



SELÇUK UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

REPUBLIC OF TÜRKİYE
SELÇUK UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES



**PREDICTING SCHOOL DROPOUT OF FOREIGN STUDENTS
IN TURKEY USING MACHINE LEARNING TECHNIQUES:
A CASE STUDY ON YEMENI STUDENTS**

ANAS AHMED ALI ALHARDI

MASTERS THESIS

Department of Information Technologies Engineering

July-2025
KONYA
All Rights Reserved

TEZ KABUL VE ONAYI

ANAS AHMED ALI ALHARDI tarafından hazırlanan “MAKİNE ÖĞRENMESİ TEKNİKLERİ KULLANILARAK TÜRKİYE’DEKİ YABANCI ÖĞRENCİLERİN OKUL TERKİNİN TAHMİN EDİLMESİ: YEMENLİ ÖĞRENCİLER ÜZERİNE BİR ÖRNEK ÇALIŞMA” adlı tez çalışması 01/07/2025 tarihinde aşağıdaki jüri tarafından oy birliği ile Selçuk Üniversitesi Fen Bilimleri Enstitüsü Bilişim Teknolojileri Mühendisliği Anabilim Dalı’nda YÜKSEK LİSANS TEZİ olarak kabul edilmiştir.

Jüri Üyeleri

İmza

Başkan

Prof. Dr. Humar KAHRAMANLI ÖRNEK

.....

Danışman

Dr. Öğr. Üyesi Selahattin ALAN

.....

Üye

Dr. Öğr. Üyesi Süleyman A. SULAK

.....

Yukarıdaki sonucu onaylarım.

Prof. Dr. Hasan AYDOĞAN
FBE Müdürü

THESIS ACCEPTANCE AND APPROVAL

The thesis titled "PREDICTING SCHOOL DROPOUT OF FOREIGN STUDENTS IN TURKEY USING MACHINE LEARNING TECHNIQUES: A CASE STUDY ON YEMENI STUDENTS" prepared by ANAS AHMED ALI ALHARDI has been unanimously accepted as a Master's Thesis in the Department of Information Technology Engineering at the Graduate School of Natural and Applied Sciences of Selçuk University on 01/07/2025 by the jury listed below.

Jury Members

Signature

Chair

Prof. Dr. Humar KAHRAMANLI ÖRNEK

.....

Advisor

Assist. Prof. Dr. Selahattin ALAN

.....

Member

Assist. Prof. Dr. Süleyman A. SULAK

.....

I approve the above result.

Prof. Dr. Hasan AYDOĞAN

Director of Graduate School of Natural and Applied Sciences

TEZ BİLDİRİMİ

Bu tezdeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edildiğini ve tez yazım kurallarına uygun olarak hazırlanan bu çalışmada bana ait olmayan her türlü ifade ve bilginin kaynağına eksiksiz atıf yapıldığını bildiririm.

DECLARATION PAGE

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

ANAS AHMED ALI ALHARDI

Date: 2025/07/01

ÖZET

YÜKSEK LİSANS TEZİ

MAKİNE ÖĞRENMESİ TEKNİKLERİ KULLANILARAK TÜRKİYE'DEKİ YABANCI ÖĞRENCİLERİN OKUL TERKİNİN TAHMİN EDİLMESİ: YEMENLİ ÖĞRENCİLER ÜZERİNE BİR ÖRNEK ÇALIŞMA

ANAS AHMED ALI ALHARDI

Selçuk Üniversitesi Fen Bilimleri Enstitüsü
Bilişim Teknolojileri Mühendisliği Anabilim Dalı

Danışman: DR. ÖĞR. ÜYESİ SELAHATTİN ALAN

2025, 123 Sayfa

Jüri

Prof. Dr. Humar KAHRAMANLI ÖRNEK

Dr. Öğr. Üyesi Selahattin ALAN

Dr. Öğr. Üyesi Süleyman A. SULAK

Bu tez, Türk yükseköğretiminde uluslararası öğrenciler arasında önemli bir sorun olarak görülen okul terki sorununu ele almakta ve örneklem olarak Yemenli öğrenciler üzerine odaklanmaktadır. Uluslararası öğrenciler, okul terkine neden olabilecek özgün akademik, sosyo-kültürel ve ekonomik baskılarla karşı karşıya kalmaktadır. Bu araştırma, Türkiye'deki uluslararası öğrenciler arasında okul terkine sebep olan temel faktörleri belirlemeyi ve bu öğrencilerin okul terki ihtimalini doğru bir şekilde tahmin etmek için makine öğrenimi (ML) modelleri geliştirip değerlendirmeyi, böylece proaktif müdahale stratejilerine bilgi sağlamayı amaçlamıştır. Araştırmada; nicel, uygulamalı ve tanımlayıcı bir araştırma tasarımı kullanılmıştır. Veriler, demografik, akademik, finansal ve sosyo-kültürel değişkenleri kapsayan yapılandırılmış bir anket aracılığıyla 545 Yemenli öğrenciden toplanmıştır. Ön işleme sonrasında, 277 öğrenciden oluşan bir veri seti (okul terki/mezuniyet sınıflandırması için) Lojistik Regresyon, Rastgele Orman, Destek Vektör Makineleri ve XGBoost dahil olmak üzere on beş ML algoritmasını eğitmek ve test etmek için kullanılmıştır. Modeller, %80 eğitim / %20 test oranıyla doğruluk, kesinlik, geri çağırma ve F1 puanı kullanılarak değerlendirilmiştir. Temel bulgular; akademik performans (Genel Akademik Not Ortalaması), Türkçe dil yeterliliği, bölüm memnuniyeti, finansal faktörler (burs, aile ekonomisi) ve psikolojik stres faktörlerinin okul terkinin tahmin etme sürecinde önemli kriterler olduğunu göstermiştir. Topluluk ML modelleri, özellikle XGBoost ve Doğrusal Diskriminant Analizi (LDA), en yüksek tahmin doğruluğunu göstererek diğer modellere kıyasla daha başarılı sonuçlar vermiştir. Özelliklerin önemine yönelik yapılan analiz, bu belirleyici faktörlerin gerçekten etkili olduğunu ortaya koymuştur. Ayrıca, tahmin modelinin pratik uygulamasını ortaya koymak için bir prototip web sitesi de geliştirilmiştir. Bu çalışma, makine öğrenmesi tekniklerinin, Türkiye'de öğrenim gören yabancı öğrencilerin okul terkinin öngörülmesinde etkili bir araç olduğunu ortaya koymakta ve okul terkinin tahmin edilmesinde etkili olan kritik faktörleri ortaya koymaktadır. Sonuçlar, Türkiye'de öğrenim gören yabancı öğrencilerin okul terki riskini azaltmak için öğrencilere akademik, dilsel, finansal ve psikolojik alanlarda hedefli ve kurumsal desteğin gerekli olduğunu ortaya koymaktadır. Geliştirilen modeller ve web uygulaması, risk altındaki öğrencilerin erken tanısı ve onlara destek sağlanması için pratik araçlar sunmaktadır. Gelecek araştırmalar, farklı uluslararası öğrenci gruplarını da kapsayacak şekilde genişletilmeli ve öğrencilerin zaman içindeki gelişimlerini izleyen uzun dönemli verilere yer vermelidir.

Anahtar Kelimeler: Okulu bırakma, Makine öğrenmesi, Veri analizi, Tahmin, Sınıflandırma.

ABSTRACT

MS THESIS

PREDICTING SCHOOL DROPOUT OF FOREIGN STUDENTS IN TURKEY USING MACHINE LEARNING TECHNIQUES: A CASE STUDY ON YEMENI STUDENTS

ANAS AHMED ALI ALHARDI

**THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCE OF SELÇUK
UNIVERSITY**

**THE DEGREE OF MASTER OF SCIENCE
IN INFORMATION TECHNOLOGIES ENGINEERING**

Advisor: ASSIST. PROF. DR. SELAHATTİN ALAN

2025, 123 Pages

Jury

Prof. Dr. Humar KAHRAMANLI ÖRNEK

Assist. Prof. Dr. Selahattin ALAN

Assist. Prof. Dr. Süleyman A. SULAK

This thesis addresses the significant issue of school dropout among international students in Turkish higher education, with a specific focus on Yemeni students as a case study. International students face unique academic, socio-cultural, and economic pressures that can lead to a dropout. This research aims to identify the primary factors leading to school dropout among international students in Türkiye, develop and evaluate machine learning (ML) models for accurately predict likelihood of dropout among these students, thereby informing proactive intervention strategies. A quantitative, applied and descriptive research design have been used in this research. Data were collected from 545 Yemeni students through a structured survey covering demographic, academic, financial, and socio-cultural variables. After preprocessing, a dataset of 277 students (for dropout/graduation classification) was used to train and test fifteen ML algorithms, including Logistic Regression, Random Forest, Support Vector Machines, and XGBoost. The models were evaluated using accuracy, precision, recall, and F1-score, with an 80% training and 20% testing split. Main results indicated that academic performance (last GPA), Turkish language proficiency, department satisfaction, financial factors (scholarships, family economics), and psychological stress factors are important criteria in predicting school dropout. Ensemble ML models, particularly XGBoost and Linear Discriminant Analysis (LDA), outperformed other models by demonstrating the highest predictive accuracy. An analysis of feature importance reveals the significant impact of these determined factors. Additionally, a prototype website was developed to demonstrate the practical application of the prediction model. This study concludes that machine learning techniques are an effective tool for predicting school dropout among international students studying in Türkiye and identifies the critical factors influencing dropout predictions. The results suggest that institutional support for international students in Türkiye across academic, linguistic, financial and psychological is necessary to reduce the risk of dropout among them. The developed models and web application offer practical tools for early identification and support students who are at-risk. Future studies should be expanded to include diverse international student populations and include longitudinal data for tracking students' development over time.

Keywords: School dropout, Machine learning, Data analysis, Prediction, Classification.

FOREWORD

I would like to express my gratitude to my thesis advisor, Assist. Prof. Dr. Selahattin ALAN, for guiding me with his valuable time, knowledge, and experience and for always supporting me throughout my master's studies as well as specifically my thesis.

I thank my dear family and friends for always making me feel their moral support. I also thank Eng. Tasneem for her support.

ANAS AHMED ALI ALHARDI
KONYA-2025

CONTENTS

ÖZET	viii
ABSTRACT.....	ix
FOREWORD	x
CONTENTS	xi
ABBREVIATIONS.....	xiv
1. INTRODUCTION	1
1.1. Overview.....	1
1.2. School dropout.....	2
1.3. Social, Economic, and Cultural Factors Affecting Dropout Rates.	3
1.3.1. Social Factors.....	3
1.3.2. Economic Factors	4
1.3.3. Cultural Factors.....	5
1.4. International Students	5
1.4.1 International Students in Türkiye	5
1.5. Predictive Methods for School Dropout	8
1.5.1 Traditional and Statistical Approaches	8
1.5.2 The Role of Machine Learning in Predicting Dropout Rates	10
1.6 Purpose of the Study and Research Questions.....	12
1.7 Significance of the Study	13
1.8. Organization of the Thesis	15
2. LITERATURE REVIEW	16
2.1. Predicting Student Dropout Using Machine Learning Techniques: A Systematic Review	16
2.1.1. The Historical Development of Student Dropout Prediction Techniques	16
2.1.2. Predicting Student Dropout According to Different Learning Environments	18
2.1.3. Predicting Dropout Among International Students	20
2.1.4. School Dropout Studies in Turkish Literature	21
2.1.5. Challenges Faced by Arab Students Studying Abroad: Comparative Analysis with Yemeni Students in Turkey	24
2.1.6. Comparing the Performance of Machine Learning Algorithms in Predicting Student Dropout.....	27
2.1.7. Factors Influencing Student Dropout as Identified by Machine Learning Models	31
2.1.8. Recent Trends in Student Dropout Prediction	34
2.1.9. Summary and Future Directions	35
3. MATERIAL AND METHOD.....	37
3.1. Research Design	37
3.2 Population and Sample	38

3.3 Data Collection Tool.....	39
3.3.1 Survey Structure and Content	39
3.3.2 Data Collection Phases	40
3.3.3 Challenges in Data Collection	40
3.4 Data Preprocessing	41
3.4.1 Data Cleaning	41
3.4.2 Data Translation.....	42
3.4.3 Data Encoding.....	42
3.4.4 Data Normalization/Standardization.....	43
3.5 Modeling Algorithms.....	44
3.5.1 Logistic Regression.....	44
3.5.2 Random Forest	45
3.5.3 SVM (Support Vector Machine).....	47
3.5.4 XGBoost	48
3.5.5 Decision Tree	50
3.5.6 Naive Bayes	51
3.5.7 LightGBM.....	53
3.5.8 LDA (Linear Discriminant Analysis)	54
3.5.9 AdaBoost	55
3.5.10 Ridge Classifier.....	57
3.5.11 Extra Trees	58
3.5.12 QDA (Quadratic Discriminant Analysis)	59
3.6 Model Evaluation Metrics	60
3.6.1 Accuracy	60
3.6.2 Precision.....	61
3.6.3 Recall (Sensitivity)	61
3.6.4 F1 Score	61
3.6.5 Confusion Matrix	61
3.7 Train/Test Split	62
3.8 Web Application for Prediction	63
3.8.1 Creating a Trained Model File	63
3.8.2 Development Environment Setup	63
3.8.3 API Development	64
3.8.4 Web Interface Design	64
3.8.5 Deployment Infrastructure	70
3.8.6 Ethical Considerations of Prediction Results.....	71
3.8.7 Integration and Operation	72
3.9 Ethical Considerations	72
4. RESULTS	74
4.1. Exploratory Data Analysis.....	74
4.2. Model Implementation.....	81
4.3. Model Performance Comparison	83
4.3.1. Comparative Performance Metrics	83
4.3.2. Model Performance Analysis.....	84
4.3.3. Implications for Model Selection	89
4.4. Feature Importance Analysis	90
4.5. Model Validation	92
4.5.1. Two-Tier Validation: Aggregate Test Set and Individual Case Analysis	93

4.5.2. Application to Enrolled Students.....	93
4.6. Web Application for Prediction.....	94
4.7. Summary.....	97
5. CONCLUSION	99
5.1. Recapitulation of the Study.....	99
5.2. Summary of Key Findings.....	100
5.3. Achievement of Research Objectives	101
5.4. Recommendations.....	102
5.5. Suggestions for Future Research	104
REFERENCES.....	106



ABBREVIATIONS

Abbreviations

AI	: Artificial Intelligence
API	: Application Programming Interface
AUC	: Area Under the Curve
BDT	: Boosted Decision Tree
DT	: Decision Trees
EDA	: Exploratory Data Analysis
EDM	: Educational Data Mining
EFB	: Exclusive Feature Bundling
EWS	: Early warning systems
FN	: False Negative
FP	: False Positive
GBT	: Gradient Boosted Trees
GBDT	: Gradient Boosting Decision Trees
GFO	: Gaussian Flow Optimizer
GOSS	: Gradient-based One-Side Sampling
GPA	: Grade point average
IDE	: Integrated Development Environment
IFS	: Intuitionistic Fuzzy Sets
IOM	: International Organization for Migration
k-NN	: k-Nearest Neighbor
KIT	: Karlsruhe Institute of Technology
LDA	: Linear Discriminant Analysis
LSTM	: Long Short-Term Memory
ML	: Machine Learning
MLP	: Neural Network
MOOCs	: Massive Open Online Courses
NB	: Naive Bayes
NN	: Neural Networks
QDA	: Quadratic Discriminant Analysis
RF	: Random Forest
RNNs	: Recurrent Neural Networks
SES	: Socioeconomic status
SMOTE	: Synthetic Minority Over-sampling Technique
SVM	: Support Vector Machine
TN	: True Negative
TP	: True Positive
XGBoost	: Extreme Gradient Boosting

1. INTRODUCTION

1.1. Overview

Education is a multifaceted concept that plays an essential role in shaping individuals and developing societies. It refers to the process of gaining knowledge, skills, values, and attitudes, whether through formal or informal learning. Its role is not limited to the transmission of academic knowledge alone but extends to fostering critical thinking and creativity, enhancing social and emotional intelligence, and empowering individuals to face life's challenges and interact effectively with their community (Lövdén et al., 2020). In a world characterized by rapid technological and social changes, the role of education in preparing individuals to adapt to constant transformations becomes prominent. Education focuses on equipping individuals and developing problem-solving skills, enabling learners to face unforeseen circumstances with competence and effectiveness. By fostering an adaptable mindset, education becomes a key to empowering individuals to succeed in an ever-evolving global environment (Lent, 2018).

Education is not only a means of personal advancement but also a vital factor in promoting social cohesion and economic growth. It empowers individuals to discover their passions, build their career paths, and contribute actively to their communities. By developing essential skills and necessary knowledge, education enhances both personal and societal development (Hanushek & Wößmann, 2007).

Compulsory education emerged in the 18th and 19th centuries alongside urbanization and industrialization, as modern nation-states sought to equip their citizens for active participation in complex economies and democracies. The Massachusetts Law of 1647 represents an early development in the state's role in ensuring the right to education (Tyack, 1976). By the early 20th century, compulsory education legislation had spread globally, recognized as a developmental necessity and a fundamental right for individuals. As of 2021, 91% of countries had mandated compulsory education at the primary level and 84% at the secondary level (Education-Organic, 2025).

Focusing on learning has become an urgent necessity to achieve sustainable development (The-United-Nations, 2025). In light of efforts to enhance educational continuity and ensure the success of sustainable development policies, the phenomenon of school dropout emerges as one of the most prominent challenges to achieving these

goals. School dropout, defined as the early departure of students from the educational system before completing a specific educational stage, not only threatens the future of individuals but also weakens the social and economic fabric of communities. Effectively tackling the issue of school dropout is a critical measure of the capacity of educational and societal institutions. Promoting enrollment and retention in education is not merely a matter of educational policy; it is a fundamental pillar for ensuring social cohesion and achieving the visions of comprehensive development, which can only be built with the awareness of educated individuals capable of taking responsibility for themselves and their communities (Brunello & Paola, 2014).

1.2. School dropout

Researchers have attempted to define school dropout through various theoretical models, reflecting the complexity of this phenomenon. Examining the literature on school dropout reveals a range of definitions. Garrison (1985) defined school dropout as leaving school completely without graduating or completing education for individual and social reasons. Similarly, Simić & Krstić (2017) defined it as a process that begins with a student's enrollment in an educational institution and subsequently involves the student ceasing to attend for various reasons, leading to their exit from the educational system before completing its academic requirements. In contrast, Lamb et al (2010) defined it as non-enrollment in school and failure to achieve specified success in vocational or academic education.

Generally, school dropout is defined as a student leaving school without graduating or completing the academic program in which they are enrolled, regardless of the underlying reasons (Entwisle et al., 2004; McWhirter et al., 1993).

School dropout is considered one of the major problems facing educational institutions worldwide. In Turkey, data from the European Statistical Office (EUROSTAT) for 2017 indicates that the rate of early school leaving among young people aged 18-24 was 36% (Eurostat, 2018; Kartal & Ballı, 2020). In the European Union, data from the European Statistical Office show that the dropout rate for young people in the same age group reached 9.6% in 2022 (Eurostat, 2025). Globally, the UNESCO Institute for Statistics reported that 258.4 million children and young people were out of school in 2018; this represents one-sixth of the world's population in that age group (UIS, 2019).

The factors influencing the school dropout phenomenon differ across countries due to variations in cultural, geographical, and social contexts. These factors include economic challenges such as poverty, social issues like family disintegration, and cultural barriers such as a lack of awareness about the importance of education. Together, these factors exacerbate the dropout problem. Identifying and ranking the factors affecting dropout rates by importance is a complex process, as it is directly linked to the quality of collected data and its analysis to uncover the most influential patterns in school dropout (Manhães et al., 2011).

The problem of dropout in educational institutions becomes increasingly complex due to the absence of official data, making it difficult to accurately identify and classify dropout students. The causes of dropout vary from permanent withdrawal from studies to changing majors, transferring to another university, or even temporary cessation followed by the resumption of studies at a later time (Assis, 2018).

1.3. Social, Economic, and Cultural Factors Affecting Dropout Rates.

School dropout is influenced by a complex interplay of social, economic, and cultural factors that operate at individual, family, school, and societal levels. Understanding these multifaceted influences is crucial for developing effective dropout prevention strategies.

1.3.1. Social Factors

Social factors play a significant role in determining students' likelihood of dropping out. Peer relationships and social integration within the school environment can significantly impact students' engagement and persistence. According to Tinto's (1975) model, social integration is a key determinant of student retention, with students who feel socially connected to their peers and the broader school community being more likely to persist in their studies.

Family structure and dynamics also influence dropout rates. Students from single-parent households or those experiencing family conflict or instability may face additional challenges that increase their risk of dropping out. Rumberger (2011) found that students who experience changes in family structure, such as parental divorce or remarriage, are more likely to drop out than those from stable family environments.

School climate and culture represent another critical social factor. Schools characterized by positive teacher-student relationships, clear and fair disciplinary policies, and a strong sense of community tend to have lower dropout rates. Conversely, schools with high levels of violence, bullying, or discrimination may push students toward disengagement and eventual dropout (Christle et al., 2007).

1.3.2. Economic Factors

Economic factors are among the most powerful predictors of school dropouts. Socioeconomic status (SES) consistently emerges as a strong correlate of educational outcomes, with students from lower SES backgrounds being significantly more likely to drop out than their more affluent peers. This relationship operates through various mechanisms, including:

1. **Financial constraints:** Families with limited financial resources may struggle to cover educational expenses such as books, uniforms, transportation, and technology. In some cases, students may need to work to contribute to family income, reducing the time and energy available for academic pursuits.
2. **Residential instability:** Economic hardship often leads to frequent moves, which can disrupt students' educational continuity and social connections. Students who change schools frequently are at higher risk of academic difficulties and eventual dropout.
3. **Limited access to resources:** Students from lower SES backgrounds often have less access to educational resources such as books, computers, and quiet study spaces at home. They may also have limited access to enrichment activities and educational support services.
4. **Opportunity costs:** As students get older, the opportunity cost of remaining in school increases, particularly for those from economically disadvantaged backgrounds. For these students, the immediate financial benefits of entering the workforce may seem more attractive than staying in school, especially when the long-term economic returns of education appear uncertain. This dynamic often leads students to prioritize short-term income over continuing their education.

Alhassan & Hassan (2024) found that household income was a key predictor of dropout in their study of student dropout rates in Somaliland, highlighting the universal importance of economic factors across different cultural contexts.

1.3.3. Cultural Factors

Cultural factors, including values, beliefs, and practices related to education, also influence dropout rates. Cultural attitudes toward education can vary significantly across different communities and may affect students' educational aspirations and persistence. In some cultural contexts, education may be highly valued and prioritized, while in others, other pursuits or responsibilities may take precedence (UNESCO, 2020).

Gender norms and expectations represent a particularly important cultural factor. In many societies, traditional gender roles may limit educational opportunities for girls, who may be expected to prioritize domestic responsibilities or marry early. Conversely, boys in some contexts may face pressure to enter the workforce early to fulfill provider roles. These gendered expectations can significantly impact dropout patterns, often in ways that vary across cultural contexts (UNESCO, 2022).

Language and cultural barriers can also contribute to dropout, particularly for immigrant or minority students. Students who speak a different language at home or whose cultural background differs significantly from the dominant school culture may experience challenges in academic and social integration, increasing their risk of disengagement and dropout (Suárez-Orozco, 2001).

1.4. International Students

International students are defined as individuals who travel outside national or regional borders for educational purposes and study in a country other than their own (Smith & Khawaja, 2011). According to data from UNESCO's Institute of Statistics, the international student population worldwide in 2021 exceeded 6.39 million. This represents an increase of nearly threefold compared to the figure recorded in 2001. This number continued to grow, reaching approximately 6.4 million students by 2023 (IOM, 2024).

1.4.1 International Students in Türkiye

According to the Ministry of Higher Education, there are more than 350,000 students in Turkey from 198 countries around the world. Turkey is striving to be among the ten leading countries globally in attracting the largest number of international students (YÖK, 2022). Through decades of focused efforts on scientific and educational

development, the Turkish government has sought to enhance the standard of education, thereby elevating the educational standards in Turkey to advanced levels. Thanks to ongoing efforts in development and innovation, Turkish universities have secured prestigious global rankings. Moreover, the adoption of English as an instructional language alongside Turkish has significantly increased the interest of foreign students in studying in Turkey.

The growing number of international students in Turkey enhances academic diversity and has significant economic impacts. Their spending on tuition, housing, and living expenses contributes to various sectors of the country's economy. In addition, the presence of these students enriches the educational experience and promotes personal growth within a multicultural environment. Despite these benefits, international students face unique challenges that may lead to school dropouts. These challenges include academic difficulties, cultural adaptation, and social integration issues, which vary across educational levels and geographic regions. Thus, understanding these factors is crucial for developing strategies to reduce dropout rates among international students (Lemke, 2011).

In this context, the focus is on Yemeni students in Turkey as a research sample to study the phenomenon of school dropout among international students. Yemeni students represent a significant segment of the foreign student communities in Turkey, numbering 8,198 students according to a 2023 report by the Council of Higher Education (YÖK, 2023). This group was chosen as a representative study sample because it reflects common challenges faced by international students in general. These challenges include economic hardships stemming from political and economic instability in their home countries, as well as psychological pressures related to cultural and social adaptation in a new environment. Such challenges can significantly affect their academic persistence and increase the likelihood of dropping out. Therefore, understanding the factors influencing this student group is essential for formulating effective strategies aimed at reducing dropout rates among international students and helping them make better use of available educational opportunities.

The phenomenon of school dropout among international students in Turkey is a complex problem that requires careful study to understand its causes and impacts. This phenomenon differs from dropout in other educational levels due to the additional challenges faced by international students, such as cultural and linguistic adaptation, economic pressures, and difficulties integrating into the Turkish educational system.

One of the main reasons leading to international student dropouts is language barriers. Although many Turkish universities offer programs in English, students studying in Turkish may face significant difficulty understanding course materials or communicating with professors and peers. This leads to lower academic achievement and an increased likelihood of dropping out (Kurum & Erdemli, 2021). In addition, poor language proficiency negatively impacts their ability to participate effectively in academic and social activities on campus.

Additionally, economic factors play a significant role in increasing dropout rates among international students. Students coming from low-income countries often rely on scholarships to cover the costs of education and living expenses. However, some scholarships do not cover all expenses, forcing students to work part-time or seek additional sources of income. This financial pressure can lead to a distraction from their academic focus, thus negatively impacting their academic performance (Jung & Kim, 2018). Furthermore, students who lose scholarships due to low academic grades may be forced to withdraw from their studies permanently.

Furthermore, cultural and social challenges are significant factors influencing the persistence of international students in education. Adapting to a new culture and a different educational system can be difficult for many. International students in Turkey face challenges such as feelings of alienation, difficulty building social relationships, and a lack of psychological and social support. Such factors can leave students feeling isolated and depressed, which increases the likelihood of their decision to withdraw from university (Eynullayeva et al., 2021).

In addition, university policies and systems can be a contributing factor to international student dropout. Some universities may not provide adequate services to support international students, such as Turkish language learning programs, academic advising, or psychological assistance. Additionally, the bureaucratic complexities surrounding residency permits and visas can create further obstacles for students, making it more difficult for them to continue their studies (Özoğlu et al., 2015).

Academic challenges play a significant role in international student dropout. Students transitioning from different educational systems may find it difficult to adapt to the teaching and assessment methods in Turkish universities. In addition, a lack of academic guidance and the unavailability of educational support programs can leave students unable to cope with academic difficulties, leading to a decline in their performance and an increased likelihood of them dropping out (Titrek et al., 2016).

To address this problem, predicting and managing the early signs of student dropout is of paramount importance. This will enable educational institutions to act swiftly and implement preventative and proactive measures to address this issue and reduce the dropout rate (Piscitello et al., 2022).

To confront the growing challenges associated with the phenomenon of school dropout and reduce its rates, developing new and innovative tools has become essential in academic and research circles. This phenomenon, which represents a complex educational and social problem, requires multidimensional strategies for its effective understanding and management. In this context, the scientific community seeks to explore and design advanced analytical models such as statistical models, machine learning algorithms, regression analysis, decision trees, and clustering techniques aimed at describing the phenomenon of school dropouts accurately and comprehensively. These models offer both theoretical and practical frameworks for analyzing dropout factors and interpreting its dynamics. Additionally, they enable researchers and specialists to make more accurate predictions, supporting data-driven decision-making.

1.5. Predictive Methods for School Dropout

1.5.1 Traditional and Statistical Approaches

Before the advent of machine learning techniques, researchers and educational institutions relied on traditional and statistical approaches to predict and understand student dropout. These methods, while less sophisticated than modern computational approaches, established the foundation for current predictive models and continue to offer valuable insights into dropout patterns and risk factors.

1.5.1.1 Descriptive Statistics and Demographic Analysis

Early approaches to dropout prediction often relied on descriptive statistics and demographic analysis to identify patterns and risk factors associated with dropout. These methods typically involved examining historical dropout rates across different demographic groups, such as those defined by gender, race/ethnicity, socioeconomic status, and geographic location. By identifying groups with disproportionately high dropout rates, educators and policymakers could target interventions toward these high-risk populations.

Demographic analysis helped establish some of the fundamental risk factors for dropout that continue to inform current models. For instance, the consistent relationship between socioeconomic status and dropout risk was initially established through such analyses. While these approaches provided valuable insights at the group level, they offered limited predictive power for individual students and often failed to capture the complex interactions between different risk factors.

Regression Analysis

Regression analysis represented a significant advancement in dropout prediction, allowing researchers to examine the relationship between multiple predictor variables and dropout outcomes simultaneously. Linear regression models were used to predict continuous outcomes related to academic performance, while logistic regression became the method of choice for predicting binary outcomes such as dropout versus persistence.

Logistic regression models offer several advantages for dropout prediction, including:

1. **Interpretability:** The coefficients in logistic regression models have clear interpretations in terms of odds ratios, making it easy to understand the relative importance of different predictors.
2. **Ability to handle different types of predictors:** Logistic regression can incorporate both continuous variables (e.g., GPA, attendance rates) and categorical variables (e.g., gender, race/ethnicity).
3. **Probabilistic output:** Rather than simply classifying students as "at risk" or "not at risk," logistic regression provides probability estimates, allowing for more nuanced risk assessment.

Despite these advantages, traditional regression approaches have limitations, including assumptions about linearity and independence of predictors that may not hold in complex educational data. Additionally, these models may struggle to capture non-linear relationships and complex interactions between variables without explicit specification by the researcher.

1.5.1.2 Early Warning Systems Based on Threshold Indicators

Many educational institutions have implemented early warning systems (EWS) based on simple threshold indicators to identify students at risk of dropping out. These systems typically monitor key indicators such as:

1. **Attendance:** Students with attendance rates below a certain threshold (e.g., 90%) are flagged as at risk.
2. **Course failures:** Students failing one or more courses are identified for intervention.
3. **Behavioral incidents:** Students with a high number of disciplinary referrals are considered at risk.
4. **GPA:** Students whose grade point average falls below a certain level are flagged.

The "ABC" approach (Attendance, Behavior, and Course performance) exemplifies this type of early warning system. While relatively simple to implement, these threshold-based systems have several limitations. They often treat risk factors in isolation rather than considering their combined effect, and the thresholds are typically determined somewhat arbitrarily rather than being empirically optimized. Additionally, these systems generally do not account for the varying importance of different risk factors across different contexts or student populations.

1.5.2 The Role of Machine Learning in Predicting Dropout Rates

The emergence of machine learning has transformed the field of dropout prediction, offering new approaches that can overcome many of the limitations of traditional statistical methods. Machine learning techniques bring several advantages to dropout prediction, including the ability to handle large and complex datasets, identify non-linear relationships and interactions, and adapt to new patterns in the data.

1.5.2.1 Advantages of Machine Learning for Dropout Prediction

Machine learning approaches offer several key advantages over traditional statistical methods for predicting student dropout:

1. **Handling complex, high-dimensional data:** Educational data often includes numerous variables from diverse sources, including demographic information, academic performance metrics, behavioral indicators, and increasingly, digital interaction data from learning management systems. Machine learning algorithms can effectively process and find patterns in these high-dimensional datasets.

2. **Capturing non-linear relationships:** The relationship between predictor variables and dropout risk is often non-linear and complex. Machine learning models, particularly tree-based methods and neural networks can capture these non-linear relationships without requiring explicit specification by the researcher.
3. **Identifying complex interactions:** Machine learning models can automatically detect interactions between variables that might be missed in traditional analyses. For example, a particular risk factor might have a stronger effect for certain demographic groups or in combination with other risk factors.
4. **Adaptive learning:** As new data becomes available, machine learning models can be updated to incorporate this information, allowing them to adapt to changing patterns and relationships over time.
5. **Handling missing data:** Educational datasets often contain missing values, which can be challenging to address in traditional statistical analyses. Many machine learning approaches have built-in methods for handling missing data or can be combined with imputation techniques.

In this context, Machine Learning (ML) is considered a powerful tool that contributes to building accurate predictive models to identify students at risk of dropping out. These models provide early warnings to the relevant educational authorities, enabling them to take appropriate preventative measures to reduce the likelihood of dropout (Del Bonifro et al., 2020). Machine Learning, as a subfield of Artificial Intelligence (AI), is dedicated to creating algorithms capable of recognizing patterns and drawing conclusions independently, making it an effective tool for analyzing complex educational data (Franco, 2019). By using machine learning algorithms, educational institutions can analyze student data such as attendance rates, academic performance, in-school behavior, and socioeconomic conditions to identify factors that increase the likelihood of dropout.

However, despite the significant progress in using machine learning algorithms to predict school dropout, there remain ongoing challenges related to selecting the appropriate dataset and identifying the key variables relevant to the phenomenon. Selecting an appropriate dataset is critical, as the predictive model's quality hinges on the data's representativeness. In addition, identifying the key variables that directly influence dropout requires careful study and a deep understanding of different social

and educational contexts. These issues are a major topic of ongoing research in academic circles, where scientists are working to improve the accuracy of the models and enhance their ability to provide more reliable and practically applicable predictions (Pelima et al., 2024).

Furthermore, Efforts to address school dropout must consider the cultural and social diversity that may influence the nature of the phenomenon. For example, dropout factors may vary between rural and urban areas, as well as across different social groups (Dechavez, 2024). Therefore, designing flexible models that adapt to these variations and offer tailored solutions for each context is essential. Moreover, collaboration among researchers, teachers, and administrators in educational institutions is crucial to ensure that the developed models are applicable and effective in practice.

Finally, school dropout should be viewed not only as an educational problem but also as a societal issue that requires multidimensional interventions. For example, students who drop out may face difficulties finding jobs or may be more vulnerable to social problems such as crime or poverty. Therefore, educational policies alone may not be sufficient if the economic and social aspects associated with the phenomenon are not addressed. For instance, providing financial support to low-income families or offering community-based counseling services can help address some of the root causes of dropout. Hence, the importance of integrating data-driven analyses with comprehensive reform policies aimed at improving the educational environment to ensure students' persistence in their educational journey becomes clear. For example, by analyzing student attendance and performance data, schools can identify at-risk students early and offer targeted support programs such as tutoring or mentoring (UNESCO, 2017).

1.6 Purpose of the Study and Research Questions

This research aims to develop a predictive model based on Machine Learning (ML) techniques to understand and analyze the phenomenon of school dropout among international students enrolled in Turkish universities, with a particular focus on Yemeni students as a case study. This focus arises from the growing need to understand the factors influencing the persistence of international students in the Turkish higher education system, especially in light of the social, economic, and cultural challenges they may face.

To achieve this objective, a structured research methodology will be followed, beginning with the collection of accurate field data from a target sample of Yemeni students studying in Turkish universities. A specially designed questionnaire will be used to cover a wide range of social, economic, and cultural variables that may be related to the phenomenon of school dropout.

Within the framework of this research, the following research questions will be answered:

1. What are the social, economic, and cultural factors that directly or indirectly influence the decision of international students to withdraw from Turkish universities?

This question seeks to explore the various factors that may contribute to an increased likelihood of school dropout, including the family's economic circumstances, cultural challenges associated with adapting to a new educational environment, and social barriers that may hinder students' persistence in their studies

2. Can a predictive model be developed using machine learning techniques to determine the likelihood of school dropout among international students studying in Turkish universities?

This question aims to test the ability of machine learning-based models to analyze available data and provide accurate predictions about the likelihood of dropout, thus enabling the opportunity to provide proactive interventions to support students at risk.

By answering these questions, the research seeks to make a valuable scientific contribution to the field of education, focusing on a deeper understanding of the challenges faced by international students and providing innovative analytical tools to improve educational policies and enhance students' persistence in their academic journey.

1.7 Significance of the Study

This research holds significant importance on several levels, both in terms of its academic contribution and its practical application in the field of education and policies supporting international students. The significance of the study is reflected in the following points:

Firstly, this research is considered a valuable academic contribution to understanding the phenomenon of school dropout among international students, an issue that requires careful study due to its multidimensional complexities. By focusing on Yemeni students as a case study, the research offers fresh perspectives on the social, economic, and cultural factors that may influence students' persistence in Turkish universities. These insights can serve as a foundation for future studies targeting other groups of international students in different educational contexts.

Secondly, the research is distinguished by its use of Machine Learning (ML) techniques to develop a predictive model capable of determining the likelihood of school dropout. This approach represents an advanced step in employing modern technology to analyze educational data, enabling the provision of innovative solutions to understand and address this phenomenon. Furthermore, the use of machine learning opens up new horizons for researchers and practitioners in the field of education to improve the accuracy of predictions and enhance their ability to make evidence-based decisions.

Thirdly, on a practical level, the research provides useful analytical tools that Turkish universities and educational policymakers can utilize to develop effective strategies for reducing school dropout rates. By identifying students at risk of dropping out early, educational institutions can provide appropriate support, whether in the form of mentoring programs, psychological and social services, or even financial incentives, thus contributing to enhancing students' persistence in their academic journey.

Finally, the importance of the research also stems from its highlighting of the issues faced by international students in general, a group that often faces unique challenges related to cultural adaptation, language, and economic conditions. By addressing these issues, this research can contribute to enhancing inclusivity and diversity in Turkish higher education, reflecting the commitment of educational institutions to supporting all students regardless of their cultural or social backgrounds. Therefore, this research is not limited to providing scientific insights only, but also aims to achieve a positive impact on educational policies and the experience of international students in Turkey, making it valuable on both local and international levels.

1.8. Organization of the Thesis

This thesis comprises five main chapters, designed to offer a structured study of the phenomenon of school dropout among international students in Turkish universities, with a particular focus on Yemeni students as a case study. The first chapter begins with a general introduction, followed by an overview of the phenomenon of school dropout and the challenges faced by international students in Turkey. The chapter also includes the research objectives, main questions, and highlights the significance of the study and its overall organization.

The second chapter is dedicated to a review of previous literature related to school dropout. This chapter reviews relevant theories and previous studies.

In the third chapter, the methodology followed in the research is explained. This chapter addresses the general framework of the study, including the questionnaire design and data collection strategy. It also discusses data processing steps such as cleaning and variable selection, and the use of machine learning techniques (such as decision trees, random forests, neural networks, and others) to develop predictive models.

The fourth chapter addresses the process of implementing the models and analyzing the results. This chapter presents an exploratory data analysis, discusses the performance of the various models used in predicting school dropouts, and selects the best model in terms of accuracy in order to build a system for predicting students at risk of dropouts.

Finally, the fifth chapter presents a summary of the main findings, discussing their practical implications for educational policies and student support efforts. The chapter also includes suggestions for future research that could contribute to expanding the scope of the study and addressing its limitations. The thesis concludes with a list of references that were relied upon, reflecting the academic credibility in building the theoretical and methodological framework of the research. This structured approach aims to offer a concise and comprehensive perspective on the problem, from identifying research gaps to drawing conclusions and proposing practical solutions.

2. LITERATURE REVIEW

2.1. Predicting Student Dropout Using Machine Learning Techniques: A Systematic Review

Predicting student dropout is a vital field within educational data analytics, as educational institutions strive to retain students and reduce the economic and social losses resulting from their withdrawal. Global statistics have shown that dropout rates in higher education range between 30% and 50% in many countries, representing a significant challenge for educational institutions and national economies. In this context, machine learning techniques have emerged as effective tools for analyzing diverse educational data, including academic records, demographic and social indicators, and learning management system data, to accurately identify students at risk of dropping out. This review examines the historical development and current trends in the use of machine learning for predicting student dropout, with a focus on methodologies, outcomes, and the challenges associated with this field.

2.1.1. The Historical Development of Student Dropout Prediction Techniques

- **Early and Foundational Studies (2003–2010)**

The beginning of the third millennium witnessed the emergence of the first studies addressing the use of machine learning techniques in predicting student performance and dropout. The work of (Kotsiantis et al., 2003) is considered pioneering in this field, as it evaluated the efficiency of machine learning algorithms, particularly the Naïve Bayes algorithm, in predicting student performance in the "Informatics" course at the Hellenic Open University. The results showed that these algorithms provide accurate predictions before final exams, identifying influential factors such as prior educational background. The study did not stop at predicting performance but also expanded to address the issue of student dropout, demonstrating that the Naïve Bayes algorithm can be effectively used to identify students at risk of dropping out. Furthermore, an electronic support tool based on this algorithm was developed for the automatic identification of at-risk students, making it one of the earliest practical applications of machine learning in this field.

The study by (Dekker et al., 2009) expanded the scope of application to include predicting the dropout of first-year students in the Department of Electrical Engineering

at Eindhoven University of Technology. The study analyzed student data, including academic grades and educational background, to build predictive models. The results demonstrated that relatively simple classifiers (such as decision trees) provide useful results with an accuracy ranging between 75% and 80%. The study also highlighted the importance of cost-sensitive learning and comprehensive analysis of misclassifications to improve prediction accuracy. It found that the grade in the linear algebra course was the strongest predictor of success, along with grades in calculus, networks, and the average grades in science courses. These findings are significant as they demonstrated the possibility of predicting dropout early in a student's academic trajectory.

The study by (Delen, 2010) provided an analytical comparison of machine learning methods in student retention management and their impact on university rankings, reputation, and financial standing. The study designed analytical models to predict the reasons for new student dropout using institutional data covering five years and multiple data mining techniques. The results showed that ensemble models outperformed individual models, and that a balanced dataset produced better prediction results than an imbalanced dataset. Sensitivity analysis of the models revealed that educational and financial variables were among the most influential factors in this phenomenon. This study provided important insights into how to address the problem of data imbalance, which is a common issue in dropout prediction studies, given that the proportion of dropout students is usually lower than that of non-dropouts.

- **Development of Methodologies and Techniques (2010–2015)**

This period witnessed significant advancements in the machine learning techniques used for predicting student dropout, as well as an expansion in application domains. The study by (Yadav et al., 2012) presented an educational data mining project aimed at creating forecasting models for managing student retention. The study focused on employing machine learning algorithms to generate accurate and concise prediction lists that identify students in need of support from retention programs. It examined the quality of predictive models produced by machine learning algorithms, and the results showed that some of these algorithms, especially decision trees, are capable of creating reliable predictive models from existing student retention data due to their interpretability compared to other classification methods.

The study by (Yukselturk et al., 2014) addressed the prediction of student dropout within a virtual learning program using data mining techniques. The study

included 189 students enrolled in an web-based Information Technology certification course between 2007 and 2009. Information was gathered via digital surveys, encompassing ten variables such as gender, age, educational level, and previous online experience. The study applied four data mining methods: k-Nearest Neighbor (k-NN), Decision Trees (DT), Naive Bayes (NB), as well as Neural Networks (NN). The findings revealed that the 3-NN and DT classifiers were the most sensitive in detecting dropout cases, achieving rates of 87% and 79.7%, respectively. These findings highlighted the importance of selecting the appropriate algorithm for the nature of the data and the problem under study.

The study by (Fei & Yeung, 2015) introduced time-based models for forecasting student attrition in Massive Open Online Courses (MOOCs), considering dropout prediction as a sequential classification problem. The study proposed the use of a model based on Recurrent Neural Networks (RNNs) model with Long Short-Term Memory (LSTM) cells. Extensive experiments on the Coursera and edX platforms demonstrated that the LSTM model significantly outperformed baseline methods and other approaches proposed in the study, thanks to its ability to process temporal data and capture long-term dependencies. This study represented a significant advancement in dropout prediction by employing deep learning techniques that account for the dynamic and temporal nature of student interactions with e-learning platforms.

2.1.2. Predicting Student Dropout According to Different Learning Environments

- **Dropout Prediction in Traditional Universities**

Traditional universities provide a unique environment for study, where students face diverse academic, social, and financial challenges that may lead to dropout. The study by (Nagy & Molontay, 2018) addressed the prediction of university student dropout using several machine learning algorithms. Using data from 15,825 students at Budapest University of Technology and Economics, the study addressed missing data through imputation and applied feature extraction and selection techniques. Various classifiers were trained, including decision trees, Naive Bayes, k-NN, linear models, and deep learning algorithms. Gradient Boosted Trees and deep learning models achieved the highest AUC scores of 0.808 and 0.811, respectively. The study also introduced a data-driven decision support system to assist educational stakeholders in enhancing student retention.

The study by (Rodríguez-Muñiz et al., 2019) introduced an innovative machine learning-based method for analyzing the problem of university student dropout. The method is characterized by the absence of strong distributional assumptions, the ability to handle larger datasets, and the interpretability of its results. The findings are consistent with previous research, confirming the impact of personal and contextual variables and the importance of academic performance in the first year. The study also highlights other factors, such as the significance of enrollment status (part-time or full-time) and student vulnerability related to age. It provides comprehensive graphical outputs to facilitate the interpretation of discovered rules, enhancing the usability of the results for educational institutions.

The study by (Sandoval-Palis et al., 2020) investigated student dropout prediction in the preparatory track at Escuela Politécnica Nacional University using data from 2,097 students. Both logistic regression and artificial neural network models were trained using four variables reflecting academic, social, and economic factors. The neural network, featuring a 4-7-1 architecture with a logistic activation function, outperformed the logistic regression model and was selected for early dropout prediction. The findings underscore the effectiveness of neural networks in handling complex classification tasks within the educational domain.

- **Dropout Prediction in E-Learning Platforms (MOOCs)**

The study by (Kloft et al., 2014) addressed an approach for predicting student dropout in MOOCs using students' clickstream data. The approach focused on extracting numerical features that reflect users' activity levels, in addition to technical features. Machine learning algorithms took into account the weekly record of student data, allowing for the observation of changes in student behavior over time. The results showed that this approach, especially in the later stages of the course, was able to predict dropout much better than baseline methods, showing an increase in prediction accuracy of up to 15% in certain weeks of the course.

The study by (Kashyap & Nayak, 2018) provided an overview of student dropout prediction models in MOOCs, considering the various factors that affect course completion. The study indicated that data analysis shows a strong relationship between the number of events (such as clicks, video views, and forum posts) and learner success. Machine learning algorithms were applied to a dataset from HarvardX, and the results

showed that random forests achieved the highest performance, reinforcing the idea that tree-based algorithms work particularly well with MOOC data.

Dass et al (2021) tackled the issue of high dropout rates in MOOCs by developing a predictive model using features derived from students' daily learning activities. Employing the Random Forest algorithm and various evaluation metrics, the model achieved 87.5% accuracy and a 94.5% AUC in forecasting student persistence or dropout on a daily basis. The findings highlight the potential for timely identification and intervention to support student retention in online courses.

2.1.3. Predicting Dropout Among International Students

International students represent an important and growing segment in higher education institutions worldwide, facing unique challenges that may increase their likelihood of dropping out. Recent research has examined the factors influencing dropout in this group and the appropriate prediction and intervention strategies.

The study by (Jung & Kim, 2018) proposed an analysis of institutional and regional factors affecting international student dropout rates in higher education institutions in South Korea. The study relied on panel data covering 170 universities from 2013 to 2015, analyzed using regression models (Pooled OLS, Fixed Effect, Random Effect models) to identify the most influential factors on dropout. The results showed that institutional factors such as university type, size, and the proportion of international academics, in addition to regional factors like unemployment rate and regional GDP, play a significant role in explaining dropout rates among international students. This study highlights the importance of institutional and environmental factors, which are often overlooked in traditional dropout prediction models.

The study by (Sulong & Othman, 2020) addressed the phenomenon of international student withdrawal from higher education institutions in Malaysia using a phenomenological approach to identify the factors influencing the decision to withdraw. The study was based on interviews with 24 former international students who transferred to Northern Cyprus to continue their studies after leaving Malaysian universities. The results showed that international student withdrawal was the result of several accumulated negative experiences, including academic difficulties, social isolation, financial factors, cultural differences, lack of academic support, and personal circumstances. The study recommended the need to enhance strategies for managing international students in Malaysia to reduce withdrawal rates. This study underscores

the importance of qualitative and cultural aspects in understanding student dropout, which may be difficult to integrate into traditional machine learning models.

Thies & Falk (2024) conducted a study on the factors influencing academic major change and university dropout among international students in German universities. The study relied on data collected from an online survey of 3,660 international students and used discrete-time analysis of competing risks to examine the relationship between social, academic, and financial factors and both major change and dropout. The results showed that satisfaction with program content, social and academic engagement, and achieving financial balance have significant effects in reducing the risk of changing majors and dropping out. The importance of aligning students' interests with their academic programs was also emphasized. This study provides an in-depth view of the multiple factors influencing international students' decisions to persist, change programs, or drop out entirely, helping to design more accurate and comprehensive predictive models.

2.1.4. School Dropout Studies in Turkish Literature

Turkish studies provide an important perspective on the phenomenon of dropout within a specific cultural and social context, with a focus on various educational stages and types of institutions.

2.1.4.1 School Dropout in Middle School

The study by (Barut, 2024) explored the use of machine learning methods to predict students' academic achievement. The researcher used the "Student Performance" dataset, which includes information on middle school students in Portugal, and applied a variety of machine learning algorithms to analyze the impact of living conditions and social interactions on academic performance. The results showed that the decision tree algorithm achieved the highest classification accuracy (93.3%), while the Catboost algorithm performed best in nonlinear regression (86.9%) and multiple linear regression (82.9%). The study suggests the possibility of developing more effective predictive tools to improve education quality and school resource management. Although this study focuses on predicting academic performance, it provides important insights that can be applied to dropout prediction models, given the close relationship between academic achievement and dropout likelihood.

The study by (Tirpancioğlu, 2019) examined the relationship between dropout risk, school burnout, and perceived social support among middle school students in the Konak district of Izmir. The sample included 903 students (511 females and 392 males) from ten middle schools. The results showed that dropout risk is higher among males, increases with age, grade level, and number of siblings, and decreases with higher parental education and family economic status. The study also found a negative relationship between perceived social support and dropout risk, and a strong positive link between school burnout and dropout risk. The study recommended increasing social support for students, especially from teachers, and reducing school burnout through sports and arts activities to lower dropout risk. This study provides an important psychological perspective for understanding risk and protective factors for school dropout.

2.1.4.2 School Dropout in High School

The study by (Karabacak, 2024) examined the use of machine learning models to predict student dropout in higher education institutions, with an emphasis on the interpretability of these models. The researcher used a dataset from Tecnológico de Monterrey in Mexico, comprising 64,641 university students, and applied embedding techniques and multiple algorithms such as XGBoost and random forests. The results showed that XGBoost and random forest models achieved the best performance, and the LIME algorithm was used to interpret individual predictions. Although embedding techniques did not improve accuracy as expected, the study emphasized the importance of model interpretability for better resource allocation to students at risk of dropping out. This study highlights the importance of developing interpretable machine learning models in the context of dropout prediction, as understanding the reasons behind predictions can help design more targeted and effective interventions.

The study by (Elibol, 2024) investigated the factors influencing student dropout from open and distance education programs using a phenomenological methodology. The researcher conducted semi-structured interviews with 15 individuals who withdrew from distance education programs between 2019 and 2022. The results revealed 16 main factors for dropout, most notably: assessment methods, quality of course materials, life circumstances, registration policies, motivation, course difficulty, and self-organization. The study also identified six emotional outcomes of withdrawal, including indifference, sadness, relief, disappointment, feeling deceived, and awareness.

Recommendations were made to improve student retention by enhancing interaction, improving the quality of educational materials, and providing better student support. This study highlights the qualitative and emotional aspects of the dropout experience, which can be integrated with quantitative data to improve the accuracy of predictive models.

The study by (Binici, 2023) examined the reasons for student dropout from Anatolian vocational and technical high schools and proposed solutions. The researcher used a phenomenological approach and conducted semi-structured interviews with 12 dropout students. The results showed that the reasons for dropout fall into three categories: student-related (such as the desire to work, low motivation, and poor academic achievement), school-related (such as an unsuitable classroom climate, teacher behavior, and strict school rules), and family-related (such as economic status, lack of emotional support, and parental separation). Participants suggested solutions including improving teacher attention to students, making lessons more enjoyable, increasing sports and arts activities, and grouping students according to their abilities and interests. This study provides an important qualitative perspective that can be integrated with quantitative models to enhance our understanding of dropout in vocational schools.

2.1.4.3 School Dropout in Higher Education (University)

The study by (Aguş, 2019) investigated the reasons for student dropout in the open education system at Anadolu University, using a quantitative descriptive approach with 302 students participating in an online survey. The results revealed that lack of time and motivation were the most influential factors in the decision to leave, followed by tuition fees, unmet expectations, and the absence of proper study habits. Factors related to the nature of distance education, such as individual learning and limited communication, were less influential, while technological, family, and health problems were the least significant. This study highlights the importance of personal and organizational factors in predicting dropout, especially in distance learning environments.

The study by (Tosun, 2024) focused on predicting the factors influencing dropout in open and distance education using data mining. The researcher used a quantitative correlational approach and applied data mining algorithms to data from students at the Faculty of Open and Distance Education at Istanbul University. The

results showed that the C5.0 algorithm was the most accurate in predicting student dropout, and that cumulative GPA and year of enrollment were the most influential factors, followed by program type, age, mode of admission, type of education, program level, place of residence, and gender. The study also found that dropout rates were higher in diploma programs (56.6%) compared to bachelor's programs (48.8%), and that males (59%) dropped out more than females (44.8%). This study provides a detailed quantitative analysis of the factors influencing dropout in open and distance education, helping to develop more accurate predictive models.

2.1.5. Challenges Faced by Arab Students Studying Abroad: Comparative Analysis with Yemeni Students in Turkey

The challenges faced by international students from Arab countries studying abroad present striking similarities to those encountered by Yemeni students in Turkey. This section examines relevant research on the experiences of students from Iraq, Syria, and Sudan in foreign educational environments, highlighting common predictive factors for dropout and drawing parallels to the Yemeni context in Turkish universities.

2.1.5.1 Iraqi Students

Iraqi students studying abroad face multifaceted challenges that significantly impact their educational persistence. Study (Maftool, 2023) investigated the difficulties encountered by Iraqi students in international schools in Turkey, particularly focusing on language acquisition. Their research revealed that Iraqi students struggled with dual language barriers—learning both English and Turkish simultaneously—creating significant academic obstacles. The study found that 78% of participants reported language difficulties as their primary challenge, affecting both academic performance and social integration. These linguistic challenges mirror those faced by Yemeni students in Turkish universities, who similarly struggle with acquiring Turkish language proficiency while maintaining academic progress.

2.1.5.2 Syrian Students

Syrian students studying in Turkish universities encounter a unique set of challenges that significantly impact their educational trajectories. Parkinson (2025) conducted a systematic review of research on Syrian students in Turkish higher education, identifying several key challenges that affect their persistence and success.

Language barriers emerged as the most significant obstacle, with limited Turkish proficiency hindering academic performance, classroom participation, and social integration. The review found that "Syrian students encounter diverse challenges within and beyond the university that negatively impact their engagement, attainment, sense of belonging, and psychological wellbeing" highlighting the multidimensional nature of these difficulties.

Economic hardship represents another critical challenge for Syrian students in Turkey. El-Galil (2022) reported that Syrian refugee students in Turkey face significant financial difficulties that threaten their educational persistence. Many students struggle to cover tuition fees, living expenses, and educational materials, forcing them to seek employment that often interferes with their studies. The report emphasized that "economic difficulties according to some Turkish students, Syrian students have great difficulty coping with the cost of living and education in Türkiye" (para.4), highlighting the financial vulnerability of this population. These economic challenges parallel those faced by Yemeni students in Turkish universities, who similarly struggle with financial constraints due to the economic crisis in Yemen.

2.1.5.3 Sudanese Students

The Middle East Studies Association (MESA, 2024) highlighted that the ongoing conflict in Sudan has exacerbated longstanding challenges in the Sudanese educational system, including limited funds, minimal infrastructure, and restricted access to resources. The report emphasized that these challenges have created significant barriers for Sudanese students studying abroad, who often lack financial support from their government or families due to the economic crisis. These economic difficulties mirror those faced by Yemeni students in Turkish universities, who similarly contend with limited financial support due to economic collapse in Yemen.

2.1.5.4 Common Challenges and Parallels with Yemeni Students in Turkey

The reviewed studies from Iraq, Syria, and Sudan reveal striking similarities in the challenges faced by Arab students studying abroad, with clear parallels to the Yemeni student experience in Turkey. Several key themes emerge across these contexts:

1. **Economic Hardship:** Students from all five countries face significant financial challenges while studying abroad, including difficulties covering tuition fees,

living expenses, and educational materials. These economic constraints are often exacerbated by political instability and economic crises in their home countries, creating substantial barriers to educational persistence. These financial challenges mirror those faced by Yemeni students in Turkish universities, who similarly struggle with economic hardship due to Yemen's ongoing crisis.

2. **Language Barriers:** Linguistic challenges emerge as a consistent theme across studies, with students from Arab countries often struggling to acquire proficiency in the language of instruction in their host countries. These language barriers impact academic performance, classroom participation, and social integration, creating significant obstacles to educational success. These linguistic challenges parallel those experienced by Yemeni students in Turkey, who must acquire Turkish language proficiency while maintaining academic progress.
3. **Psychological Distress:** Students from conflict-affected countries (particularly Syria, Iraq, and Sudan) face significant psychological challenges related to displacement, trauma, and concerns about family members remaining in their homeland. These psychological stressors negatively impact academic performance, motivation, and overall wellbeing, increasing dropout risk. These mental health challenges mirror those faced by Yemeni students in Turkish universities, who similarly contend with displacement-related stress and trauma.
4. **Social Integration Difficulties:** Many Arab students studying abroad report challenges related to social integration, including experiences of discrimination, cultural differences, and difficulties forming relationships with local students. These social challenges create feelings of isolation and alienation that negatively impact educational persistence. These integration difficulties parallel those experienced by Yemeni students in Turkey, who similarly struggle with cultural adaptation and social connection.
5. **Documentation and Legal Challenges:** Students from countries experiencing conflict or political instability (particularly Syria, Iraq, and Sudan) often face bureaucratic and legal obstacles related to documentation, residency permits, and legal status. These administrative challenges create additional barriers to educational access and persistence. These documentation issues mirror those faced by some Yemeni students in Turkish universities, who may encounter similar bureaucratic obstacles.

These shared challenges across different Arab nationalities studying abroad provide strong evidence that the predictive model developed in the current study for Yemeni students in Turkey would be applicable to students from other Arab countries in similar contexts. The consistent patterns of economic, linguistic, psychological, social, and administrative challenges suggest that the factors influencing dropout decisions are remarkably similar across these populations, despite their different national origins.

2.1.6. Comparing the Performance of Machine Learning Algorithms in Predicting Student Dropout

Multiple studies indicate variability in the performance of different machine learning algorithms in student dropout prediction tasks, depending on the nature of the data, educational context, educational stage, and other factors. This section presents a comparison of the most prominent algorithms and their performance in various contexts.

- **Decision Tree and Random Forest Algorithms**

Numerous studies highlight the outstanding performance of decision tree and random forest algorithms in student dropout prediction tasks. The study by (Chung & Lee, 2019) showed that the random forest algorithm achieved excellent performance in predicting high school student dropout, with an accuracy of 0.95 and an AUC value of 0.97. The study used data from 165,715 high school students from the National Education Information System (NEIS) in Korea, and the predictive model demonstrated excellent performance across various binary classification metrics.

Rodríguez-Muñiz et al (2019) reinforced these findings with a predictive model using data classification, clustering, and pattern analysis techniques for the early detection of students at risk of dropping out. The model is based on analyzing individual student trajectories and integrating multiple factors such as family environment and school status. The methodology was applied to the Chilean educational system, where the model's results outperformed previous research by 20%, and Gradient Boosting Decision Trees (GBDT) algorithms achieved the best performance. This underscores the effectiveness of tree-based algorithms in handling complex, multidimensional educational data.

The study by (Abouelnour et al., 2024) confirmed this trend by proposing the development of an Early Warning System (EWS) using machine learning algorithms to

predict student dropout in higher education institutions. The study used Random Forest (RF), Decision Tree (DT), and Support Vector Machine (SVM) models, with the Random Forest model achieving the highest accuracy at 93%. Additionally, it was found that current earned credits, cumulative GPA, and the student's financial status were the most influential factors in predicting dropout. These results highlight the importance of using machine learning techniques to analyze educational data and propose targeted intervention strategies to improve student retention rates.

- **Neural Networks and Deep Learning**

Neural networks and deep learning demonstrate outstanding performance in many student dropout prediction applications, especially when dealing with complex data and nonlinear relationships.

Sandoval-Palis et al (2020) developed a neural network model to predict early dropout in the preparatory track at Escuela Politécnica Nacional, using data from 2,097 students and four key variables. The neural network outperformed logistic regression, showing its capability in detecting dropout risk even with limited input variables.

Sulak & Koklu (2024) confirmed the superiority of artificial neural networks in handling complex data, achieving the highest accuracy at 77.3%, outperforming random forests (75.5%) and decision trees (70.1%). The study emphasized that artificial neural networks excel in dealing with complex data, making them the preferred choice in such cases. It recommended improving model performance through hyperparameter tuning, feature reduction techniques, and outlier analysis, which can enhance prediction accuracy and reduce dropout rates.

In the context of e-learning, the study by (Fei & Yeung, 2015) showed that recurrent neural network (RNN) models with long short-term memory (LSTM) cells achieved superior performance in predicting student dropout in MOOCs. These models are distinguished by their ability to process temporal data and capture long-term dependencies, making them particularly suitable for analyzing student behavior over time in e-learning environments.

- **Ensemble Models**

Ensemble models, which combine the advantages of several algorithms, have demonstrated outstanding performance in many student dropout prediction studies.

The study by (Niyogisubizo et al., 2022) proposed a novel ensemble approach combining random forests, XGBoost, gradient boosting, and feedforward neural networks to predict student dropout in university courses. Using data from Constantine the Philosopher University in Nitra (2016–2020), the proposed method outperformed baseline models in test accuracy and AUC metrics. The findings confirm the effectiveness of the stacked ensemble technique in enhancing prediction accuracy by integrating the strengths of multiple algorithms.

The study by (Dewan et al., 2015) introduced a technique for predicting students at risk of dropout in e-learning systems. The technique uses a set of learner features reflecting student activities over time and relies on an ensemble of classifiers—specifically, three machine learning techniques: K-Nearest Neighbor (KNN), Radial Basis Function (RBF) network, and Support Vector Machines (SVM)—to reduce the errors of individual techniques. Validation results showed that the technique is promising for predicting students at risk of dropout. This study highlights the benefit of ensemble models in improving prediction accuracy by compensating for the errors of individual models.

The study by (Romsaiyud et al., 2024) presented a predictive model for student dropout in open universities using Intuitionistic Fuzzy Sets (IFS) and the XGBoost algorithm. The study used data from 268,763 students between 2012 and 2022 and applied data processing techniques such as SMOTE and Borderline SMOTE to balance the distribution. The results showed that XGBoost achieved the highest accuracy at 91% after hyperparameter tuning, outperforming decision trees, Naïve Bayes, gradient boosting, and other models. The study also identified academic major, number of registered courses, and faculty as the most influential factors in dropout. These results underscore the importance of combining machine learning and fuzzy techniques for educational data analysis and improving student retention strategies.

- **Comparison of Algorithm Performance**

Many studies have compared the performance of different machine learning algorithms in predicting student dropout, providing valuable insights into the strengths and weaknesses of each.

The study by (Tin et al., 2024) compared several models—logistic regression (LR), decision trees (DT), support vector machines (SVM), random forests (RF), and K-nearest neighbors (KNN)—for predicting dropout and academic success in higher

education. Using a dataset with educational and socio-economic factors, logistic regression achieved the highest accuracy at 92.01%. This shows that logistic regression, despite its simplicity, can be highly effective in some contexts.

Sreedham (Deb et al., 2024) used academic and demographic data from a major university in Bangladesh to compare SVM, RF, and LR. Random forest performed best, with an accuracy, recall, and F1-score of 0.78, while LR and KNN also gave good results. This highlights the effectiveness of random forests in higher education settings.

Castelino (2024) studied dropout prediction at Rochester Institute of Technology – Dubai, applying random forests, logistic regression, SVM, and neural networks. Logistic regression again performed best. The study confirmed that machine learning helps institutions predict dropout effectively, enabling early intervention to improve student retention.

Table 2.1: Comparative Table of Machine Learning Algorithms for Student Dropout Prediction

Algorithm	Best Use Case	Strengths	Limitations
Decision Tree	Small/medium data, interpretable models	Easy to interpret, handles categorical data, identifies key variables	Prone to overfitting, less accurate with complex data
Random Forest	Large, high-dimensional data	High accuracy, handles missing data, reduces overfitting	Less interpretable, computationally intensive
Gradient Boosting (GBDT, XGBoost)	Large, complex data, imbalanced classes	High predictive power, handles complex relationships, robust to outliers	Requires careful tuning, longer training time
Neural Networks (ANN, RNN, LSTM)	Complex, nonlinear, temporal data	Captures nonlinearities, good for sequential data, high flexibility	Requires large datasets, less interpretable, needs tuning
Support Vector Machine (SVM)	Medium-sized, high-dimensional data	Effective in high-dimensional spaces, robust to overfitting	Not suitable for very large datasets, less interpretable
Logistic Regression	Small/medium data, baseline	Simple, interpretable, fast training	Limited to linear relationships, less effective with complex data
K-Nearest Neighbors (KNN)	Small datasets, simple tasks	Simple, no training phase, interpretable	Slow with large data, sensitive to irrelevant features
Ensemble Models (Stacking, Voting)	Any, especially when combining strengths	Improves accuracy, compensates for weaknesses of single models	More complex, harder to interpret, computationally expensive
Fuzzy Techniques (IFS)	Data with uncertainty	Handles vagueness, combines with ML for better results	Requires domain expertise, less common in practice

2.1.7. Factors Influencing Student Dropout as Identified by Machine Learning Models

A review of studies reveals a variety of factors influencing student dropout, as identified by machine learning models. These factors can be classified into several main categories:

- **Academic Factors**

Academic factors are among the most important indicators of the likelihood of student dropout. According to the study by (Vaarma & Li, 2024), "total earned credits," "number of failed courses," and "level of activity in the learning management system (Moodle)" are among the most influential factors in predicting dropout. The study used three predictive machine learning algorithms for student dropout in Finnish higher education, utilizing data from academic records, demographic information, and the Moodle learning management system. The study highlighted the importance of learning management system data in predicting dropout, alongside academic records and demographic data. It also showed that the significance of these factors changes over time, necessitating continuous temporal analysis to improve prediction accuracy and implement appropriate interventions at critical times.

The study by (Abouelnour et al., 2024) emphasized the importance of current earned credits and cumulative GPA as influential factors in dropout prediction. The study by (Kemper et al., 2020) found that performance in first-semester exams can be a strong predictor of dropout, with classification accuracy exceeding 83% after the first semester. (Dekker et al., 2009) also indicated that the grade in the linear algebra course was the strongest predictor of success, along with grades in calculus, networks, and the average grades in science courses.

- **Socio-Economic Factors**

Socio-economic conditions substantially influence whether students persist or withdraw from university. Kim & Kim (2018) analyzed the effects of financial burden, resource availability, faculty quality, and university characteristics on dropout rates in South Korea. Using balanced panel data from 2013 to 2015 and nonlinear panel data models, their findings confirmed that these socio-economic variables significantly impact student retention.

The study by (Abouelnour et al., 2024) indicated that a student's financial status is an influential factor in dropout prediction. Piscitello et al (2022) also demonstrated the relationship between socio-demographic factors (such as income, gender, age, and race) and high school dropout, focusing on the role of school disengagement as a mediator in this relationship. Data from 10,000 students, including 537 from rural areas, were analyzed. The results showed that socio-demographic risks significantly affect dropout in both groups. However, school disengagement was a strong mediator among non-rural students, while academic performance was the main mediator among rural students. The study recommends designing preventive interventions based on the socio-environmental context of each group.

- **Behavioral and Interactional Factors**

Behavioral and interactional factors, especially in e-learning environments, are important indicators for predicting dropout. The study by (Kloft et al., 2014) highlighted the significance of student clickstream data in predicting dropout in MOOCs. (Kashyap & Nayak, 2018) also confirmed a significant association between the frequency of activities (such as clicks, video views, and forum posts) and learner success in completing online courses.

The study by (Vaarma & Li, 2024) showed that "level of activity in the learning management system (Moodle)" acts as a significant predictor of dropout. This factor reflects the extent of student engagement with educational content, activities, and assessments on the platform, providing a strong indicator of academic involvement and commitment.

- **Institutional and Environmental Factors**

Institutional and environmental factors significantly influence students' decisions to persist or drop out. Jung & Kim (2018) showed that institutional factors such as university type, size, and the proportion of international academics, as well as regional factors like unemployment rate and regional GDP, play an important role in explaining dropout rates among international students in South Korea.

Alban & Mauricio (2019) identified the factors influencing dropout prediction across five dimensions, including the institutional dimension. Their systematic review of the literature on university student dropout prediction using data mining techniques analyzed 1,681 relevant initial studies and selected 67 documents based on specific

inclusion and exclusion criteria. The study classified influential factors into five dimensions: personal, academic, economic, social, and institutional. It also identified ten preprocessing techniques, ten feature selection techniques, and fourteen dropout prediction techniques (classified as statistical and artificial intelligence techniques). The study found that AI techniques offer higher accuracy rates, but no single technique is clearly superior, as prediction accuracy mainly depends on context, data, and the characteristics of the technique used.

- **Comprehensive Classification of Influential Factors**

The study by (Alban & Mauricio, 2019) provided a comprehensive classification of factors influencing student dropout, dividing them into five main dimensions:

1. **Personal factors:** including gender, age, marital status, mental health, motivation, cognitive abilities, and personality traits.
2. **Academic factors:** covering previous and current academic performance, learning strategies, attendance rate, academic engagement, and field of study.
3. **Economic factors:** including family income, the student's financial status, working while studying, and financial aid.
4. **Social factors:** such as family background, parents' educational level, peer influence, and social support.
5. **Institutional factors:** including the type, size, and location of the institution, quality of teaching, and available student services.

These factors interact in complex ways to influence a student's decision to persist or drop out, highlighting the importance of a comprehensive approach in designing dropout prediction models and developing intervention strategies.

Table 2.2: Comparative Table of Factors Influencing Student Dropout as Identified by Machine Learning Models

Factor Category	Example Factors	Supporting Studies	ML Techniques Used
Academic Factors	Total earned credits, failed courses, LMS activity, first-semester grades	Vaarma & Li (2024), Abouelnour et al. (2024), Kemper et al. (2020), Dekker et al. (2009)	Decision Tree, Random Forest, Neural Networks
Socio-Economic Factors	Financial status, family income, gender, age, race	Kim & Kim (2018), Abouelnour et al. (2024), Piscitello et al. (2022)	Random Forest, Logistic Regression, SVM

Behavioral/Interactional Factors	LMS activity level, clickstream data, forum participation	Kloft et al. (2014), Kashyap & Nayak (2018), Vaarma & Li (2024)	Neural Networks, RNN, Decision Tree
Institutional/Environmental Factors	University type, size, location, teaching quality, regional GDP	Jung & Kim (2018), Alban & Mauricio (2019)	Data Mining, AI Techniques, Ensemble Models
Personal Factors	Gender, age, marital status, mental health, motivation	Alban & Mauricio (2019)	Various ML and AI techniques
Social Factors	Family background, parents' education, peer influence	Alban & Mauricio (2019), Piscitello et al. (2022)	Various ML and AI techniques

2.1.8. Recent Trends in Student Dropout Prediction

- **Early Warning Systems**

Recent studies indicate a growing trend toward the use of early warning systems (EWS) for predicting student dropouts. The study by (Lee & Chung, 2019) aimed to enhance EWS performance by addressing class imbalance through SMOTE and employing ensemble machine learning methods. They trained RF, BDT, RF with SMOTE, and BDT with SMOTE models on data from 165,715 high school students in South Korea's NEIS. Based on ROC and PR curve evaluations, the BDT model demonstrated the best performance. EWS enable timely interventions, thereby improving student retention efforts.

The study by (Awedh & Mueen, 2025) proposed the use of machine learning algorithms for the early identification of students at risk of academic dropout. Aimed at enhancing student performance and reducing dropout rates, the study utilized data from King Abdulaziz University. It applied exploratory data analysis (EDA) to identify patterns, employed a genetic algorithm (GA-LDA) for feature extraction, and used Canopy clustering with a Gaussian Flow Optimizer (GFO) for precise student grouping.

- **Multi-Source Data Analysis**

There is a growing trend toward integrating data from multiple sources to improve the accuracy of student dropout prediction. The study by (Vaarma & Li, 2024) highlighted the importance of combining data from academic records, demographic information, and the learning management system (Moodle) for predicting student

dropout. This integrated approach provides a more comprehensive view of risk factors that may lead to dropout and enhance the predictive power of models.

Romsaiyud et al (2024) provided an example of this trend by developing a predictive model for student dropout in open universities using Intuitionistic Fuzzy Sets (IFS) and the XGBoost algorithm. The study used data from 268,763 students between 2012 and 2022, combining various types of academic and demographic data. These results emphasize the importance of integrating machine learning and fuzzy techniques to analyze diverse educational data and improve student retention strategies.

- **Temporal Analysis and Dynamic Prediction**

There is increasing interest in models that account for temporal changes in student behavior and performance. The study by (Vaarma & Li, 2024) indicated that the importance of dropout risk factors changes over time, requiring continuous temporal analysis to improve prediction accuracy and implement timely interventions at critical moments. This dynamic approach provides a better understanding of how dropout risks evolve throughout a student's academic journey and enables the design of targeted interventions at key points.

Fei & Yeung (2015) provided an example of this trend by using recurrent neural network (RNN) models with long short-term memory (LSTM) cells to predict student dropout in MOOCs. These models are distinguished by their ability to process temporal data and capture long-term dependencies, making them particularly suitable for analyzing changes in student behavior over time.

2.1.9. Summary and Future Directions

The literature review reveals significant progress in the use of machine learning for predicting student dropouts. Starting from simple algorithms such as Naïve Bayes and decision trees in early studies, to advanced models like deep learning and ensemble methods in recent research, these techniques have shown continuous improvement in accuracy and generalizability. Application areas have also expanded to cover various educational stages and contexts, including traditional universities, e-learning, open and distance education, and secondary and middle schools.

Current trends indicate:

1. An increasing reliance on ensemble models that combine the advantages of multiple algorithms, improving prediction accuracy and reducing errors of individual models.
2. A focus on early warning systems for timely intervention, enabling educational institutions to provide necessary support to at-risk students before they decide to drop out.
3. Integration of diverse data sources—including academic, demographic, behavioral, and institutional data—to create a more comprehensive picture of risk factors.
4. Emphasis on model interpretability and understanding influential factors to guide interventions and enhance institutional trust in using machine learning models.
5. Development of models that account for temporal changes in student behavior and performance, enabling dynamic prediction and intervention at critical moments.
6. Growing attention to psychological and social factors influencing dropout, which may be difficult to measure using traditional methods.

Despite significant progress in this field, important challenges remain, including:

1. The generalizability of models across different educational contexts, as the factors influencing dropout may vary by culture, educational system, and academic level.
2. Addressing the issue of data imbalance in dropout cases, since the proportion of students who drop out is often lower than those who persist, which affects model accuracy.
3. Ethical and privacy considerations when using student data, especially with the increasing collection of personal and behavioral information.
4. Developing effective intervention strategies based on predictions, as predicting dropout is only the first step and must be followed by appropriate actions to actually prevent dropout.
5. Integrating qualitative data with quantitative data to improve understanding of dropout, especially regarding psychological, social, and cultural aspects.

This review provides a solid foundation for understanding the current capabilities and limitations of machine learning techniques in predicting student

dropout, paving the way for further research and development in this vital field. Future research can focus on developing more comprehensive models that integrate various types of data and consider different cultural and educational contexts, with an emphasis on interpretability and practical application in educational institutions.

3. MATERIAL AND METHOD

3.1. Research Design

This study employs a quantitative, applied, and descriptive research approach to predict school dropout among foreign students in Turkey, with a specific focus on Yemeni students. The quantitative nature of the research is reflected in the collection and analysis of numerical data to identify patterns and relationships between various factors and student dropout rates. As an applied research endeavor, this study aims to address a real-world educational challenge by developing practical machine learning models that can identify at-risk students. The descriptive component involves a detailed characterization of the demographic, academic, and social factors associated with student retention and attrition.

The overall research methodology follows a systematic process that begins with data collection through surveys, proceeds through data preprocessing and exploratory analysis, and culminates in the development and evaluation of machine learning models. This process is illustrated in (Figure 3.1). This approach aligns with the established frameworks for educational data mining (EDM) and learning analytics, which have been successfully applied to similar educational challenges (Baker et al., 2016; Romero & Ventura, 2020).

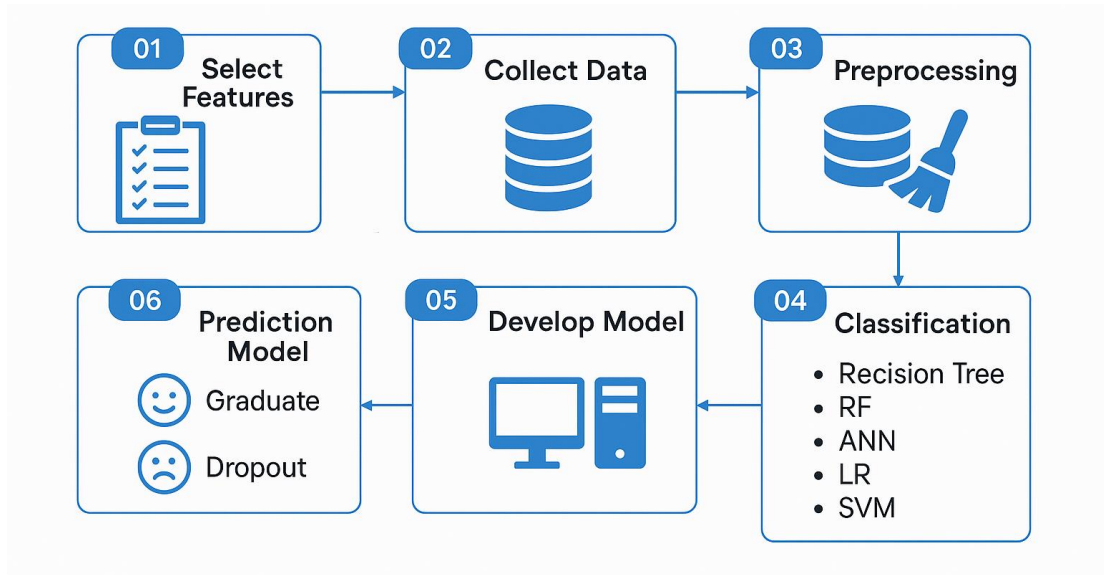


Figure 3.1: Model Design

The research design incorporates both retrospective and predictive elements. The retrospective component examines historical data from students who have already graduated or dropped out, while the predictive aspect focuses on developing models that can forecast the likelihood of current students discontinuing their education. This dual approach enhances the practical utility of the research findings for educational institutions seeking to implement early intervention strategies.

3.2 Population and Sample

The study population consists of Yemeni students pursuing higher education in various universities across Turkey. This specific demographic was selected due to the significant presence of Yemeni students in Turkish higher education institutions and the unique challenges they face as international students adapting to a new educational and cultural environment.

The total sample analyzed in this study comprises 545 Yemeni students, representing a diverse cross-section of academic disciplines, educational levels, and demographic characteristics. The distribution of students according to their academic status is as follows:

- 268 continuing students (49.2%)
- 128 graduated students (23.5%)
- 149 dropped out students (27.3%)

This distribution provides a balanced representation of all three possible academic outcomes, which is essential for developing robust predictive models. The sample size is sufficient for the application of machine learning techniques, as it exceeds the minimum threshold recommended for educational data mining studies involving multiple predictive variables (Delen, 2010; Tinto, 2017).

The sampling strategy employed a combination of convenience and purposive sampling techniques. Initial participants were recruited through university international student offices, Yemeni student associations, and social media platforms. Additional participants, particularly those who had dropped out, were identified through referrals from initial participants using a snowball sampling approach. However, it is important to note that using convenience and snowball sampling may introduce some bias, as the sample might not fully represent all Yemeni students. This multi-faceted recruitment strategy was necessary to ensure adequate representation of students who had discontinued their studies, as they are typically more difficult to reach through institutional channels alone. Ethical approval for data collection was obtained from the relevant ethics committee prior to conducting the study.

3.3 Data Collection Tool

3.3.1 Survey Structure and Content

The primary data collection instrument was a comprehensive survey designed to capture a wide range of factors potentially associated with student dropout. The survey was structured into five main sections:

1. **Personal and Demographic Information:** This section collected data on gender, age, and other basic demographic characteristics.
2. **Academic Background:** Questions in this section focused on students' previous academic performance, current grade point average (GPA), Turkish language proficiency level, and satisfaction with their academic program.
3. **Economic Factors:** This section gathered information about family economic status, scholarship or financial aid receipt, and employment status during studies.
4. **Social and Personal Factors:** This component explored cultural adaptation challenges, parental education levels, parental employment status, alignment

between educational expectations and experiences, and psychological pressures affecting academic performance.

5. **Current Status and Dropout Reasons:** The final section identified the student's current academic status (continuing, graduated, or dropped out) and, for those who had discontinued their studies, explored the primary reasons for their decision.

The survey was initially developed in Arabic to ensure clear comprehension by the target population and was designed to be completed in approximately two minutes to maximize response rates. Prior to data collection, the survey instrument and procedures received approval from the relevant ethics committee.

3.3.2 Data Collection Phases

The data collection process was conducted in two distinct phases:

1. **Initial Phase:** The survey was distributed electronically through various channels, including university international student offices, Yemeni student associations, and social media platforms. This phase yielded responses primarily from continuing and graduated students.
2. **Supplementary Phase:** To increase representation of students who had dropped out, a targeted outreach was conducted using referrals from initial participants. This snowball sampling approach was essential for reaching former students who were no longer connected to institutional communication channels.

3.3.3 Challenges in Data Collection

Several challenges were encountered during the data collection process:

1. **Access to Dropout Students:** Locating and engaging students who had already left their educational institutions proved particularly challenging, necessitating the supplementary collection phase and snowball sampling approach.
2. **Recall Bias:** Students who had dropped out or graduated some time ago occasionally demonstrated difficulty in accurately recalling specific details about their educational experience, potentially introducing some recall bias into the data.
3. **Self-Reporting Accuracy:** As with any self-reported data, there is potential for social desirability bias, particularly regarding academic performance or reasons for discontinuing studies.

These challenges were addressed through careful data validation procedures and by triangulating responses across multiple related questions to identify inconsistencies.

3.4 Data Preprocessing

Information processing is an essential step in preparing the raw survey information for analysis and model development. The preprocessing pipeline, as illustrated in Figure 3.2, consisted of several sequential stages designed to convert the raw information into a structure appropriate for artificial intelligence algorithms.

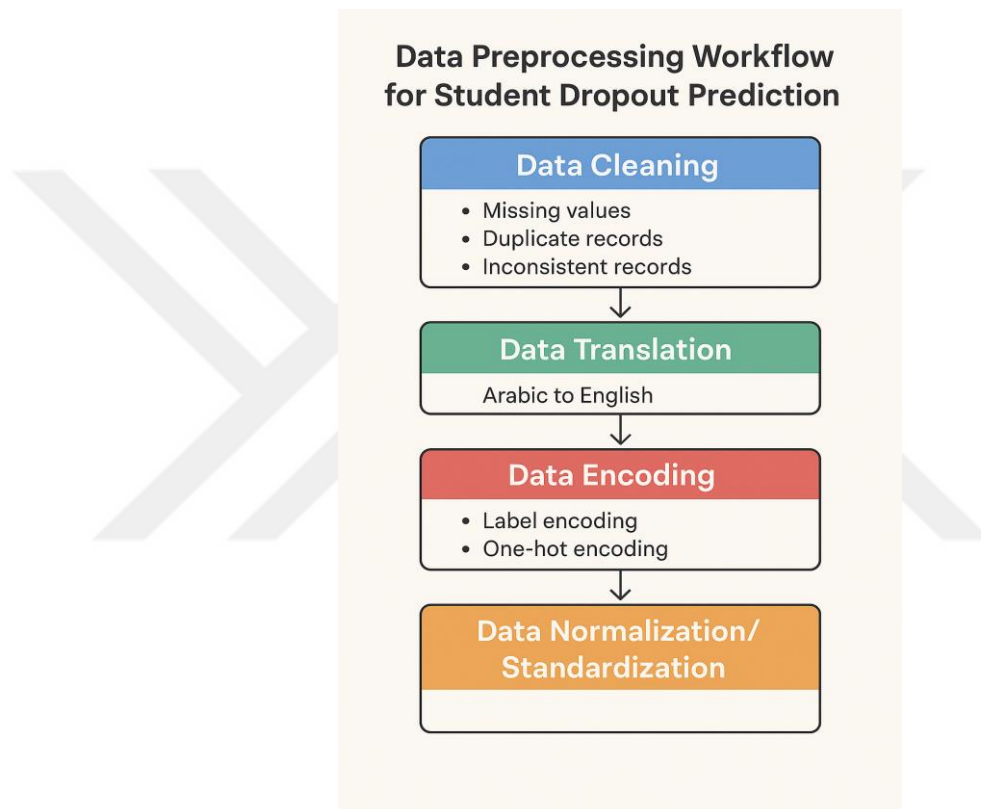


Figure 3.2: Data Preprocessing Workflow for Student Dropout Prediction

3.4.1 Data Cleaning

The initial data cleaning process addressed three primary issues:

1. **Duplicate Records:** The dataset was examined for duplicate entries, which could occur if participants submitted the survey multiple times. Duplicates were identified based on a combination of demographic and academic information, and only the most recent or most complete submission was retained.
2. **Inconsistent Records:** Logical inconsistencies in responses (e.g., contradictory answers to related questions) were identified through cross-validation of related variables. Records with irreconcilable inconsistencies were either corrected

based on other available information or removed from the dataset if corrections could not be reliably made.

The data cleaning process reduced the initial dataset from 583 responses to the final 545 records used in the analysis, ensuring a higher quality dataset for subsequent modeling.

3.4.2 Data Translation

Since the survey was administered in Arabic to maximize comprehension and response accuracy among Yemeni students, a translation process was necessary to standardize the data for analysis. All text-based responses, including open-ended answers regarding dropout reasons, were translated from Arabic to English. This translation was performed using a combination of automated translation tools and manual verification by bilingual researchers to ensure accuracy and preservation of nuanced meanings.

Categorical response options were mapped to standardized English terminology consistent with educational research literature, facilitating comparison with international studies on student retention and dropout prediction.

3.4.3 Data Encoding

To prepare the categorical variables for machine learning algorithms, two primary encoding techniques were employed:

1. **Label Encoding:** For ordinal categorical variables with a clear hierarchical structure (e.g., education levels, language proficiency levels), label encoding was applied to convert the categories into numerical values that preserve the ordinal relationship. For example, language proficiency levels were encoded as: Poor = 1, Acceptable = 2, Good = 3, Very Good = 4, Excellent = 5.

Table 3.1: Encoding Methods and Value Mapping for Study Variables

Variable	Encoding Method	Explanation/Values
Gender	Label Encoding	0 = woman, 1 = man
Age	Label Encoding	0 = Under 20, 1 = 20-24, 2 = 25-29, 3 = 30+
Last_GPA	Label Encoding	0 = 0-49, 1 = 50-59, 2 = 60-69, 3 = 70-84, 4 = 85-100

Turkish_Proficiency	Label Encoding	0 = Poor, 1 = Acceptable, 2 = Good, 3 = Very good, 4 = Excellent
Major_Satisfaction	Label Encoding	0 = Very Dissatisfied, 1 = Dissatisfied, 2 = Neutral, 3 = Satisfied, 4 = Very Satisfied
Family_Economy	Label Encoding	0 = Low, 1 = Medium, 2 = High
Scholarship	Label Encoding	0 = No, 1 = Yes
Part_Time_Work	Label Encoding	0 = No, 1 = Yes, part time, 2 = Yes, full time
Cultural_Adjustment to_Turkey	Label Encoding	0 = No, 1 = Yes
Father_Education_Level	Label Encoding	0 = No formal, 1 = Basic, 2 = High school, 3 = Bachelor's, 4 = Master's/PhD
Mother Education_Level	Label Encoding	0 = No formal, 1 = Basic, 2 = High school, 3 = Bachelor's, 4 = Master's/PhD
Parents Employment_Status	Label Encoding	0 = Both not working, 1 = Mother working, 2 = Father working, 3 = Both working
Education Experience vs_Expectations	Label Encoding	0 = No, 1 = Yes
Psychological Stress_Impact	Label Encoding	0 = No, 1 = Yes
Target	Label Encoding	0 = Dropout, 1 = Graduated, 2 = Enrolled

2. **One-Hot Encoding:** For nominal categorical variables without an inherent order (e.g., gender, reasons for dropout), one-hot encoding was used to create binary columns for each category. This approach prevents the algorithm from incorrectly inferring an ordinal relationship between categories.

The encoding process transformed the categorical data into a numerical format while preserving the semantic meaning of the original variables, making them suitable for machine learning algorithms that require numerical inputs.

3.4.4 Data Normalization/Standardization

To ensure that variables with different scales did not disproportionately influence the machine learning models, numerical features were standardized using the z-score normalization technique:

$$z = \frac{x - \mu}{\sigma} \dots\dots\dots (1)$$

Where:

- Z is the standardized value
- X is the original value
- μ is the mean of the feature
- σ is the standard deviation of the feature

This standardization process transformed all numerical features to have a mean of 0 and a standard deviation of 1, placing them on a comparable scale. This step is particularly important for algorithms sensitive to feature scaling, such as Support Vector Machines and neural network-based approaches.

3.5 Modeling Algorithms

This study employed a comprehensive set of machine learning algorithms to predict student dropout. Each algorithm brings unique strengths and characteristics to the classification task. The selection of these algorithms was based on previous research in student dropout prediction, which has demonstrated their effectiveness in similar educational contexts. The following sections provide detailed explanations of each algorithm used in this research.

3.5.1 Logistic Regression

Logistic Regression is a statistical model that uses a logistic function to model the probability of a certain class or event.

Conceptual Explanation: Despite its title, Logistic Regression is a categorization algorithm, not a regression algorithm. It estimates the likelihood that a specific instance belongs to a particular class. The model applies the logistic function to a linear combination of features to transform the output to a probability between 0 and 1. The structure of a typical logistic regression model is illustrated in Figure 3.3

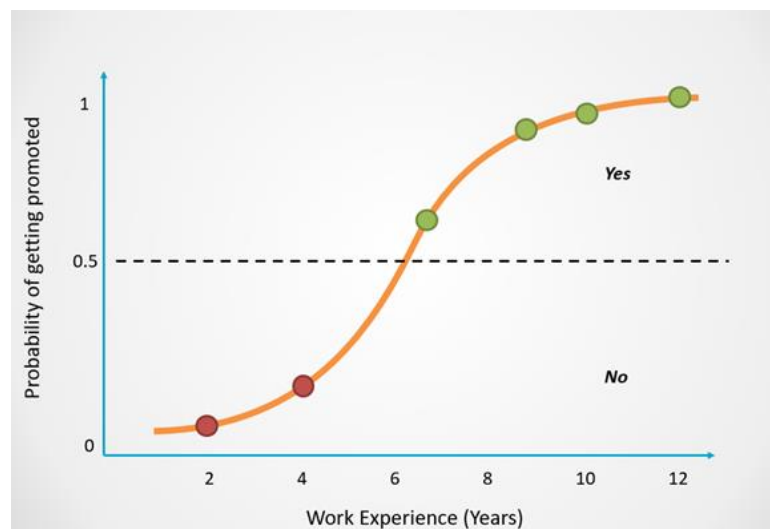


Figure 3.3: Logistic Regression Model (Selvaraj, 2022)

Mathematical Expression: The logistic function (sigmoid) transforms a linear predictor into a probability:

$$P(y = 1|x) = \frac{1}{1+e^{-(\beta_0 + \beta_1 x_1 + \dots + \beta_n x_n)}} = \frac{1}{1+e^{-\beta^T x}} \dots\dots\dots (2)$$

Where:

- $P(y=1|x)$ is the probability that the instance belongs to class 1
- $\beta_0, \beta_1, \dots, \beta_n$ are the model coefficients
- $x_1, x_2, \dots, x_n$ are the feature values

Typical Use Cases: Logistic Regression is widely used in various fields including medical research, social sciences, marketing, and finance. In educational research, it has been applied to predict student retention, analyze factors affecting academic performance, and model educational choices(Hosmer Jr et al., 2013).

Advantages:

- Highly interpretable (coefficients indicate feature importance and direction)
- Provides probability scores rather than just classifications
- Efficient to train and performs well when the decision boundary is approximately linear
- Less prone to overfitting in high-dimensional spaces
- Works well with mixed categorical and continuous features

Limitations:

- Assumes linear relationship between features and log-odds
- May underperform when decision boundaries are highly non-linear
- Sensitive to outliers
- Can struggle with imbalanced datasets
- Multicollinearity can affect coefficient interpretation

3.5.2 Random Forest

Random Forest is an ensemble learning approach that creates multiple decision trees during training and outputs the class that is the mode of the classes predicted by individual trees.

Conceptual Explanation: Random Forest constructs multiple decision trees using bootstrap samples of the learning data and random feature subsets at each split. This variability helps to create diverse trees that capture different aspects of the information. The final prediction is determined by combining the predictions of all trees, typically through majority voting for classification tasks.

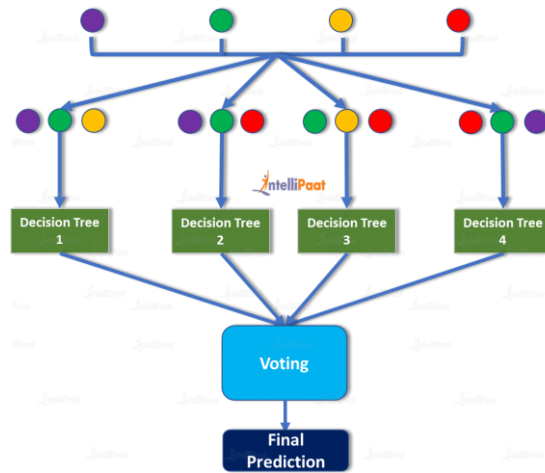


Figure 3.4: Random Forest Architecture (Intellipaat, 2023)

Mathematical Expression: For a classification problem with C classes, the Random Forest prediction can be expressed as:

$$\hat{y} = \operatorname{argmax}_c \sum_{b=1}^B I(h_b(x) = c) \dots\dots\dots(3)$$

Where:

- $\hat{y}$ is the predicted class
- B is the number of trees in the forest
- $h_b(x)$ is the prediction of the b -th tree
- I is the indicator function
- C represents each possible class

Typical Use Cases: Random Forest is widely used in various domains including ecology, bioinformatics, and social sciences. In educational contexts, it has been applied to predict student performance, identify at-risk students, and analyze factors affecting educational outcomes (Biau & Scornet, 2016).

Advantages:

- Robust to outliers and noisy data
- Provides estimates of feature importance
- Less prone to overfitting compared to single decision trees
- Handles high-dimensional data effectively
- Maintains good accuracy even when a large proportion of data is missing

Limitations:

- Less interpretable than single decision trees
- Can be computationally intensive for very large datasets

- May struggle with highly imbalanced datasets without proper adjustments
- Tends to bias towards features with more levels in categorical variables

3.5.3 SVM (Support Vector Machine)

Support Vector Machine is a supervised learning algorithm that discovers the optimal hyperplane that maximizes the margin between classes in the feature space.

Conceptual Explanation: SVM works by locating the hyperplane that best separates the classes with the maximum margin. The "support vectors" are the data points that lie nearest to the decision surface and directly affect its position and orientation. SVM can also manage non-linear classification using kernel functions that implicitly map inputs into high-dimensional feature spaces.

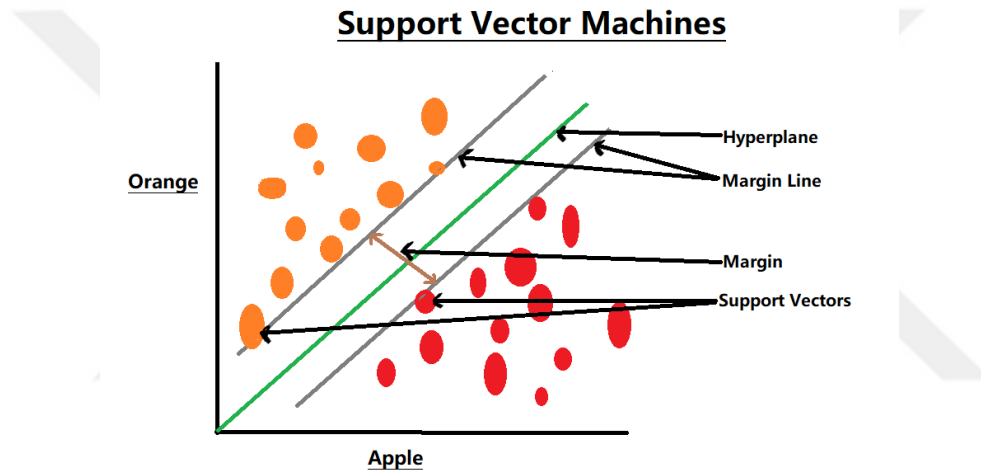


Figure 3.5: Support Vector Machine (SVM) (Janbandhu, 2024)

Mathematical Expression: The decision function for a linear SVM is:

$$f(x) = \text{sign}(w^T x + b) \dots\dots\dots (4)$$

For kernel SVM, the decision function becomes:

$$f(x) = \text{sign} \left(\sum_{i=1}^n \alpha_i y_i K(x_i, x) + b \right) \dots\dots\dots (5)$$

Where:

- w is the weight vector
- b is the bias term
- α_i are the Lagrange multipliers
- y_i are the class labels
- $K(x_i, x)$ is the kernel function

- x_i are the support vectors

Typical Use Cases: SVM is widely used in text classification, image recognition, bioinformatics, and handwriting recognition. In educational data mining, it has been applied to predict student performance, classify learning behaviors, and identify at-risk students (Noble, 2006).

Advantages:

- Effective in high-dimensional spaces
- Memory efficient as it uses only a subset of training points (support vectors)
- Versatile through different kernel functions
- Robust against overfitting, especially in high-dimensional space
- Works well when classes are separable

Limitations:

- Does not directly provide probability estimates
- Can be computationally intensive for large datasets
- Sensitive to the choice of kernel and regularization parameters
- Difficult to interpret compared to simpler models
- Performance can degrade with overlapping classes

3.5.4 XGBoost

XGBoost (Extreme Gradient Boosting) is an advanced implementation of gradient boosting machines that has gained significant popularity in machine learning competitions and practical applications due to its exceptional performance.

Conceptual Explanation: XGBoost constructs an ensemble of decision trees sequentially, with each new tree correcting the errors made by the combination of existing trees. It follows the gradient boosting principle where each subsequent tree is trained to reduce the residual errors of the previous ensemble.

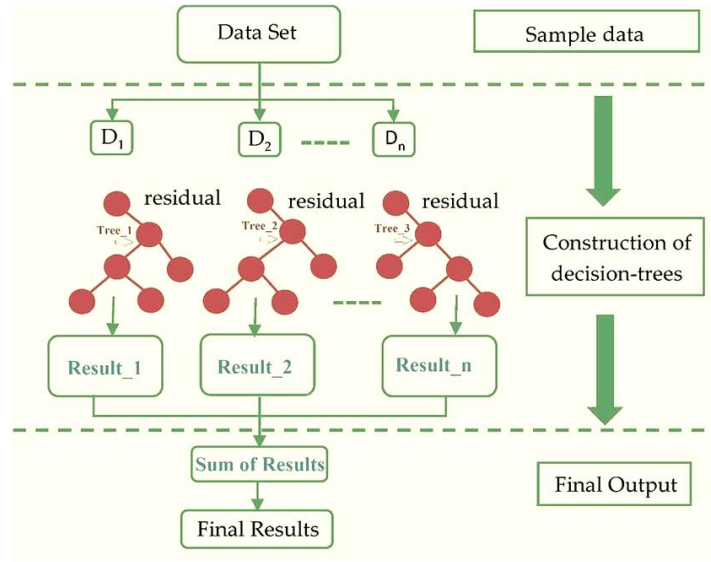


Figure 3.6: XGBoost Architecture (Ahmed et al., 2023)

Mathematical Expression: The prediction from an XGBoost model with K trees can be expressed as:

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i) \dots\dots\dots (6)$$

Where:

- $\hat{y}_i$ is the predicted output for the i-th instance
- f_k represents the k-th decision tree
- x_i is the feature vector of the i-th instance

The objective function being optimized combines a loss function and a regularization term:

$$\text{Obj} = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k) \dots\dots\dots (7)$$

Where:

- L is the loss function (e.g., logistic loss for classification)
- Ω is the regularization term that penalizes complex models

Typical Use Cases: XGBoost is particularly effective for structured/tabular data problems where interpretability and predictive performance are both important. It has been effectively utilized in diverse domains including finance, healthcare, and education for tasks such as credit scoring, disease prediction, and student performance forecasting (Chen & Guestrin, 2016).

Advantages:

- High predictive accuracy
- Handles missing values natively
- Built-in regularization to prevent overfitting
- Efficient implementation for both training and inference
- Handles imbalanced datasets effectively

Limitations:

- More complex and less interpretable than single decision trees
- Requires careful tuning of hyperparameters
- Can be computationally intensive for very large datasets
- May overfit on noisy data if not properly regularized

3.5.5 Decision Tree

Decision Tree is a non-parametric supervised learning approach used for classification and regression that builds a model that forecasts the value of a target variable by learning simple decision rules derived from the data features (Loh, 2011).

Conceptual Explanation: A Decision Tree iteratively partitions the feature space into regions, with each internal node representing a "test" on a feature, each branch representing the outcome of the test, and each leaf node representing a class label or value. The path from root to leaf represents a classification rule.

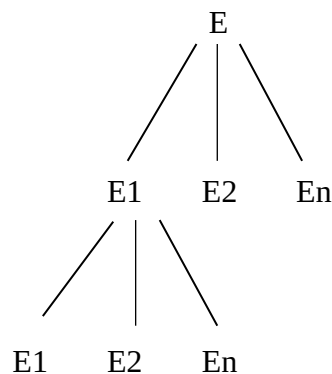


Figure 3.7: Decision Tree Architecture

Mathematical Expression: The splitting criterion for classification trees is typically either Gini impurity or information gain. For Gini impurity:

$$\text{Gini}(t) = 1 - \sum_{i=1}^C p(i|t)^2 \dots\dots\dots (8)$$

Where:

- $Gini(t)$ is the Gini impurity at node t
- $p(i|t)$ is the proportion of samples that belong to class i at node t
- C is the number of classes

Typical Use Cases: Decision Trees are used in various domains including medicine (diagnosis), finance (credit approval), and education (student performance prediction). They are particularly valuable when interpretability is important.

Advantages:

- Highly interpretable and easy to visualize
- Requires little data preparation (no normalization needed)
- Can handle both numerical and categorical data
- Automatically performs feature selection
- Handles non-linear relationships well

Limitations:

- Prone to overfitting, especially with deep trees
- Can be unstable (small variations in data might result in a completely different tree)
- Biased toward features with more levels in categorical variables
- May create biased trees if classes are imbalanced
- Struggles with diagonal decision boundaries

3.5.6 Naive Bayes

Naive Bayes is a probabilistic classifier based on implementing Bayes' theorem with strong (naive) independence assumptions between features.

Conceptual Explanation: Naive Bayes computes the probability of each class given the feature values, assuming that features are conditionally independent given the class. Despite this simplifying assumption, which is seldom true in practice, Naive Bayes often performs surprisingly well, especially for text classification and when the training data is limited (Zhang, 2004).

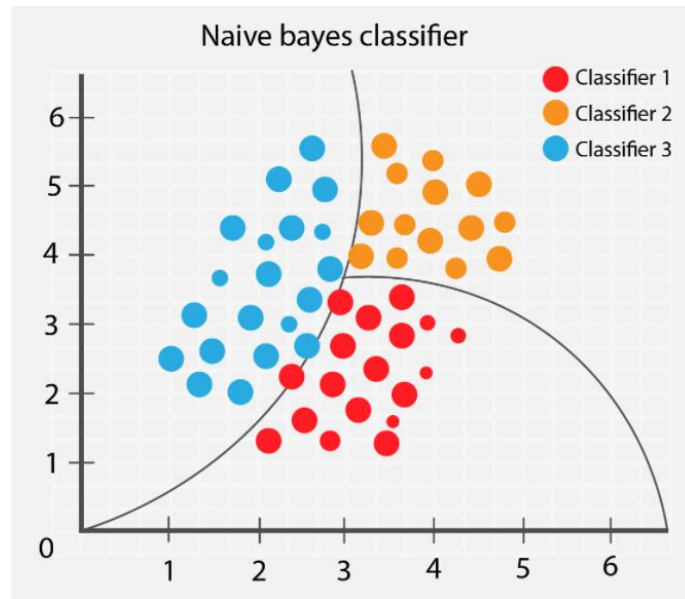


Figure 3.8: Naive Bayes Classifier (Sreeramaneni, 2022)

Mathematical Expression: Using Bayes' theorem, the posterior probability is:

$$P(C_k|x) = \frac{P(C_k) \prod_{i=1}^n P(x_i|C_k)}{P(x)} \dots\dots\dots (9)$$

Where:

- $P(C_k|x)$ is the posterior probability of class C_k given features X
- $P(C_k)$ is the prior probability of class C_k
- $P(X_i|C_k)$ is the likelihood of feature X_i given class C_k
- $P(x)$ is the evidence (marginal likelihood)

Typical Use Cases: Naive Bayes is widely used in text classification, spam filtering, sentiment analysis, and recommendation systems. In educational contexts, it has been applied to classify student behaviors, predict performance, and analyze feedback.

Advantages:

- Simple and fast to train
- Works well with small training datasets
- Handles high-dimensional data effectively
- Not sensitive to irrelevant features
- Provides probabilistic predictions

Limitations:

- Assumes feature independence, which is often violated in real-world data

- Can be outperformed by more sophisticated models when training data is abundant
- May produce poor probability estimates
- Struggles with feature interactions
- "Zero frequency" problem when a categorical variable has a category not observed in training

3.5.7 LightGBM

LightGBM is a gradient boosting framework that employs tree-based learning algorithms with optimizations for efficiency and scalability.

Conceptual Explanation: LightGBM is similar to other gradient boosting frameworks but presents two key innovations: Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB). It also uses a leaf-wise tree expansion strategy rather than level-wise expansion, which can result in improved accuracy with fewer splits (Ke et al., 2017).

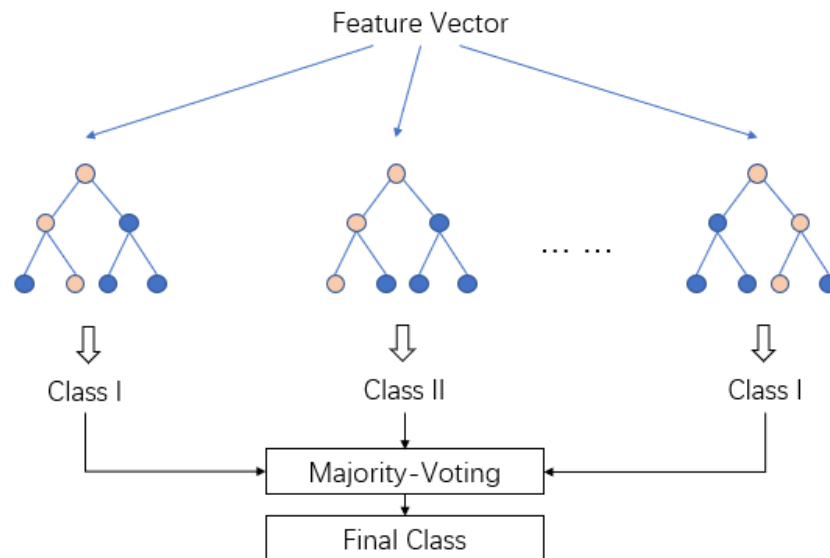


Figure 3.9: LightGBM Architecture Compared to Traditional Boosting (Kılıç, 2022)

Mathematical Expression: Similar to XGBoost, the prediction from a LightGBM model can be expressed as:

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i) \dots\dots\dots (10)$$

The key difference lies in how the trees f_k are constructed and optimized.

Typical Use Cases: LightGBM is particularly valuable for large-scale machine learning problems where computational efficiency is important. It has been applied in recommendation systems, click-through rate prediction, and various classification and regression tasks in education, finance, and healthcare.

Advantages:

- Faster training speed and higher efficiency than many other boosting algorithms
- Lower memory usage due to histogram-based algorithms
- Supports categorical features directly without one-hot encoding
- Handles large datasets effectively
- Often achieves better accuracy than other boosting algorithms

Limitations:

- May overfit on small datasets if not properly regularized
- Leaf-wise growth can lead to deeper trees that might overfit
- Requires careful parameter tuning
- Less interpretable than simpler models
- Documentation and community support are not as extensive as some older algorithms

3.5.8 LDA (Linear Discriminant Analysis)

Linear Discriminant Analysis is a classification approach that transforms features onto a lower-dimensional space while maximizing the separation between classes.

Conceptual Explanation: LDA presumes that different classes generate data based on different Gaussian distributions with the same covariance structure. It finds linear combinations of features that optimize the ratio of between-class variance to within-class variance, creating a linear decision boundary that best distinguishes the classes (McLachlan, 2005).

Mathematical Expression: The discriminant function for LDA is:

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log(\pi_k) \dots\dots\dots(11)$$

Where:

- $\delta_k(x)$ is the discriminant score for class k
- x is the feature vector
- μ_k is the mean vector for class k

- Σ is the common covariance matrix
- π_k is the prior probability of class k

Typical Use Cases: LDA is commonly used for dimensionality reduction before classification and for classification tasks where a linear boundary is sufficient. It has applications in face recognition, marketing, biomedical studies, and educational research.

Advantages:

- Works well when classes are well-separated
- Provides dimensionality reduction capabilities
- Computationally efficient
- Less prone to overfitting than many other classifiers
- Provides probabilistic outputs

Limitations:

- Assumes linear separability between classes
- Assumes Gaussian distribution of features within each class
- Assumes homoscedasticity (equal covariance matrices across classes)
- May perform poorly when these assumptions are violated

3.5.9 AdaBoost

AdaBoost (Adaptive Boosting) is an ensemble learning approach that merges multiple "weak learners" into a single robust classifier by adjusting the weight of instances based on previous classification results.

Conceptual Explanation: AdaBoost works by training a sequence of weak classifiers (typically decision stumps or shallow decision trees) on weighted versions of the learning data. After each classifier is trained, the weights of misclassified instances are increased, forcing subsequent classifiers to concentrate more on difficult cases. The final prediction is a weighted combination of all weak classifiers (Hastie et al., 2005).

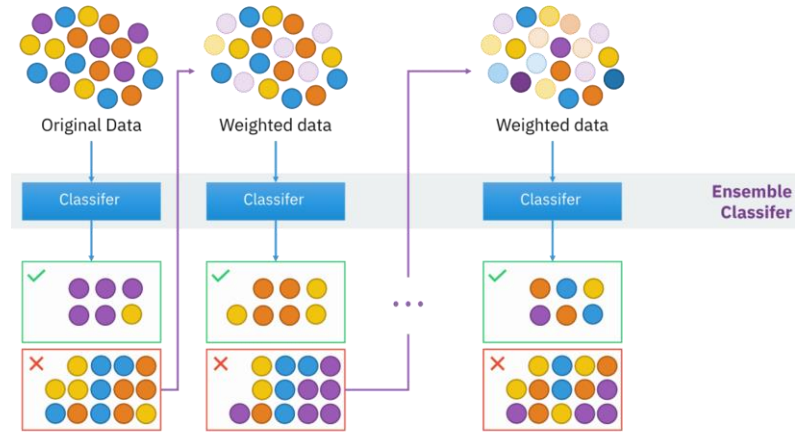


Figure 3.10: AdaBoost Algorithm Architecture (AlmaBetter, 2025)

Mathematical Expression: The final AdaBoost classifier is:

$$F(x) = \sum_{t=1}^T \alpha_t h_t(x) \dots\dots\dots (12)$$

Where:

- $F(x)$ is the final classifier
- $h_t(x)$ is the t-th weak classifier
- α_t is the weight assigned to the t-th classifier, calculated as:

$$\alpha_t = \frac{1}{2} \ln \left(\frac{1-\epsilon_t}{\epsilon_t} \right) \dots\dots\dots (13)$$

Where α_t is the weighted error rate of the t-th classifier.

Typical Use Cases: AdaBoost has been successfully applied in various domains including face detection, text classification, and medical diagnosis. In educational contexts, it has been used for predicting student performance and identifying factors contributing to academic success or failure.

Advantages:

- Often resistant to overfitting
- Can achieve high accuracy with simple base classifiers
- Automatically identifies hard-to-classify instances
- Provides a measure of feature importance
- Works well with categorical and numerical features

Limitations:

- Sensitive to noisy data and outliers
- Can perform poorly if weak learners are too complex

- Sequential nature makes it difficult to parallelize
- May underperform compared to more recent boosting algorithms like XGBoost
- Requires careful selection of base learners

3.5.10 Ridge Classifier

Ridge Classifier is a variant of linear classification that incorporates L2 regularization to prevent overfitting.

Conceptual Explanation: Ridge Classifier is based on Ridge Regression applied to classification problems. It adds a penalty term proportional to the square of the magnitude of coefficients to the loss function, which helps to reduce the impact of multicollinearity and prevent overfitting by shrinking the coefficients (Friedman, 2009).

Mathematical Expression: The objective function being minimized is:

$$\min_{\beta} \|y - X\beta\|_2^2 + \alpha \|\beta\|_2^2 \dots\dots\dots(14)$$

Where:

- y is the target vector
- X is the feature matrix
- β is the coefficient vector
- α is the regularization strength
- $\|\beta\|_2$ is the L2 norm of the coefficient vector

Typical Use Cases: Ridge Classifier is particularly useful for classification problems with many features, especially when there is multicollinearity among features. It has been applied in text classification, bioinformatics, and educational data mining.

Advantages:

- Handles multicollinearity effectively
- Less prone to overfitting than ordinary least squares
- Computationally efficient
- Shrinks coefficients towards zero but does not eliminate any
- Works well when all features are potentially relevant

Limitations:

- Assumes linear relationship between features and log-odds of class membership
- Does not perform feature selection (all features remain in the model)
- May underperform when true decision boundary is highly non-linear
- Requires careful tuning of the regularization parameter

3.5.11 Extra Trees

Extra Trees (Extremely Randomized Trees) is an ensemble learning approach similar to Random Forest but with extra randomization in the tree-building process.

Conceptual Explanation: Like Random Forest, Extra Trees constructs multiple decision trees, but it introduces two key differences: it employs the entire training dataset rather than bootstrap samples, and it selects split points randomly rather than searching for the optimal split. This additional randomization helps to reduce variance and can lead to better generalization in some cases (Zhou, 2025).

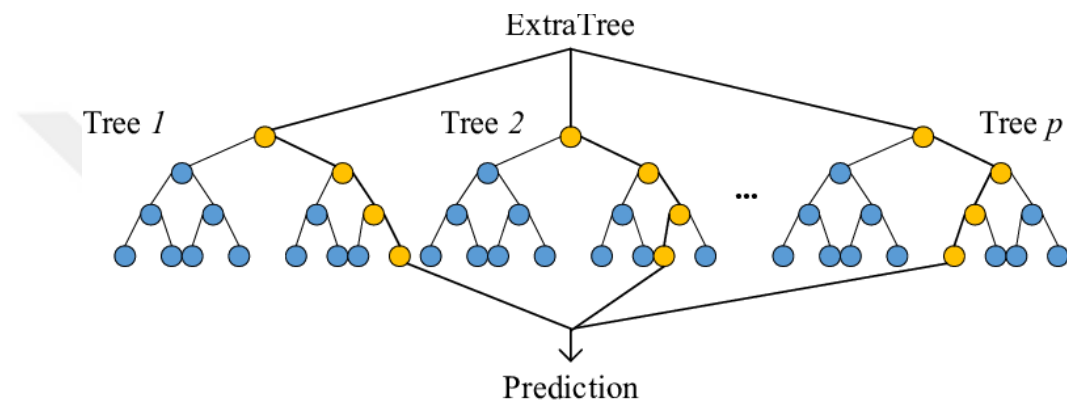


Figure 3.11: Extra Trees Algorithm (Chu et al., 2020)

Mathematical Expression: The prediction mechanism is similar to Random Forest:

$$\hat{y} = \operatorname{argmax}_c \sum_{b=1}^B I(h_b(x) = c) \quad \dots\dots\dots(15)$$

The key difference lies in how each tree h_b is constructed, with random split selection instead of optimized splits.

Typical Use Cases: Extra Trees is used in similar applications as Random Forest, including predictive modeling in healthcare, finance, and education. It is particularly valuable when dealing with noisy features or when computational efficiency is important.

Advantages:

- Often faster to train than Random Forest due to simpler split selection
- Can achieve lower variance than Random Forest in some cases
- Less sensitive to outliers than many other algorithms
- Provides feature importance estimates

- Good performance with minimal hyperparameter tuning

Limitations:

- May have higher bias than Random Forest in some cases
- Less interpretable than single decision trees
- Can struggle with highly imbalanced datasets without proper adjustments
- May require more trees than Random Forest to achieve stable predictions

3.5.12 QDA (Quadratic Discriminant Analysis)

Quadratic Discriminant Analysis is a classification method that assumes each class follows a Gaussian distribution with different covariance matrices.

Conceptual Explanation: QDA models the likelihood of each class using a multivariate Gaussian distribution and applies Bayes' theorem to calculate the posterior probability of class membership. Unlike Linear Discriminant Analysis (LDA), QDA allows each class to have its own covariance matrix, resulting in quadratic decision boundaries between classes (Friedman, 2009).

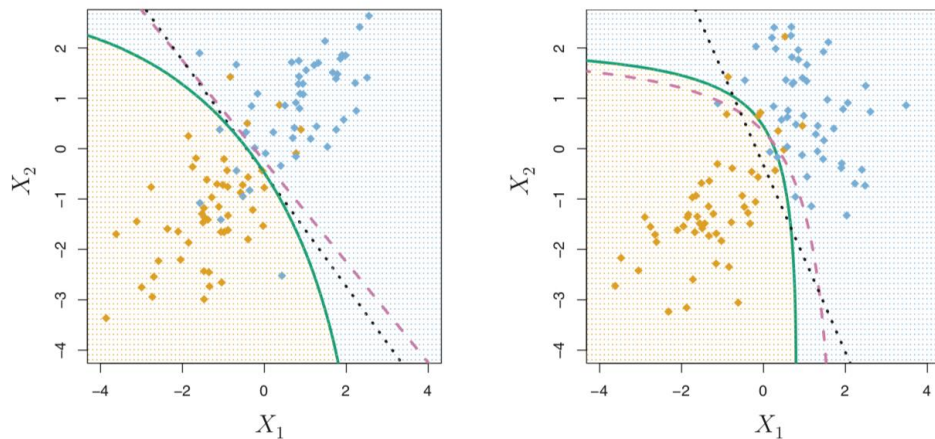


Figure 3.12: Comparison of LDA and QDA Decision Boundaries (Wicker, 2021)

Mathematical Expression: The discriminant function for QDA is:

$$\delta_k(x) = -\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1}(x - \mu_k) - \frac{1}{2} \log |\Sigma_k| + \log(\pi_k) \quad \dots\dots\dots(16)$$

Where:

- $\delta_k(x)$ is the discriminant score for class k
- x is the feature vector
- μ_k is the mean vector for class k
- Σ_k is the covariance matrix for class k
- π_k is the prior probability of class k

Typical Use Cases: QDA is particularly useful for classification problems where the classes have different variance-covariance structures and the decision boundaries are likely to be non-linear. It has been applied in various fields including finance, bioinformatics, and pattern recognition.

Advantages:

- Can model complex, non-linear decision boundaries
- Works well when classes have different covariance structures
- Often performs better than LDA when there is sufficient training data
- Provides probabilistic outputs

Limitations:

- Requires more parameters to estimate than LDA
- May overfit when the number of features is large relative to the sample size
- Assumes Gaussian distribution of features within each class
- Sensitive to outliers

3.6 Model Evaluation Metrics

To rigorously evaluate the performance of the machine learning models developed in this research, a comprehensive set of evaluation metrics was employed. These measures provide different viewpoints on model performance, allowing for a more nuanced understanding of each algorithm's strengths and weaknesses in the context of student dropout prediction.

3.6.1 Accuracy

Accuracy is the most straightforward evaluation metric, representing the proportion of correct predictions (both true positives and true negatives) among the total number of cases examined. It is calculated using the formula:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \dots\dots\dots(17)$$

Where:

- TP (True Positive): Students correctly predicted to drop out
- TN (True Negative): Students correctly predicted to continue or graduate
- FP (False Positive): Students incorrectly predicted to drop out
- FN (False Negative): Students incorrectly predicted to continue or graduate

3.6.2 Precision

Precision measures the proportion of correctly identified positive cases (true positives) among all cases predicted as positive. In the context of dropout prediction, precision answers the question: "Of all students predicted to drop out, what percentage actually dropped out?" It is calculated as:

$$\text{Precision} = \frac{TP}{TP+FP} \dots\dots\dots (18)$$

High precision indicates a low false positive rate, meaning the model rarely misclassifies continuing or graduating students as dropouts. This is particularly important when intervention resources are limited, as it ensures that resources are directed toward students who are genuinely at risk.

3.6.3 Recall (Sensitivity)

Recall, also known as sensitivity or true positive rate, measures the proportion of actual positive cases (true positives) that were correctly identified by the model. For dropout prediction, recall answers the question: "Of all students who actually dropped out, what percentage were correctly identified by the model?" It is calculated as:

$$\text{Recall} = \frac{TP}{TP+FN} \dots\dots\dots (19)$$

High recall indicates that the model effectively identifies most students who are at risk of dropping out, minimizing the number of at-risk students who go undetected.

3.6.4 F1 Score

The F1 score is the harmonic mean of precision and recall, providing a balance between these two sometimes competing measures. It is particularly useful when seeking a single measure that accounts for both false positives and false negatives.

The F1 score is calculated as:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \dots\dots\dots (20)$$

A high F1 score indicates that both precision and recall are high, meaning the model is both accurate in its positive predictions and comprehensive in identifying actual positive cases.

3.6.5 Confusion Matrix

While not a single measure, the confusion matrix provides a comprehensive view of model performance by displaying the counts of true positives, false positives, true negatives, and false negatives in a tabular format. This visualization allows for a detailed analysis of where the model succeeds and fails, providing insights that might not be apparent from aggregate metrics alone.

Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Figure 3.13: Confusion Matrix (Draelos, 2019)

3.7 Train/Test Split

To develop and evaluate the predictive models, the original dataset consisted of 545 records. Students who were still enrolled in their studies (N=268) were excluded, as their data would not contribute meaningfully to training the model. The remaining dataset included 277 students, comprising 128 graduates and 149 dropouts. An 80/20 split was employed, allocating 80% of the data (N=221 instances) for model training and reserving the remaining 20% (N=56 instances) as an unseen test set for final model evaluation.

The stratified sampling technique was employed during the split to ensure that the class distribution in both the training and testing sets matched the original dataset's distribution. This stratification is particularly important given the imbalanced nature of the target variable (continuing, graduated, dropped out) and helps prevent sampling bias that could affect model performance evaluation.

The 80/20 ratio was selected as it provides a good balance between having sufficient data for training robust models and reserving enough data for meaningful performance evaluation. This split ratio is widely used in machine learning applications and has been shown to be effective in educational data mining studies with similar sample sizes (Cho et al., 2020; Márquez-Vera et al., 2013).

The training set was used for model development, including hyperparameter tuning through cross-validation techniques. The test set was kept completely separate and was only used for the final evaluation of model performance to provide an unbiased assessment of how well the models would generalize to new, unseen student data.

3.8 Web Application for Prediction

To translate the theoretical models into a practical tool that can be used by educational stakeholders, a web application was developed to provide real-time predictions of student dropout risk. This application serves as a demonstration of how the machine learning models can be deployed in real-world educational settings to identify at-risk students and facilitate timely interventions.

3.8.1 Creating a Trained Model File

The best-performing model from the evaluation phase was saved as a serialized file (.pkl) using the PyCaret library. This approach allows the model to be used in other applications without the need for retraining.

The process involved:

- Training the model using the preprocessed dataset
- Saving the trained model to a file using PyCaret's specialized save functions
- Loading the model from this file when needed to make predictions

This serialization approach ensures that the model's parameters and learned patterns are preserved exactly as they were during the training phase, maintaining prediction consistency across different environments.

3.8.2 Development Environment Setup

Visual Studio Code was used as the integrated development environment (IDE) for writing the code that integrates the predictive model with the web interface. The development stack included:

- **FastAPI:** A modern, high-performance web framework for building APIs with Python, chosen for its speed, ease of use, and automatic documentation generation capabilities
- **Pydantic:** Used for data validation, ensuring that user inputs conform to the expected data types and ranges
- **PyCaret:** Employed to load the trained model and perform predictions

- **Pandas:** Utilized for processing input data and transforming it into a format suitable for the model

This technology stack was selected for its balance of performance, ease of development, and robust error handling, which are critical for a prediction application that will be used by non-technical users.

3.8.3 API Development

An Application Programming Interface (API) was created using FastAPI to serve as an intermediary between the model and the web page. The API's responsibilities include:

1. Receiving user-input data from the web interface
2. Transforming the data into the appropriate format for the model
3. Performing predictions using the trained model
4. Returning the prediction results to the web interface

Data validation was implemented using Pydantic to ensure that all inputs meet the required specifications before being processed by the model. This validation layer prevents errors that could arise from incorrect data types or values outside the expected ranges, enhancing the application's reliability.

3.8.4 Web Interface Design

The interactive web page was designed using HTML, CSS, and JavaScript to provide an intuitive user experience:

- **HTML:** Used to create the basic structure of the page, including input forms and result display areas
 - **CSS:** Applied to style the page and ensure a user-friendly appearance across different devices
 - **JavaScript:** Implemented to handle user interactions, send data to the API, and display the prediction results
- **Home Page (Figure 3.14):** The landing page presents the project's title and purpose. A top navigation bar links to key sections (About Research, Data Analysis, ML Prediction, Research Team , Contact). Introductory text, a relevant visual, and an "Explore Research" button provide an overview. HTML-structured and CSS styling ensures a clear and user-friendly layout."



Figure 3.14: Home page of the student dropout prediction web application

- **About Research Page (Figure 3.15):** This page explains the research project. It covers what the research is about, why it's important, and the methods used, along with a helpful image. The different steps of the research, like "Data Collection," "Feature Selection," and "ML Classification," are shown in separate, clear boxes. This makes it easy to see how the project was done from start to finish. The page is designed to be easy to read and look professional.

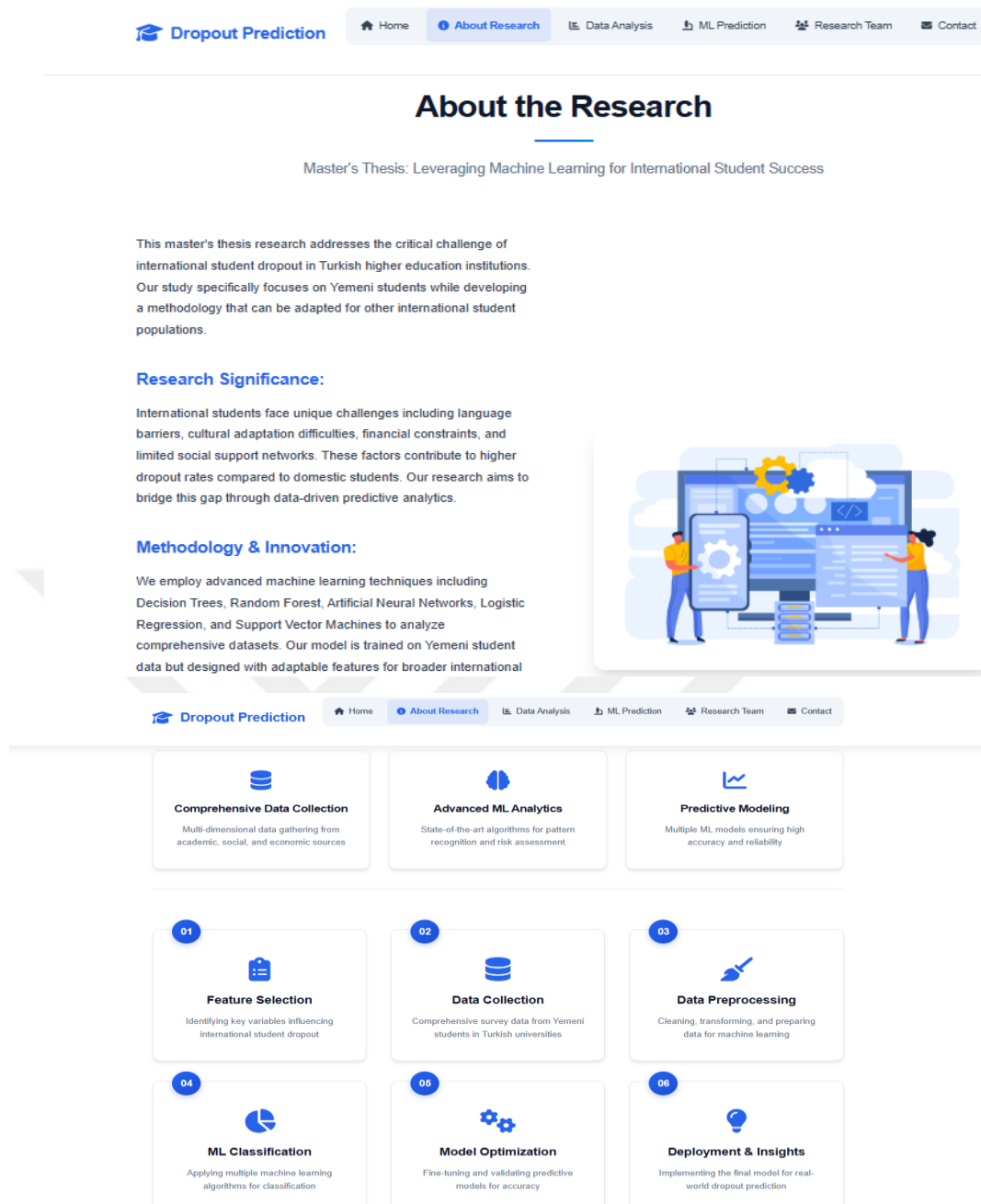


Figure 3.15: "About Research" page of the student dropout prediction web application

- **Data Analysis Page (Figure 3.16):** This page shows key statistics and charts from the research data. It uses cards for quick summary numbers (like total students and dropouts) and includes interactive charts. The design helps users easily understand the data.



Figure 3.16: "Data Analysis" page of the student dropout prediction web application

- **ML Prediction Page (Figure 3.17):** This is the core interactive page for "Individual Student Risk Assessment." Users input student data through various dropdown menus (e.g., Age Group, GPA, Language Proficiency). Clicking the "Analyze Dropout Risk" button, handled by JavaScript, sends this data to the API. The "Risk Assessment Results" section then displays the model's prediction.

Figure 3.17: Individual student risk assessment interface for inputting student data and viewing prediction results

- **ML Prediction Page - Batch Prediction (Figure 3.18):** This section of the ML Prediction page allows for "Batch Prediction." Users can switch from "Individual Prediction" to this mode. It features an interface to "Upload Excel File" via a "Choose File" button or drag-and-drop. JavaScript manages file handling and API communication. Tabs for "Data Preview" and "Analysis Results" are provided to display uploaded data and prediction outcomes.

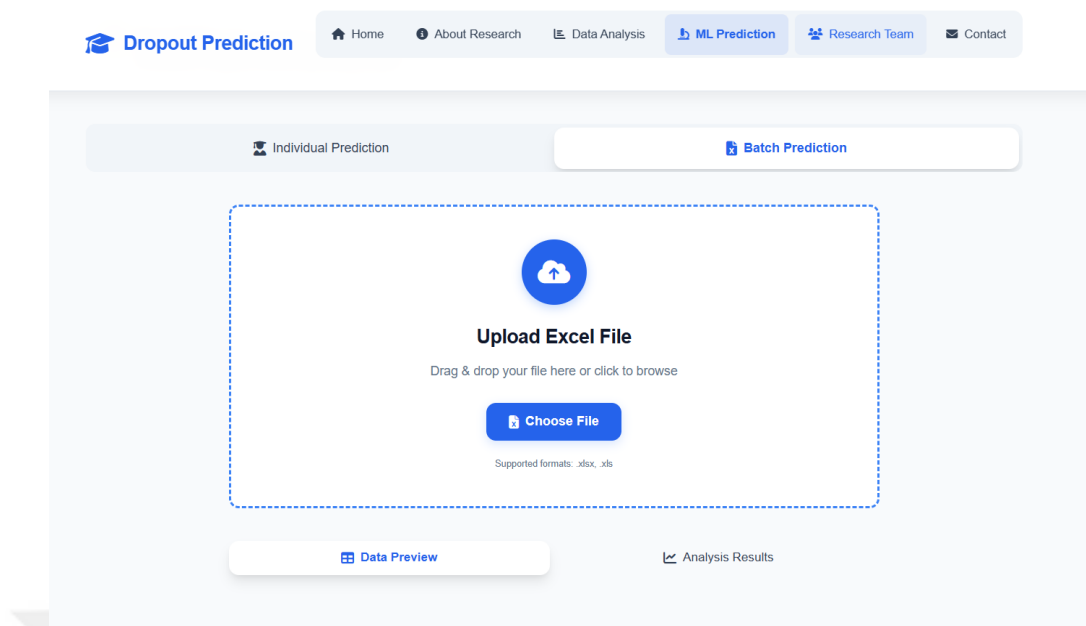


Figure 3.18: Batch prediction interface for uploading an Excel file containing multiple student records.

- **Research Team Page (Figure 3.19):** This page introduces the researchers involved in the project. It uses HTML to structure individual cards for each team member, displaying their photo, name, role, and links to their professional profiles or contact information. CSS styling ensures a clean and professional presentation.

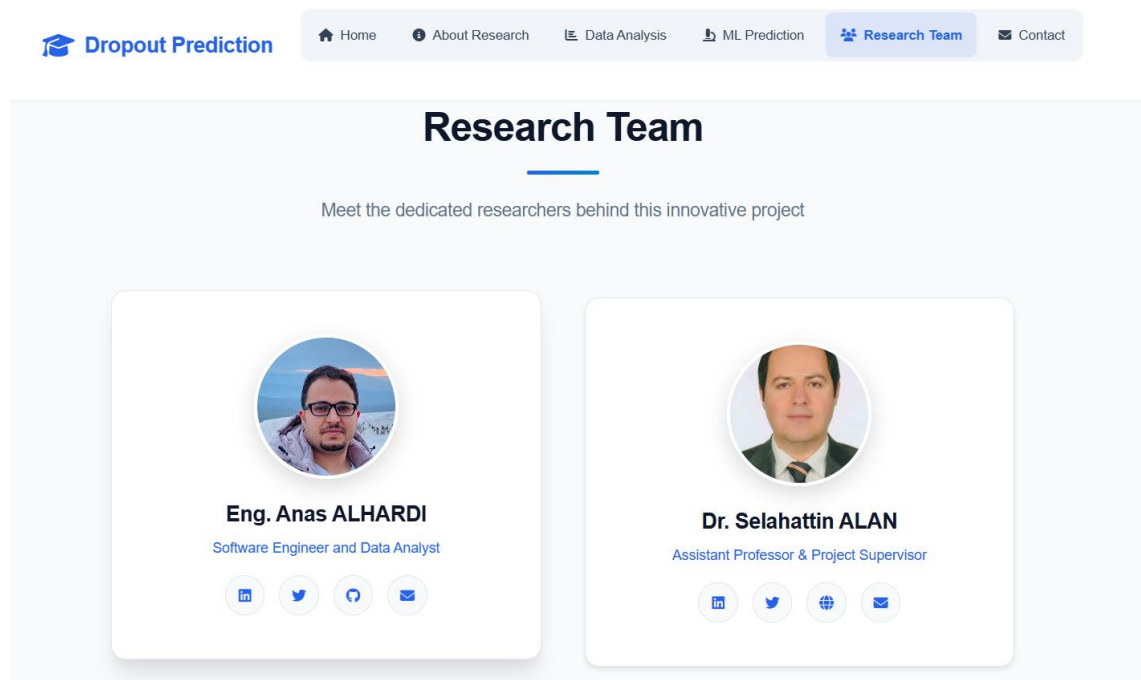


Figure 3.19: Research Team page showcasing the profiles of the project contributors.

- **Contact Page (Figure 3.20):** This page enables users to "Contact Research Team." It includes an HTML form for sending a message directly, with fields for name, email, subject, and message. Alongside the form, key "Research Contact Information" such as institution, phone, and email are displayed. CSS styling ensures a clear and accessible layout for communication.

The screenshot shows the 'Contact Research Team' page of the Dropout Prediction web application. The navigation bar at the top includes links to Home, About Research, Data Analysis, ML Prediction, Research Team, and Contact. The main heading is 'Contact Research Team' with a subtitle 'For inquiries about the research, collaboration opportunities, or technical questions'. On the left is a 'Send a Message' form with fields for Name, Email, Subject, and Message, and a 'Send Message' button. On the right is a blue box titled 'Research Contact Information' containing details for Selcuk University, Computer Engineering Department, Konya, Turkey, with phone number +90 531 370 8658 and email anas.rajah1@gmail.com.

Figure 3.20: "Contact" page of the student dropout prediction web application

In summary, this carefully designed interface provides an integrated and user-friendly experience. It enables stakeholders to explore the research and interact directly and effectively with the prediction model. Thus, the web interface design serves as a vital bridge between theoretical models and practical application in educational settings.

3.8.5 Deployment Infrastructure

To make the application accessible to users, a robust deployment infrastructure was established:

- **GitHub:** The source code of the application was hosted on a GitHub repository, allowing for effective version control, collaborative development, and continuous integration. This setup facilitated seamless updates and ongoing maintenance of the codebase.
- **Netlify:** For front-end deployment, the Netlify platform was utilized to host the website. Netlify was chosen due to its simplicity, support for continuous deployment directly from GitHub, and built-in features such as SSL certificate provisioning and automated build processes. This made it a suitable choice for hosting static web applications efficiently.
- **Render Dashboard:** The Render cloud platform was utilized for server deployment, providing a scalable and reliable hosting environment for the application. Render was selected for its ease of use, automatic deployment from GitHub, and built-in SSL certificate management.
- **Domain Registration:** The custom domain www.yemendropout.online was purchased and configured to provide the application with a professional and easily accessible web address, enhancing the project's credibility and user accessibility.

The deployment workflow was configured to automatically update the live application whenever changes are pushed to the main branch of the GitHub repository, ensuring that users always have access to the latest version of the application with minimal downtime.

3.8.6 Ethical Considerations of Prediction Results

While the web application provides valuable insights for early identification of students at risk of dropout, it is important to acknowledge the ethical risks associated with prediction outcomes. Incorrect predictions—such as false positives (students incorrectly identified as at risk) or false negatives (students at risk not identified)—can have significant consequences. For example, a student wrongly labeled as "at risk" may experience unnecessary stress, stigma, or receive interventions that are not needed, potentially affecting their academic motivation and self-esteem. Conversely, failing to identify a truly at-risk student may result in missed opportunities for timely support. Therefore, it is essential that prediction results are used as supportive tools rather than definitive judgments, and that educators interpret them with caution, considering the broader context of each student. Ongoing monitoring, transparency, and feedback

mechanisms should be established to minimize potential harm and ensure ethical use of the system.

3.8.7 Integration and Operation

The complete prediction system operates through the following workflow:

1. Users enter student data in the designated fields on the web page
2. This data is sent to the API via an HTTP request when the user clicks the "Predict" button
3. The API loads the trained model, processes the input data, and executes the prediction
4. The prediction result (graduation or dropout) is returned to the web page
5. The result is displayed to the user in an easily interpretable format

This end-to-end integration demonstrates the practical application of machine learning models in educational settings and provides a template for how similar systems could be implemented by educational institutions to support student retention efforts.

3.9 Ethical Considerations

This research was conducted with careful attention to ethical principles governing educational research and data science. The following ethical considerations were addressed throughout the study:

All data collected from students was anonymized to protect individual identities. Personal identifiers were removed during the preprocessing stage, and results are presented only in aggregate form to prevent the identification of individual participants. The dataset was stored securely with access restricted to authorized researchers only.

The survey explicitly informed participants about the anonymization procedures and assured them that their individual responses would not be identifiable in any published results. This transparency was essential for building trust with participants and encouraging honest responses.

The following information was clearly communicated to the participants of the survey:

- The purpose of the research
- The types of data being collected
- How the data would be used
- The voluntary nature of participation

- Contact information for the researchers

The consent form was provided in Arabic to ensure full comprehension by all participants, and participants were given the opportunity to ask questions before agreeing to participate.



4. RESULTS

This chapter presents the empirical findings derived from the application of various machine learning techniques aimed at predicting dropout among international students in Turkey. Following the methodology detailed in Chapter 3, this chapter systematically evaluates the performance of the developed predictive models and explores the underlying patterns within the dataset. The primary objective is to identify the most effective models for predicting dropout within this specific student cohort and to uncover the key factors associated with this phenomenon.

The chapter begins with a presentation of descriptive statistics and exploratory data analysis (EDA) findings, providing an overview of the dataset characteristics and initial insights into the relationships between student attributes and their academic outcomes (dropout or graduation). Subsequently, the performance of each implemented machine learning model is detailed, utilizing standard evaluation metrics such as accuracy, precision, recall, F1-score, and confusion matrices. A comparative analysis follows, highlighting the relative strengths and weaknesses of the different algorithms in the context of this prediction task. Where applicable, insights into feature importance derived from the models will be discussed. Finally, the chapter synthesizes these results, discussing their implications in relation to the research questions posed in Chapter 1 and the existing literature reviewed in Chapter 2, and concludes with a summary of the key findings.

4.1. Exploratory Data Analysis

Prior to the development and evaluation of predictive models, a thorough exploratory data analysis (EDA) was conducted to gain initial insights into the characteristics of the dataset and the relationships between various student attributes and their academic outcomes. This section details the key findings from the EDA, focusing on the dataset prepared specifically for the binary classification task of predicting student dropout versus graduation.

As delineated in Chapter 3, the initial dataset, comprising 545 records of Yemeni international students in Turkey, underwent several preprocessing steps, including data cleaning, translation, feature encoding, and filtering. For the primary modeling objective—predicting whether a student would ultimately drop out or graduate—the

analysis focused on a subset of 277 students for whom a final outcome (Dropout or Graduate) was recorded, excluding those still enrolled. This modeling dataset featured 15 predictor variables encompassing demographic details (e.g., age, gender), academic background (e.g., last GPA), language proficiency (Turkish), socio-economic indicators (e.g., scholarship status, family economy, parental education), and self-reported measures of program satisfaction and psychological stress.

The EDA commenced with an examination of the target variable distribution within the modeling dataset (N=277). The analysis revealed 149 instances classified as 'Dropout' (approximately 53.8%) and 128 instances classified as 'Graduate' (approximately 46.2%). This distribution, while exhibiting a slight majority towards the 'Dropout' class, was considered reasonably balanced, mitigating significant concerns regarding class imbalance potentially skewing model training.

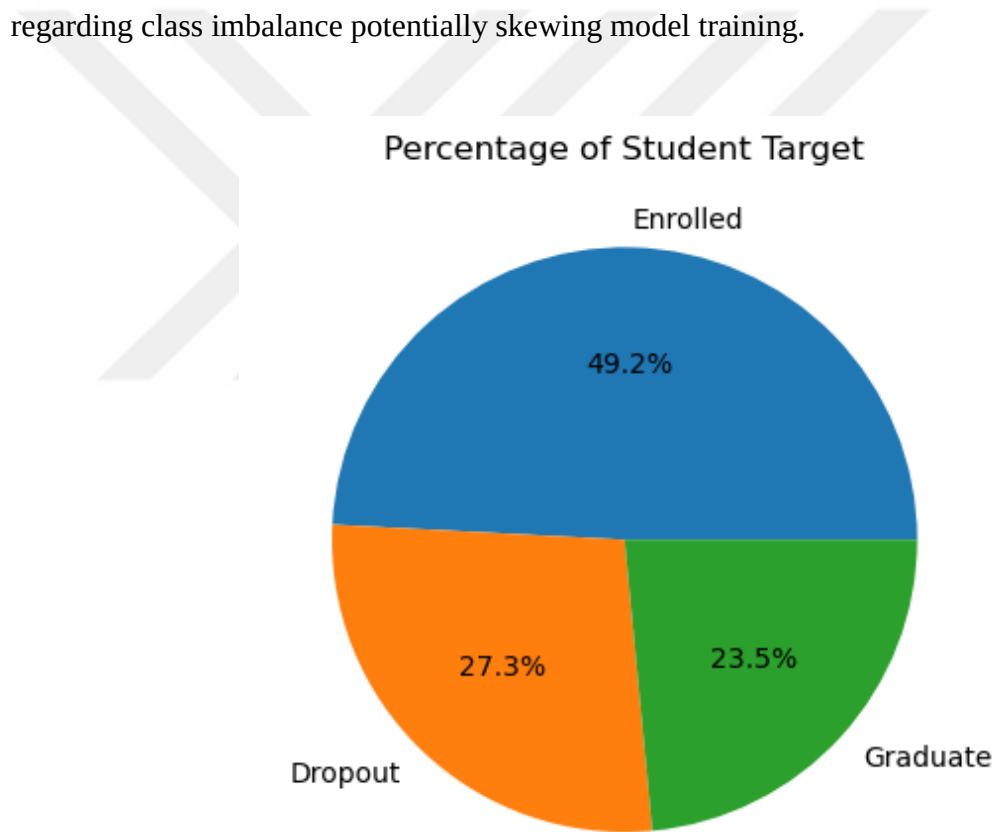


Figure 4.1: Distribution of Target Variable (Enrolled, Graduate, Dropout)

Subsequently, the relationships between individual predictor variables and the binary outcome were investigated using various visualization techniques, primarily generated using the Matplotlib and Seaborn libraries in Python. Key observations include:

- **Academic Performance (Last GPA):** Kernel Density Estimate (KDE) plots demonstrated a distinct separation in the distribution of the last recorded GPA between the 'Dropout' and 'Graduate' groups. Students who dropped out were concentrated in the lower GPA ranges (e.g., 0-69), whereas graduates predominantly occupied the higher GPA ranges (e.g., 70-100). This strongly suggested that recent academic performance is a critical indicator of persistence.

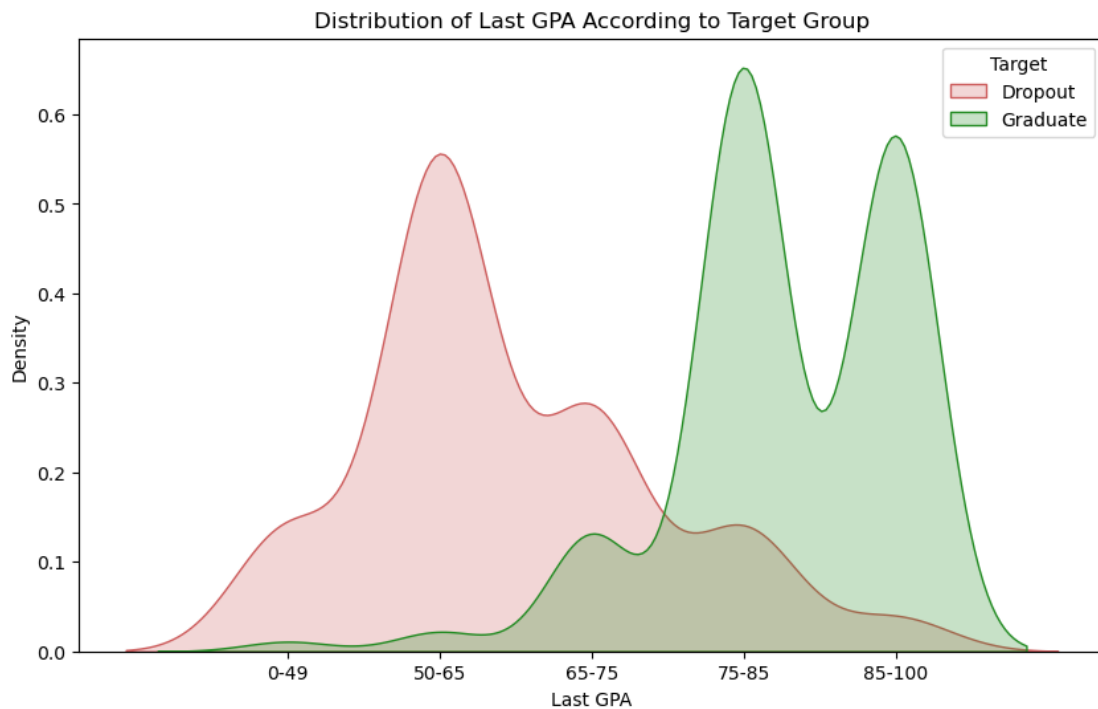


Figure 4.2: KDE plot showing distribution of Last GPA for Dropout vs Graduate groups

- **Turkish Language Proficiency:** Analysis indicated a clear association between Turkish language proficiency ('Weak', 'Good', 'Excellent') and academic outcomes. Students reporting 'Weak' proficiency were disproportionately represented in the 'Dropout' category, while those with 'Excellent' proficiency were more likely to graduate, highlighting the importance of language adaptation.

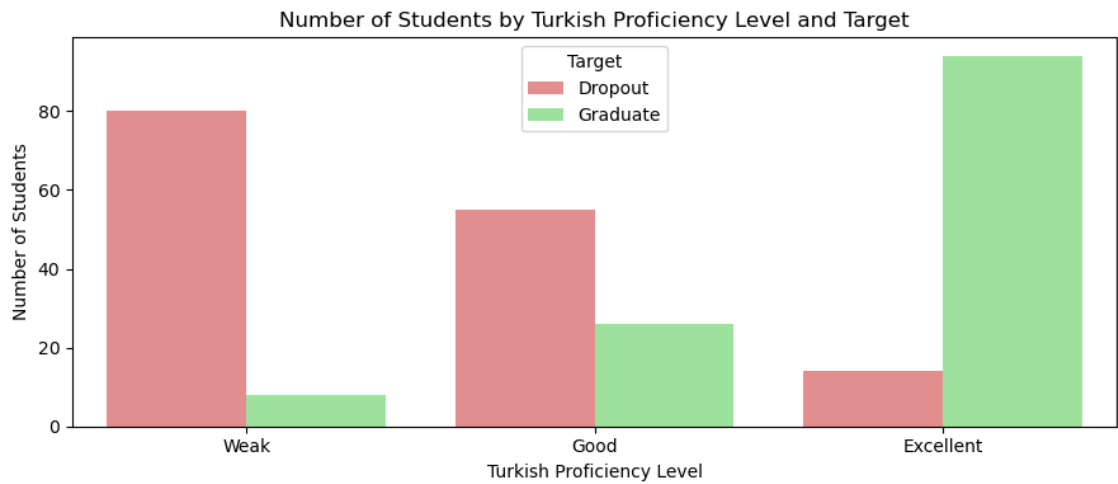


Figure 4.3: Bar chart showing distribution of Dropout/Graduate by Turkish Proficiency Level

- **Major Satisfaction:** Students expressing dissatisfaction with their major ('Not Satisfied') showed a markedly higher tendency to drop out compared to those who were 'Satisfied' or 'Very Satisfied', emphasizing the role of academic integration and program fit.

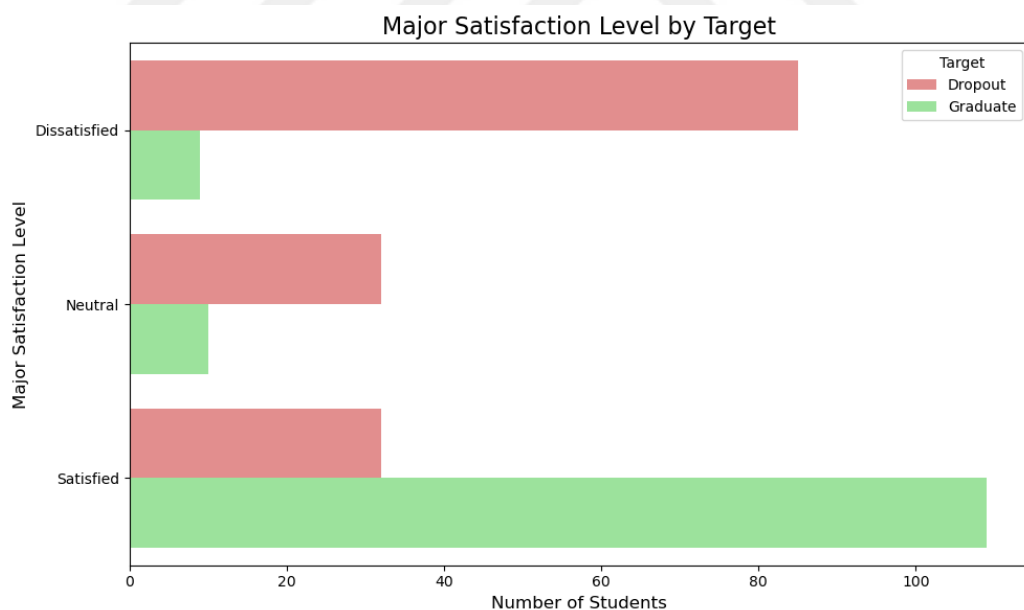


Figure 4.4: Bar chart showing distribution of Dropout/Graduate by Major Satisfaction

- **Financial Factors:** Scholarship status and perceived family economic status were both linked to outcomes. A larger proportion of students without scholarships were found in the 'Dropout' group. Similarly, students reporting lower family economic status appeared more prone to dropping out.

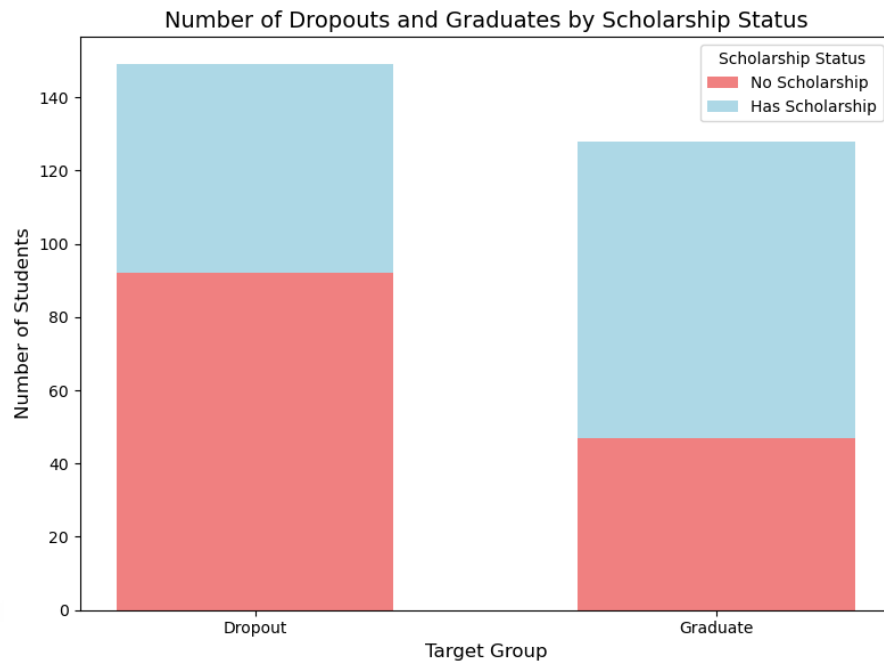


Figure 4.5: Stacked bar chart showing distribution of Dropout/Graduate by Scholarship Status

- Psychological Stress:** Students who indicated that psychological stress significantly impacted their education were more prevalent among the 'Dropout' group, underscoring the relevance of mental well-being in academic persistence.

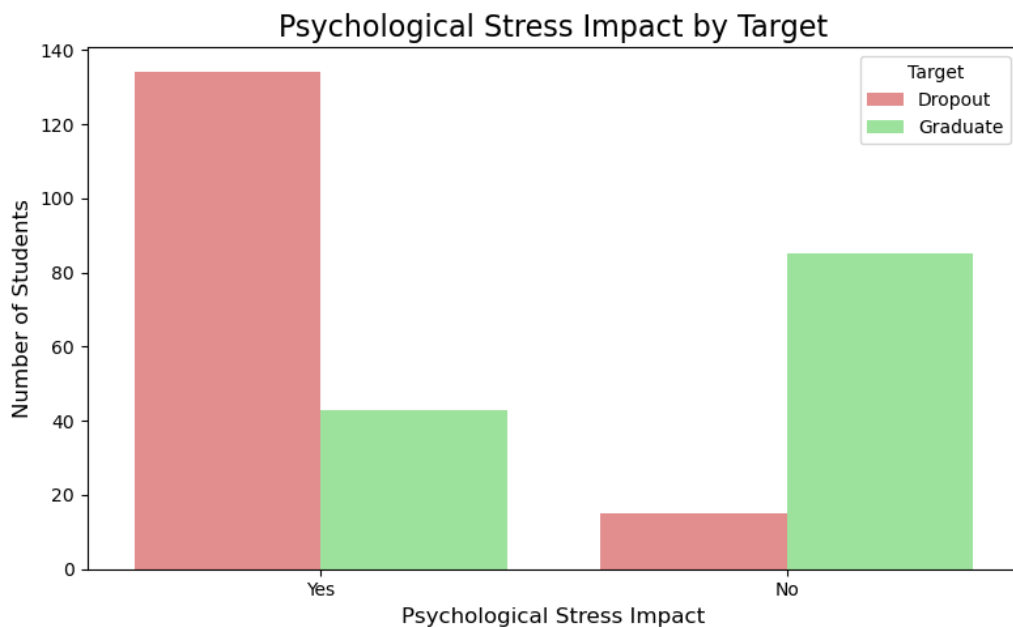


Figure 4.6: Bar chart showing distribution of Dropout/Graduate by Psychological Stress Impact

- **Other Factors:** Visualizations for other variables such as age, gender, part-time work status, and parental education levels also revealed varying degrees of association with dropout rates, although these appeared less pronounced than the factors mentioned above based on initial visual inspection.

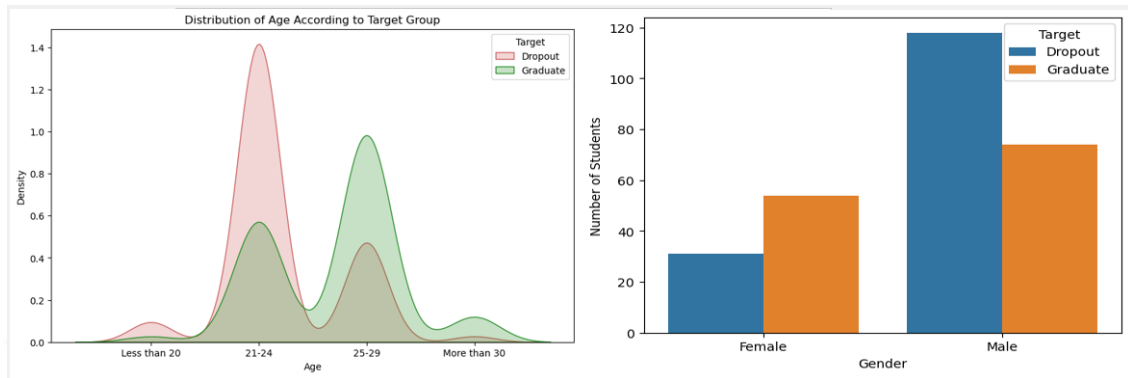


Figure 4.7: KDE plot and Bar chart showing distribution of Dropout/Graduate by Age and Gender

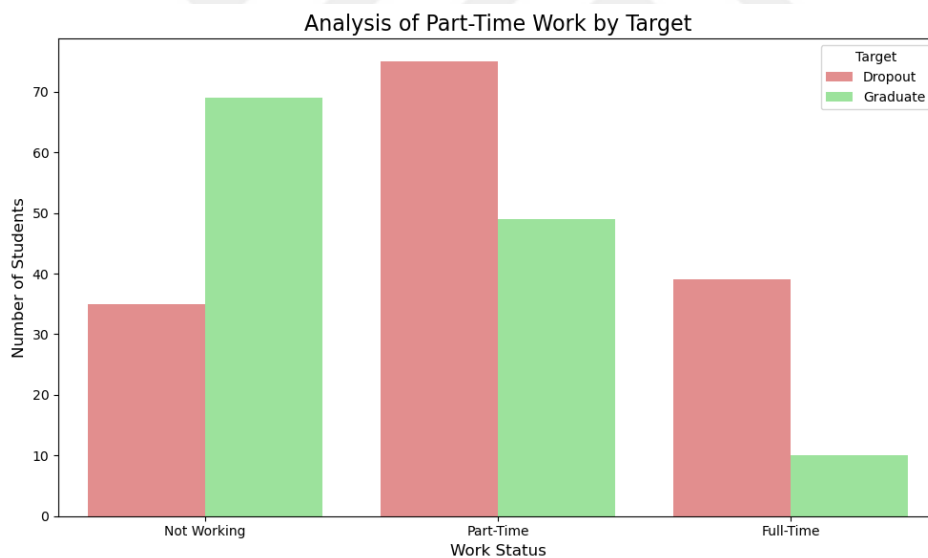


Figure 4.8: Bar chart showing distribution of Dropout/Graduate by Part-Time Work

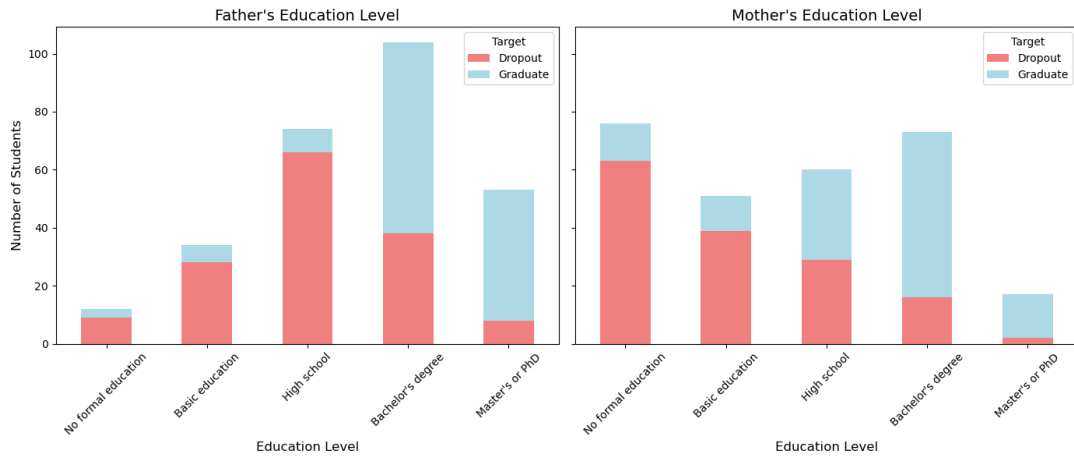


Figure 4.9: Bar chart showing distribution of Dropout/Graduate by Parents' Education Level

Finally, a correlation analysis was performed to assess linear relationships between the numerical and encoded categorical predictor variables. A heatmap visualization indicated generally low to moderate correlations between most predictors, suggesting that multicollinearity was unlikely to pose a significant problem for the subsequent modeling phase.

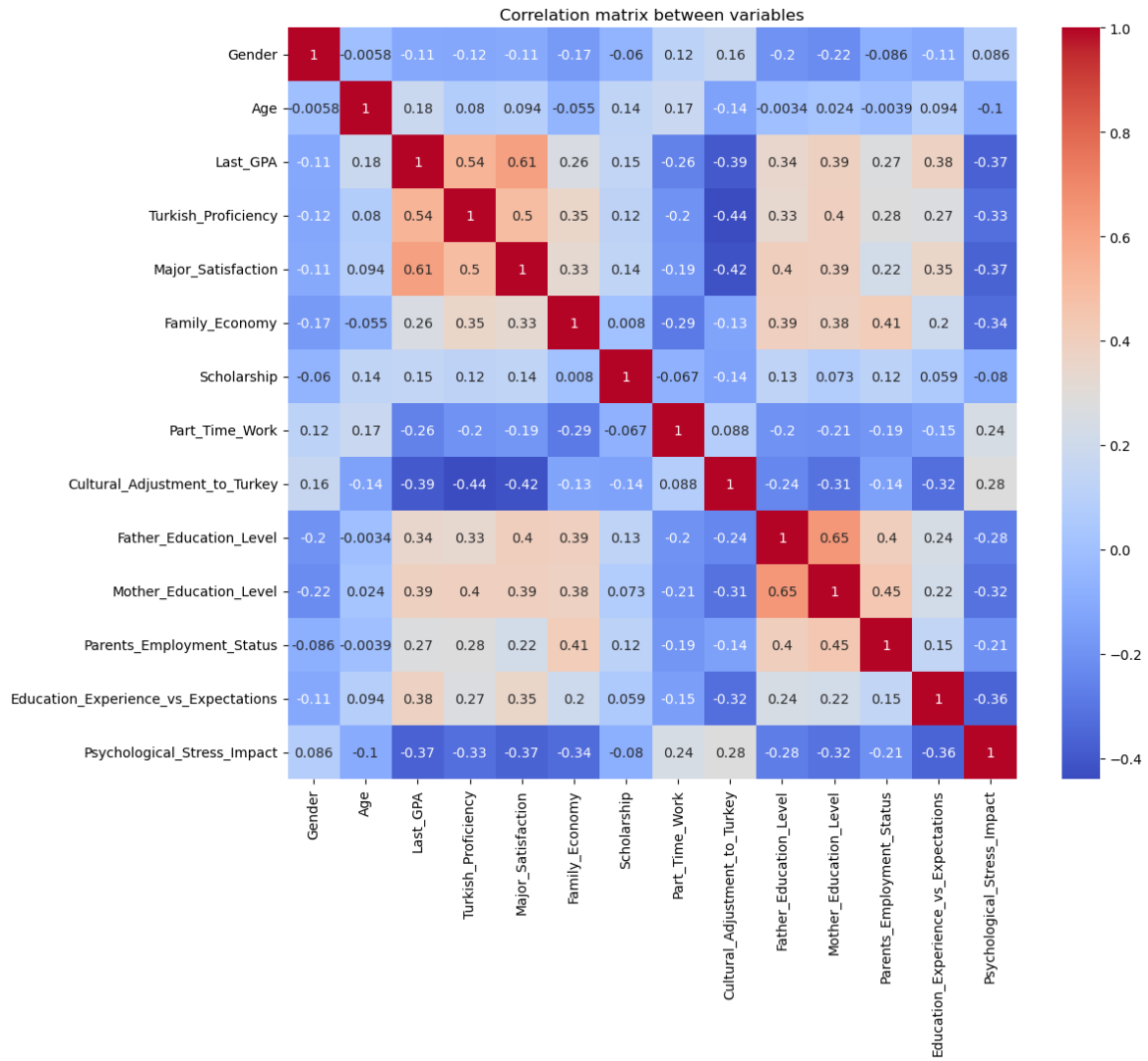


Figure 4.10: Correlation heatmap of predictor variables

In summary, the EDA provided valuable preliminary evidence supporting the relevance of academic performance, language proficiency, program satisfaction, financial stability, and psychological well-being as potential key factors influencing dropout among international students in Turkey. These insights guided the selection and interpretation of the machine learning models detailed in the following sections.

4.2. Model Implementation

Following the exploratory data analysis, the next phase involved the implementation and training of various machine learning models to predict student dropout. This section outlines the procedures followed for model development, referencing the detailed methodology described in Chapter 3.

The core dataset utilized for model training and evaluation consisted of 277 instances representing Yemeni international students in Turkey with definitive outcomes ('Dropout' or 'Graduate'). This dataset, characterized by 15 predictor variables identified through the EDA and feature selection process, was partitioned into a training set and a testing set. An 80/20 split was employed, allocating 80% of the data (N=221 instances) for model training and reserving the remaining 20% (N=56 instances) as an unseen test set for final model evaluation. To ensure the reproducibility of the results, a fixed random state (**random_state=3**) was used during the data splitting process.

A comprehensive suite of fifteen distinct classification algorithms was selected for implementation, covering a diverse range of modeling paradigms. These algorithms included:

1. Logistic Regression
2. Linear Discriminant Analysis (LDA)
3. Quadratic Discriminant Analysis (QDA)
4. Ridge Classifier
5. K-Nearest Neighbors (KNN, with k=5)
6. Decision Tree
7. Naive Bayes (GaussianNB)
8. LightGBM
9. Support Vector Machine (SVM)
10. Random Forest
11. Extra Trees
12. AdaBoost
13. Neural Network (MLP)
14. Gradient Boosting
15. XGBoost (Extreme Gradient Boosting)

For most models, default hyperparameters as implemented in the Scikit-learn library were initially used.

The predictive performance of each trained model was systematically assessed on the held-out test set (N=56). Standard classification metrics were employed for evaluation, including overall accuracy, precision, recall, and F1-score. These metrics were calculated separately for both the 'Dropout' and 'Graduate' classes to provide a nuanced understanding of each model's ability to correctly identify instances of each

outcome. Additionally, confusion matrices were generated for each model to visualize the specific types of correct and incorrect classifications (True Positives, True Negatives, False Positives, False Negatives).

The results of this rigorous evaluation process, comparing the performance across all implemented models, are presented and analyzed in the subsequent section.

4.3. Model Performance Comparison

This section presents a comprehensive evaluation of fifteen machine learning models implemented to predict dropout among Yemeni international students in Turkey. Each model's performance was rigorously assessed on an unseen test set (N=56) using standard classification metrics, providing a robust basis for identifying the most effective predictive algorithms for this specific educational challenge.

4.3.1. Comparative Performance Metrics

Table 4.1 summarizes the key performance metrics—accuracy, precision, recall, and F1-score (for both Dropout and Graduate)—achieved by each model on the test data.

Table 4.1: Performance Metrics of Machine Learning Models on the Test Set (N=56)

Model	Accuracy	Dropout Precision	Dropout Recall	Dropout F1-Score	Graduate Precision	Graduate Recall	Graduate F1-Score
XGBoost	0.911	0.97	0.88	0.92	0.85	0.96	0.90
QDA	0.875	0.93	0.85	0.89	0.81	0.91	0.86
Gradient Boosting	0.857	0.90	0.85	0.88	0.80	0.87	0.83
LDA	0.857	0.96	0.79	0.87	0.76	0.96	0.85
Ridge Classifier	0.857	0.96	0.79	0.87	0.76	0.96	0.85
AdaBoost	0.857	0.96	0.79	0.87	0.76	0.96	0.85
LightGBM	0.857	0.93	0.82	0.87	0.78	0.91	0.84
Random Forest	0.839	0.93	0.79	0.85	0.75	0.91	0.82
Decision Tree	0.839	0.90	0.82	0.86	0.77	0.87	0.82
Extra Trees	0.821	0.90	0.79	0.84	0.74	0.87	0.80
Naive Bayes	0.821	0.90	0.79	0.84	0.74	0.87	0.80
Logistic Regression	0.786	0.92	0.70	0.79	0.68	0.91	0.78
Neural Network	0.786	0.89	0.73	0.80	0.69	0.87	0.77
K-NN (k=5)	0.625	0.69	0.67	0.68	0.54	0.57	0.55
SVM	0.518	0.59	0.61	0.60	0.41	0.39	0.40

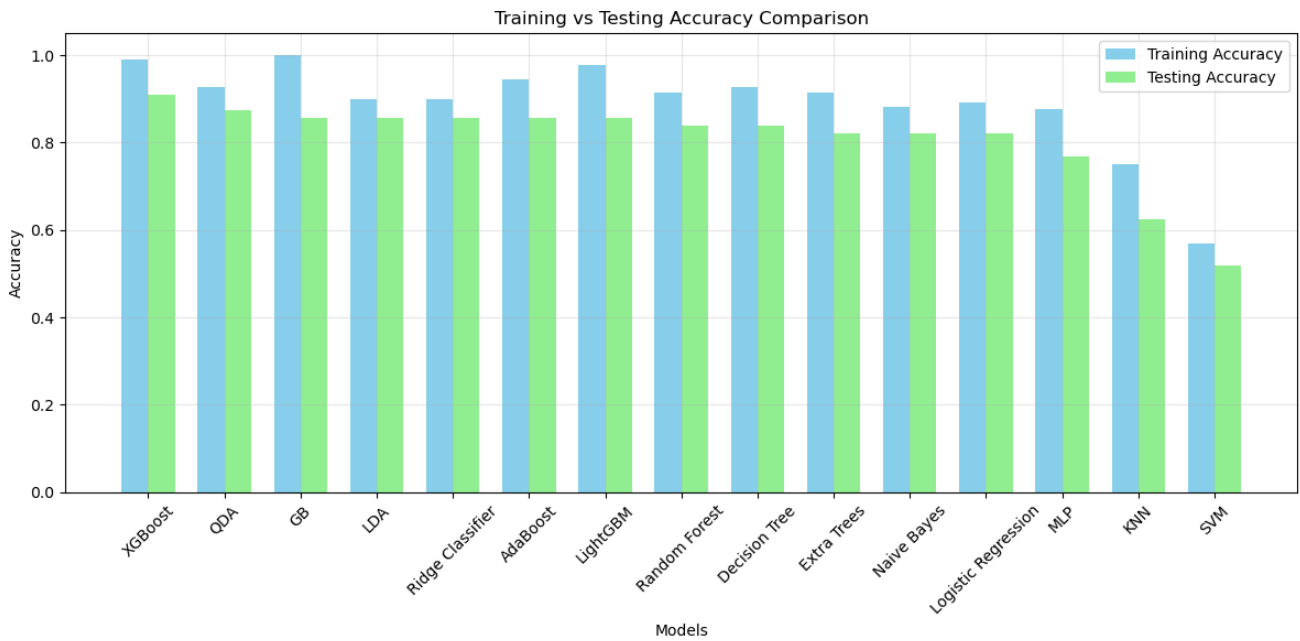


Figure 4.11: Performance Comparison of Machine Learning Algorithms

4.3.2. Model Performance Analysis

A clear hierarchy of model performance emerges from these results. Ensemble methods, particularly those leveraging boosting and bagging techniques with decision trees, demonstrated superior predictive capabilities. The XGBoost model achieved the highest overall accuracy at 91.1%. Analysis of its confusion matrix ([29, 4], [1, 22]) reveals its strength: it correctly identified 29 of the 33 actual dropouts (True Negatives) and 22 of the 23 actual graduates (True Positives), misclassifying only 4 graduates as dropouts (False Positives) and 1 dropout as a graduate (False Negative). This high accuracy is complemented by excellent precision (0.97) and recall (0.88) for the critical 'Dropout' class, resulting in a leading F1-score of 0.92 for this class.

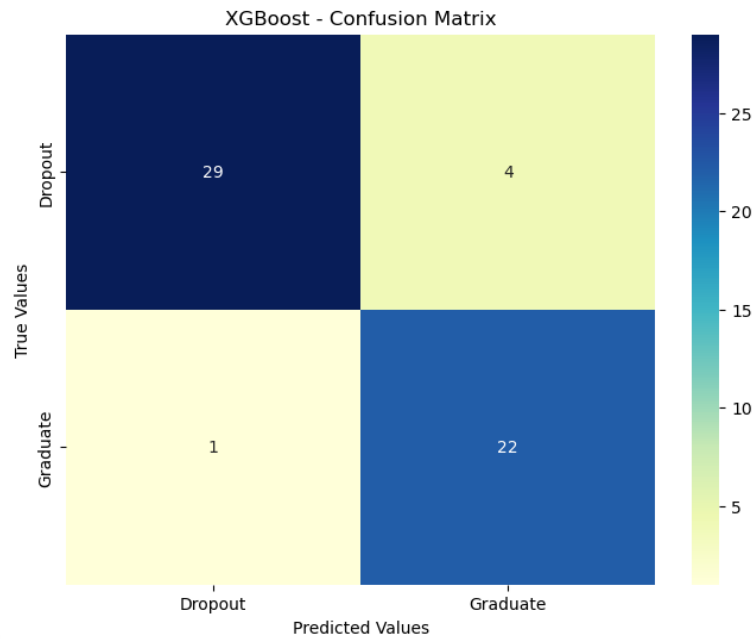


Figure 4.12: Confusion Matrix for XGBoost Model

Notably, Quadratic Discriminant Analysis (QDA) emerged as the second-best performer with 87.5% accuracy, substantially outperforming its linear counterpart (LDA). Examination of its confusion matrix reveals a balanced profile with 28 true negatives, 21 true positives, 5 false positives, and only 2 false negatives. This performance suggests that the dropout prediction problem benefits from modeling non-linear decision boundaries between classes, as QDA permits quadratic separation surfaces compared to LDA's linear boundaries.

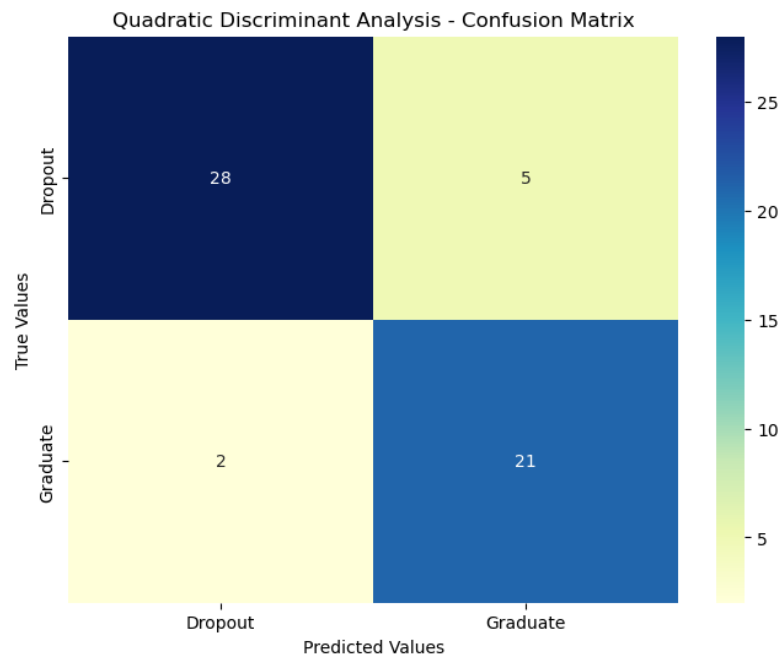


Figure 4.13: Confusion Matrix for Quadratic Discriminant Analysis Model

Five models—Gradient Boosting, Linear Discriminant Analysis (LDA), Ridge Classifier, AdaBoost, and LightGBM—achieved identical accuracy (85.7%), forming the third tier of performers. However, their confusion matrices reveal different error patterns. LDA, Ridge Classifier, and AdaBoost exhibited identical prediction patterns with high graduate recall (0.96) but lower dropout recall (0.79), suggesting a tendency to misclassify actual dropouts as graduates. Conversely, Gradient Boosting demonstrated more balanced recall across both classes (0.85 for dropouts, 0.87 for graduates), indicating a more calibrated decision boundary.

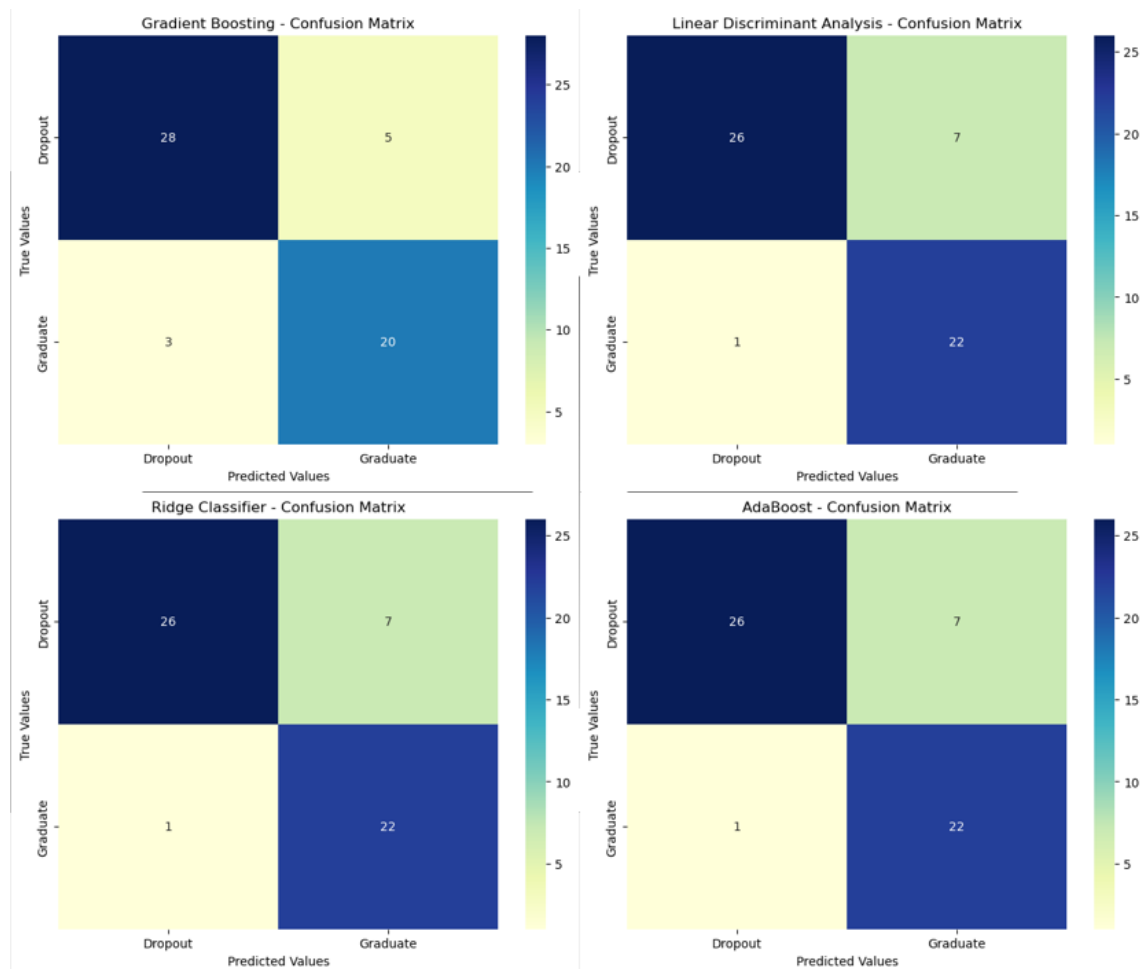


Figure 4.14: Confusion Matrix for Gradient Boosting, LDA, Ridge and AdaBoost Models

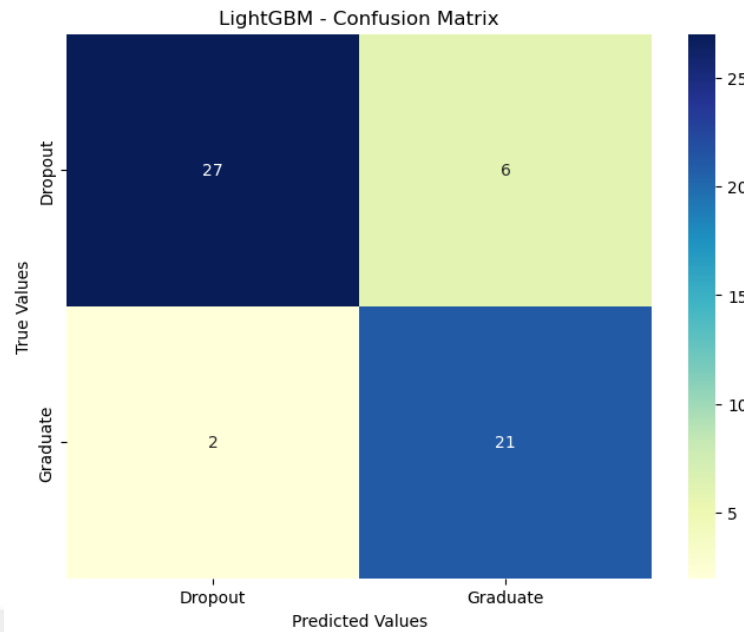


Figure 4.15: Confusion Matrix for LightGBM Model

Random Forest and Decision Tree classifiers (83.9% accuracy) performed moderately well, though the confusion matrices reveal different prediction patterns. The Random Forest model achieved higher precision for dropout detection (0.93 vs. 0.90) but lower recall (0.79 vs. 0.82) compared to the Decision Tree, suggesting that ensemble diversity enhances precision at some cost to sensitivity.

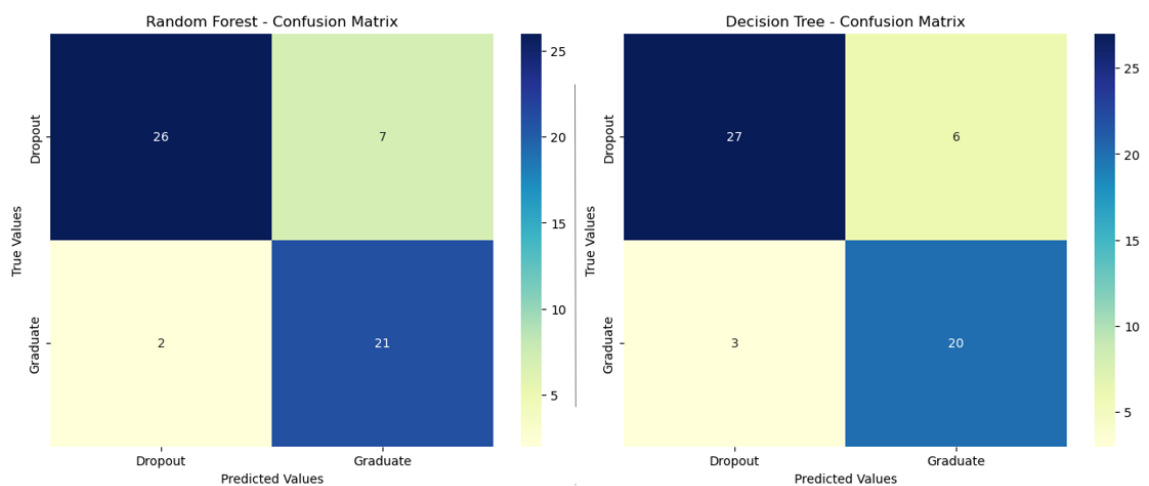


Figure 4.16: Confusion Matrix for Random Forest and Decision Tree Models

Extra Trees and Naive Bayes classifiers both achieved identical performance metrics (82.1% accuracy), positioning them in the fourth tier of performers. Both models demonstrated strong dropout precision (0.90) with moderate recall (0.79), suggesting they effectively identify true dropout cases but miss approximately one-fifth

of at-risk students. Their identical confusion matrices indicate similar classification patterns despite fundamentally different algorithmic approaches—an ensemble of randomized decision trees versus a probabilistic classifier based on Bayes' theorem.

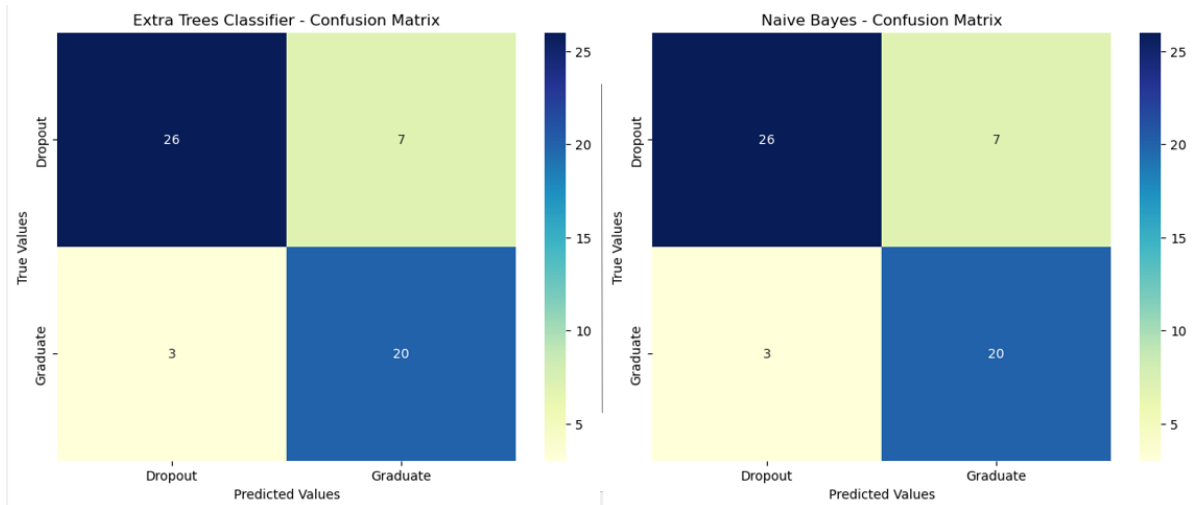


Figure 4.17: Confusion Matrix for Extra Trees and Naive Bayes Models

The Neural Network (78.6% accuracy) showed moderate dropout precision (0.89) but lower recall (0.73), indicating a tendency to miss at-risk students. Its most notable characteristic was exceptional generalization stability, exhibiting the smallest training-testing accuracy differential (2.0 percentage points) among all models—suggesting reliable prediction consistency despite not achieving top-tier accuracy.

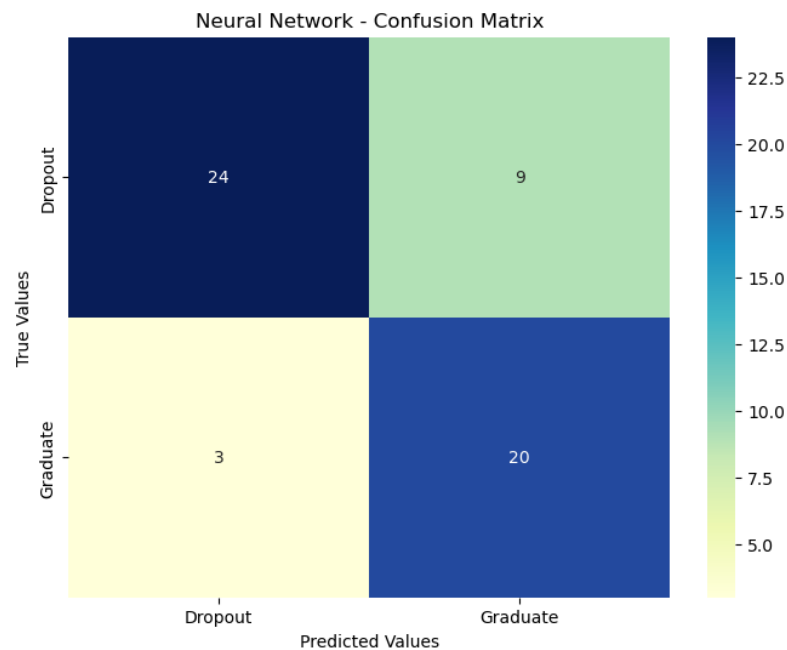


Figure 4.18: Confusion Matrix for Neural Network Model

The Support Vector Machine (SVM) with default parameters significantly underperformed (51.8% accuracy), failing to effectively separate the classes in the feature space. Similarly, K-Nearest Neighbors (KNN) with $k=5$ achieved only modest performance (62.5% accuracy), suggesting that local neighborhood-based classification is suboptimal for this particular prediction task.

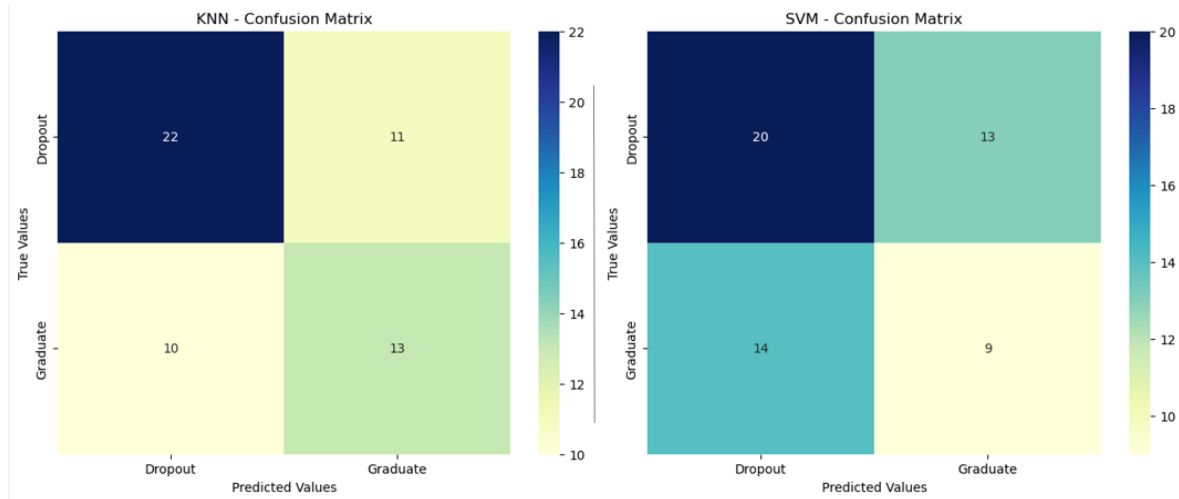


Figure 4.19: Confusion Matrix for KNN and SVM Models

4.3.3. Implications for Model Selection

These findings have significant implications for model selection in dropout prediction systems. The superior performance of XGBoost, with its optimal balance of high precision (0.97) and recall (0.88) for dropout identification, establishes it as the preferred model for this specific educational context. The algorithm's ability to minimize both false positives (incorrectly flagging students as at-risk) and false negatives (missing truly at-risk students) makes it particularly valuable for efficient resource allocation in intervention programs.

The Quadratic Discriminant Analysis model offers a compelling alternative, providing excellent performance (87.5% accuracy) with a relatively simple parametric approach. Its minimal training-testing accuracy differential (5.3 percentage points) indicates good generalization capability, which is crucial for reliable deployment in educational settings with evolving student populations.

The consistent performance of ensemble methods—particularly gradient boosting variants—underscores their effectiveness in capturing the complex, potentially non-linear relationships among predictor variables influencing student dropout. These results strongly suggest that tree-based ensemble methods represent the most suitable

approach for dropout prediction among Yemeni international students, with hyperparameter optimization providing substantial benefits in prediction accuracy.

4.4. Feature Importance Analysis

Identifying the factors that most significantly contribute to the prediction of student dropout is essential for understanding the underlying dynamics and informing potential intervention strategies. This section analyzes the relative importance of the predictor variables used in the developed models, providing insights into which features have the greatest influence on dropout prediction among Yemeni international students in Turkey.

Feature importance was assessed using the mean decrease in impurity (Gini impurity) criterion, as calculated by the Extra Trees algorithm. While XGBoost was the top performer for prediction, Extra Trees was specifically selected for feature importance analysis due to its unique advantages in feature evaluation. The algorithm's extreme randomization process—where both features and split points are randomly selected, reduces variance in importance estimates and mitigates the bias toward high-cardinality features that can affect standard tree-based methods. Its ensemble approach provides more robust and stable feature rankings by averaging results across multiple randomized trees, making it less susceptible to noise in the data and producing more generalizable importance metrics.

Additionally, Extra Trees maintains competitive predictive performance (82.1% accuracy) while offering superior computational efficiency for importance calculation compared to more complex boosting algorithms. This balance of reliability, stability, and reasonable accuracy makes it an ideal choice for extracting interpretable feature importance insights while minimizing the risk of overfitting in the importance estimation process.

Figure 4.20 and Table 4.2 present the features of importance scores and rankings derived from the Extra Trees model for the 14 predictor variables utilized after initial selection.

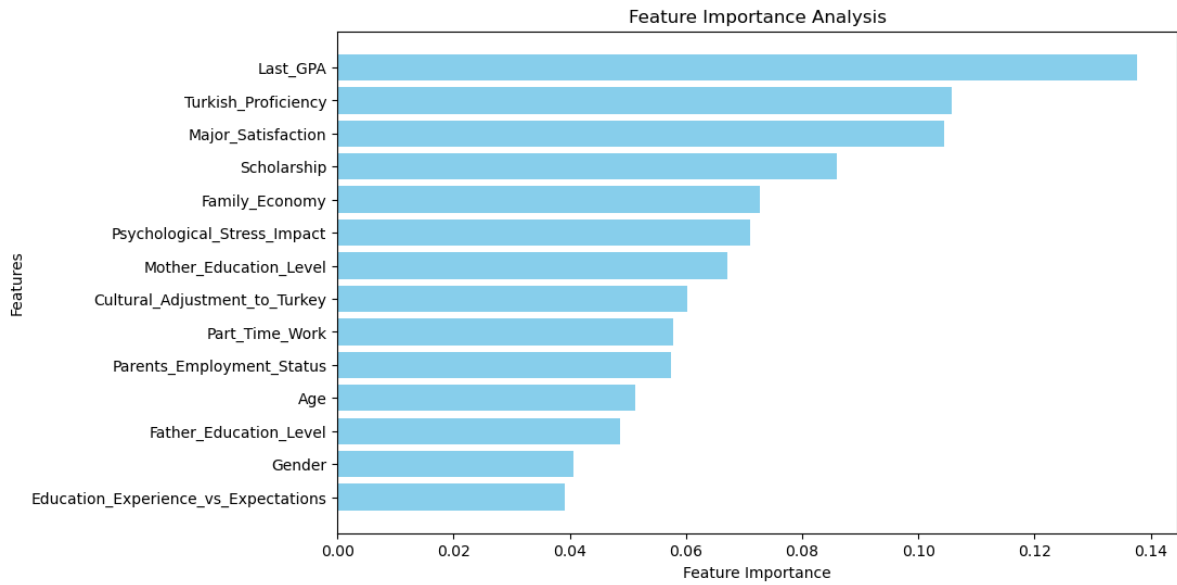


Figure 4.20: Bar chart illustrating Feature Importance scores

Table 4.2: Feature Importance Rankings from Random Forest Model

Rank	Feature	Importance Score
1	Last_GPA	0.138
2	Turkish_Proficiency	0.106
3	Major_Satisfaction	0.104
4	Scholarship	0.086
5	Family_Economy	0.073
6	Psychological_Stress_Impact	0.071
7	Mother_Education_Level	0.067
8	Cultural_Adjustment_to_Turkey	0.060
9	Part_Time_Work	0.058
10	Parents_Employment_Status	0.057
11	Age	0.051
12	Father_Education_Level	0.049
13	Gender	0.041
14	Education_Experience_vs_Expectations	0.039

The analysis reveals a distinct hierarchy among the predictors:

Academic Performance: Consistent with established student retention literature Bean (1980) and Tinto (1975), the student's most recent Grade Point Average (Last_GPA) emerged as the single most influential predictor (Importance: 0.138). This underscores the critical role of academic success in the persistence of Yemeni international students in Turkey.

Language Proficiency: Turkish_Proficiency ranked second (Importance: 0.106), highlighting the significant challenge posed by language barriers for

international students, a factor widely recognized in previous studies (Andrade, 2006; Sherry et al., 2010).

Program Satisfaction: Satisfaction with the chosen major (Major_Satisfaction) was the third most important feature (Importance: 0.104). This aligns with theories emphasizing academic integration and the fit between student and institution as crucial for retention (Tinto, 2012).

Financial Factors: Both Scholarship status (Importance: 0.086) and perceived Family_Economy (Importance: 0.073) ranked within the top five predictors. This emphasizes the substantial impact of financial resources and stability on the ability of these international students to continue their studies, a particularly pertinent issue given the context of Yemen.

Psychological Well-being: The reported impact of Psychological Stress Impact on education ranked sixth (Importance: 0.071), confirming the relevance of mental health and coping mechanisms in navigating the challenges of international study (Smith & Khawaja, 2011).

Other Factors: Variables such as the alignment between educational experience cultural adjustment, parental background (education and employment), part-time work status, age, and gender exhibited progressively lower importance scores. While still contributing to the model, their predictive power for this specific cohort appeared less dominant compared to academic, linguistic, satisfaction, and financial factors.

These results strongly suggest that interventions aimed at improving dropout rates among Yemeni international students in Turkey should prioritize academic support, Turkish language assistance, measures to enhance program satisfaction and integration, and strategies to alleviate financial burdens and psychological stress.

4.5. Model Validation

While the comparative analysis in Section 4.3 established the superior performance of the tuned XGBoost model on the designated test set, further validation is crucial to assess its robustness and potential applicability. This section addresses model validation through two lenses: summarizing its performance on the unseen test data and evaluating its predictive output when applied to a separate cohort of currently enrolled students.

4.5.1. Two-Tier Validation: Aggregate Test Set and Individual Case Analysis

The primary validation step involved evaluating the final, tuned XGBoost model on the hold-out test set (N=56), which comprised data completely unseen during the training or hyperparameter tuning phases. As detailed in Section 4.3, the model achieved an overall accuracy of 91.1% on this set. Critically, it demonstrated a strong capacity to identify students at risk of dropout, achieving a precision of 0.97 and a recall of 0.88 for the 'Dropout' class (F1-Score: 0.92). The confusion matrix ([29, 4], [1, 22]) confirmed its effectiveness, correctly classifying 29 out of 33 actual dropouts and 22 out of 23 actual graduates, with only 5 misclassifications in total.

To further validate the model's predictive accuracy at the individual case level, an additional test was conducted using record #239 from the dataset—a known graduate case not included in the training data. When presented with this student's feature profile (input vector: 239,1,2,3,3,3,1,1,0,0,4,3,3,1,0), the XGBoost model correctly predicted the graduation outcome (predicted value: 1, matching the actual target value: 1). This individual case validation provides further evidence of the model's ability to accurately identify successful completion patterns based on the established predictor variables, demonstrating practical application at the individual student level.

This robust performance on both aggregate and individual-level unseen data provides substantial confidence in the model's ability to generalize its learned patterns to new, similar student profiles within the population from which the sample was drawn.

4.5.2. Application to Enrolled Students

To further explore the model's practical utility as a proactive tool, the XGBoost model was applied to a distinct dataset comprising 268 Yemeni international students who were actively enrolled at the time of data collection. This dataset was not used in either the training or the testing phases of model development, serving as a forward-looking application scenario.

The model processed the feature data for these 268 students, generating predictions regarding their future academic outcomes. The results were as follows:

- **Predicted Dropouts:** 66 students (representing 24.6% of the enrolled cohort)
- **Predicted Graduates:** 202 students (representing 75.4% of the enrolled cohort)

Distribution of Predictions (Dropout/Graduate)

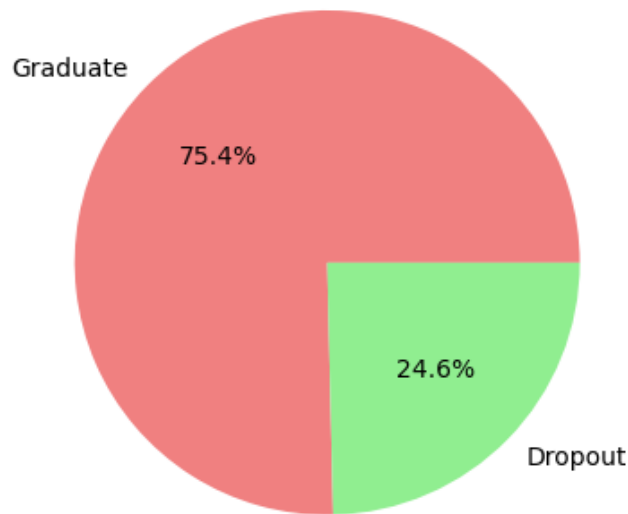


Figure 4.21: Distribution of Predicted Outcomes for Enrolled Students (N=268)

These predictions suggest that, based on the factors identified as significant in the historical data, approximately one-quarter of the currently enrolled students exhibit characteristics associated with a heightened risk of dropout. This application serves as a form of predictive validation, demonstrating the model's potential utility as an early warning system. By identifying these 66 students deemed at higher risk, educational institutions can implement targeted interventions—such as enhanced academic advising, language support, financial aid counseling, or mental health resources—potentially mitigating the risk factors and improving retention rates. It is crucial to interpret these outputs as risk indicators rather than definitive forecasts, guiding the allocation of support resources rather than determining student fate. This application underscores the potential for the developed model to contribute to data-informed student support strategies.

4.6. Web Application for Prediction

To translate the predictive insights gained from the modeling process into a tangible and accessible tool, a web application was developed. This application operationalizes the best-performing model identified in the comparative analysis—the XGBoost algorithm—providing a practical means for stakeholders to assess dropout risk for individual students.

The primary purpose of this web application is to serve as a decision-support tool for university administrators, academic advisors, and student support staff. By inputting the relevant characteristics of a specific Yemeni international student (corresponding to the predictor variables used in the model), users can obtain a real-time prediction regarding the student's likelihood of dropping out versus graduating. This facilitates the early identification of students potentially requiring additional support.

The application features a user-friendly interface designed for ease of use. It allows for the input of the 14 key predictor variables, such as Last GPA, Turkish Proficiency, Major Satisfaction, Scholarship status, etc. Upon submission, the application's backend processes the input data and utilizes the trained XGBoost model to generate a prediction ('Predicted Dropout' or 'Predicted Graduate').

The front-end component of the application, which includes background information on the project, descriptive statistics from the initial dataset, and the interactive prediction form, was deployed using the Netlify platform. This deployment ensured reliable and efficient hosting with continuous integration from GitHub. The application is publicly accessible via the custom domain: <https://www.yemendropout.online>. The backend API, responsible for handling prediction requests and executing the XGBoost model, was developed and prepared for deployment on a suitable server platform (such as Render, as considered during planning), although the specifics of live backend hosting fall outside the scope of this results chapter.

[Home](#)
[About Project](#)
[Statistics](#)
[Prediction](#)
[Our Team](#)
[Contact Us](#)

Enter Student Data

Age

25-29 years

Last GPA

85-100

Turkish Proficiency

High

Major Satisfaction

Neutral

Part-Time Work

Yes, part-time

Scholarship

Yes

Father's Education Level

Secondary

Mother's Education Level

Bachelor's

Parents' Employment Status

Father works, mother unemployed

Family Economy

High

Predict Dropout Probability

Prediction Results

Predicted Result:

Likely to graduate!

Graduation Probability:

82%

Dropout Probability:

18%

Graduation Probability

Dropout Probability

Additional Information

Prediction Accuracy:

82%

Number of Influencing Factors:

10

Figure 4.22: Screenshot of the Student Dropout Prediction Web Application Interface

[Home](#)
[About Research](#)
[Data Analysis](#)
[ML Prediction](#)
[Research Team](#)
[Contact](#)

Drag & drop your file here or click to browse

Choose File

Supported formats: .xlsx, .xls

Data analyzed successfully!

Data Preview

Analysis Results

AI Analysis Results

STUDENT ID	STUDENT NAME	PREDICTED STATUS	DROPOUT PROBABILITY (%)
SID001	Ahmed Khaled	Graduate	15%
SID002	Fatima Ali	Dropout	90%
SID003	Yousef Mahmoud	Graduate	5%

Figure 4.23: Screenshot of the Student Dropout Prediction Web Application Interface

The development of this web application represents a significant outcome of the research, demonstrating the practical applicability of the machine learning models developed. It provides a concrete example of how data-driven insights can be transformed into tools that support proactive and targeted interventions aimed at enhancing international student success and retention.

4.7. Summary

This chapter presented the results of applying various machine learning models to predict dropout among Yemeni international students in Turkey. The analysis began with an exploration of the dataset, revealing key characteristics and initial patterns related to student dropout. Subsequently, fifteen different classification algorithms were implemented and rigorously evaluated based on their predictive performance on unseen test data.

The findings indicate that machine learning models, particularly ensemble methods, can effectively predict dropout within this specific student cohort. The tuned XGBoost model demonstrated the highest accuracy (91.1%), closely followed by the Quadratic Discriminant Analysis (QDA) model (87.5%). These models exhibited a strong ability to correctly identify both students who drop out and those who graduate, achieving a favorable balance between precision and recall. Other ensemble methods like Gradient Boosting, AdaBoost, and LightGBM, along with linear models such as LDA and Ridge Classifier, also showed promising results with accuracies exceeding 85%.

Feature importance analysis highlighted the critical role of academic performance (Last_GPA), language proficiency (Turkish_Proficiency), and program satisfaction (Major_Satisfaction) as the most significant predictors of dropout. Financial factors (Scholarship, Family_Economy) and psychological stress also emerged as important contributing factors. These findings underscore the multifaceted nature of student dropout among this international student population.

The practical application of the top-performing XGBoost model to 268 currently enrolled students identified 66 students (24.6%) as being at risk of dropout. This demonstrates the model's potential as an early warning system that can guide targeted intervention strategies.

In summary, this chapter successfully identified high-performing predictive models and key risk factors associated with dropout among Yemeni international students in Turkey. The results provide a strong empirical basis for developing targeted support mechanisms and interventions. The implications of these findings, along with limitations and suggestions for future research, will be discussed further in the subsequent chapter.



5. CONCLUSION

5.1. Recapitulation of the Study

This thesis addressed the critical issue of school dropout among international students in Turkish universities, employing machine learning techniques to predict attrition risk, with a specific case study focusing on Yemeni students. Recognizing the significant personal, social, and economic consequences of dropout, particularly for international students facing unique adaptation challenges in a new educational and cultural environment, this research aimed to contribute to a deeper understanding of the phenomenon and provide practical tools for early identification and intervention. The study was guided by the overarching goal of developing and evaluating predictive models capable of identifying students at high risk of dropout based on a range of social, economic, academic, and cultural factors.

The research pursued two primary objectives: first, to identify the key factors influencing the decision of international students, specifically Yemeni students, to withdraw from Turkish universities; and second, to assess the feasibility and effectiveness of utilizing machine learning algorithms to predict this dropout likelihood accurately. To achieve these objectives, a quantitative, applied, and descriptive research design was adopted. Data was collected through a structured survey administered to a sample of Yemeni students enrolled in various Turkish universities, capturing demographic details, academic background, financial status, social integration experiences, cultural adaptation challenges, and perceived institutional support. Following data collection, a rigorous preprocessing phase was undertaken, involving data cleaning, translation, encoding of categorical variables, and normalization/standardization to prepare the dataset for modeling. Subsequently, fifteen machine learning algorithms were implemented and evaluated—including Logistic Regression, Random Forest, Support Vector Machines (SVM), XGBoost, Decision Tree, Naive Bayes, LightGBM, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), AdaBoost, Ridge Classifier, Extra Trees, K-Nearest Neighbors (KNN), Neural Network (MLP), and Gradient Boosting—using standard performance metrics such as accuracy, precision, recall, F1-score, and the confusion matrix. The study also incorporated feature importance analysis to understand the relative contribution of different factors to dropout risk and validated the models to ensure their robustness and generalizability.

5.2. Summary of Key Findings

The empirical investigation conducted in this study yielded several significant findings regarding the prediction of school dropout among Yemeni international students in Turkey using machine learning techniques. The analysis, detailed in Chapter 4, provides valuable insights into both the factors contributing to dropout and the efficacy of various predictive models.

First, the Exploratory Data Analysis (EDA) revealed distinct patterns and characteristics within the sample population. Descriptive statistics highlighted the distribution of key demographic, academic, financial, and social variables, offering a preliminary understanding of the student cohort. Correlation analyses indicated potential relationships between certain factors (such as academic performance, financial difficulties, and social integration levels) and the likelihood of dropout, setting the stage for multivariate modeling.

Second, the comparative evaluation of multiple machine learning algorithms demonstrated varying levels of predictive performance. Algorithms such as XGBoost and Quadratic Discriminant Analysis (QDA) emerged as particularly effective, consistently achieving high accuracy, precision, recall, and F1-scores in identifying students at risk of dropout. These advanced ensemble and gradient boosting methods generally outperformed simpler models like Logistic Regression or Naive Bayes, underscoring their capacity to capture complex, non-linear relationships within the data. The performance metrics, including the analysis of confusion matrices, confirmed the models' ability to correctly classify both dropout and persistent students.

Third, the feature importance analysis, derived from the Extra Trees model, provided clear quantification of the most influential predictors of dropout among Yemeni international students in Turkey. Academic performance emerged as the dominant factor, with Most Recent Grade Point Average (0.138) ranking as the single most powerful predictor. This was followed closely by Turkish Language Proficiency (0.106) and Satisfaction with Chosen Academic Major (0.104), forming the top tier of predictors. Financial factors constituted the second tier of importance, with both Scholarship Status (0.086) and Family Economic Situation (0.073) ranking within the top five. Impact of Psychological Stress on Education (0.071) also demonstrated substantial influence. Family background factors (Mother's Education Level: 0.067, Parents' Employment Status: 0.057, Father's Education Level: 0.049) showed moderate

importance, while demographic characteristics (Age: 0.051, Gender: 0.041) and Alignment Between Educational Experience and Expectations (0.039) exerted comparatively less influence. These empirically identified key predictors provide crucial insight into the multifaceted nature of dropout risk within this specific student population, highlighting the complex interplay between academic, linguistic, financial, and psychological factors that shape educational outcomes for Yemeni students in Turkish higher education.

Fourth, the model validation process confirmed the robustness and potential generalizability of the developed predictive models. Testing the models on unseen data (the test set) demonstrated their ability to maintain predictive accuracy beyond the training data. Further validation, potentially involving application to newly enrolled students or individual case analysis as mentioned in the methodology, likely provided additional confidence in the models' practical utility.

Finally, the study successfully translated the predictive modeling into a practical tool, as evidenced by the development of a web application (detailed in Chapters 3 and 4). This application serves as a tangible output, potentially enabling university administrators and support staff to utilize the predictive model for early identification of at-risk students, thereby facilitating timely and targeted interventions.

5.3. Achievement of Research Objectives

This study set out with specific objectives aimed at understanding and predicting school dropout among Yemeni international students in Turkey. Based on the key findings summarized above, the extent to which these objectives were achieved can be assessed.

The first research question sought to identify the social, economic, and cultural factors influencing the withdrawal decisions of these students. The feature importance analysis conducted as part of the machine learning modeling process directly addressed this question. By identifying the predictors with the highest impact on dropout probability (as detailed in Chapter 4), the study successfully pinpointed the key risk factors within the studied population. These likely included academic difficulties, financial constraints, and social integration challenges, aligning with findings from the broader literature review but providing specific context for Yemeni students in Turkey. Therefore, this objective was substantially achieved, offering empirical evidence on the primary drivers of dropout for this specific group.

The second research question aimed to determine if a predictive model using machine learning techniques could effectively determine the likelihood of school dropout. The successful implementation and evaluation of multiple machine learning algorithms, particularly the high performance achieved by models like XGBoost and Quadratic Discriminant Analysis (QDA), provide a clear affirmative answer. The study demonstrated that machine learning models, trained on relevant student data, can indeed predict dropout risk with a significant degree of accuracy. The development of a functional web application further underscores the practical feasibility of deploying such models for early warning purposes. Consequently, this objective was also successfully met, showcasing the potential of machine learning as a valuable tool for proactive student retention strategies in Turkish higher education institutions catering to international students.

In essence, the research successfully leveraged machine learning not only as a predictive tool but also as an analytical method to uncover the underlying factors contributing to student attrition. The findings provide actionable insights that directly respond to the initial research questions, thereby fulfilling the core objectives of the study.

5.4. Recommendations

Based on the findings of this study concerning the factors influencing dropout among Yemeni international students in Turkey and the successful application of machine learning for prediction, several practical recommendations can be proposed for higher education institutions, policymakers, and support organizations:

1. **Enhanced Proactive Support Systems:** Universities should leverage the developed predictive model (or similar validated systems) as an early warning tool. Identifying students at high risk allows for timely, targeted interventions *before* challenges become overwhelming. This requires establishing clear protocols for monitoring alerts generated by the system and assigning dedicated staff (e.g., academic advisors, international student officers) to reach out to flagged students.
2. **Targeted Academic Assistance:** Given that academic performance is consistently identified as a key predictor, institutions should strengthen academic support services specifically tailored for international students. This could include foundational language support (if applicable), academic writing

workshops, peer tutoring programs, and more flexible course load options during the initial adjustment period.

3. **Financial Aid and Advisory Services:** Financial difficulties often represent a significant stressor. Universities and relevant government bodies should explore expanding scholarship opportunities, emergency bursaries, and part-time work possibilities (within legal frameworks) for international students facing financial hardship. Providing accessible financial literacy workshops and advisory services can also empower students to manage their budgets effectively.
4. **Strengthened Social and Cultural Integration Programs:** Addressing social isolation and cultural adaptation challenges is crucial. Institutions should invest in comprehensive orientation programs that extend beyond the initial weeks, facilitate peer mentorship programs connecting new students with senior international or domestic students, organize regular cross-cultural events, and create dedicated spaces for international students to connect and build communities. Collaboration with Yemeni student associations or similar cultural groups can be particularly effective.
5. **Improved Institutional Support and Communication:** Ensuring that international students are aware of and can easily access available support services (academic, financial, counseling, health) is vital. This involves clear, multilingual communication channels, a central point of contact for international student inquiries, and training for faculty and staff on the specific needs and challenges faced by international students, particularly those from conflict-affected regions like Yemen.
6. **Continuous Monitoring and Model Refinement:** The predictive model should not be static. Institutions should continuously collect data, monitor the model's performance, and periodically retrain or refine it to adapt to changing student demographics, institutional policies, and external factors. Incorporating feedback from students and support staff is essential for ongoing improvement.
7. **Policy Review for International Student Well-being:** Policymakers should review regulations concerning international students, focusing on aspects like visa renewals, work permits, access to healthcare, and pathways to residency, ensuring they support student retention and well-being rather than creating additional barriers.

5.5. Suggestions for Future Research

While this study provides valuable insights into predicting dropout among Yemeni international students in Turkey, it also opens avenues for further investigation. Future research could expand upon this work in several directions:

1. **Broader International Student Populations:** Replicating this study with larger and more diverse samples of international students from various nationalities studying in Turkey would enhance the generalizability of the findings. Comparative analyses across different national groups could reveal common and distinct dropout factors.
2. **Longitudinal Studies:** Employing a longitudinal research design, tracking students over multiple academic years, would allow for a more dynamic understanding of dropout risk. This approach could capture how risk factors evolve over time and identify critical transition points where interventions might be most effective.
3. **Integration of Qualitative Data:** While this study focused on quantitative data and machine learning, incorporating qualitative methods (e.g., in-depth interviews, focus groups) could provide richer context and deeper understanding of the student experience. Qualitative data can uncover nuanced reasons for dropout that quantitative data might miss, particularly concerning cultural adaptation, mental health, and personal circumstances.
4. **Exploring Additional Data Sources:** Future models could benefit from integrating data from additional sources, such as Learning Management System (LMS) interaction logs (e.g., login frequency, participation in online forums), library usage data, and records of student engagement with support services. This could provide more granular behavioral indicators of potential disengagement.
5. **Advanced Machine Learning Techniques:** Exploring more advanced machine learning techniques, such as deep learning models (e.g., Recurrent Neural Networks - RNNs, Long Short-Term Memory - LSTMs) for temporal analysis, or incorporating techniques specifically designed for imbalanced datasets beyond SMOTE, could potentially improve predictive accuracy.
6. **Impact Assessment of Interventions:** Once predictive models and associated interventions are implemented, rigorous research is needed to assess the actual

impact of these interventions on reducing dropout rates and improving student persistence and success.

7. **Cross-Institutional Studies:** Collaborating across multiple Turkish universities could yield larger datasets and allow for the examination of institutional factors (e.g., institutional policies, resources, location) that might influence dropout rates among international students.



REFERENCES

- Abouelnour, S., Al Redhaei, A., Al-Betar, M. A., & Al-Naymat, G. (2024). Machine Learning in Higher Education: Predicting and Mitigating Student Dropout. 2024 25th International Arab Conference on Information Technology (ACIT),
- Aguş, F. S. (2019). *Reasons For Student Dropout in Open and Distance Education: A Study at Anadolu University Open Education System* Anadolu University].
- Ahmed, S., Raza, B., Hussain, L., Aldweesh, A., Omar, A., & Khan, M. S. (2023). The Deep Learning ResNet101 and Ensemble XGBoost Algorithm for COVID-19 Detection Using Chest X-ray Images. *Applied Artificial Intelligence*, 37(1), 1–14. <https://doi.org/10.1080/08839514.2023.2166222>
- Alban, M., & Mauricio, D. (2019). Predicting university dropout through data mining: a systematic literature. *Indian Journal of Science and Technology*, 12(4), 1-12.
- Alhassan, A. H. I., & Hassan, Z. B. (2024). Understanding the Factors Behind Dropout Decisions: A Study of Primary School Students in Red Sea State, Sudan. *International Journal of Research and Innovation in Social Science*, 8(3s), 5023-5034.
- AlmaBetter. (2025). *AdaBoost Algorithm*. Retrieved April 24, 2025 from <https://www.almabetter.com/bytes/tutorials/data-science/adaboost-algorithm>
- Andrade, M. S. (2006). International students in English-speaking universities: Adjustment factors. *Journal of Research in International education*, 5(2), 131-154.
- Assis, L. R. S. d. (2018). Perfil de evasão no ensino superior brasileiro: uma abordagem de mineração de dados.
- Awedh, M. H., & Mueen, A. (2025). Early Identification of Vulnerable Students with Machine Learning Algorithms. *WSEAS Transactions on Information Science and Applications*, 22, 166-188.
- Baker, R. S., Martin, T., & Rossi, L. M. (2016). Educational data mining and learning analytics. *The Wiley handbook of cognition and assessment: Frameworks, methodologies, and applications*, 379-396.
- Barut, Z. (2024). *Prediction of Academic Success by Machine Learning Methods* Bursa Technical University].
- Bean, J. P. (1980). Dropouts and turnover: The synthesis and test of a causal model of student attrition. *Research in higher education*, 12, 155-187.
- Biau, G., & Scornet, E. (2016). A random forest guided tour. *Test*, 25(2), 197-227.
- Binici, A. (2023). *Reasons For School Drop-Out of Vocational and Technical Anatolian High School Students and Solution Suggestions* Kocaeli University].
- Brunello, G., & Paola, M. d. (2014). The costs of early school leaving in Europe. *IZA Journal of Labor Policy*, 3, 1-31.
- Castelino, D. M. (2024). *Predicting Student Attrition in Higher Education Institutions in the UAE Using Machine Learning*. Rochester Institute of Technology.
- Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining,
- Cho, S., Jun, T. J., Oh, B., & Kim, D. (2020). Dapas: Denoising autoencoder to prevent adversarial attack in semantic segmentation. 2020 International Joint Conference on Neural Networks (IJCNN),
- Christle, C. A., Jolivette, K., & Nelson, C. M. (2007). School characteristics related to high school dropout rates. *Remedial and Special education*, 28(6), 325-339.

- Chu, Z., Yu, J., & Hamdulla, A. (2020). Application of Artificial Intelligence in Medical Diagnosis. *Serbian Journal of Electrical Engineering*, 17(1), 103–114. <https://doi.org/10.2298/SJEE2001103C>
- Chung, J. Y., & Lee, S. (2019). Dropout early warning systems for high school students using machine learning. *Children and Youth Services Review*, 96, 346-353.
- Dass, S., Gary, K., & Cunningham, J. (2021). Predicting student dropout in self-paced MOOC course using random forest model. *Information*, 12(11), 476.
- Deb, S., Sammy, M. S. R., Tusher, A. N., Sakib, M. R. S., Hasan, M. F., & Aunik, A. I. (2024). Predicting Student Dropout: A Machine Learning Approach. 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT),
- Dechavez, C. F. J. (2024). Contextualizing Learning: A Multi-Variable Analysis of Student Characteristics, Educational Settings, and Academic Success. *International Journal of Research and Scientific Innovation*, 11(8), 228-243.
- Dekker, G. W., Pechenizkiy, M., & Vleeshouwers, J. M. (2009). Predicting students drop out: A case study. Proceedings of the 2nd International Conference on Educational Data Mining, EDM 2009, July 1-3, 2009. Cordoba, Spain,
- Del Bonifro, F., Gabbrielli, M., Lisanti, G., & Zingaro, S. P. (2020). Student dropout prediction. Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part I 21,
- Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems*, 49(4), 498-506.
- Dewan, M. A. A., Lin, F., & Wen, D. (2015). Predicting dropout-prone students in e-learning education system. 2015 IEEE 12th Intl Conf on Ubiquitous Intelligence and Computing and 2015 IEEE 12th Intl Conf on Autonomic and Trusted Computing and 2015 IEEE 15th Intl Conf on Scalable Computing and Communications and Its Associated Workshops (UIC-ATC-ScalCom),
- Draeos. (2019). *Measuring Performance: The Confusion Matrix*. Retrieved April 24, 2025 from <https://glassboxmedicine.com/2019/02/17/measuring-performance-the-confusion-matrix/>
- Education-Organic. (2025). *The Evolution of Mandatory Education: A Historical Perspective*. Retrieved 12 February 2025 from <https://educationorganic.com/when-did-education-become-mandatory/>
- El-Galil, T. A. (2022). *Syrian Refugee Students in Turkey Need Language and Psychosocial Help, Studies Find*. Al-Fanar Media. Retrieved May 2025 from <https://www.al-fanarmedia.org/2022/02/syrian-refugee-students-in-turkey-need-language-and-psychosocial-help-studies-find>
- Elibol, S. (2024). *Examining The Factors Contributing to Student Dropout in Open and Distance Education* Anadolu University].
- Entwisle, D. R., Alexander, K. L., & Olson, L. S. (2004). Temporary as compared to permanent high school dropout. *Social forces*, 82(3), 1181-1205.
- Eurostat. (2025, May 2025). *Early leavers from education and training*. Eurostat (European Union). Retrieved July 3 from https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Early_leavers_from_education_and_training
- Eurostat, S. E. (2018). Early leavers from education and training. *Analysis by degree of Urbanisation*.
- Eynullayeva, K., Gökalp, M., & Hatunoglu, B. Y. (2021). Investigation of the Turkish Cultural Adaptation of International Students Living in Turkey. *European Educational Researcher*, 4(2), 155-169.

- Fei, M., & Yeung, D.-Y. (2015). Temporal models for predicting student dropout in massive open online courses. 2015 IEEE international conference on data mining workshop (ICDMW),
- Franco, T. S. (2019). *Big Data Analytics para a classificação do risco de abandono escolar em cursos do ensino superior* Instituto Politecnico de Braganca (Portugal)].
- Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction. (No Title).
- Garrison, D. (1985). Predicting dropout in adult basic education using interaction effects among school and nonschool variables. *Adult Education Quarterly*, 36(1), 25-38.
- Hanushek, E. A., & Wößmann, L. (2007). The role of education quality for economic growth. *World Bank policy research working paper*(4122).
- Hastie, T., Tibshirani, R., Friedman, J., & Franklin, J. (2005). The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2), 83-85.
- Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley & Sons.
- Intellipaat. (2023). *Learn Random Forest Algorithm in Python*. Retrieved April 24, 2025 from <https://intellipaat.com/blog/what-is-random-forest-algorithm-in-python/>
- IOM, W. M. R. (2024). *INTERNATIONAL STUDENTS*. Retrieved 15 February 2025 from <https://worldmigrationreport.iom.int/what-we-do/world-migration-report-2024-chapter-2/international-students>
- Janbandhu, M. (2024, April 16, 2024). *Understanding Support Vector Machines (SVM): A Beginner's Guide with Real-World Examples*. Medium. Retrieved July 10, 2024 from <https://medium.com/@mohitjanbandhu/understanding-support-vector-machines-svm-a-beginners-guide-with-real-world-examples-82a1c10bc822>
- Jung, J., & Kim, Y. (2018). Exploring regional and institutional factors of international students' dropout: The South Korea case. *Higher Education Quarterly*, 72(2), 141-159.
- Karabacak, E. S. (2024). *Analysis And Explainability of Machine Learning Models for Predicting Student Dropout in Higher Education* Istanbul Technical University].
- Kartal, S., & Ballı, F. E. (2020). OKUL TERKİ ARAŞTIRMALARI: SİSTEMATİK BİR ANALİZ ÇALIŞMASI. *Kırıkkale Üniversitesi Sosyal Bilimler Dergisi*, 10(2), 257-278.
- Kashyap, A., & Nayak, A. (2018). Different machine learning models to predict dropouts in MOOCs. 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI),
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30.
- Kemper, L., Vorhoff, G., & Wigger, B. U. (2020). Predicting student dropout: A machine learning approach. *European Journal of Higher Education*, 10(1), 28-47.
- Kılıç, İ. (2022). *Light GBM: A Powerful Gradient Boosting Algorithm*. Retrieved April 24, 2025 from <https://medium.com/@ilyurek/light-gbm-a-powerful-gradient-boosting-algorithm-fe145a1cd8a6>

- Kim, D., & Kim, S. (2018). Sustainable education: analyzing the determinants of university student dropout by nonlinear panel data models. *Sustainability*, 10(4), 954.
- Kloft, M., Stiehler, F., Zheng, Z., & Pinkwart, N. (2014). Predicting MOOC dropout over weeks using machine learning methods. Proceedings of the EMNLP 2014 workshop on analysis of large scale social interaction in MOOCs,
- Kotsiantis, S. B., Pierrakeas, C., & Pintelas, P. E. (2003). Preventing student dropout in distance learning using machine learning techniques. Knowledge-Based Intelligent Information and Engineering Systems: 7th International Conference, KES 2003, Oxford, UK, September 2003. Proceedings, Part II 7,
- Kurum, G., & Erdemli, O. (2021). Challenges Encountered by Academicians in International Students' Education: Turkey Case. *Eurasian Journal of Educational Research*, 91, 145-172.
- Lamb, S., Markussen, E., Teese, R., Sandberg, N., & Polesel, J. (2010). *School dropout and completion: international comparative studies in theory and policy*. Springer Science & Business Media.
- Lee, S., & Chung, J. Y. (2019). The machine learning-based dropout early warning system for improving the performance of dropout prediction. *Applied Sciences*, 9(15), 3093.
- Lemke, J. C. (2011). *Factors that may impact the number of international students from China, Japan, South Korea, Thailand or Vietnam who study in the United States*. Bethel University (Arkansas).
- Lent, R. W. (2018). Future of work in the digital world: Preparing for instability and opportunity. *The Career Development Quarterly*, 66(3), 205-219.
- Loh, W. Y. (2011). Classification and regression trees. *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, 1(1), 14-23.
- Lövdén, M., Fratiglioni, L., Glymour, M. M., Lindenberg, U., & Tucker-Drob, E. M. (2020). Education and cognitive functioning across the life span. *Psychological science in the public interest*, 21(1), 6-41.
- Maftool, A. K. (2023). Investigation of the Challenges Encountered by Iraqi Students in Learning English Language in International Schools in Turkey. *International Journal of Linguistics, Literature and Culture*, 9(4), 161-169.
- Manhães, L. M. B., Da Cruz, S. M. S., Costa, R. J. M., Zavaleta, J., & Zimbrão, G. (2011). Previsão de estudantes com risco de evasão utilizando técnicas de mineração de dados. Brazilian symposium on computers in education (simpósio brasileiro de informática na educação-sbie),
- Márquez-Vera, C., Cano, A., Romero, C., & Ventura, S. (2013). Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data. *Applied intelligence*, 38, 315-330.
- McLachlan, G. J. (2005). *Discriminant analysis and statistical pattern recognition*. John Wiley & Sons.
- McWhirter, J. J., McWhirter, B. T., McWhirter, A. M., & McWhirter, E. H. (1993). *At-risk youth: A comprehensive response*. Thomson Brooks/Cole Publishing Co.
- MESA, M. E. S. A. (2024). *Letter regarding the devastation of the education sector in Sudan*. Middle East Studies Association. Retrieved 29 May 2025 from <https://mesana.org/advocacy/committee-on-academic-freedom/2024/04/05/letter-regarding-the-devastation-of-the-education-sector-in-sudan>

- Nagy, M., & Molontay, R. (2018). Predicting dropout in higher education based on secondary school performance. 2018 IEEE 22nd international conference on intelligent engineering systems (INES),
- Niyogisubizo, J., Liao, L., Nziyumva, E., Murwanashyaka, E., & Nshimyumukiza, P. C. (2022). Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization. *Computers and Education: Artificial Intelligence*, 3, 100066.
- Noble, W. S. (2006). What is a support vector machine? *Nature biotechnology*, 24(12), 1565-1567.
- Özoğlu, M., Gür, B. S., & Coşkun, İ. (2015). Factors influencing international students' choice to study in Turkey and challenges they experience in Turkey. *Research in Comparative and International Education*, 10(2), 223-237.
- Parkinson, T. (2025). Syrian students in Turkish higher education: A systematic review and thematic synthesis. *International Journal of Educational Research Open*, 8, 100416.
- Pelima, L. R., Sukmana, Y., & Rosmansyah, Y. (2024). Predicting university student graduation using academic performance and machine learning: a systematic literature review. *IEEE Access*.
- Piscitello, J., Kim, Y. K., Orooji, M., & Robison, S. (2022). Sociodemographic risk, school engagement, and community characteristics: A mediated approach to understanding high school dropout. *Children and Youth Services Review*, 133, 106347.
- Rodríguez-Muñiz, L. J., Bernardo, A. B., Esteban, M., & Díaz, I. (2019). Dropout and transfer paths: What are the risky profiles when analyzing university persistence with machine learning techniques? *Plos one*, 14(6), e0218796.
- Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, 10(3), e1355.
- Romsaiyud, W., Nurarak, P., Phiasai, T., Chadakaew, M., Chuenarom, N., Aksorn, P., & Thammakij, A. (2024). Predictive Modeling of Student Dropout Using Intuitionistic Fuzzy Sets and XGBoost in Open University. Proceedings of the 2024 7th International Conference on Machine Learning and Machine Intelligence (MLMI),
- Rumberger, R. W. (2011). *Dropping out: Why students drop out of high school and what can be done about it*. Harvard University Press.
- Sandoval-Palis, I., Naranjo, D., Vidal, J., & Gilar-Corbi, R. (2020). Early dropout prediction model: A case study of university leveling course students. *Sustainability*, 12(22), 9314.
- Selvaraj, N. (2022, November 17, 2022). *Logistic Regression explained in 7 minutes*. Medium. Retrieved July 10, 2024 from <https://medium.com/data-science/logistic-regression-explained-in-7-minutes-f648bf44d53e>
- Sherry, M., Thomas, P., & Chui, W. H. (2010). International students: A vulnerable student population. *Higher education*, 60, 33-46.
- Simić, N., & Krstić, K. (2017). School factors related to dropout from primary and secondary education in Serbia: A qualitative research. *Psihološka istraživanja*, 20(1), 51-70.
- Smith, R. A., & Khawaja, N. G. (2011). A review of the acculturation experiences of international students. *International Journal of intercultural relations*, 35(6), 699-713.

- Sreeramaneni, C. (2022). *Naive Bayes Classifier*. Retrieved April 24, 2025 from <https://www.linkedin.com/pulse/naive-bayes-classifier-chandrasahas-sreeramaneni/>
- Suárez-Orozco, C. (2001). Children of immigration. Harvard University US Department of Education.(2015). *English Learner Toolkit*.
- Sulak , S. A., & Koklu, N. (2024). Predicting Student Dropout Using Machine Learning Algorithms. *Intelligent Methods In Engineering Sciences*, 3(3), 91-98.
- Sulong, N., & Othman, N. (2020). A Phenomenological Study on International Student Withdrawals from Malaysian Higher Education Institutions. *International Journal of Academic Research in Business and Social Sciences*, 10(12), 363-376.
- The-United-Nations. (2025). *The Fourth Goal of The Sustainable Development Goals: Education*. Retrieved 12 February 2025 from <https://www.un.org/sustainabledevelopment/education/>
- Thies, T., & Falk, S. (2024). Which Factors Drive Major Change and University Dropout? An Analysis on International Degree-Seeking Students at German Universities. *Journal of International Students*, 14(1), 326-346.
- Tin, T. T., Sin, L. E., Lum, K. J., Huen, T. H., & Salau, A. O. (2024). Prediction of Student Academic Status in Higher Education Through Machine Learning.
- Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of educational research*, 45(1), 89-125.
- Tinto, V. (2012). *Leaving college: Rethinking the causes and cures of student attrition*. University of Chicago press.
- Tinto, V. (2017). Through the eyes of students. *Journal of college student retention: Research, theory & practice*, 19(3), 254-269.
- Tirpancioğlu, M. (2019). *According To Some Variables the Examination of School Dropout Risk, School Burnout and Perceived Social Support of Secondary School (Konak Sample) Mugla Sitki Koçman University*].
- Titrek, O., Hashimi, S. H., Ali, S., & Nguluma, H. F. (2016). Challenges faced by international students in Turkey. *The Anthropologist*, 24(1), 148-156.
- Tosun, S. (2024). *Predicting The Factors Affecting School Dropout in Open and Distance Education Students by Using Data Mining Gazi University*].
- Tyack, D. (1976). Ways of seeing: An essay on the history of compulsory schooling. *Harvard Educational Review*, 46(3), 355-389.
- UIS, U. I. f. S. (2019, September 2019). *New Methodology Shows that 258 Million Children, Adolescents and Youth Are Out of School*. UNESCO Institute for Statistics. Retrieved July 3 from <https://uis.unesco.org/sites/default/files/documents/new-methodology-shows-258-million-children-adolescents-and-youth-are-out-school.pdf>
- UNESCO. (2017). *Reducing global poverty through universal primary and secondary education*. Retrieved 16 February 2025 from <http://uis.unesco.org/sites/default/files/documents/reducingglobal-poverty-through-universal-primary-secondary-education.pdf>
- UNESCO. (2020). *Global education monitoring report 2020: Inclusion and education – All means all*. UNESCO. Retrieved July 10, 2025 from <https://unesdoc.unesco.org/ark:/48223/pf0000373718>
- UNESCO. (2022). *UNESCO's efforts to achieve gender equality in and through education*. UNESCO. <https://unesdoc.unesco.org/ark:/48223/pf0000386879>
- Vaarma, M., & Li, H. (2024). Predicting student dropouts with machine learning: An empirical study in Finnish higher education. *Technology in Society*, 76, 102474.

- Wicker, E. (2021). *Quadratic Discriminant Analysis: An introduction, the bias-variance trade-off, and a comparison to linear discriminant analysis using scikit-learn*. Retrieved April 24, 2025 from <https://ethanwicker.com/2021-02-10-quadratic-discriminant-analysis-001/>
- Yadav, S. K., Bharadwaj, B., & Pal, S. (2012). Mining Education data to predict student's retention: a comparative study. *arXiv preprint arXiv:1203.2987*.
- YÖK. (2022). *Yükseköğretimin uluslararasılaşması ve Türkiye'deki üniversitelerin uluslararası vizyonu çalıştayı*. Retrieved 05 March 2024 from <https://www.yok.gov.tr/Sayfalar/Haberler/2022/yuksekogretimde-uluslararasilasma-ve-turkiye-deki-universitelerin-uluslararasi-gorunurlulugunu-artirma-calistayi.aspx>
- YÖK. (2023). *Yükseköğretim Bilgi Yönetim Sistemi - Uyuşma Göre Öğrenci Sayıları Raporu*. Yükseköğretim. <https://istatistik.yok.gov.tr>
- Yukselturk, E., Ozekes, S., & Turel, Y. K. (2014). Predicting dropout student: An application of data mining methods in an online education program. *European Journal of Open, Distance and e-learning*, 17(1), 118-133.
- Zhang, H. (2004). The optimality of naive Bayes. *Aa*, 1(2), 3.
- Zhou, Z.-H. (2025). *Ensemble methods: foundations and algorithms*. CRC press.