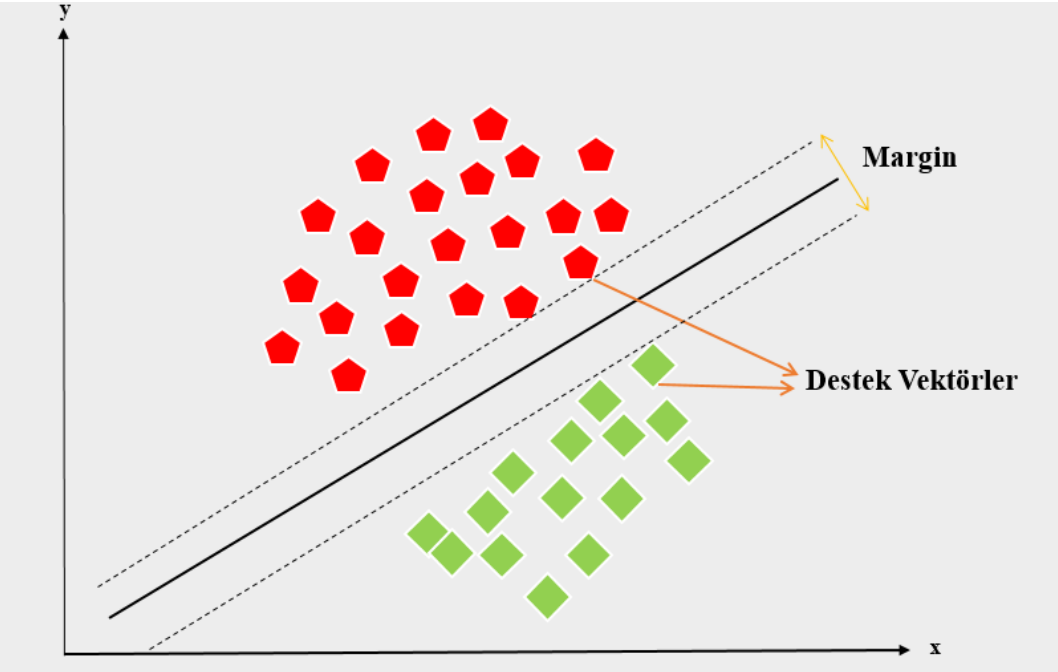
Şekil 2.2 Karar ağaçları örnek gösterim

2.3 Destek Vektör Makineleri

Destek Vaktör makineleri [45, 46] yöntemi çoğunlukla sınıflandırma yapmak için kullanılan makine öğrenmesi modellerindendir. Düzlem yüzeyine yerleştirilmiş noktaları bölmek için bir doğru oluşturur. Bu doğrunun böldüğü sağ ve sol tarafı arasındaki uzaklığa Margin adı verilir. Bu doğrunun, farklı iki sınıfın noktaları için de arasındaki uzaklığın en yüksek seviyede olması hedeflenir. Bir başka deyişle margin genişliği arttıkça iki ve daha fazla sınıf daha iyi farklılaşır. Şekil 2.3'te bir destek vektör gösterimi bulunmaktadır. Şekilde verilen koyu renk çizginin üstünde kalan alan +1, altında kalan alan -1 gibi düşünülebilir.

Şekilde kırmızı altıgeler ve yeşil kareler olarak iki farklı sınıf bulunmaktadır. Amaç gelecek verinin hangi gruba dahil olacağını tahmin edebilmektir. Aşağıda verilen formülden de anlaşılacağı gibi, $\hat{t}$ tahmin edilen değer, a ağırlık vekörünü temsil etmektedir. b girdi vektörünü temsil etmektedir. c değeri ise sapmayı ifade etmektedir. Örnek ile açıklamak gerekir ise sonuç sıfıra eşit veya büyükse ilgili veri kırmızı altıgen grubuna dahil olacaktır. Eğer sonuç sıfırdan küçük çıkar ise ilgili veri yeşil kare grubuna dahil olacaktır.

$$\hat{t} = \begin{pmatrix} 0 \text{ eğer} & a^T.b + c < 0, \\ 1 \text{ eğer} & a^T.b + c \geq 0 \end{pmatrix} \quad (2.6)$$

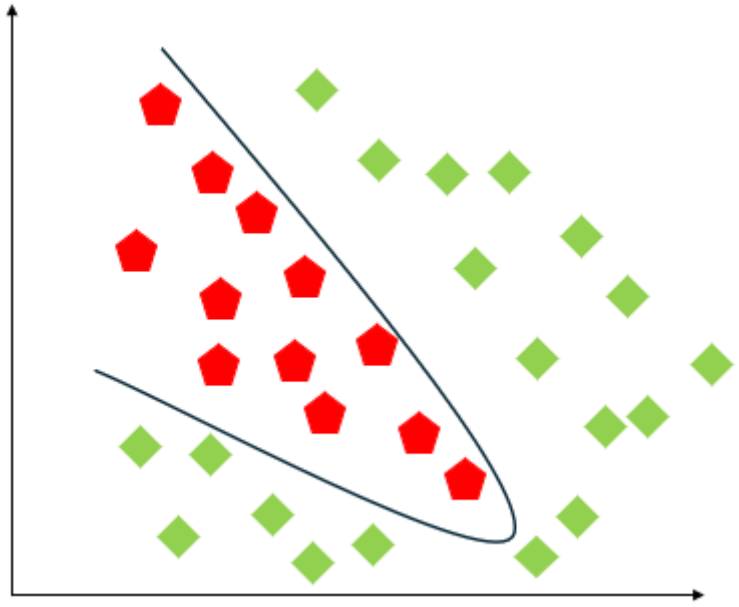


Şekil 2.3 Destek vektör makineleri örnek gösterim

Margin hep bahsedildiği gibi olmayabilir. Bazı durumlarda veriler margin bölgesine girebilmektedir, bu durum Soft Margin olarak isimlendirilir. Eğer veri doğrusal olarak bölünebiliyorsa bu durum da Hard Margin olarak isimlendirilir. Hard Margin sıra dışı verilere karşı çok hassastır. Böyle örneklerle çalışıldığı zaman Soft Margin'i seçmek gerekir.

Soft Margin ve Hard Margin arasındaki denge C hiperparametresi ile kontrol edilebilir. Burada C değeri, Margin genişliğini ifade eder. Modelin veriyi ezberlemesi durumunda da C değeri değiştirilerek, azaltılarak ilerlenebilir.

Daha kompleks verilerde lineer bir doğru ile verileri kümelemek mümkün olmayabilir. Bu durumda ise Kernel adı veriler hiperparametreler kullanılabilir. Kerneller, lineer bir doğru yerine iki boyutta yada üç boyutta, iki ve üzeri dereceden denklemlerle ifade edilebilen ayrım eğrileri ile verileri kümelemeye yarar. Şekil 2.4'te ikinci dereceden bir denklemle ifade edilen, bir başka deyişle bir polinomla ayrıştırılmış örneklemler görülmektedir.



Şekil 2.4 Destek vektör makineleri, ikinci derece polinom ile kümeleme

2.4 Naive Bayes

Naive Bayes yöntemi ismini Matematikçi Thomas Bayes'den almaktadır. Yöntemin temeli olasılık kurallarına dayanmaktadır. Model, bir örnek için tüm olasılık değerlerini hesap eder ve değeri en büyük çıkan sonuca göre diğer örnekleri sınıflandırır. Naive Bayes algoritma ile önem derecesi düşük değişkenler de dikkate alınarak, çok sayıda değişken kullanılarak tahmin yapabilmek mümkündür. Az sayıda veya çok sayıda veri ile başarılı sonuçlar verebilir.

Aşağıda Naive Bayes [47, 48] algoritmasının temelini oluşturan Bayes Teoremi bir değişken için verilmiştir.

$$P(A|B) = \frac{P(B|BA)P(A)}{P(B)} \tag{2.7}$$

$P(A|B)$: B durumunun gerçek olup A'nın gerçekleşme ihtimali;

$P(B|A)$; A durumunun gerçek olup B'nın gerçekleşme ihtimali;

$P(A)$ ve $P(B)$: A ve B olaylarının önsel ihtimalidir.

Üsste paylaşılan formülasyon tekil bir değişken için iken, formül tüm veriye uygulandığında aşağıdaki şekli alır.

$$P(C|F_1, \dots F_n) = \frac{P(C)p(F_1, \dots F_n|C)}{p(F_1, \dots F_n)} \quad (2.8)$$

C verilen hedef değişkeni, F ise diğer değişkenleri (örneklemin özellikleri) temsil etmektedir. Algoritma bütün koşullu olasılık çarpımını sonuç olarak alır.

Çoklu sınıf tahminlemelerinde, öncelikle bir kategorinin/değişkenin gerçek olma olasılığı hesaplanır. Bu örneklem, eğitim veri setinde ne kadar çok rastlanılırsa olasılığı o kadar artar. Ardından, her bir özelliğin hedef değişkenin herhangi bir sınıfına ait olma olasılığı hesaplanır. Daha sonra hesaplanan olasılıklar karşılaştırılır ve en yüksek olasılığa sahip olan sınıf seçilir. Bir başka deyişle, örneklemin hangi sınıfa ait olma olasılığı daha fazla ise, o sınıf için etiketlenir.

Bu yöntemde sıfır frekans ismi verilen durum ile karşılaşılabilir. Bu durum test verisinde yer alan bir değerin, eğitim verisinde gözlenmemesi durumunda olasılık değeri 0 olarak kabul edilip, tahmin yapılmasıdır. Karşılaşılan bu sorunu ortadan kaldırmak amacıyla Laplace yöntemi kullanılabilir. Laplace yönteminde tüm verilere minimum değer eklenir, bu değer genellikle 1 olmaktadır. Bu şekilde sıfır frekans sorunu ihtimali olmamaktadır.

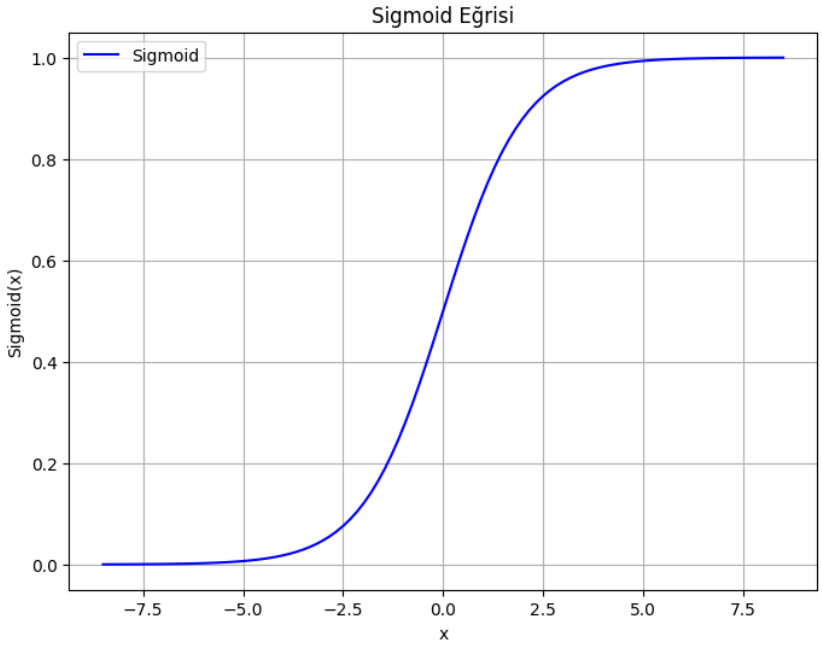
Naive Bayes algoritması Çok Sınıflı Naive Bayes, Bernoulli Naive Bayes ve Gauss Naive Bayes olarak üçe ayrılmaktadır. Algoritma genellikle çoklu duygu analizinde, spam filtrelemede, öneri sistemlerinde vb. kullanılmaktadır.

2.5 Lojistik Regresyon

Lojistik regresyon [49] iki veya daha fazla sınıfa ait sınıflandırma problemlerinde sıkça tercih edilen bir makine öğrenmesi yöntemidir. Lojistik regresyonun üç türü vardır. Binominal, multinominal ve sıralı lojistik regresyon olmak üzere. En basit haliyle, binominal lojistik regresyona, elimizdeki verilerin bir yetişkine mi bir çocuğa mı ait olduğunun tahmin edildiği veya bir görseldeki bitkinin gül mü orkide mi olduğunun tahmin edildiği ikili sınıflandırmaları örnek verebiliriz.

Lojistik Regresyon, sınıflandırma yaparken Sigmoid (Lojistik) fonksiyonunu kullanmaktadır. Sigmoid fonksiyonu S harfine benzeyen bir eğridir. Sigmoid fonksiyonun temel görevi değerleri 0-1 arasına sıkıştırmasıdır. Bu sayede

sınıflandırma yapılabilmektedir. Aşağıda Şekil 2.5 örnek bir Sigmoid fonksiyonu çizimi bulunmaktadır.



Şekil 2.5 Sigmoid eğrisi gösterimi

Aşağıda Lojistik Regresyonu formülüne yer verilmiştir. Formülü kısaca özetlemek gerekirse, p İstenilen durumun gerçekleşme olasılığını temsil etmektedir. β_0, sabit değeri temsil etmektedir. $\beta_{i=1,2,..n}$, her bir bağımsız değikenin$(X_{i=1,2,..n})$ katsayını göstermektedir.

$$log\frac{p}{(1-p)} = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \cdots + \beta_n * X_n \qquad (2.9)$$

Bu formüle göre elde edilen değer nihai tahmin edilen değerdir.

Lojistik regresyon veri setini doğrusal olarak ayırma ihtimali var ise yüksek performans gösterebilmektedir. Lojistik Regresyon, makinelerde parça arızası tahmini yapmak için, sağlık alanında hastalık tahmini yapmak için, finans alanında dolandırcılık, finansal işlemlerin analiz edilmesi, kredi başvuruları ve sigorta uygulamalarında risklerin belirlenmesinde, pazarlama alanında reklamların tıklanıp tıklanılmayacağının tahmin edilmesi gibi pek çok alanda tercih edilmektedir.

2.6 Rastgele Orman

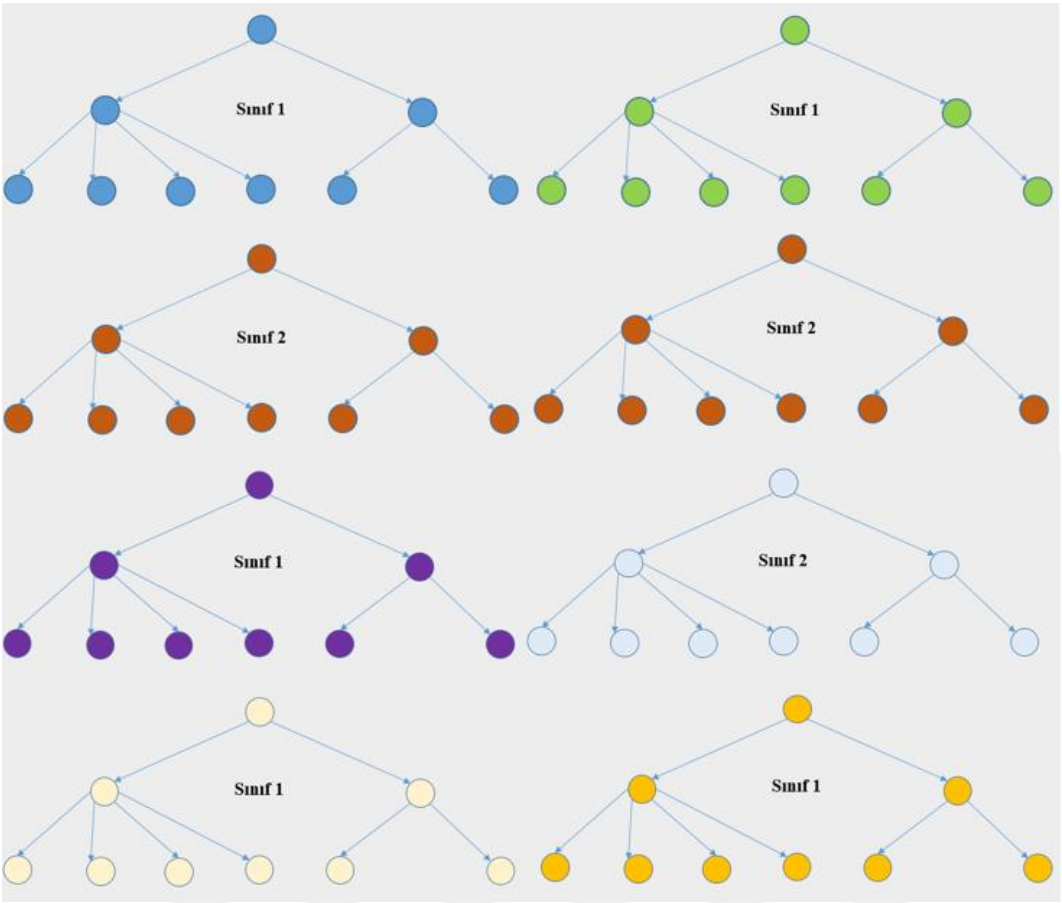
Rastgele Orman [50] Yöntemi hem regresyon hem sınıflandırma problemlerinin çözümünde kullanılan bir algoritmadır. Basit olarak birden fazla karar ağacının yaptığı tahminlerin toplanıp, analiz edilmesine dayanmaktadır. Karar ağaçlarının en büyük sorunlarından biri olan aşırı öğrenmeyi engellemek için rassal olarak çok sayıda farklı alt grup seçilir ve bunlar eğitilir. Bu sayede çok sayıda karar ağacı oluşturulup, her bir karar ağacı bireysel tahminde bulunur.

Veri setinde bulunan tüm değişkenlerden rassal olarak belirli sayıda değişken seçilir. Her bir örneklem için karar ağaçları oluşturulur ve her karar ağacının bir çıktısı olur. Ardından, Çoğunluk Oylama metoduyla sınıflandırma nihayete erdirilir. Buradaki Çoğunluk Oylama tabiri, en çok oyu alan tahminin nihai tahmin olacağını söylemektedir.

Aşağıda Rastgele Orman metodunun çalışma prensibini anlatan bir sözdekod (pseude-code) paylaşılmıştır.

- Veri setinden rassal olarak değişkenleri seç
- Karar ağaçları oluştur
- Yeni bir örneklemi karar ağaçlarında test et
- Çoğunluk oylama ile nihai tahminlemeyi yap

Şekil 2.6 bu durumu gösteren örnek bir rastgele orman modeli gösterimi bulunmaktadır. Modelde farklı alt veri setleri için eğitim gerçekleştği için varyans azalır yani aşırı öğrenmenin önüne geçilmiş olur.



Şekil 2.6 Örnek karar ağacı gösterimi

Rasgtele orman algoritması, bankacılık, finans, sağlık vb. bir çok farklı alanda kullanılmaktadır. Finans alanında dolandırcılık tespiti, bir hissenin gelecekteki durumun tahmin edilmesi, sağlık alanında bir hastanın tıbbi geçmişini analiz etmek amacıyla kullanılabilmektedir.

2.7 GBM

Gradient Boosting Machine (Gradyan Artırma Makineleri) [51] algoritması hem sınıflandırma hem regresyon problemlerinde kullanılmaktadır. Algoritmada, bir adet karar ağacının istenen ölçüde güçlü olmadığı düşünüldüğü için, bir karar ağacı yöntemi yapılır, oluşan hata ile yeni bir karar ağacı yapılır, bu döngü hata en küçük haline gelene kadar sürdürülür.

Algoritmadaki ana düşünce hataların minimize edilmesidir. Algoritma başlangıcında ilk ağaç Rastgele Orman'daki gibi oluşturulur. En büyük farkı ise ilk ağaç oluşturuluktan sonra, onun hatasını minimize edecek ikinci bir ağaç oluşturulmasıdır. Ağaçlar oluşturulur ve ardışık olarak algoritmaya eklenir.

Oluşturulan her ağacın hatasına farklı ağırlık verilerek bir sonraki adıma geçilir. Bu işlem çok defa tekrarlanır.

Modele öncül bir tahmin ile, $F_0(x)$, başlanır. Ardından her bir sonraki adımda, bir önceki adımda yapılan tahmin ile aradaki fark ölçülür. Bu ölçüm için bir hata metriği kullanılır. Burada hata fonksiyonu, yada bir başka deyişle kayıp fonksiyonu olarak minimum kareler hatası (Mean squared error) kullanılabilir.

$$Kayıp(y_i, \gamma) = \sum (y_i - \gamma)^2 \tag{2.10}$$

Formulde y_i gözlemlenen değeri, γ ise tahmini sembolize etmektedir. Tahminleme işlemi iteratif bir şekilde devam eder.

$$F_{tahmin} = F_{tahmin-1} + \varphi\, h_{tahmin} \tag{2.11}$$

h_{tahmin}, karar ağacı gibi relatif olarak zayıf bir öğrenici modelinin yaptığı tahmindir. φ öğrenme oranıdır ve bu zayıf modellerin katkısını kontrol eder. İterasyonlar bu şekilde devam eder ve amaç oluşan hatayı minimize eden bir makine öğrenmesi modeli elde etmektir.

Gradyan arttırma metodunun sözdekodu (pseudocode) aşağıda paylaşılmıştır.

- Karar ağacı eğit ve tahminleme yap
- Kayıp fonksiyonu ile hata hesapla
- Hatayı minimize etmek için karar ağaçları eğitmeye devam et
- Nihai tahmini yap

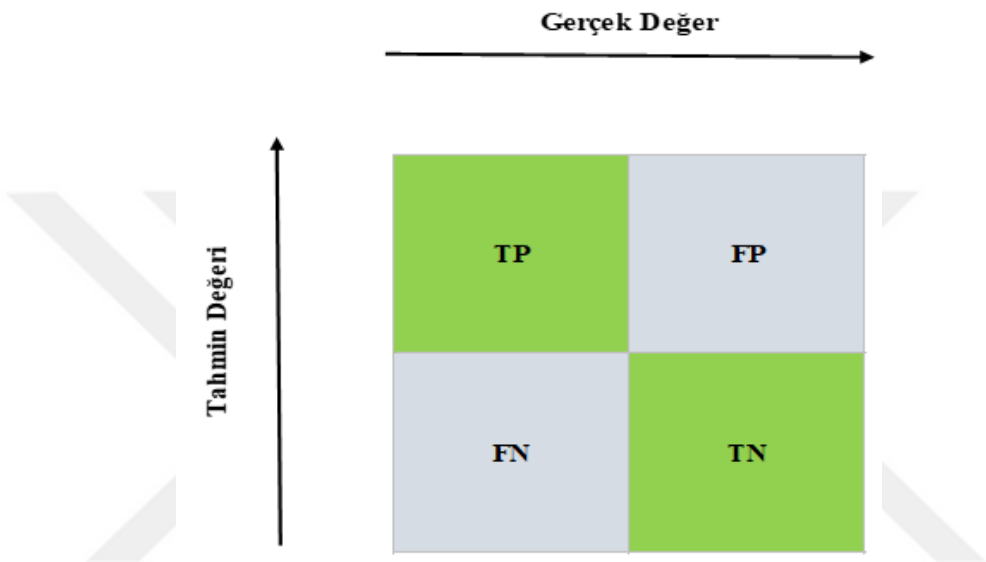
2.8 Topluluk Öğrenmesi

Tahminler, tekil makine öğrenmesi modeli kullanmak yerine birden fazla öğrenici kullanmayı sağlayan topluluk öğrenmesi (ensemble learning) [52] ile de desteklenmiştir. Bu metotta, birçok farklı tekil model bir araya gelir, farklı özellikleri ele alınır ve günün sonunda daha genelleştirilebilir ve daha başarılı bir model hedeflenir. Bu sayede, farklı tekil modellerin farklı yaklaşımları ve arka planındaki matematiksel hesaplamalar dikkate alınmış olur. Tekil öğrenicilerin zayıflıkları da topluluk öğrenmesi ile elimine edilmiş olur. Topluluk öğrenmesinde kullanılan yöntemler bölüm 3.2.2'de detaylı bir şekilde açıklanmıştır.

2.9 Çalışmada Kullanılan Anahtar Performans Göstergeleri

Geliştirilen modellerin başarısını ölçmek için bir dizi anahtar performans göstergesi kullanılmıştır. Kullanılan bu göstergeler [53] aşağıda paylaşılmıştır.

Karmaşıklık matrisi makine öğrenmesinde tahmin sonuçlarının özetini sunan, tahmin hatası ile ilgili bilgi veren, modelin performasını ölçmek için hedef değişkene ait tahminlerin ve gerçek değerlerin karşılaştırıldığı bir yapıdır.



Şekil 2.7 Karmaşıklık matrisi örnek gösterim

Karmaşıklık matrisini anlamak ve analiz etmek için yukarıdaki Şekil 2.7'deki 4 kavram sıklıkla kullanılmaktadır. Bu kavramların açıklaması, aşağıdaki gibi yapılabilir:

- Doğruyu doğru tahmin etmek. (True Positive-TP)
- Yanlışı yanlış tahmin etmek. (True Negative-TN)
- Doğruyu yanlış tahmin etmek. (False Positive-FP)
- Yanlışı doğru tahmin etmek. (False Negative-FN)

Karmaşıklık matrisine dayanılarak yukarıda bulunan tahmin sonuçlarının ve gerçek değerlerinin kıyaslanması ile üretilen Kesinlik, Duyarlılık, Doğruluk, F_1 gibi kriterlerdir. Model performansını değerlendirmek için kullanılmaktadırlar. Bu kriterlerin açıklaması aşağıda formülleri ile kısaca verilmiştir.

Kesinlik: Doğru sınıflandırılan verilerin oranını vermektedir.

$$Kesinlik(Precision) = \frac{TP}{TP + FP} \tag{2.12}$$

Duyarlılık: Sadece doğru değerlerden doğru sınıflandırılanların oranını vermektedir.

$$Duyarlılık(Recall) = \frac{TP}{TP + FN} \tag{2.13}$$

Doğruluk: Doğruluk değerinin kendisidir ve aşağıdaki gibi hesaplanmaktadır.

$$Doğruluk(Accuracy) = \frac{TP + TN}{TP + FP + TN + FN} \tag{2.14}$$

F_1 Skoru: Kesinlik ve Duyarlılık değerlerinin harmonik ortalamasını vermektedir.

$$F_1\ Skoru = 2 * \frac{Kesinlik * Duyarlılık}{Kesinlik + Duyarlılık} \tag{2.15}$$

3
MATERYAL VE METOT

3.1 Veri Setinin İncelenmesi, Temizliği ve Hazırlığı

Sağlıklı bir makine öğrenmesi modelinin temelinde nitelikli, eğer varsa sınıfları mümkün olduğunca net bir şekilde ayrıklaştırılmış ve modelin öğrenmesini güçleştirecek etmenlerden uzak bir veri seti bulunmaktadır. Bununla birlikte, veri seti ne modelin veriyi ezberleyeceği kadar kolay bir trende sahip olmalıdır, ne de modellerin örneklemler arası ilişki kurmasını zorlayacak kadar komplike olmalıdır. Bu hususta veri temizliği ile hazırlığı verinin niteliği, her makine öğrenmesi modelinin kendi içindeki hiperparametrelerinin optimizasyonu da modellerin öğrenim başarısı açısından kritik önem arz etmektedir.

Çalışma başlangıcında 2016 yılında oluşturulan, Türkiye İstatistik Kurumundan alınan Yetişkin Eğitimi Araştırması Mikro Veri Seti ham halde incelenmiştir. Veri seti oluşturulurken, Türkiye Cumhuriyeti sınırları içinde yer alan, nüfusu 100 kişi ve altında olan yerleşim yerleri, yurtlar, hastaneler ve benzeri kurumlarda bulunan nüfus kapsam dışı tutulmuştur. Veri setinde 18 yaş ve üstü bireylerin bilgileri bulunmaktadır.

Veri setinde 23609 bireyin ankete vermiş oldukları yanıtlar mevcuttur. Veride 150 adet kategorik, 4 adet de sürekli değişken gözlemlenmiştir. Veride bireylerin bir ve/veya iki yabancı dil bilgisi de sorgulanmıştır. Çalışmanın hedefi bireylerin yabancı dil bilgisini ölçmek olduğundan, ham veride bulunan ikinci yabancı dilin bilgisi ve düzeyi çıkarılmıştır, zira ikinci yabancı dil bilgisi olan bir birey, çalışmanın ilk isterini halihazırda karşılamaktadır. Ek olarak, hedef değişken ile ilgisi olmadığı açık olan anketin yapıldığı aynı yıl, ankete katılım sırası, birey referans numarası benzeri kategoriler de veri setinden kaldırılmıştır. Veri temizliğinin ardından, 148 değişken ile çalışmaya devam edilmiştir.

Öğrenimi ve modelin doğruluğunu etkileyecek, hatalı sonuçlara sebebiyet vermeye yol açacak boş hücreler ve bulundukları satırlar da temizlenmiştir. Örneğin, veri setinde “Gelir” değişkeninin 9975 anketçi tarafından yanıtlanmadığı gözlemlenmiştir. Literatürde boş hücre varlığında ortalama alma metodunun da tercih edilebildiği gözlemlenmiştir. Lakin, bu metodun ortalamayı ve verinin

trendini değiştirebileceği düşünüldüğünden, bu çalışmada boş hücreler ve ait oldukları satırlar silinmiştir.

Çalışmaya 13611 kişinin verisi ile devam edilmiştir. Yapılan değerlendirmeler neticesinde, kişilerin yabancı dil bilgi seviyeleri Tablo 3.1'de görüldüğü üzere 5 sınıfa ayrılmış olup, bu sınıfların dağılımı da incelenmiştir.

Tablo 3.1 Yabancı dil bilgi düzeyleri sınıf açıklaması, örneklem sayıları ve yüzdelik dilimleri

Sınıf	Sınıf Açıklaması	Örneklem Sayısı	Yüzdelik Dilim (%)
-2	Yabancı dil bilmiyor.	10549	77.50
0	Sadece anlayabiliyorum ve birkaç kelime ve cümle kullanabiliyorum.	899	6.60
1	Dili tanıdık şeyler ve durumlarla ilgili olduğunda kullanıyorum.	1099	8.07
2	Dilin net bir biçimde kullanıldığı durumlarda özünü anlayabiliyor ve basılı metinler oluşturup deneyim ve olayları aktarabiliyorum.	565	4.15
3	Dile neredeyse tamamen hakimim.	499	3.67

Veri setindeki yabancı dil bilgi seviyesi "-2 sınıfı" olan grubun ağırlığı %77.5 olup, "0" seviyesinin ağırlığı %6.6, "1" seviyesinin ağırlığı %8.07, "2" seviyesinin ağırlığı %4.15 ve "3" seviyesinin ağırlığı ise %3.67'dir. Dolayısıyla, veri setinin dengesizliğinden söz edilebilir. Bu durumun makine öğrenmesi modelleri üzerindeki olumsuz etkisini ortadan gidermek için literatürde de sıkça kullanılan SMOTE (Synthetic Minority Oversampling Technique) [54] sentetik veri üretme metodu kullanılmıştır. SMOTE metodu, özellikle azınlıkta kalan sınıfın örnek sayısını arttırma odaklıdır. Azınlık sınıfındaki veriler arasında benzerlikler bulur, bu benzerlikleri kullanarak da yeni veri noktaları oluşturur. Bu durum da modelin daha doğru bir şekilde öğrenmesine yardımcı olur.

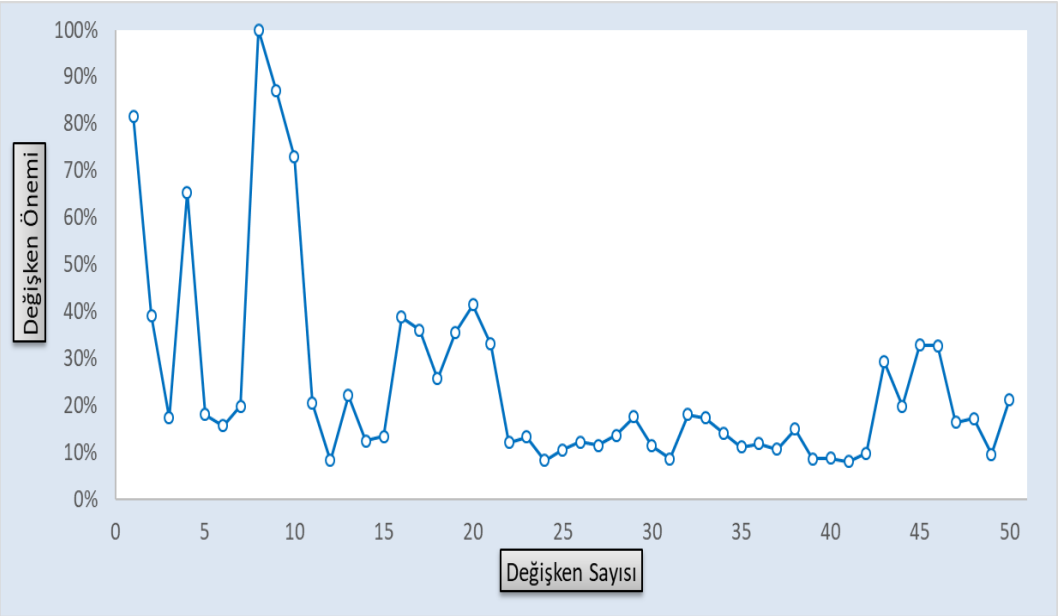
Her bir azınlık sınıfı veri örnekleminin 5 komşusu ile sentetik örnek oluşturmasına karar verilmiş olup, sentetik veri üretime ardından oluşan sınıf dağılımı Tablo 3.2'de paylaşılmıştır.

Tablo 3.2 Sentetik veri üretimi sonrası yabancı dil bilgi düzeyleri sınıf ve örneklem sayıları

Sınıf	Örneklem Sayısı
-2	10549
0	10549
1	10549
2	10549
3	10549

Hedef değişken olan yabancı dil bilgisi düzeyi haricinde veri setinde 147 değişken bulunmaktadır. Her değişkenin hedef değişken üzerinde aynı oranda etkili olamayacağı aşikardır. Bu nedenle, çalışmanın başında en baskın değişkenleri bulma analizi yürütülmüştür. Bu ön analiz için öncül bir rastgele orman modeli kurulup değişken önemi çalışması (variable importance) yürütülmüştür.

Ön analizler neticesinde çıkan sonuçlar, 0-100 aralığında normalize edilmiştir. En yüksek etkiye sahip parametre %100 değerine ulaşmış olup, %8'den daha küçük etkiye sahip değişkenler makine öğrenimini sağlıklı ve verimli bir hale getirebilmek maksadıyla modellere dahil edilmemiştir. Şekil 3.1'de de yüzdesel etki değerlerinin görülebileceği gibi 50 adet değişkenin %8'den daha yüksek bir etkiye sahip olduğu, bir başka deyişle en baskın olduğu tespit edilmiş olup, çalışmada uyarlanan makine öğrenmesi modellerinde kullanılmıştır.



Şekil 3.1 Değişkenin önemi çalışması grafiği

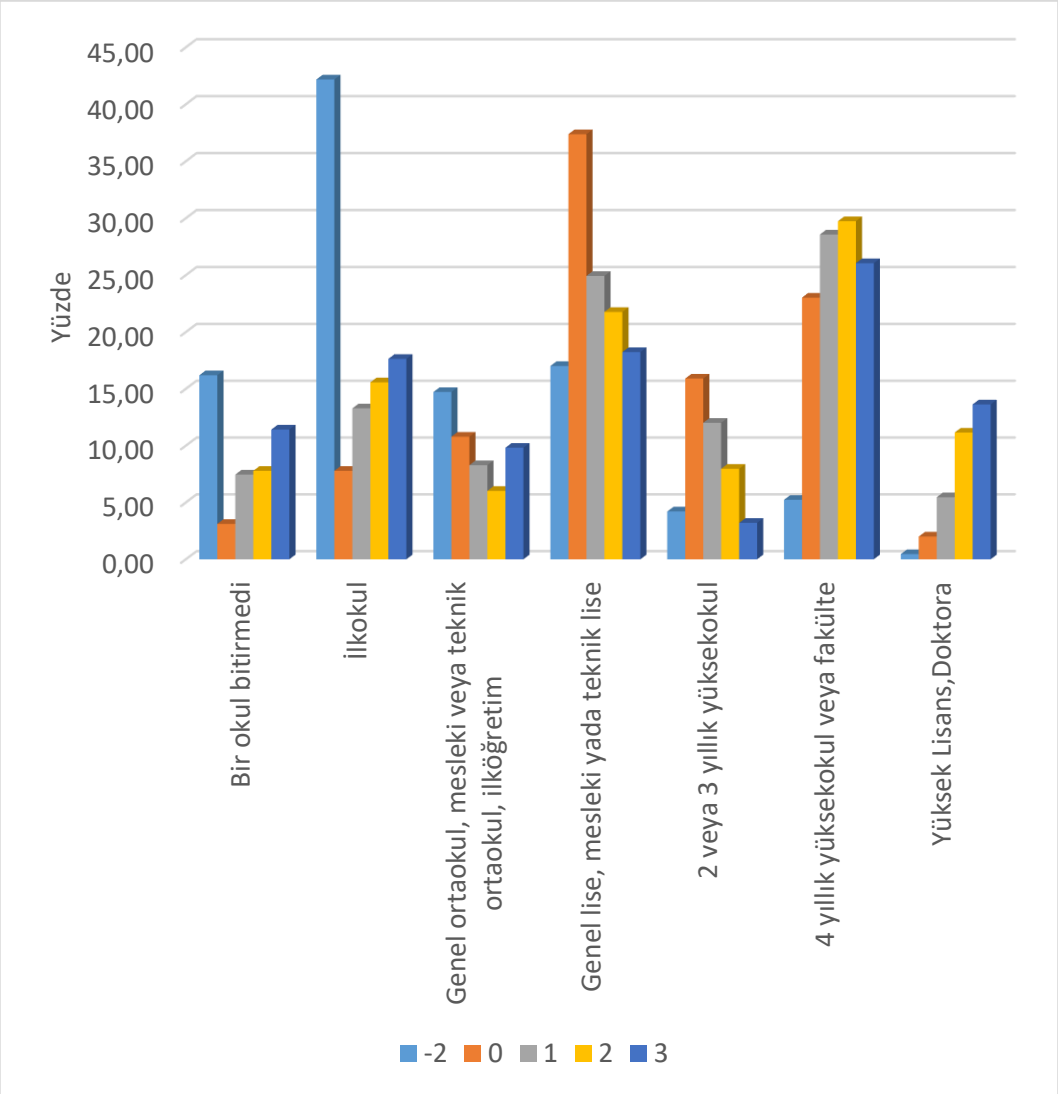
Belirlenen değişkenler arasında en baskın olan ilk 10 değişken Tablo 3.3'te verilmiştir.

Tablo 3.3 En baskın ilk on değişken

Baskınlık Sırası	Değişken
1	Eğitim durumu (Mezun olunan son eğitim kurumu)
2	Mezun olunan bölüm
3	Gelir (Şahsi)
4	Mezuniyet yılı
5	Yaş
6	Babanın eğitim durumu
7	Hane gelir seviyesi
8	Çalışılan meslek kodu
9	Çalışılan faaliyet alanı (NACE2)
10	İşe başlama yılı

Tablo 3.3 dikkatlice incelendiğinde, kişilerin yabancı dil bilgisi ile sosyokültürel ve ekonomik düzeyi ile eğitim seviyesi arasında bir ilişkiden söz etmek mümkün olabilir. Bununla birlikte, kişinin kendisinden bağımsız olarak ailevi durumunun da hedef değişken üzerinde rol oynadığı görülmektedir.

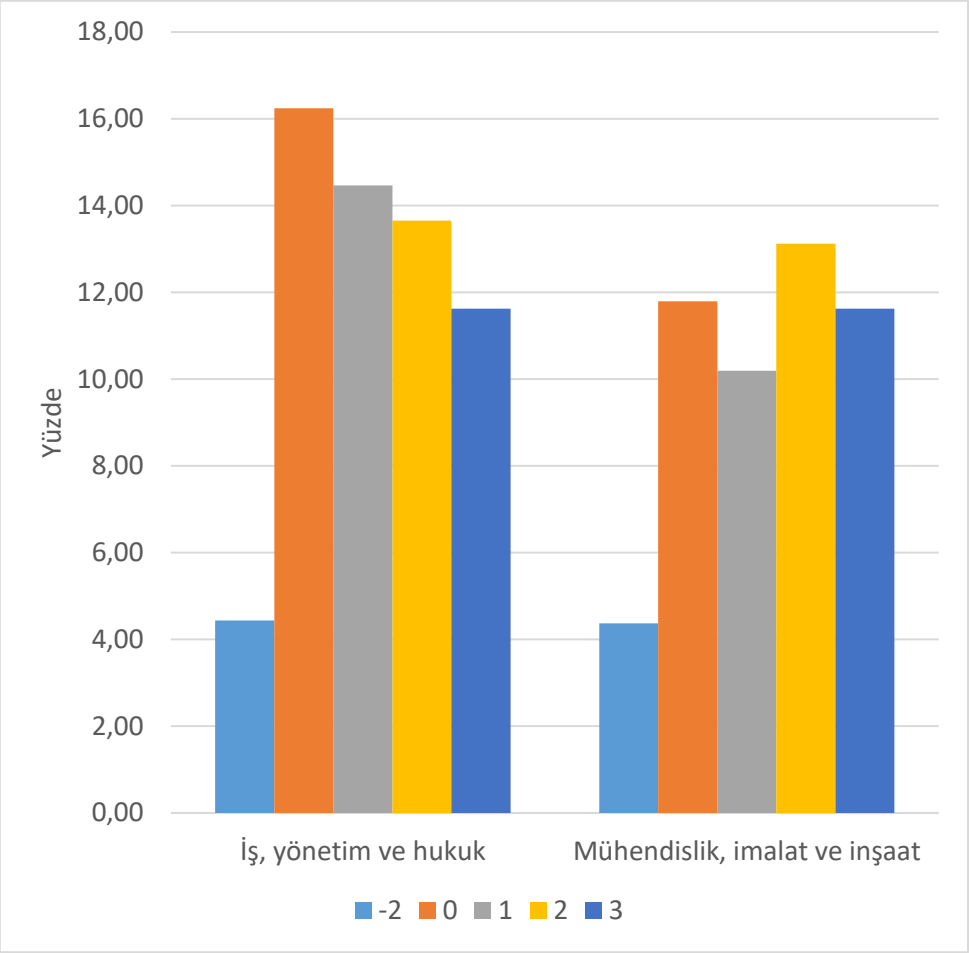
Buna karşın, değişkenler kendi içlerinde incelendiğinde de hedef değişken ile aralarında korelasyonlar yakalanabilir olsa da, bu nihai kararı verme konusunda yanıltıcı olabilir. Bu durum Şekil 3.2'de daha derinlemesine incelenmiştir.



Şekil 3.2 Eğitim düzeyi ile yabancı dil bilgisi ilişkisi

Şekil 3.2, eğitim düzeyi ile yabancı dil bilgisi seviyesini görselleştirmektedir. Örneğin, yabancı dil bilgisi olmayan grup (-2 numaralı sınıf) çoğunluğunun herhangi bir okul bitirmediği veya ilkokul-ortaokul düzeyinde mezuniyeti olduğu görülmektedir. İlaveten, lise, dört yıllık fakülte yada lisansüstü seviyede mezuniyet durumu da yüksek yabancı dil bilgisi ile koreledir. Lakin, dört yıllık bölüm bitirenler arasında da düşük yabancı dil seviyesi olabildiği de bir görülmektedir. Bu ilişkiler bütününe bakılarak, yabancı dil bilgisinin yüksek eğitim mezuniyeti gerektirdiği iddia edilebilir olsa da, herhangi bir okul bitirmeyenler yada ilkokul

seviyesinde mezun olan bireyler arasında da çok iyi düzeyde yabancı dil bilgisi olanlar (3 numaralı sınıf) mevcuttur.



Şekil 3.3 Mezun olunan bölüm ile yabancı dil bilgisi ilişkisi

Örneğin, bir başka değişken olan mezun olunan bölümün etkisi de incelenmiştir. Şekil 3.3'te mühendislik bölümlerinden ve iş-yönetim-hukuk ile ilgili bölümlerden mezun olan bireylerin, yabancı dil düzeyleri grafikleştirilmiştir. -2 numaralı sınıfın içinde olan, yani yabancı dil bilgisi olan bireylerin sayısı iki bölüm arasında neredeyse aynıdır. Benzer şekilde, çok iyi düzeyde yabancı dil bilgisi olan bireylerin sayısı da birbirine çok yakın olarak görülmüştür. Farklılaşma ise yalnızca 0 ve 1 numaralı sınıflarda gözlemlenmektedir.

Özetle, tekil değişkenleri incelemek doğrudan bir çıkarım yapmaya engel olabilmektedir. Böylesine bir tahminleme yapabilmek için çok farklı alandaki verilerin incelenmesi gereklidir. Değişkenler arasında gerçek ilişkileri bütüncül bir

bakış açısı ile ortaya çıkarabilmek, bireylerin demografik ve sosyokültürel verilerinden yola çıkarak yabancı dil seviyelerinin tahminlemesi için bu çalışmada makine öğrenmesi modellerinden faydalanılmıştır.

3.2 Makine Öğrenmesi Modelleri

3.2.1 Tekil Modeller

Veri temizliği ve en baskın parametrelerin belirlenmesinin ardından tekil makine öğrenmesi modellerinin kurulumuna geçilmiştir. K En Yakın Komşu, Karar Ağaçları, Lojistik Regresyon, Destek Vektör Makineleri, Naive Bayes, Rastgele Orman ve Gradient Boosting Machine (GBM) modelleri kurulmuştur. Hedef değişken olan yabancı dil bilgisi düzeyi ayrıklaştırılmıştır. Veri seti tüm modellerde rastgeleliğe önem vererek, %80 eğitim verisi ve %20 test verisi olacak şekilde bölünmüştür.

Modeller kurulmaya başlanmadan önce SMOTE metodu ile azınlıkta kalan sınıfların örneklem sayılarını arttırma maksadıyla sentetik veri üretilmiş ve verideki dengesizlik giderilmiştir. Ardından tüm modellerde, eğitilen bir modelin, modelin öğrenmemiş olduğu bir veri setinde ne derece başarılı olacağını belirlemeye yarayan ve modelin yeteneğini test eden çapraz doğrulama [55] metotları arasından 5 katlı çapraz doğrulama uygulanmıştır. 5 katlı çapraz doğrulamada, eğitim verisi rastgelelik ilkesi göz önünde bulundurularak 5'e bölünür. Her bir kat sırayla test bloğu olarak kullanılırken, diğer katlar eğitim verisi olarak kullanılır. Bu işlem, sırayla her kat için ayrı ayrı yapılır. Bir başka deyişle, her kat en az bir defa doğrulama verisi olarak kullanılır ve modelin başarısı ortaya konulur. Çapraz doğrulama yapılmadığında modellerin başarısı daha yüksek çıkabilir olsa da, bu durum yanıltıcıdır. Dolayısıyla, veri setini çapraz doğrulama gibi metotlarla test ediyor olmak elzemdir.

3.2.1.1 Hiperparametre Optimizasyonları

Makine öğrenmesi modellerinde, modelin en yüksek başarıyı verebilmesi için optimizasyonlara olanak tanıyan bazı parametreler mevcuttur, ve bunlara hiperparametre ismi verilir. Geliştirilmiş olan K En Yakın Komşu, Karar Ağaçları, Lojistik Regresyon, Destek Vektör Makineleri, Naive Bayes, Rastgele Orman ve

GBM modellerinin her birinde sırayla belirlenmiş bazı parametreler optimizasyona tabi tutulmuştur. Belirlenen hiperparametreler arasından öncelikle herhangi biri ile model sonucu iyileştirilmiş, ardından diğer parametreler iyileştirme sürecine dahil edilmiştir. Örneğin, rastgele orman modelinde ağaç sayısı her adımda arttırılırken minimum yaprak ve maksimum derinlik parametreleri sabit tutulmuş, en iyi sonucu veren ağaç sayısı belirlendikten sonra minimum yaprak sayıları test edilmeye başlanmıştır. Bu noktada da en iyi sonuç yakalandıktan sonra maksimum derinlik parametresi sürece dahil edilmiştir.

Burada dikkat edilmesi gereken bir husus, hiperparametrelerin sayısal değerleri arttıkça yada azaldıkça model iyileşmemektedir, lineer bir ilişki söz konusu değildir. Optimizasyon çalışması ihtiyacı da buradan gelmektedir. Özetle, en iyileme mantığı ile model başarısını arttırma hedefi güdülmüştür.

Rastgele orman modelinde ağaç sayısı, minimum yaprak sayısı ve maksimum derinlik hiperparametreler [56] olarak seçilmişken, karar ağaçları modelinde ise minimum yaprak, maksimum derinlik ve minimum bölme optimizasyona tabi tutulacak hiperparametreler [57] olarak belirlenmiştir.

K en yakın komşu modelinde komşu sayısı, örneklem ağırlıkları ve komşuların yakınlığını belirlemede kullanılan "metric" [58] parametresi sürece dahil edilmiştir. Destek vektör makineleri modelinde C değeri, gamma değeri ve kernel [59] kullanılmış olup, lojistik regresyonda ise C değeri ve penaltı (penalty) [60] parametreleri test edilmiştir. Naive Bayes modelinin hiperparametresi süreçte bulunmamıştır.

Bahsedilen minimum yaprak sayısı bir karar ağacı oluşurken yaprağının oluşmasında gerekli olan örneklem sayısını, ağaç sayısı ise temelde örneklemin bir özelliğinin sınıfını ifade etmektedir.

Maksimum derinlik terimi bir karar ağacının dallanma seviyesi için kullanılırken, her bir düğüm veriyi bir özelliğe göre böler, yeni alt gruplar oluşturur. Minimum bölme tanımı ise bir karar ağacının bir düğümde bölünmeden önce en az kaç örneklem barındırması gerektiği anlamına gelmektedir.

Komşu sayısı ise K en yakın komşu modellerinde bir parametre olup, bir veri noktasının sınıflandırılmasında kaç komşu veri noktasının incelenip benzer özellikler bulunması neticesinde sınıflandırılacağını açıklar. Yine aynı modele ait

Ağırlık parametresinde ise uniform ve mesafe değişkenleri test edilmiştir. Uniform parametresi, her bir komşunun tahminlemede eşit ağırlığa sahip olduğunu kabul etmektedir. Bununla birlikte mesafe isimli ağırlık parametresi ise herhangi bir veri noktasına daha yakın olan komşuların tahminlemede daha yüksek ağırlıklara sahip olacağını açıklar.

K en yakın komşu modelinde test edilen son parametre ise uzaklık metrikleridir. Uzaklık metrikleri [61], iki veri noktası arasındaki uzaklığı inceler. Çalışmada kullanılan öklid uzaklığı bir düzlemdeki iki veri noktası arasındaki bir düz çizginin mesafesidir. Manhattan uzaklığı, iki vektörün mutlak farklarının toplamı olarak ifade edilebilir. Chebyshev mesafesi ise, uzaydaki bir boyutta en uzun mesafeyi oluşturan uzaklığı belirtir. Tüm bu mesafe hesabı yaklaşımları, en yakın komşular arasındaki uzaklığı ya da benzerlikleri hesaplamak için kullanılır.

$$\left(\sum_{i=1}^{n}|x_i - y_i|^p\right)^{1/p} \begin{cases} p = 1, Manhattan\ mesafesi \\ p = 2, \text{Öklid}\ mesafesi \\ p = 3, Chebyshev\ mesafesi \end{cases} \quad (3.1)$$

Lojistik regresyon modelindeki hiperparametre olan C parametresi modelin ezberlemeye karşı hassasiyeti arttırmak için kullanılır. C değeri arttıkça ezberleme riski artar. Lojistik regresyondaki bir diğer parametre ise Ceza (Penaltı)'dır. L1 (lasso) penaltı, ortalama kareler hatasına (MSE) eklenen sayıların mutlak değerlerinin toplamını kullanır. L2 (ridge) ise ortalama kareler hatasına (mean square error: MSE) eklenen sayıların karelerinin toplamını ifade eder. Bahsedilen bu ceza parametreleri, temelde regresyon modelinin hatasını indirgemek için kullanılır.

GBM metodunda ise, modelin öğrenme oranı, tahmin edici sayısı ve maksimum derinlik parametreleri [62] değerlendirmeye alınmıştır. Öğrenme oranı modelin ne kadar hızlı yada yavaş öğrendiğini belirleyen bir parametredir. Bu değer, makine öğrenmesi modelinin ne kadar sık güncelleneceğini belirtir. Çok yüksek bir öğrenme oranı, modelin öğrenimini çok hızlandırır ve yeterli öğrenimi sağlayamayabilir; model başarısız tahminler yapar. Çok düşük olması da modelin öğrenimini yavaşlatır ve model eğitim sürecini uzatır, modeli hatalı eğitimlere yönlendirebilir. Dolayısıyla optimum bir öğrenme hızı bulunmalıdır. Tahmin edici ya da bir başka deyişle öğrencisi sayısı ise temelde ağaç sayısına benzerdir. Yüksek

bir öğrenici sayısı modeli karmaşıklaştırır ve eğitim süresini uzatır. Düşük sayıda bir öğrenici kullanıldığında ise model yeteri kadar genelleştirilemeyebilir, ve model başarısı azalabilir.

Vektör destek makinelerindeki C değeri ise sınıf hatalarına karşı hoşgörüyü gösterir. Yine benzer şekilde yüksek bir C değeri, modelin veriyi ezberlemesi ile sonuçlanabilir. Destek vektör makineleri modellerindeki bir diğer parametre olan Kernel ise, veri setinin lineer ayrıştırılamadığı durumlarda VDM'nin esnekliğini arttırır. Son olarak, Gamma değişkeni vektör destek makineleri modelinin non-lineer sınıflandırma yeteneğini arttıran bir parametredir. Gamma değeri arttıkça, model eğitim verisine daha fazla uyum sağlayabilir, bir başka deyişle ezberleyebilir. Bununla birlikte, gamma arttıkça modelin daha karmaşık bir karar sınırı da oluşturabilir. Tam tersi şekilde, düşük gamma değeri ise modelin fazla genelleme yapmasına ve başarısının düşmesine yol açabilir. Dolayısıyla, diğer tüm parametrelerde olduğu gibi burada da bir optimizasyon şarttır.

3.2.2 Topluluk Öğrenmesi

K En Yakın Komşu, Karar Ağaçları, Lojistik Regresyon ve Rastgele Orman olmak üzere seçilen 4 modelin topluluk öğrenmesinde kullanılmak üzere kombinasyonları oluşturulmuş ve Tablo 3.4'te paylaşılmıştır. Buradaki dört model, tekil modeller arasından F_1 skoru en iyi olan üç, ve kötü olan bir model şeklinde belirlenmiştir. Kötü, yada bir başka deyişle başarısız olan modelin seçim kriteri Bölüm 4'te açıklanmaktadır. Yapılan bu seçimlerin amacı, hem başarısız olan bir modelin dahi topluluk öğrenmesi ile kümüle başarısının arttırılıp arttırılamayacağını test etmek, hem de halihazırda tekil iken başarılı sonuçlar veren modellerin daha da iyileştirilip iyileştirilemeyeceğini analiz etmektir.

Tablo 3.4 Topluluk öğrenmesi kombinasyonları

Kombinasyon 1	**Kombinasyon 2**	**Kombinasyon 3**	**Kombinasyon 4**
Rastgele Orman	Rastgele Orman	Karar Ağaçları	Rastgele Orman
Karar Ağaçları	K En Yakın Komşu	K En Yakın Komşu	Karar Ağaçları
Lojistik Regresyon	Lojistik Regresyon	Lojistik Regresyon	K En Yakın Komşu

Tüm bu kombinasyonlar, Tablo 3.5'te verilen topluluk öğrenmesi metotlarına tabi tutulmuştur.

Tablo 3.5 Topluluk öğrenmesi metotları

Çoğunluk oylama
Yumuşak oylama
Yığma metodu
Torbalama
Harmanlama (blending)
Ağırlıklı ortalama
Adaboosting

Çoğunluk oylama [63] metodunda her tekil model ayrı ayrı eğitilir ve final tahminlemeyi yaparken en fazla oyu alan tahminin, topluluğun tahmini olduğu kabul edilir. Örneğin, üç sınıf tahminleme durumunda, üç farklı tekil makine öğrenmesi modelinin en çok/sık tahmin ettiği değişkeni nihai tahmin olarak kabul eder. Bu yöntem yapılan tahminlerin olasılıklarına dair herhangi bir sonuç sunmaz, direkt olarak nihai sınıf tahmini sonucu verir. Dolayısıyla, yapılan tahminlerin doğruluk olasılıklarının da göz önünde bulundurulma ihtiyacı olan durumlarda yumuşak oylama metodu bir tercih sebebi olabilir.

Literatürde yumuşak oylama (soft voting) [64] olarak tanımlanan metotta ise, modellerin tahminlerinin güvenirliği arttıkça o modelin topluluk öğrenmesindeki ağırlığı da artar. Yumuşak oylama farklı tekil modellerin verdiği tahminlerin olasılıklarını da dikkate alarak sınıf tahminleri yapar. Bir başka deyişle, farklı modellerin sınıf tahminlerinin ortalamalarını alarak finalize edilmiş bir tahmin yapar.

Kullanılan diğer metot olan yığma [65] yönteminde ise, önce tekil modeller ayrı ayrı eğitilirken, daha sonra yeni bir birleştirilmiş meta model, kırılan alt veri gruplarıyla çalışan tekil modellerin güçlü yönlerini alarak beslenir. Böylelikle, tekil modellerin akümüle başarısından faydalanılır.

Torbalama [66] yönteminde ise orijinal veri alt kümelerinden türetilen farklı alt veri kümelerinde çalışılır, her bir alt veri kümesi için birer model çalıştırılır. Bu modellerin birbiri arasında herhangi bir etkileşim yoktur. Bu adımın ardından tüm tahminler birleştirilir.

Harmanlama [67] metodunda ise öncelikle veri seti farklı alt veri gruplarına kırılır, her alt veri grubu için bir model kurulur ve eğitilir. Ardından ortaya çıkan tahminler birleştirilir. Birleştirilmiş bu yeni tahminler, yeni bir meta model içinde yeniden eğitilir.

Ağırlıklı ortalama metodu ise her bir modelin önem derecesini dikkate alarak, farklı modellerin farklı derecedeki katkılarını ele alır. Model tahminlerine sıfır ila bir arasında bir ağırlık katsayısı verilebilirken, bu çalışmada daha önce bahsi geçen kombinasyonlar içerisindeki en yüksek F_1 skoruna sahip model en yüksek ağırlığı almıştır. Bu sayede, daha başarılı modellerin topluluk öğrenmesindeki katkısı artarken, görece daha başarısız modellerin etkisi azaltılmıştır. Dolayısıyla, daha sağlam (robust) bir öğreniciden bahsetmek mümkün olmaktadır.

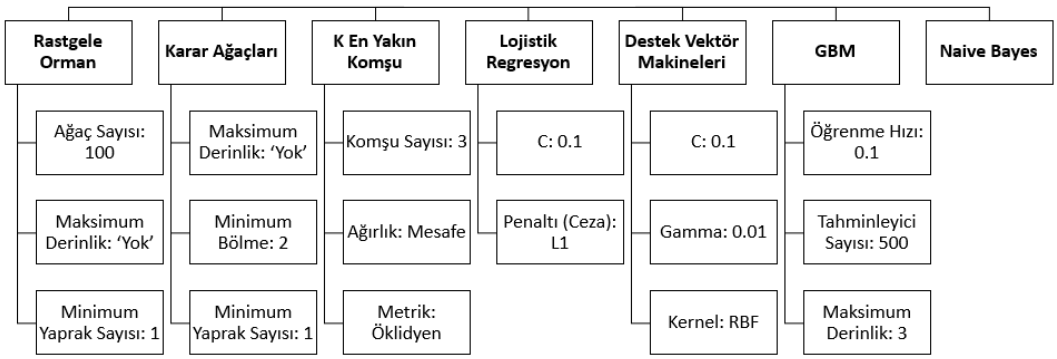
Adaboosting [68] metodu ise yükseltme (boosting) algoritma sınıfına girmektedir. Bu modelin iç yapısında, her bir iterasyondaki son modeldeki hataları düzelten yeni modeller oluşturur. Bir başka deyişle, yanlış tahmin edilen gözlemlere daha fazla ağırlık verir ve bunları her bir sonraki adımda/iterasyonda gidermeye çalışır. Bu süreç, belirtilen maksimum iterasyon sayısına kadar devam eder.

Tablo 3.5'te de paylaşılan topluluk öğrenmesi metodu, Tablo 3.4'teki tekil model kombinasyonları ile çalışıkdıktan sonra, elde edilen F_1 skorları, kesinlik, duyarlılık, doğruluk ve kappa değerleri incelenerek tekil modeller ile karşılaştırılarak nihai ve en başarılı model konfigürasyonu tespit edilmiştir.

4
SONUÇLAR

4.1 Tekil Model Sonuçları

Bir önceki bölümde de bahsedildiği üzere, Yetişkin Eğitimi Araştırması Mikro Veri Seti öncelikle tekil makine öğrenmesi modellerinde kullanılarak bireylerin sosyoekonomik düzeylerinin yabancı dil bilgilerine olan etkisini incelenmiştir. Test edilen tekil makine öğrenmesi modelleri K En Yakın Komşu, Karar Ağaçları, Lojistik Regresyon, Destek Vektör Makineleri, Naive Bayes, Rastgele Orman ve Gradient Boosting Machine(GBM)'dir. Bu modeller, öncelikle varsayılan ayarları ile eğitilmiş olup, ardından parametrik bir çalışma ile hiperparametreleri optimize edilmiştir. Belirlenen parametrelerden öncelikle biri seçilerek optimize edilmiş, en iyi sonucu veren değer yakalandıktan sonra diğer parametreye geçilmiştir. Böylelikle, parametreler arasında rastgelelikle seçim yaparak çalışan optimizasyon kütüphanelerinden ziyade, en iyileme metodu tercih edilmiştir.



Şekil 4.1 Test edilen parametreler ve en iyi sonuçları veren değerleri

Şekil 4.1'de her bir tekil modele ait, bahsedilen en iyileme metoduna tabi tutulan parametreler ve bunların değerleri paylaşılmıştır. Tablo 4.1'de ise, hiperparametreleri optimize edilmiş tekil makine öğrenmesi modellerinin sonuçları, elde edilen F_1 skorları, kesinlik, duyarlılık, doğruluk ve kappa değerleri açısından incelenip karşılaştırılmıştır.

Tablo 4.1 Tekil makine öğrenmesi modellerinin performans değerleri

Model/Hata	Kesinlik	Duyarlılık	Doğruluk	Kappa	F_1
K En Yakın	0,90	0,90	0,90	0,88	0,90
Karar Ağaçları	0,85	0,85	0,85	0,82	0,85
Lojistik Regresyon	0,44	0,46	0,47	0,34	0,42
Destek Vektör	0,81	0,73	0,73	0,66	0,74
Naive Bayes	0,37	0,37	0,38	0,23	0,33
Rastgele Orman	0,94	0,94	0,94	0,93	0,94
GBM	0,78	0,77	0,78	0,72	0,77

Tekil makine öğrenmesi metotlarının anahtar performans göstergelerinin ortalama değerleri incelendiğinde, kesinlik, duyarlık, doğruluk, kappa ve F_1 skorunda rastgele orman modelinin en yüksek başarıya sahip olduğu gözlemlenmiştir. Tablo 4.2 Rastgele Orman modelinin ortalamadaki karmaşıklık matrisini göstermektedir.

Tablo 4.2 Rastgele orman karmaşıklık matrisi

Karar Sınıfı	-2	0	1	2	3
-2	1931	82	64	16	17
0	71	1999	20	13	9
1	86	43	1946	16	17
2	39	10	12	2039	9
3	33	7	10	6	2054

Tablodan da görülebileceği üzere Rastgele Orman modelinin doğru tahmin ettiği sınıf sayısı, yanlış tahminlere göre oldukça fazladır. Model, -2 sınıfına ait örneklemlerden, 1931 adetini doğru tahmin etmiş olup, 82 adetini 0 numaralı sınıf, 64 adetini 1 numaralı sınıf, 16 adetini 2 numaralı sınıf ve 17 adetini de 3 numaralı sınıf olarak tahminlemiştir. 0 sınıfına ait örneklemler arasından 1999 adetini, 1 sınıfına ait örneklemlerden 1946 adetini, 2 sınıfına ait 2039 adetine ve 3 sınıfına ait 2054 örneklemi doğru tahminlemiştir. Oransal olarak doğru tahmin miktarı, yanlış tahminlerin çok üstündedir.

Rastgele Orman metodunun entropi yöntemiyle hesaplamış olduğu en baskın parametreler ise gelir, bitirilen okul, yaş ve mezun olunan yıl olduğu görülmüştür. Tablo 3.3'teki permütasyon bazlı hesaplanan baskın parametrelerle uyumlu olduğu gözlemlenmiştir. Başarılı tahminlemede Rastgele Orman metodunu K en yakın komşu, karar ağaçları ve GBM izlemektedir. Dolayısıyla, kullanılan veri setine en uygun yöntemlerin ağaç bazlı yöntemler olduğunu söylemek mümkündür. Bu başarı, ağaç bazlı modellerin doğrusal olmayan ilişkilere sahip verileri öğrenme konusunda daha başarılı olmaları ile açıklanabilir. Ağaç bazlı modellerdeki düğüm sayısını ifade eden ağaç sayısı arttıkça bu doğrusal olmayan ilişkiler daha da rahat kurulabilir, öğrenici başarısı arttırılabilir. Bununla birlikte, yine ağaç bazlı modellerin çok sayıda özelliğe sahip veri setleri ile başarılı bir şekilde çalışabiliyor olması, nihai başarıyı arttıran etmenlerdendir. En başarısız modellerin ise Naive Bayes ve Lojistik Regresyon olduğu görülmektedir. Naive Bayes modeli değişkenler arasında tam bağımsızlık kabul eder, ve değişkenlerin kendi iç dinamikleri arasındaki korelasyonları yok sayar. Dolayısıyla, birbirini etkileyen yada birbirleri ile uyumlu ve model başarısını arttıracak parametrelerin etkisini dikkate almaz. Sonucun kötü çıkıyor olması, Naive Bayes'in çalışma prensipleri ile açıklanabilir. Bununla birlikte, her ne kadar lojistik regresyonun değişkenler arası bağlantıyı ve ilişkiyi dikkate alabiliyor olsa da doğrusal olmayan, bir başka deyişle, non-lineer ilişkileri modelleyemez, analiz edemez ve non-lineer ilişkiye sahip değişkenleri olan veri setlerinde başarısız olabilir. Bu iki sebep, Naive Bayes ve Lojistik Regresyon modellerinin diğerlerine kıyasla başarısız sonuçlanmasının kaynağı olarak değerlendirilebilir.

4.1.1 Aşırı Uyum Testi Sonuçları

Aşırı uyumu engellemek için kullanılan çapraz doğrulama metoduna ek olarak, en iyi sonucu veren Rastgele Orman makine öğrenmesi modelinde, test veri seti yerine eğitim veri seti kullanılarak tahminleme yapılmıştır.

Tablo 4.3 Eğitim veri seti ile çalıştırılan rastgele orman modelinin anahtar performans gösterge sonuçları

Model/Hata	Kesinlik	Duyarlılık	Doğruluk	Kappa	F_1
Rastgele Orman	0,84	0,84	0,81	0,80	0,84

Tablo 4.3, eğitim veri seti kullanılarak çalıştırılan bu modelin sonuçlarını göstermektedir. Tablo 4.1'de paylaşılan tekil model sonuçlarındaki Rastgele Orman başarısı ile kıyaslandığında, anahtar performans göstergelerindeki düşüş açıktır. Bu durum, modelin aşırı uyum göstermediğinin bir nevi ikincil bir ispatıdır.

4.2 Topluluk Öğrenmesi Modelleri Sonuçları

Ardından, topluluk öğrenmesi metotları uygulanmıştır. Topluluk öğrenmesinde, tek bir öğrenici yerine birden fazla modelin farklı faydalarının kümüle edilerek daha yüksek başarı ile tahminleme hedeflenir.

Bu hususta, topluluk öğrenmesinde uygulanmak üzere Tablo 3.4'te verilen tekil modellerin kombinasyonları seçilmiştir. Bu seçimlerde F_1 skoru açısından en yüksek başarıya sahip ilk üç model, ve düşük F_1 skoruna sahip modellerden biri tercih edilmiştir.

Tablo 4.1'de görülebileceği üzere Naive Bayes ve Lojistik Regresyon modellerinin F_1 skorları özellikle diğer modellere göre düşük kalmıştır. İkisi arasından topluluk öğrenmesinde kullanılmak üzere yapılan tercih ise modellerin çalışma biçimlerine göre yapılmıştır. Naive Bayes modeli, bir veri setinde birbiri ile ilişkili olan parametrelerin bağlantılarını dikkate almayıp her birini ayrı birer örneklem olarak alırken, Lojistik Regresyon setler arası ilişkileri yakalar ve bunları değerlendirir [69]. Veri setinde bilgileri bulunan bireylerin farklı alanlardaki sosyoekonomik değerlerinin yabancı dil düzeyine etkisi birbiriyle bağlantılı olabileceği düşünüldüğünden, topluluk öğrenmesine Naive Bayes yerine Lojistik Regresyon dahil edilmiştir.

Örneğin, veri setinde yer alan Gelir değişkeni ile Hane Geliri değişkenleri arasında korelasyondan söz edilebilir. Veyahut ebeveyn eğitim durumu ile kişinin eğitim durumu arasında da bir ilişki ihtimali mevcuttur.

Bu gibi ilişki, korelasyon ihtimalleri düşünüldüğünde, değişkenler arası ilişkileri gözardı eden Naive Bayes metodu hatalı tahminlerde bulunabileceği sebebiyle topluluk öğrenmesine dahil edilmemiştir. Tablo 4.4'de topluluk öğrenmesi modellerinin F_1 skorlarına yer verilmiştir.

Tablo 4.2 Topluluk öğrenmesi F_1 skorları

	Kombinasyonlar			
	#1	**#2**	**#3**	**#4**
Çoğunluk Oylama	0,90	0,87	0,82	0,94
Yumuşak Oylama	0,86	0,92	0,86	0,92
Yığma	0,956	0,956	0,90	0,956
Torbalama	0,93	0,80	0,32	0,93
Harmanlama	0,92	0,91	0,93	0,94
Ortalama Ağırlık	0,86	0,94	0,86	0,92
Adaboosting	0,95	0,954	0,84	0,95

Yüksek F_1 skorlarının, modelin hem hassasiyet hem de duyarlılık açısından başarısını gösterdiği söylenebilir ve bu iki metriğin harmonik ortalaması olarak hesaplanır. Yapılan çalışmada en yüksek başarının yığma metodu ve 1, 2 ve 4. kombinasyonlar ile geldiği gözlemlenmiştir. Tekil modellerde en yüksek F_1 skoru 0,94 ile Rastgele Orman metoduyla elde edilmişken, topluluk öğrenmesi metotlarında bir çok yöntemde bu değerin üstüne çıkılabilmiştir. Örneğin, yığma metodunun ardından kombinasyonlar arası ortalama alındığında en yüksek başarıyı gösteren yöntemin harmanlama olduğu kabul edilebilir. Ardındından bunu sırasıyla adaboosting, ortalama ağırlık, yumuşak oylama, çoğunluk oylama, ve torbalama izlemektedir. Bazı topluluk öğrenmesi metotları arasında aynı metotla farklı tekil makine öğrenmesi modelleri test edildiğinde, F_1 skorları arasında önemli farklar gözlemlenmiştir.

Örneğin, torbalama metodunda 1, 2 ve 4. kombinasyonlarda F_1 skoru sırasıyla 0,93, 0,80 ve 0,93 iken 3 numaralı kombinasyonda, yani Karar Ağaçları, K En Yakın Komşu ve Lojistik Regresyon kombinasyonunda F_1 skoru 0,32 seviyesine düşmüştür ve tekil modellerin bile altında kalmıştır. Bu durum, lojistik regresyonun düşük varyanslı bir model ve torbalama metodunun da yüksek varyanslı modellerle daha iyi çalışıyor olması ile açıklanabilir [70].

Tablo 4.5 Topluluk öğrenmesi doğruluk değerleri

	Kombinasyonlar			
	#1	**#2**	**#3**	**#4**
Çoğunluk Oylama	0,89	0,87	0,82	0,94
Yumuşak Oylama	0,86	0,92	0,88	0,92
Yığma	0,95	0,95	0,90	0,95
Torbalama	0,94	0,81	0,35	0,94
Harmanlama	0,93	0,91	0,93	0,94
Ortalama Ağırlık	0,86	0,85	0,86	0,92
Adaboosting	0,95	0,95	0,86	0,95

Tablo 4.5'te, kullanılan metotların doğruluk değerleri paylaşılmıştır. Sınıflandırma problemlerinde doğruluk değeri, doğru olarak sınıflandırılan örneklerin toplam örnek sayısına oranıdır. Bir başka deyişle, herhangi bir modelin doğru tahmin yüzdesini belirtir. Tablo incelendiğinde yine yığma metodundaki başarı, ve benzer şekilde 3. kombinasyonu kullanan torbalama metodundaki başarısızlık dikkat çekmektedir.

Ortalama doğruluk değerleri üzerinden bir karşılaştırma yapmak gerekirse en yüksek başarıyı gösteren yığma metodunun arından harmanlama, adaboosting, yumuşak oylama, çoğunluk oylama, ortalama ağırlık ve torbalama metodu gelmektedir. F_1 skorları ile aynı sıraya sahiptir. Dolayısıyla, özellikle F_1 skoru ve doğruluk değerleri dikkate alındığında veri setine en uygun metodun ise yığma metodu olduğu dikkat çekmektedir.

Topluluk öğrenmesi metotlarının avantajları, F_1 skorunda olduğu gibi burada da ortaya çıkmıştır. Tekil modellerde en yüksek doğruluk değeri 0,94 olarak gözlenmişken, yığma metodu sayesinde bu değer az da olsa 0,95 seviyesine arttırılabilmiştir.

Tablo 4.3 Topluluk öğrenmesi kappa değerleri

	Kombinasyonlar			
	#1	**#2**	**#3**	**#4**
Çoğunluk Oylama	0,87	0,83	0,78	0,92
Yumuşak Oylama	0,82	0,90	0,85	0,90
Yığma	0,94	0,94	0,88	0,94
Torbalama	0,92	0,76	0,18	0,92
Harmanlama	0,91	0,89	0,92	0,93
Ortalama Ağırlık	0,82	0,81	0,82	0,90
Adaboosting	0,94	0,94	0,83	0,94

Tablo 4.6'te paylaşılan Kappa değerlerine bakıldığında, torbalama metodu ve üçüncü kombinasyonun yine başarısız tahminler yaptığı görülmektedir. Kappa değerleri de, F_1 skorları ve doğruluk değerlerine paralel sonuçlar vermiştir.

Tablo 4.7 Topluluk öğrenmesi kesinlik ve duyarlılık değerleri

	Kesinlik				**Duyarlılık**			
	Kombinasyonlar				**Kombinasyonlar**			
	#1	**#2**	**#3**	**#4**	**#1**	**#2**	**#3**	**#4**
Çoğunluk Oylama	0,86	0,88	0,84	0,94	0,90	0,87	0,84	0,94
Yumuşak Oylama	0,86	0,92	0,88	0,92	0,84	0,92	0,88	0,92
Yığma	0,95	0,95	0,92	0,95	0,96	0,96	0,91	0,96
Torbalama	0,93	0,81	0,33	0,93	0,94	0,81	0,35	0,94
Harmanlama	0,92	0,91	0,93	0,94	0,93	0,91	0,93	0,94
Ortalama Ağırlık	0,86	0,94	0,86	0,93	0,86	0,94	0,86	0,93
Adaboosting	0,95	0,95	0,86	0,95	0,95	0,95	0,86	0,95

Tablo 4.7'te ise kesinlik ve duyarlılık performans göstergeleri paylaşılmıştır. Tekil modellerde hem kesinlik hem de duyarlılık için en yüksek değerler rastgele orman

metoduyla sırasıyla 0,94 ve 0,94 olarak gözlemlenmişken, topluluk öğrenmesinde tekrardan yığma metodunda en yüksek başarı elde edilmiştir. Yine daha önce paylaşılan anahtar performans göstergelerine benzer bir trend gözlemlenmiştir.

Tablo 4.8 Topluluk öğrenmesi 2. kombinasyon ile yığma metodunun karmaşıklık matrisi

Karar Sınıfı	**-2**	**0**	**1**	**2**	**3**
-2	1946	63	72	29	21
0	55	2063	17	3	3
1	74	12	1996	6	4
2	20	4	6	2060	7
3	28	2	8	12	2038

Topluluk öğrenmesinin anahtar performans göstergelerine bakılarak en başarılı sonuçlardan birini verdiği tespit edilen 2. Kombinasyon ile Yığma Metodunun karmaşıklık matrisi sonuçları Tablo 4.8'de paylaşılmıştır. Tablo, 5 katlı çapraz doğrulama metodunun ortalaması olarak paylaşılmıştır. Her kat 10549 örnekleme bölündüğünden, tabloda paylaşılan toplam örneklem sayısı da virgülden sonraki hane yuvarlamaları ile birlikte 10549 çıkmaktadır.

Tablo incelendiğinde bu yöntemin -2 sınıfının 1946 adet örneklemini doğru tahminlediği görülmektedir. Buna ek olarak, 0 sınıfının 2063, 1 sınıfının 1996, 2 sınıfının 2060 ve 3 sınıfının da 2038 adet örneklemi doğru tahminlenmiştir. Özellikle en başarılı tekil öğrenme sonucunu veren Rastgele Orman modelinin karmaşıklık matrisi ile karşılaştırıldığında, doğru tahminlenen örneklem sayısının arttığı, hatalı yapılan tahminlerin ise azaldığı göze çarpmaktadır. Sınıf tahminlerindeki en yüksek iyileşme ise, 1 numaralı sınıfın 0 numaralı sınıfa ait olduğuna dair yapılan hatalı tahminlerde olmuştur. Rastgele Orman tekil modelinde, bu sınıflar arası yapılan tahmin 43 iken, topluluk öğrenmesi sayesinde 12'ye düşmüştür. En az hatalı ve en doğru tahmin edilen sınıfın ise bu metotta 37 hatalı tahmin ile 2 numaralı sınıf, tekil Rastgele Orman modelinde ise 56 hatalı tahmin ile 3 numaralı sınıf olduğu görülmüştür. Her ne kadar, yer yer Rastgele Orman modelinin yaptığı hatalı tahminler Yığma metodu kullanan topluluk öğrenmesine göre daha az olsa da, karmaşıklık matrislerine, anahtar performans

göstergelerine ve de ortalamaya bakıldığında 2. kombinasyonu kullanan Yığma metodu daha başarılı görünmektedir.

Tüm bu bilgilerin ışığında rahatlıkla söylenebilir ki, tekil modellerin başarısını arttırma hedefi güden topluluk öğrenmesi metotları makine öğrenmesi modellerinin öğrenme ve tahminleme başarısını arttırmaktadır.

4.2.1 Topluluk Öğrenmesi Modelinde Aşırı Uyum Testi Sonuçları

Tıpkı tekil makine öğrenmesi modellerinde olduğu gibi, en iyi sonucu veren topluluk öğrenmesi modelinde de aşırı uyum ihtimalini test etmek maksadıyla model test veri seti yerine eğitim veri seti ile tekrar çalıştırılmıştır. Tablo 4.9 bu durumdaki sonuçları paylaşmaktadır.

Tablo 4.9 En başarılı sonucu veren topluluk öğrenmesi modelinin eğitim verisi ile anahtar performans gösterge sonuçları

Model/Hata	Kesinlik	Duyarlılık	Doğruluk	Kappa	F_1
Yığma Metodu & 2. Kombinasyon	0,81	0,76	0,76	0,71	0,7

Anahtar performans göstergeleri incelendiğinde, Tablo 4.4-4.7'de paylaşılan sonuçlarla karşılaştırıldığında, kesinlik, duyarlılık, doğruluk, kappa ve F_1 skorlarındaki düşüş açıkça görülebilmektedir. Bu durum ise, modelin aşırı uyum göstermediğini, bir başka deyişle modelin verileri ezberlemediğini göstermektedir.

KAYNAKÇA

[1] TÜİK, Yetişkin Eğitimi Araştırması Mikro Veri Seti, 2016.

[2] L., Tilahun, B. Sekeroglu, “An intelligent and personalized course advising model for higher educational institutes”. Discover applied sciences, vol. 2, pp 1635, 2020.

[3] L. Chen, P. Chen and Z. Lin, "Artificial intelligence in education: A Review," IEEE access, vol. 8, pp. 75264-75278, 2020.

[4] Y. Kim, T. Soyata, and R. F. Behnagh, ‘‘Towards emotionally aware AI smart classroom: Current issues and directions for engineering and education,’’ IEEE access, vol. 6, pp. 5308–5331, 2018.

[5] M.Tsiamaki, G. Kostopoulos, S. Kotsiantis, and O. Ragos, “Implementing autoML in educational data mining for prediction tasks, ’’ Applied sciences, vol.10, pp. 90, 2020.

[6] Y. Hu, C. Lo, and S. Shih, “Developing early warning systems to predict students’ online learning performance,” Computers in human behavior, vol. 36, pp. 469-478, 2014.

[7] R. Trakunphutthirak, V. C. S. Lee, “Application of educational data mining approach for student academic performance prediction using progressive temporal data,” Journal of educational computing research, vol.60, pp. 742-776, 2022.

[8] H. Lakkaraju, E.Aguiar, C. Shan, D. Miller, N. Bhanpuri, R. Ghani, and K.L. Addison, “A machine learning framework to identify students at risk of adverse academic outcomes,” in Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, ACM, 2015, pp. 1909–1918.

[9] H. Al-Shehri, A. Al-Qarni, L. Al-Saati, A. Batoaq, H. Badukhen, S. Alrashed, J. Alhiyafi, and S. O. Olatunji, ”Student performance prediction using support vector machine and k-nearest neighbor,” in Proceedings of the IEEE 30th

canadian conference on electrical and computer engineering, CCECE, 2017, pp. 1-4.

[10] N. Tomasevic, N. Gvozdenovic, and S. Vranes, “An overview and comparison of supervised data mining techniques for student exam performance prediction,” Computers & education, vol. 143, pp. 103676, 2020.

[11] S. Şahin, “Makine öğrenmesi yöntemleri ile ortaokul öğrenci başarılarının tespiti ve bir uygulama,” Yüksek Lisans Tezi, İstanbul Üniversitesi, Fen Bilimleri Enstitüsü, İstanbul, 2021.

[12] M. Mohammadi, M. Dawodi, W. Tomohisa, and N. Ahmadi, “Comparative study of supervised learning algorithms for student performance prediction,” in Proceedings of the international conference on artificial intelligence in information and communication, ICAIIC, 2019, pp. 124-127.

[13] N. Poh, I. Smythe, “To what extend can we predict students’ performance? A case study in colleges in South Africa,” in Proceedings of IEEE symposium on computational intelligence and data mining, CIDM, 2014, pp. 416-421.

[14] R. R. Halde, “Application of machine learning algorithms for betterment in education system,” in Proceedings of the international conference on automatic control and dynamic optimization techniques, ICACDOT, 2016, pp. 1110-1114.

[15] M. V. Martins, D. Tolledo, J. Machado, L. M. T. Baptista, and V. Realinho, “Early prediction of student's performance in higher education: A case study,” in Proceedings of the Trends and applications in information systems and technologies, AISC, 2021, pp. 166-175.

[16] M. Gök, “Makine öğrenmesi yöntemleri ile akademik başarının tahmin edilmesi,” Gazi Üniversitesi fen bilimleri dergisi part C: tasarım ve teknoloji, vol. 5, pp. 139-148, 2017.

[17] S. T. Canibey, O. Sevli, “Bireylerin gelir dağılım seviyelerinin makine öğrenmesi teknikleri ile belirlenmesi,” Osmaniye Korkut Ata Üniversitesi fen bilimleri enstitüsü dergisi, vol. 5, pp. 753 – 766, 2022.

[18] V.L. Miguéis, A. Freitas, P. J. V. Garcia, and A. Silva, “Early segmentation of students according to their academic performance: A predictive modelling approach,” Decision support systems, vol. 115, pp. 36-51, 2018.

[19] A. M. Shahiri, W. Husain, and N. A. Rashid, “A review on predicting student's performance using data mining techniques,” Procedia computer science, vol. 72, pp. 414-422, 2015.

[20] B. Albreiki, N. Zaki, and H. Alashwal, “A systematic literature review of student’ performance prediction using machine learning techniques,” Education sciences, vol. 11, pp. 552, 2021.

[21] H. Pallathadka, A. Wenda, E. Ramirez-Asís, M. Asís-López, J. Flores-Albornoz, and K. Phasinam, “Classification and prediction of student performance data using various machine learning algorithms,” in Proceedings of the materials today, 2023, pp. 3782-3785.

[22] L. Yan, Y. Liu, ”An ensemble prediction model for potential student recommendation using machine learning,” Symmetry 12, vol. 5, pp. 728, 2020.

[23] O. W. Adejo, T. Connolly, “Predicting student academic performance using multi-model heterogeneous ensemble approach,” Journal of applied research in higher education, vol. 10, pp. 61-75, 2018.

[24] J. Yang, S. DeVore, D. Hewagallage, P. Miller, Q. X. Ryan, and J. Stewarthys, “Using machine learning to identify the most at-risk students in physics classes,” Physical review physics education research, vol. 16, pp. 020130, 2020.

[25] J. Figueroa-Cañas, T. Sancho-Vinuesa, “Early prediction of dropout and final exam performance in an online statistics course,” in IEEE revista iberoamericana de tecnologias del aprendizaje, vol. 15, pp. 86-94, 2020.

[26] B. K. Yousafzai, S. A. Khan, T. Rahman, I. Khan, I. Ullah, A. U. Rehman, M. Baz, H. Hamam, and O. Cheikhrouhou. “Student-performulator: student academic performance using hybrid deep neural network,“ Sustainability, vol. 13, pp. 9775, 2021.

[27] F. Aman, A. Rauf, R. Ali, F. Iqbal, and A. M. Khattak, “A predictive model for predicting students scademic performance,” in Proceedings of the 10th international conference on information, IISA, 2019, pp. 1-4.

[28] B. Şekeroğlu, K. Dimililer, and K. Tuncal, “Student performance prediction and classification using machine learning algorithms,” in Proceedings of the 8th international conference on educational and information technology, ICEIT, 2019, pp. 7-11.

[29] A. Kukkar, R. Mohana, A. Sharma, and A. Nayyar, “Prediction of student academic performance based on their emotional wellbeing and interaction on various e-learning platforms,” Education and information technologies, vol. 28, pp. 9655–9684, 2023.

[30] R. R. Halde, A. Deshpande, and A. Mahajan, “Psychology assisted prediction of academic performance using machine learning,” in Proceedings of the IEEE international conference on recent trends in electronics, RTEICT, 2016, pp. 431-435.

[31] M. Tadese, A. Yeshaneh, and G. B. Mulu, “Determinants of good academic performance among university students in Ethiopia: A cross-sectional study,” BMC medical education, vol. 22, pp. 395, 2022.

[32] M. M. Z. Eddin, N. A. Khodeir, and H. A. Elnemr, “A comparative study of educational data mining techniques for skill-based predicting student performance,” International journal of computer science and information security, vol. 16, pp. 56–62, 2018.

[33] K. Okoye, A. Arrona-Palacios, C. Camacho-Zuñiga, J. A. G. Achem, J. Escamilla, and S. Hosseini, “Towards teaching analytics: A contextual model for analysis of students’ evaluation of teaching through text mining and machine learning classification,” Education and information technologies, vol. 27, pp. 3891–3933, 2022.

[34] H. Behravan, J. M. Hartikainen, M. Tengström, V. M. Kosma, and A. Mannermaa, “Predicting breast cancer risk using interacting genetic and demographic factors and machine learning,” Scientific reports, vol. 10, pp. 11044, 2020.

[35] D.Taliaz, A. Spinrad, R. Barzilay, Z. Barnett-Itzhaki, D. Averbuch, O. Teltsh, R. Schurr, S. Darki-Morag, and B. Lerer, “Optimizing prediction of response to antidepressant medications using machine learning and integrated genetic, clinical, and demographic data,” Transl psychiatry, vol. 11, pp. 381, 2021.

[36] M. Grassi, N. Rouleaux, D. Caldirola, D. Loewenstein, K. Schruers, G.Perna M. Dumontier, “A novel ensemble-based machine learning algorithm to predict the conversion from mild cognitive impairment to alzheimer's disease using socio-demographic characteristics, clinical information, and neuropsychological measures,” Journal of frontiers in neurology, vol. 10, pp. 756, 2019.

[37] C. Sweeney, E. Ennis, M. Mulvenna, R. Bond, and S. O’Neill, “How machine learning classification accuracy changes in a happiness dataset with different demographic groups,” Computers, vol. 11, pp. 83, 2022.

[38] W. Luo, T. Nguyen, M. Nichols, T. Tran, S. Rana, S. Gupta, D. Phung, S. Venkatesh, and S. Allender, “Is demography destiny? Application of machine learning techniques to accurately predict population health outcomes from a minimal demographic dataset,” Plos one, vol. 10, pp. 0125602, 2015.

[39] N. T. Gildas, N. Paul, F. D. Engwali, “Do demographic structure conditions sector contribution to economic growth? A machine learning approach,” Environment, development and sustainability, 2023.

[40] L. F. Beiser-McGrath, R. A. Huber, “Assessing the relative importance of psychological and demographic factors for predicting climate and environmental attitudes,” Climatic change, vol. 149, pp. 335–347, 2018.

[41] M. S. Hajirahimova, A. S. Aliyeva, “Development of a prediction model on demographic indicators based on machine learning methods: Azerbaijan example,” International journal of education and management engineering, vol. 13, No.2, pp. 1-9, 2023.

[42] N. S.Altman, “An introduction to kernel and nearest-neighbor nonparametric regression,” The American statistician, vol. 46, pp. 175–185, 1992.

[43] Y. Freund, R. E. Schapire, “A decision-theoretic generalization of on-line learning and an application to boosting,” Journal of computer and system sciences, vol. 55, pp. 119-139, 1997.

[44] D. A. Pisner, D. M. Schnyer, Machine learning: Methods and applications to brain disorders, Chapter 6 - Support vector machine. Academic press, pp. 101-121, 2020.

[45] M. M. Sarıtaş, A. Yaşar, “Performance analysis of ann and naive bayes classification algorithm for data classification,” International journal of intelligent systems and applications in engineering, vol. 7, pp. 88–91, 2019.

[46] J.S Cramer, “The Origins of logistic regression”, Timbergen institute discussion paper, 2002.

[47] L, Breiman. Random Forests. Machine Learning vol 45, pp 5–32, 2001.

[48] J. H., Friedman, “Greedy function approximation: A gradient boosting machine”, Annals of statistics, vol 29, pp 1189-1232, 1999.

[49] B. Akyüz, “Topluluk öğrenimi ile kredi kartı dolandırıcılık tespiti: Karşılaştırmalı bir analiz,” Yüksek Lisans Tezi, Yıldız Teknik Üniversitesi, Fen Bilimleri Enstitüsü, İstanbul, 2022.

[50] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic minority over-sampling technique,” Journal of artificial intelligence research, vol. 16, pp. 321-357, 2022.

[51] R. Kohavi, “A study of cross-validation and bootstrap for accuracy estimation and model selection,” Proceedings of the fourteenth international joint conference on artificial intelligence, IJCA, 1995, pp. 1137-1143.

[52] P. Probst, M. N. Wright, and A. L. Boulesteix, “Hyperparameters and tuning strategies for random forest,” WIREs data mining knowledge discovery, vol. 9, pp. 1301, 2019.

[53] R. G. Mantovani, T. Horváth, R. Cerri, J. Vanschoren, and A. C. P. L. F. D. Carvalho, “Hyper-parameter tuning of a decision tree induction algorithm,” in Proceeding of the 5th Brazilian conference on intelligent systems, BRACIS, 2016, pp. 37-42.

[54] R. Wazirali, “An improved intrusion detection system based on knn hyperparameter tuning and cross-validation,” Arabian journal of science and engineering, vol. 45, pp. 10859–10873, 2020.

[55] J. Wainer, P. Fonseca, “How to tune the RBF SVM hyperparameters? An empirical evaluation of 18 search algorithms,” Artificial Intelligence Review, vol. 54, pp. 4771–4797, 2021.

[56] C. Foo, C. B. Do, and A. Y. Ng, “A majorization-minimization algorithm for (multiple) hyperparameter learning,” in Proceedings of the 26th annual international conference on machine learning, ICML, 2009, pp. 321–328.

[57] R. Suwanda, Z. Syahputra, and E. M. Zamzami, “Analysis of euclidean distance and manhattan distance in the k-means algorithm for variations number of centroid,” in Proceedings of the 4th international conference on computing and applied informatics, ICCAI, 2019.

[58] K. Kiatkarun, P. Phunchongharn, “Automatic hyper-parameter tuning for gradient boosting machine” in Proceedings of the 1st international conference on big data analytics and practices, IBDAP, 2020, pp. 1-6.

[59] Z. H. Zhou, Machine Learning. Springer, 2021.

[60] R. Kohavi, C. Kunz, “Option decision trees with majority votes,” in Proceedings of the fourteenth international conference on machine learning, 1998, pp. 161-169.

[61] A. Taha, “Intelligent ensemble learning approach for phishing website detection based on weighted soft voting,” Mathematics, vol. 9, pp. 2799, 2021.

[62] D. H. Wolpert, “Stacked generalization,” Neural networks, vol. 5, pp. 241-259, 1992.

[63] L., Breiman. “Bagging predictors”, Machine Learning, vol 24, pp 123–140, 1996.

[64] M. V. Arisoy, T. T. Abubakar, “A comparatıve analysis of ensemble learning methods on social media account detection”, Journal of Engineering Technology and Applied Sciences, vol 8, pp 87-105, 2023

[65] Y. Freund; R. E. Schapire, “A desicion-theoretic generalization of on-line learning and an application to boosting”, Journal of computer and system sciences, vol 55, pp. 119-139, 1997.

[66] A. Prabhat and V. Khullar, "Sentiment classification on big data using Naïve bayes and logistic regression," in Proceedings of the international conference on computer communication and informatics, ICCCI, 2017, pp. 1-5.

[67] P. Derbeko, R. El-Yaniv, R. Meir, “Variance optimized bagging”. in Proceedings of the 13th European conference on machine learning, 2002, pp 60-72.

TEZDEN ÜRETİLMİŞ YAYINLAR

Konferans Bildirileri

1. G. G. Duman Oral, S. Soner Kara, “Makine öğrenmesi ile yabancı dil düzeyi tahmini”, 12. Uluslararası mühendislik, mimarlık ve tasarım kongresi, pp 661-669. 2023.