IDENTIFYING AT-RISK STUDENTS IN HIGHER EDUCATION: A DATA-DRIVEN APPROACH USING MACHINE LEARNING

ASLIHAN BAL

Thesis for Master's Program in Industrial Engineering

Graduate School
Izmir University of Economics
İzmir
2025

IDENTIFYING AT-RISK STUDENTS IN HIGHER EDUCATION: A DATA-DRIVEN APPROACH USING MACHINE LEARNING

ASLIHAN BAL

THESIS ADVISOR: PROF. DR AHMET SERMET ANAGÜN

Master's Exam Jury Members
Prof. Dr. Ahmet Sermet ANAGÜN
Assoc. Prof. Dr. Fehmi Burçin ÖZSOYDAN
Asst. Prof. Dr. Oktay KARABAĞ

A Master's Thesis
Submitted to
the Graduate School of Izmir University of Economics
the Department of Industrial Engineering

Izmir
2025

Approval of the Graduate School

Prof. Dr. Mehmet Efe BİRESSELİOĞLU

I certify that this thesis satisfies all the requirements as a thesis for a Master's degree.

Prof. Dr Ahmet Sermet ANAGÜN

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for a Master's degree.

Prof. Dr Ahmet Sermet ANAGÜN

Master's Exam Jury Members

Prof. Dr. Ahmet Sermet ANAGÜN
İzmir University of Economics

Assoc. Prof. Dr. Fehmi Burçin ÖZSOYDAN
Dokuz Eylül University

Asst. Prof. Dr. Oktay KARABAĞ
İzmir University of Economics

ETHICAL DECLARATION

I hereby declare that I am the sole author of this thesis and that I have conducted my work in accordance with academic rules and ethical behavior at every stage from the planning of the thesis to its defense. I confirm that I have cited all ideas, information and findings that are not specific to my study, as required by the code of ethical behavior, and that not all statements cited are my own.

Name, Surname: Aslıhan Bal

Date: 24.01.2025

Signature:

ABSTRACT

IDENTIFYING AT-RISK STUDENTS IN HIGHER EDUCATION: A DATA-DRIVEN APPROACH USING MACHINE LEARNING

Bal, Aslıhan

Master's Program in Industrial Engineering

Advisor: Prof. Dr. Ahmet Sermet Anagün

January, 2025

This study focuses on identifying students at risk of attrition in various faculties of higher education institutions. Utilizing data spanning from 2017 to 2023, various demographic, academic, and enrollment-related variables were analyzed using supervised and unsupervised methods. The research examines factors such as academic performance, certain personal information, and engagement metrics to uncover the primary determinants of student attrition across different academic domains. The findings aim to provide insights into historical retention trends and enhance the ability to identify students at risk of attrition before each academic term. The primary objective of this research is to enable the early identification of at-risk students, thereby improving universities' capacity to detect this group in advance. Accordingly, the results aim to contribute to a better understanding of the dynamics behind student attrition.

Keywords: Student Attrition, Higher Education, Machine Learning, Academic Performance

ÖZET

YÜKSEKÖĞRETIMDE RISK ALTINDAKI ÖĞRENCILERIN BELIRLENMESI: MAKINE ÖĞRENMESI KULLANILARAK VERI ODAKLI BIR YAKLAŞIM

Bal, Aslıhan

Endüstri Mühendisliği Yüksek Lisans Programı

Danışman: Prof. Dr. Ahmet Sermet Anagün

Ocak, 2025

Bu çalışma, yükseköğretim kurumlarındaki çeşitli fakültelerde kayıp riski taşıyan öğrencilerin tespit edilmesine odaklanmaktadır. 2017 ile 2023 yıllarını kapsayan verilerden yararlanılarak, demografik, akademik ve kayıtla ilgili çeşitli değişkenler denetimli ve denetimsiz yöntemlerle analiz edilmiştir. Araştırmada, akademik başarı, bazı kişisel bilgiler ve katılım ölçütleri gibi faktörler ele alınarak, farklı akademik alanlarda öğrenci kaybının temel belirleyicileri ortaya konulmaya çalışılmıştır. Araştırmanın bulguları, geçmişe yönelik devamlılık eğilimleri hakkında içgörüler sunmanın yanı sıra, her dönem öncesinde kayıp riski taşıyan öğrencilerin tespit edilme yeteneğini artırmayı amaçlamaktadır. Araştırmanın temel amacı, kayıp riski taşıyan öğrencilerin erken tespitini sağlayarak, üniversitelerin bu grubu önceden belirleme yeteneğini geliştirmektir. Bu doğrultuda elde edilen bulgular, öğrenci kaybının ardındaki dinamiklerin daha iyi anlaşılmasına katkıda bulunmayı hedeflemektedir.

Anahtar Kelimeler: Öğrenci Kaybı, Yükseköğretim, Makine Öğrenmesi, Akademik Başarı

To My Family

ACKNOWLEDGMENTS

First of all, I would like to thank my advisor Ahmet Sermet Anagün, who has always helped me throughout my master's degree, guided my thesis with his ideas and whose support I have always felt, and is an expert in his field. I consulted him on every academic issue throughout my education and he kindly guided me in every way. I am very grateful to him for the encouragement he has shown me on this journey.

I would also like to thank Assoc. Prof. Dr. Fehmi Burçin Özsoydan and Dr. Oktay Karabağ, who devoted their valuable time to the jury presentation and agreed to be on the thesis jury.

Finally, I am grateful to my mother Hülya Bal, my father Mehmet Bal, my brother Bora Bal and my family, who have always been by my side throughout my educational journey. I would like to thank my friend Arda Atilay for his endless support and love.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

ML: Machine Learning

TP: True Positive

FP: False Positive

TN: True Negative

FN: False Negative

GPA: Grade Point Average

AI: Artificial Intelligence

SL: Supervised Learning

MLE: Maximum Likelihood Estimation

RF: Random Forest

LR: Logistic Regression

SVM: Support Vector Machines

ANNs: Artificial Neural Networks

K-NN: K-Nearest Neighbors Algorithm

RL: Reinforcement Learning

CHAPTER 1: INTRODUCTION

University education plays a critical role in the future careers and personal development of students. However, many students leave university before completing their studies for various reasons (Lai et al., 2024). Student retention is vital for the success of both individuals and institutions. Understanding the causes of student attrition and exploring measures to prevent it are crucial for enhancing the effectiveness of educational systems and supporting student success. Student attrition in higher education is a significant concern for educational institutions globally. The premature abandonment of academic programs not only impacts individual students' academic and career trajectories but also diminishes the overall success rates of universities (Bagabir et al. 2021). Hence, it is crucial to devise effective strategies for preventing student attrition and increasing graduation rates. Through early identification of at-risk students and personalized interventions, ML/AI algorithms aim to optimize investment in education and provide a more robust and equitable educational experience (Tapia et al, 2023). Analyzing past data and forecasting student attrition for future periods can assist educational institutions in shaping intervention strategies. University education can profoundly impact students by enhancing their skills and employment prospects, but not all students complete their studies, making retention a critical issue for both students and universities globally. High retention rates bolster universities' reputations and financial stability, whereas high attrition rates can lead to financial strain for students and institutions alike (Kelly et al., 2021).

In thesis, it is analyzed historical data collected between 2017 and 2023, including various demographic, academic, and enrollment-related variables, allowing the relationship between these variables and the probability of student attrition to be quantified. These data were obtained from a foundation university. Ethics committee approval was obtained for the anonymous use of the data. For example, variables such as gender, age, academic performance measures, enrollment patterns, and socioeconomic background can serve as independent variables, while student attrition acts as the dependent variable. By determining the strength and direction of the relationships between these variables, valuable insights can be gained into the factors affecting student attrition. The projections aim to extrapolate 2023 data from 2022

data. It analyzes historical data collected between 2017 and 2023, including various demographic, academic, and enrollment-related variables, allowing the relationship between these variables and the probability of student attrition to be quantified. For example, variables such as gender, age, academic performance measures, enrollment patterns, and socioeconomic background can serve as independent variables, while student attrition acts as the dependent variable. By determining the strength and direction of the relationships between these variables, valuable insights can be gained into the factors affecting student attrition. The projections aim to extrapolate 2023 data from 2022 data.

Furthermore, it facilitates the development of predictive models to predict student attrition for future periods based on the identified determinants (Shafiq et al., 2022). These predictive models can help educational institutions reduce the student attrition by proactively identifying at-risk students and implementing interventions to reduce attrition rates. By utilizing various methods, educational stakeholders can make informed decisions and effectively allocate resources to support student success and enrollment rates (Bagabir et al., 2021)

1.1. Problem Statement

The reasons behind students dropping out of academic programs are complex and multifaceted. Various factors such as financial difficulties, adjustment issues, academic challenges, personal reasons, and systemic factors can influence students' decisions (Ram et al. 2015). This study conducts a comprehensive analysis to understand the underlying factors contributing to student attrition and to determine the interactions among these factors. Additionally, the institutional consequences and societal impacts of student attrition are also examined.

1.2. Purpose of Study

In this study, the focus is on three main purposes related to student retention. First, the aim is to identify the factors causing student attrition. Second, it seeks to predict at-risk students at an early stage. Last is the study compares the performance of various methods. The primary objective of this study is to identify the factors influencing student attrition and to forecast student attrition for future periods. In this regard, various demographic, academic, and socio-economic variables, along with institution-

specific factors, are considered (Bagabir et al., 2015). The findings will provide a fundamental guide for educational institutions to develop intervention strategies and improve student success. By identifying at-risk students and understanding the factors influencing their decisions to leave programs, institutions can tailor interventions to address specific needs and reduce attrition rates (Bukow et al.,2023). In addition, it provides to see early dropouts so lecturers can observe the students from beginning of semester.

1.3. Structure of Thesis

This thesis comprises five main sections. The first section provides a general introduction, highlighting the purpose and significance of the research. The second section conducts a literature review to delve deeply into the phenomenon of student attrition. The third section elucidates the methodology employed and the data analysis techniques utilized. The fourth section is implementation of methods. Lastly, the fifth and final section discusses the findings obtained, conclusion, evaluates the results, and provides recommendations for future research. This structure aims to emphasize the academic and practical contributions of the study and to provide readers with a comprehensive framework.

CHAPTER 2: LITERATURE REVIEW

Literature review about student dropout from university has been examined in terms of the types of methods used, and the information and data used.

Shafiq et al. (2022) first conducted an exhaustive search across diverse databases such as academic journals and conference proceedings to find relevant articles that predicted student retention. The study explores by Shafiq et al. (2022) deeply investigates the systematic literature review based on student retention leveraging educational data mining and predictive analytics, through a well-structured methodological framework the search process starts with the application of pre-defined inclusion criteria to identify the relevant articles that address the goals of the study. Once the relevant articles are identified, a systematic approach is employed to extract key information regarding the predictive models, algorithms, and techniques utilized in predicting student retention, including decision trees, neural networks, logistic regression, or ensemble methods. The study also scrutinizes the factors considered in these predictive models, such as demographic information, academic performance data, and student engagement metrics, to provide insights into the diverse range of variables influencing student retention. Furthermore, the research incorporates a qualitative synthesis of findings to identify common themes, patterns, and gaps in the literature. Through this comprehensive analysis, the study aims to provide a holistic understanding of the current state of research in student retention prediction using educational data mining and predictive analytics techniques, contributing valuable insights that can inform future research directions and the development of effective strategies for improving student retention in academic institutions.

The article by Rezwanul Parvez et al. (2023) explores the challenges posed by the COVID-19 pandemic on student success and retention in higher education institutions. With the pandemic disrupting traditional learning environments and causing widespread uncertainty, predicting factors influencing student retention has become paramount. Through the application of data mining techniques, the study aims to empirically test various factors contributing to student retention behavior using a dataset comprising over 18,000 students enrolled at an academic institution, both pre-

pandemic and post-pandemic. Six different machine learning models, including Logistic Regression, Support Vector Machine, Decision Tree, Random Forest, and K-Nearest Neighbors, are deployed to predict student retention, with the Synthetic Minority Oversampling Technique (SMOTE) utilized to address class imbalance issues. The empirical findings reveal crucial factors such as completed credits, grade point average (GPA), college entrance age, and attempted credits significantly impacting student retention behavior. Notably, Random Forest algorithms demonstrate superior performance, achieving the highest accuracy of 0.86 with SMOTE for retention prediction. The study underscores the importance of proactive measures by academic institutions to prevent dropouts and enhance retention based on the identified metrics. Additionally, the research contributes to the literature by presenting a comprehensive assessment of student retention matrices and proposes interventions to mitigate dropout rates and improve retention, thereby supporting student success amidst the challenges posed by the COVID-19 pandemic. Future research directions include exploring the impact of diverse datasets and implementing targeted support strategies for specific student demographics to further enhance retention rates.

Predictive analysis of higher education graduation and retention in Saudi Arabia uses multinomial logistic regression in the article of Bagabir et al. (2021). The methodology involves analyzing various factors influencing graduation and retention rates. These factors may include demographic variables such as age, gender, and nationality, academic performance indicators like GPA, enrollment status, and program of study, as well as socio-economic factors such as family income and educational background. By incorporating these variables into the multinomial logistic regression model, Bagabir et al. (2021) aim to identify significant predictors of graduation and retention outcomes in Saudi Arabian higher education institutions. This analytical approach allows for a comprehensive examination of the complexities involved in student success and provides insights that can potentially inform policy and practice for improving educational outcomes in the region.

Qvortrup et al. (2022) examines the study environment factors associated with student retention in higher education through a comprehensive analysis of various elements that affect student retention in higher education.These factors include the availability and effectiveness of academic support services, the degree of social integration within

the institution, institutional characteristics such as size and resources, and demographic variables like age, gender, and socio-economic background. Utilizing surveys and advanced statistical techniques, Qvortrup et al. (2022) aim to identify significant predictors of retention and explore the intricate interplay among these factors in shaping students' decisions to continue their academic pursuits. This rigorous investigation provides valuable insights into the multifaceted nature of student retention and offers actionable recommendations for improving retention strategies within higher education settings.

In their article, Herodotou et al. (2020) found that the use of predictive learning analytics and motivational interventions played a significant role in improving student retention rates and enhancing administrative support in distance education. The study's methodology involves an in-depth analysis of various factors, including student demographics, engagement metrics, and academic performance data, meticulously designed to construct predictive models. These models are instrumental in identifying students at risk of attrition, enabling the implementation of targeted motivational interventions. By adopting a mixed-methods approach, Herodotou et al. (2020) combine quantitative analysis of student data with qualitative feedback from both students and educators, offering a comprehensive evaluation of intervention effectiveness. This approach aims to explore the untapped potential of predictive analytics and motivational strategies, ultimately contributing to meaningful advancements in student outcomes within the evolving landscape of distance education.

The study conducted by Aypay et al. (2022) investigates student retention in higher education within Turkey using qualitative methodologies. Their approach involves conducting in-depth interviews with students, faculty members, and administrators to explore the diverse factors impacting student retention, including academic support services, financial considerations, social integration, and institutional policies. Through rigorous thematic analysis of the interview data, the authors aim to uncover recurring patterns and themes pertaining to student retention. By employing a qualitative framework, the research seeks to provide a comprehensive understanding of the complexities surrounding student retention within the Turkish higher education

landscape, with the overarching goal of informing targeted strategies to enhance retention rates.

Student retention prediction at St. Cloud State University is explored through a comprehensive comparative study of predictive models. The methodological approach involves a meticulous examination of various factors known to influence student retention, including demographic information, academic performance metrics, and indicators of student engagement. These factors form the basis for the development and comparison of multiple predictive models aimed at forecasting retention rates within the university. Utilizing advanced statistical analysis techniques, Dissanayake et al. (2016) rigorously evaluate the predictive capabilities of each model, identifying the most accurate and effective approach. Additionally, the study examines model performance in detail, considering metrics such as recall, specificity, and overall predictive accuracy. This robust methodology not only enhances understanding of the complex dynamics underlying student retention but also provides actionable insights to inform strategies for improving retention rates and fostering student success at St. Cloud State University.

Predictive analysis of one-year retention at the University of Houston is conducted through a comprehensive methodological approach. Initially, a diverse range of data is gathered, including demographic information, academic performance metrics, and indicators of student engagement. These factors form the foundation for developing predictive models designed to forecast retention rates over a one-year period. Advanced statistical techniques, such as logistic regression and machine learning algorithms, are utilized to construct these models. Martinez et al. (2020) rigorously evaluate the models' effectiveness by testing them on separate datasets and analyzing performance metrics like recall, specificity, and overall predictive accuracy. Additionally, a comparative analysis of various modeling approaches is performed to identify the most effective method for retention prediction. Supplementing the quantitative analysis, qualitative insights are obtained through interviews or surveys with students, faculty, and administrators, providing context and uncovering additional factors influencing retention. This integration of qualitative and quantitative data enhances the models' comprehensiveness and accuracy.

Shafiq et al. (2022) systematically examined student retention using educational data mining and predictive analytics. The methodology includes a comprehensive search and selection process of relevant articles from various databases, adhering to predefined inclusion criteria. Once selected, the articles undergo thorough analysis to extract information on the predictive models, algorithms, and techniques used to predict student retention. These methods include a variety of approaches, such as decision trees, neural networks, logistic regression, and ensemble methods. Additionally, demographic information, academic performance data, and student engagement metrics are considered in the predictive models. Shafiq et al. (2022) employ a qualitative synthesis to identify common themes and patterns in the literature, providing insights into the diverse variables influencing student retention. Through this comprehensive approach, the study aims to offer a nuanced understanding of student retention prediction using educational data mining and predictive analytics techniques.

Matz et al. (2023) stated that in recent years, the issue of student retention in higher education has received increasing attention due to its significant implications for academic institutions, funding agencies, and students themselves. Matz et al. (2023) highlight the importance of utilizing machine learning techniques to predict student dropout, integrating socio-demographic characteristics and engagement metrics to enhance predictive accuracy. While previous studies have predominantly focused on macro-level data such as demographics and academic performance, Matz et al. emphasize the value of incorporating meso-level engagement data, including student interactions within the university community. By partnering with a mobile application facilitating student-university communication, the researchers collected comprehensive data on institutional variables and behavioral engagement, demonstrating the efficacy of this combined approach in predicting dropout after the first semester. Moreover, their findings underscored the incremental predictive power of behavioral engagement metrics, shedding light on the nuanced factors influencing student retention. Through machine learning models trained on diverse datasets from multiple universities, the study not only provided insights into predictive performance but also demonstrated the generalizability of the models across different institutional contexts. This research underscores the importance of leveraging advanced analytical techniques to address the complex challenge of student attrition, offering valuable

insights for educational institutions and policymakers striving to enhance student success and retention in higher education.

Ram et al. (2015) address the challenge of student retention through a pioneering big data approach, departing from traditional survey-based methods and conventional machine learning techniques. Their research aims to develop a predictive model for freshman retention by leveraging big data analytics, particularly focusing on the augmentation of standard institutional student datasets (ISDs) with additional behavioral insights derived from university smart card transactions. The authors argue that while survey-based approaches suffer from limitations such as time-consuming data collection and self-reporting biases, big data analytics offer a more robust and scalable solution for predicting student attrition. By integrating student demographic and academic data with implicit social networks inferred from smart card transactions, the researchers aim to capture the complex dynamics of student behavior and campus integration. They employ sequence learning algorithms to analyze students' regularity in campus activities, thereby enhancing the predictive accuracy of the model. Moreover, to address the challenge of imbalanced data classes inherent in student retention datasets, the authors propose a novel ensemble classifier that combines sampling methods with ensemble learning techniques. Through experimental evaluation on real world smart card transaction data, their study demonstrates the efficacy of the big data approach in improving prediction accuracy and recall, highlighting the potential of information systems design science research in addressing longstanding challenges in student retention.

As a result of the literature, it was found that machine learning would be a good tool in determining student loss and which data (age, gender, demographic information, academic information) is used more. 5 methods would be suitable for this thesis.

CHAPTER 3: METHODOLOGY

Machine Learning (ML) is a subset of artificial intelligence (AI), constructs analytical models by analyzing existing data. Originating from Alan Turing's mid-20th-century concept, ML has evolved into a powerful tool for various fields. Supervised, unsupervised, and reinforcement learning are its main types, each serving distinct purposes (Çelik et al.,2018). Figure 1 illustrates the different methods within Machine Learning, including supervised and unsupervised learning, along with their respective subcategories.

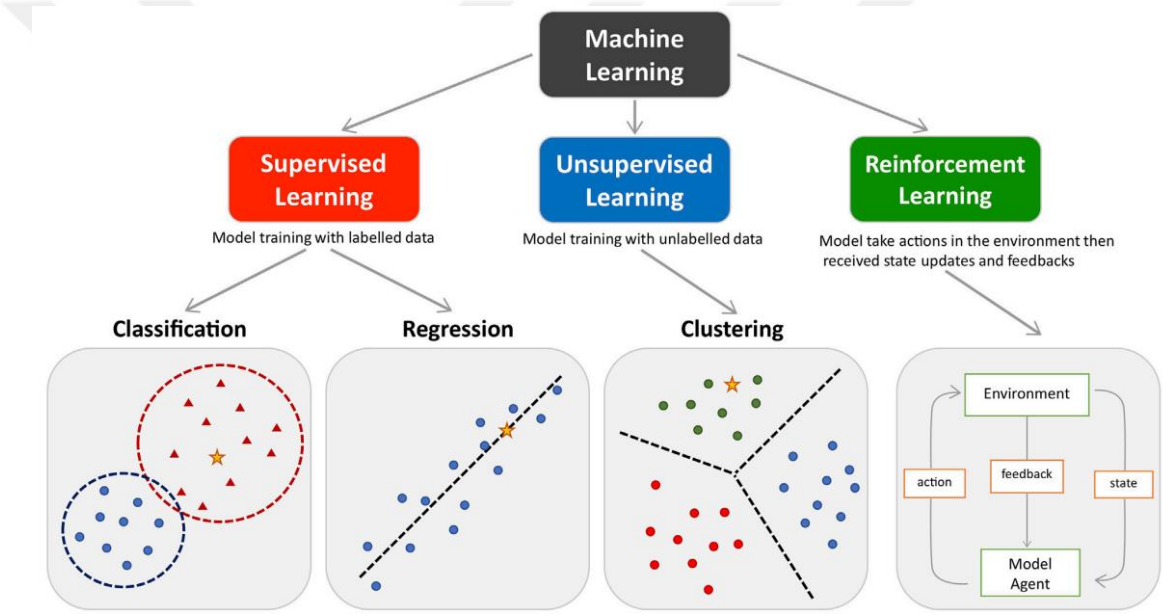

Figure 1. Machine Learning Types and Applications (Source: Peng et al., 2021)

Supervised learning (SL) serves as a fundamental framework within machine learning, where input objects, often represented as vectors of predictor variables, are paired with corresponding desired output values, forming what is known as a human-labeled supervisory signal, to train a model (Jiang et al.,2020). Through this process, the training data is analyzed and processed, ultimately constructing a function that effectively maps new input data to expected output values. The crux of supervised learning lies in its ability to generalize from the provided training data to unseen instances in a manner deemed "reasonable," as dictated by the algorithm's inductive

bias. The success of an algorithm is gauged by its ability to minimize the generalization error, reflecting its capacity to accurately predict outputs for unseen data. As the model iteratively processes input data, it refines its internal variable, adjusting weights until an optimal fit is achieved, typically as part of a cross-validation procedure. Supervised learning is pivotal in enabling organizations to address a myriad of real world challenges at scale, ranging from segregating spam emails to constructing highly accurate predictive models tailored to specific applications and domains, thereby facilitating data-driven decision-making and fostering innovation.

In this study, supervised learning methods were preferred to predict student retention. The reason for choosing the supervised method is that there is input and output data. Five-year student data were analyzed in the study and it was determined which students with which characteristics dropped out of school at a higher rate in previous years. This historical data discovered the effect of certain variables on student retention. Supervised learning was preferred because it offers the opportunity to make predictions for the future for new students using past information.

Unsupervised learning, a branch of artificial intelligence, works without human supervision and differs from supervised learning, which relies on labeled data. Instead of using predefined labels, unsupervised learning algorithms analyze unlabeled datasets independently to uncover patterns and insights. One widely used technique in unsupervised learning is clustering, which groups similar data points based on shared characteristics. Clustering algorithms automatically identify patterns or similarities in a dataset and organize the data into clusters. These clusters help reveal hidden structures or relationships in the data, which may not be immediately obvious. For example, a clustering algorithm can analyze customer data and divide it into segments based on purchasing behavior, allowing businesses to tailor marketing strategies for each group. Similarly, it can detect differences in datasets by identifying data points that don’t fit well into any cluster, making it useful for fraud detection or error analysis.

Reinforcement learning (RL) constitutes a cross-disciplinary domain within machine learning and optimal control, focusing on guiding an intelligent agent’s actions within a changing environment to maximize cumulative rewards. It stands as one of the fundamental paradigms in machine learning, alongside supervised and unsupervised

learning. Unlike supervised learning, RL works without requiring labeled input and output pairs. Its essence is to strike a balance between exploring uncharted territory and using existing knowledge to achieve long-term reward maximization, even if feedback is missing or delayed.

3.1. Machine Learning Methods Used

3.1.1. Logistic Regression

Logistic regression is selected as the primary analytical method for this study due to its suitability for modeling binary outcomes, which aligns closely with the nature of the dependent variable in the investigation of student attrition (Çelik et al.,2018). In the context of higher education research, where the focus is often on predicting whether students will persist or withdraw from academic programs, logistic regression offers a robust framework for analyzing the influence of various independent variables on this binary outcome.

Furthermore, logistic regression allows for the examination of the relationship between multiple predictors and the probability of student attrition, enabling the identification of significant factors contributing to students' decisions to leave their academic programs prematurely. The utilization of logistic regression in this research endeavor is thus well-founded, as it enables the elucidation of detailed insights into the complex phenomenon of student attrition, ultimately empowering educational institutions to adopt proactive measures aimed at identifying at-risk students and implementing targeted interventions to mitigate retention challenges effectively.

The core idea behind logistic regression is to model the probability of the binary outcome given a set of independent variables (Sperandei, 2014). Unlike linear regression, which predicts continuous outcomes, logistic regression models the probability of the outcome using a logistic function (also known as the sigmoid function), which ensures that the predicted probabilities lie between 0 and 1(Larner et al., 2002). The logistic function is defined as in Equation 1;

$$P(Y=1|X) = \frac{1}{1+e^{-z}} \quad (1)$$

Where:

(Y=1|X) is the probability of the outcome being 1 given the input features X.
e is the base of the natural logarithm.
Z is the linear combination of the input features and their corresponding coefficients.
The linear combination Z can be represented as in Equation 2:

$$Z=\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2\ x_2+...+\hat{\beta}_n x_n \quad (2)$$

Where:

$\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_n$ are the coefficients of the independent variables.
$x_1, x_2, \dots, x_n$ are independent variables.

The coefficients $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_n$ are estimated using maximum likelihood estimation (MLE) or optimization techniques like gradient descent. The goal is to find the coefficients that maximize the likelihood of observing the given data under the logistic regression model.

Once the coefficients are estimated, predictions are made by plugging the values of the independent variables into the logistic function. If the predicted probability is greater than a threshold (usually 0.5), the outcome is predicted as 1; otherwise, it's predicted as 0.

Logistic regression also allows for the assessment of the significance of independent variables and the interpretation of their effects on the probability of the outcome. This makes it a powerful tool for understanding the relationship between predictors and the binary outcome. (Larner et al., 2002).

The odds ratio (OR) is a statistical tool used in logistic regression to measure how strongly a predictor variable is associated with an outcome and whether the relationship is positive or negative. n logistic regression, Receiver Operating Characteristic (ROC) curves are utilized to determine the optimal cutoff value for

classifying new observations as either a "failure" (0) or a "success" (1).

3.1.2. Random Forest

Random Forest is an ensemble learning method utilized in machine learning for classification and prediction tasks (Peng et al. 2021). It operates by constructing multiple decision trees through a process known as bagging, where each tree is trained on a subset of the data set as seen in Figure 2. Additionally, Random Forest employs random feature selection during the construction of each decision tree, aiming to reduce correlation among the trees and enhance overall model robustness. The final prediction is determined through either a voting mechanism for classification tasks or an averaging approach for regression tasks, wherein the predictions of all trees are aggregated. Random Forest offers several advantages, including reduced overfitting, improved robustness against noise, and insights into feature importance. However, challenges such as computational complexity and interpretability remain pertinent considerations in its application. Despite these challenges, Random Forest remains widely used across various domains for its effectiveness in handling complex datasets and delivering reliable predictions.

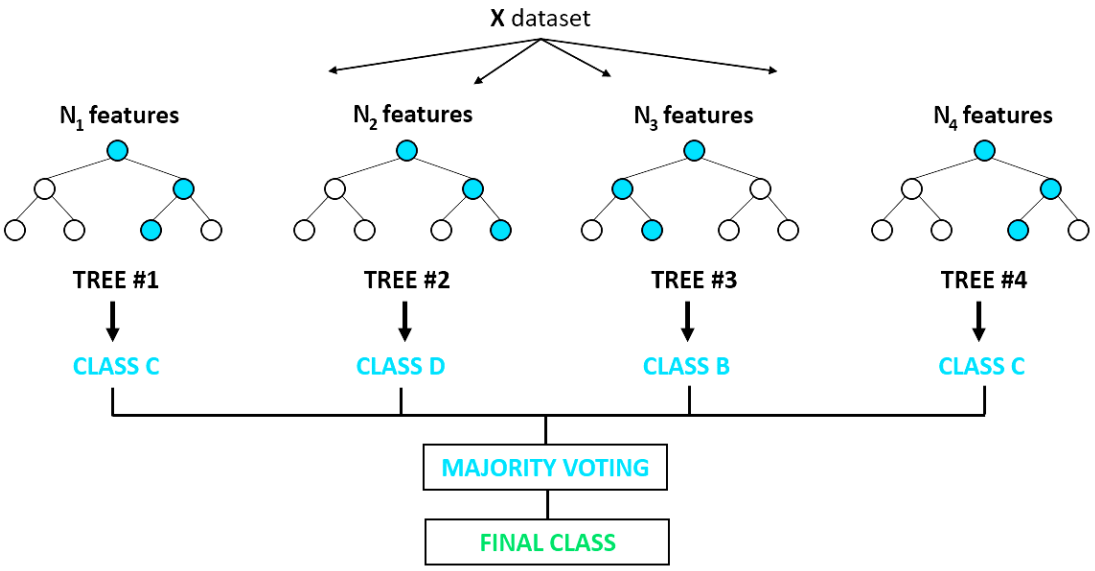

Figure *2.* Random Forest Algorithm Workflow (Source: Mbuvha et al. 2021)

3.1.3. Support Vector Machine

Support vector method, commonly known as Support Vector Machines (SVM), is a supervised learning algorithm used for classification and regression tasks (Peng et al. 2021). It operates by finding the optimal hyperplane that best separates different classes in the feature space. The key idea behind SVM is to maximize the margin between the closest data points from different classes, known as support vectors. SVM can handle both linearly separable and non-linearly separable datasets through the use of different kernel functions, such as linear, polynomial, and radial basis function (RBF) kernels (Patel et al. 2024). This method is particularly effective in high-dimensional spaces and when the number of features exceeds the number of samples. SVM offers several advantages, including robustness against overfitting, versatility in handling different types of data, and effectiveness in binary classification tasks. However, its performance may be sensitive to the choice of kernel and regularization variable, and it may not scale well to large datasets. Despite these considerations, SVM remains a widely used and well-studied algorithm in the field of machine learning, with applications spanning various domains such as image classification, text categorization, and bioinformatics.

The length of a vector x is called its norm, and is denoted as $||x||$. The Euclidean norm of a vector $x=(x_1, x_2,\ldots, x_n)$x=(x_1, x_2,…, x_n) is calculated using the Equation 3:

$$||x||= \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2} \tag{3}$$

The direction of a vector in Equation 4:

$$x=(x_1, x_2,\ldots, x_n) \tag{4}$$

is as w and is defined as the vector of unit length pointing in the same direction as x in Equation 5. The direction vector w is obtained by normalizing Equation 5.

$$w=\left(\frac{x_1}{||x||}, \frac{x_2}{||x||}, \ldots, \frac{x_n}{||x||}\right) \tag{5}$$

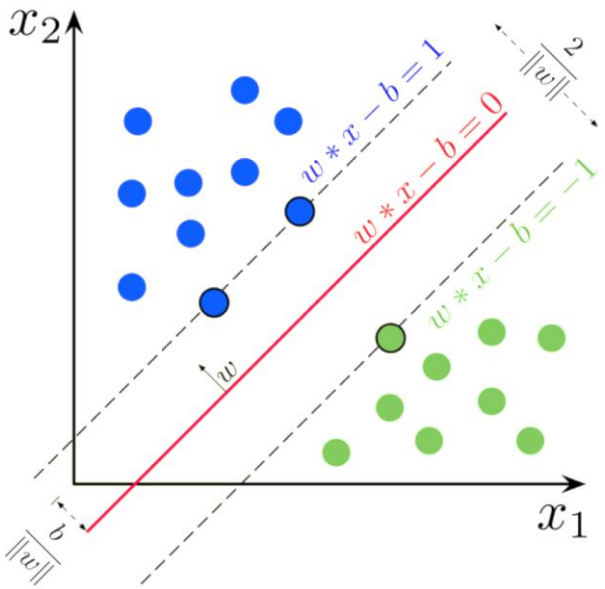

Figure 3. Support Vector Machine Decision Boundary and Margins (Source: Support Vector Machines, Medium, 2020)

In Figure 3, the graph illustrates the decision boundary and margin hyperplanes of a Support Vector Machine (SVM) classifier. The solid line represents the decision boundary in Equation 6:

$$w \cdot x - b = 0 \tag{6}$$

while the dashed lines indicate the margin hyperplanes in Equation 7

$$w \cdot x - b = 1 \tag{7}$$

and Equation 8

$$w \cdot x - b = -1 \tag{8}$$

There are four different kernel functions: Linear, Polynomial, Radial Basis Function (RBF) and Sigmoid. Linear kernel is a simple method that separates data with a straight line. Polynomial kernel is used to separate data with a more complex curve. Radial Basis Function (RBF) kernel transfers data to a higher dimensional space and analyzes complex relationships better. Sigmoid kernel is a method that separates data with an

S-shaped curve. Each offers different advantages depending on the structure of the data.

3.1.4. Artificial Neural Networks (ANNs)

Artificial Neural Networks (ANNs) represent a sophisticated computational model inspired by the intricate structure of the human brain's neural network. Comprising interconnected nodes known as neurons organized into layers, ANNs facilitate the flow of information through these nodes (Peng et al. 2021). The network refines its performance through training, adjusting the connection strengths, or weights, to recognize patterns, make predictions, and address various machine learning and artificial intelligence tasks. Fundamentally, ANNs embody a framework that mimics the biological neural networks, albeit with distinct computational characteristics tailored for processing numeric and structured data efficiently.

A typical ANN construction consists of three primary layers: the input layer, one or more hidden layers, and the output layer. This layered arrangement, often referred to as the Multi-Layer Perceptron (MLP), enables the network to undertake complex computational tasks by progressively extracting and processing relevant patterns from the input data. Particularly noteworthy is the function of the hidden layer, acting as a "distillation layer" that discerns crucial information while filtering out extraneous data, thereby enhancing the network's efficiency and efficacy in pattern recognition and analysis (Brijith, 2023)

Central to the functionality of ANNs are activation functions, pivotal in introducing non-linearity into the model and facilitating the conversion of input data into meaningful output. By capturing non-linear relationships among inputs, activation functions enable ANNs to discern intricate patterns and correlations, contributing significantly to the network's predictive capabilities. Moreover, activation functions play a critical role in the learning process, allowing the network to adapt and optimize its performance over successive iterations.

The optimization of Artificial Neural Networks (ANNs) refers to the process of adjusting the connection weights to find the best values. This is crucial for minimizing prediction errors and improving the accuracy of the model. Leveraging the

backpropagation algorithm, ANNs refine their performance by iteratively adjusting the connection weights based on observed errors, thereby transforming the network into a dynamic learning algorithm capable of continuous improvement. Through techniques such as gradient descent, the optimization process seeks to identify the most favorable weight values that minimize prediction errors, ensuring the network's robustness and efficacy in handling diverse datasets and complex tasks.

In applications, ANNs offer a myriad of advantages that render them indispensable for addressing real world challenges. Notably, ANNs excel in modeling non-linear and intricate interactions inherent in many real life phenomena enabling them to capture complex relationships between inputs and outputs accurately. Furthermore, ANNs possess the ability to generalize from learned associations, allowing them to infer unknown relationships and predict novel data—an invaluable trait in domains characterized by uncertainty and evolving conditions. Additionally, ANNs' flexibility in handling diverse input variables and their capacity to simulate heteroskedasticity make them particularly well-suited for tasks like financial time series forecasting, where data volatility and complexity pose significant challenges. Thus, ANNs emerge as a versatile and powerful tool for tackling a wide array of computational and predictive tasks across various domains. In Figure 4, the graph depicts the structure of an Artificial Neural Network (ANN) including the input units, hidden units, and output units.

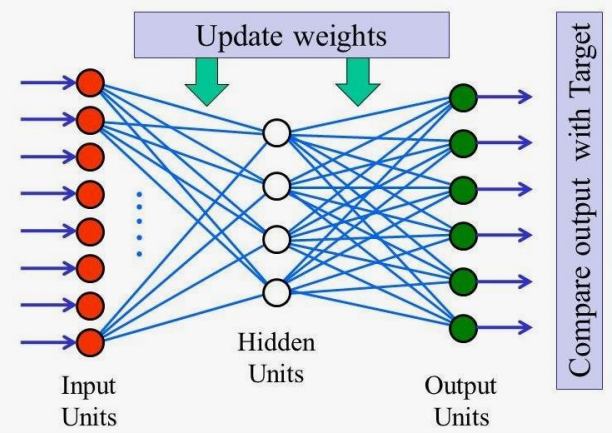

Figure 4. Artificial Neural Network Architecture: Input, Hidden, and Output Units with Weight Updates (Source: Farooq et al., 2022)

3.1.5. K-Nearest Neighbors (K-NN) Algorithm

The methodology of this study includes the k-nearest neighbors (k-NN) algorithm to examine student attrition in higher education institutions. The k-NN method is chosen for its non-parametric nature and simplicity in classification tasks, where it identifies the ‘k’ nearest data points to a query point based on a distance metric, typically Euclidean distance. This approach is suitable for analysis due to the presence of various demographic, academic, and enrollment-related variables in the large dataset spanning the years 2017-2023. Since algorithms such as k-NN can perform in large and diverse datasets, they can be effectively applied on this dataset. By leveraging k-NN, we can effectively handle multivariate data and discern local patterns in student retention across different academic domains without imposing stringent assumptions about data distribution. Hard data preprocessing, including thorough cleaning, meticulous feature selection, and rigorous normalization, ensures that the k-NN model performs optimally in predicting retention outcomes, facilitating a nuanced understanding of the determinants of student attrition in higher education.

3.2. Data Collection and Analysis Steps

Data collection comprises the systematic gathering of information through various methodologies, serving as a pivotal aspect within the realms of statistics and mathematics. Practically, data collection denotes the structured procedure involved in assembling and organizing data, often integral to initiatives aimed at process enhancement. The overarching objective of data collection resides in the growth of pertinent information for archival purposes, facilitating informed decision-making across multifarious domains, and the spreading of critical insights to relevant stakeholders. Figure 5 outlines the sequential steps of data collection, encompassing the determination of variables, organization of data, variable declaration and code assignment, analysis, and evaluation of the analysis.

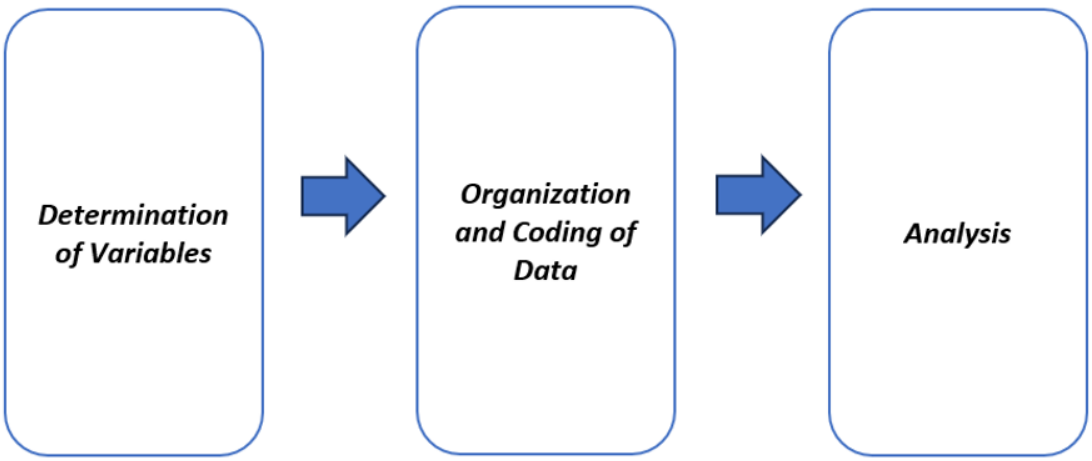

Figure 5. Data Collection Steps

3.2.1. Determination of Variables

The student data used in the study is kept confidential by the university and was requested before the research. These data were kept confidential before, during, and after the research. Since the data was anonymized, faculty or department-based analysis could not be performed.

A literature review was conducted and student observations were used to determine the factors that may affect students' decisions to leave the university. In this context, variables such as the student's age, gender, nationality, hometown, registration type (such as entrance by exam), university preference order, number of credits taken per semester, annual semester grade point average (GPA), number of "DD" and "DC" grades received each semester, high school type (Anatolian High School, Science High School, etc.), scholarship type, whether or not they received a scholarship, whether or not their scholarship was terminated, and whether or not they worked within the university were determined. These variables were selected in line with the findings in the literature and direct student observations, and it was observed that they were important in determining students' tendency to leave the university.

The data of 641 students were analyzed to guarantee data quality. During the data collection progression, it was determined there were many records with missing or incorrect entrances. To ensure the reliability of the analysis process, students with missing or incorrect data were directly excluded from the analysis.

Including more variables in the analysis (e.g., participation in student clubs, employment status outside the university, parents' level of education, etc.) could have provided more comprehensive information about students' tendency to leave the university. However, since the data on these variables were not systematically recorded, they could not be included in the study.

In this study, students who were active between 2017-2023 were evaluated. This stage played a critical role in determining the opportunity of the research and selecting the variables to be analyzed.

Variables that are used in the thesis:

Table 1. Variables, Their Corresponding Numeric Codes, and Variables Types

Symbol:	**Variables:**	**Numeric Codes:**	**Type of Variables**
X_1	Age		Numeric
X_2	Gender	Male:0 Female:1	Categorical
X_3	Region	Mediterranean Region:1 Eastern Anatolia Region:2 Aegean Region:3 Southeastern Anatolia Region:4 Central Anatolia Region:5 Black Sea Region: 6 Marmara Region:7	Categorical
X_4	Order of Preference		Ordinal
X_5	Preparatory Class Information	Preparatory Read:0 Not Read:1	Categorical
X_6	Grade Point Average		Numeric
X_7	Total Credits Received		Numeric
X_8	Total Credits Failed		Numeric
X_9	Number of FF/FD Courses		Numeric
X_{10}	Number of DD/DC Courses		Numeric
X_{11}	Scholarship Status	Receiving Scholarship:0 Not Receiving Scholarship:1	Categorical
X_{12}	Previous Year Enrollment Status	Yes:0 None:1	Categorical

Table 1 presents a comprehensive overview of the variables utilized in the thesis, including numeric codes, types of variables, and their corresponding categories,

offering a structured framework for data analysis and interpretation.

3.2.2. Organization and Coding of Data

The dataset was examined for format faults, conflicts, and missing values. Common problems identified included missing academic records, incomplete demographic information, and conflicts related to the type of record. These errors were identified and corrected by verifying them contrary to institutional archives whenever possible. However, it was not possible to verify missing or incorrect data, these students were removed from the analysis. This step was critical to ensure the reliability of the analysis process. As a result, a total of 641 student records were retained for use in the analysis.

In addition, categorical variables were converted to numerical values to comply with the numerical data format required by logistic regression analysis. For example, for the gender variable, women were coded as 1 and men as 0; and students' hometown information was represented with numerical codes according to specific categories.

Thanks to these stages, the data set was made suitable for analysis by ensuring its consistency and reliability.

3.2.3. Analysis

The obtained results were compared and analyzed with real life data to evaluate the accuracy of the model and determine how consistent the analysis results were with the real world situation. This step ensured the reliability of the analysis and facilitated the interpretation of the results.

This process laid an important foundation for understanding the factors influencing students' decisions to drop out of university and for developing effective strategies based on these factors. The model is constructed using variables such as student demographic information and academic performance.

The model is assembled using the Minitab and WEKA statistical software. It evaluates the power of variables such as age, gender, region of residence, preference order, enrollment in preparatory classes, overall GPA, total credits of courses taken, total credits of courses failed, number of failed courses, and number of conditional passes

on the probability of student dropout.

3.3. Confusion Matrix

A confusion matrix is a tabular representation that organizes the predictions made by a classification model against the actual class labels of the dataset. Typically, it consists of four quadrants delineated by the predicted and actual class labels (Larner et al., 2024).

In the domain of machine learning, the confusion matrix plays a pivotal role in evaluating the performance of classification models (Bhandari, 2024). It provides a structured framework for analyzing the predictions made by the model against the actual class labels present in the dataset.

In the thesis **0** represents students who **stay** at the university, **1** represents students who **leave** the university.

True Positive (TP):

TP denotes instances where the model correctly predicts the positive class. In other words, the classifier identifies the presence of a particular phenomenon accurately.

False Positive (FP):

FP occurs when the model erroneously predicts the positive class in instances where it should have been negative. This type of error is often referred to as a Type I error and signifies a false alarm.

True Negative (TN):

TN indicates instances where the model correctly predicts the absence of the positive class. It signifies the model's ability to discern absence or non-occurrence accurately.

False Negative (FN):

FN represents instances where the model fails to predict the positive class when it should have. This type of error, known as a Type II error, implies a missed opportunity to identify the presence of the phenomenon of interest.

Beyond the enumeration of classification outcomes, the confusion matrix facilitates

the computation of various performance metrics essential for model evaluation. Metrics such as accuracy, precision and recall are derived from the counts present in the confusion matrix, providing nuanced insights into the strengths and weaknesses of the classification model.

Table 2. Confusion Matrix: Evaluation of Classification Algorithm Performance

		Actual Values	
		Positive (0)	**Negative (1)**
Predictive Values	**Positive (0)**	TP	FP
	Negative (1)	FN	TN

In Table 2, the confusion matrix is presented, delineating the counts of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), providing a comprehensive evaluation of the classification algorithm's performance.

Accuracy:

Accuracy is a fundamental metric in classification tasks, quantifying the overall correctness of a model's predictions. It represents the proportion of correctly classified instances, encompassing both true positives and true negatives, relative to the total number of instances in the dataset. Mathematically, accuracy is expressed as the ratio of the number of correct predictions to the total number of predictions made by the model as seen in Equation 9.

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (9)$$

Precision:

Precision is a pivotal metric focusing on the accuracy of positive predictions made by the model. Equation 10 represents the proportion of instances predicted as positive that are genuinely positive, elucidating the model's precision in positive classification. Precision is computed as the ratio of true positives to the sum of true positives and false positives.

$$Precision = \frac{(TP)}{(TP+FP)} \quad (10)$$

Recall:

Recall embodies the underlying distribution of the positive class within the dataset. It quantifies the proportion of positive instances relative to the total number of instances present. Recall serves as a foundational aspect of classification evaluation, providing context for the performance metrics derived from the confusion matrix.

$$Recall = \frac{(TP)}{(TP+FN)} \quad (11)$$

F1 Score:

F1 score is a measure that balances the precision and recall values of a model. F1 score is calculated by taking the harmonic average of these two metrics. If the model focuses only on accuracy but misses some positives, or focuses only on recall but produces many false positives, F1 score will decrease. Therefore, a model that balances accuracy and recall will have a higher F1 score.

The F1 score is calculated as follows in Equation 12:

$$F1 = 2 \times \frac{\text{Precision+Recall}}{\text{Precision}\times\text{Recall}} \quad (12)$$

CHAPTER 4: IMPLEMENTATION OF METHODS

The results obtained from the discussed methods in Chapter 3 are explained under subheadings in Chapter 4. In this study, the training set was used to investigate the data. The training set permits the model to learn the relationships necessary for making predictions. The training data is used to learn the relationship between the input variables and the desired outcome (student dropout rate). This set helps the model use its variables in the appropriate way. Separating the data into training and test sets is an important step with the aim of objectively evaluate the performance of the model. The test set consists of data that the model has not seen during training and this data is used to measure the accuracy of the model.

4.1. Logistic Regression Analysis

The dataset was analyzed using Minitab, with 641 rows utilized for regression analysis. The response variable, denoted as Y, exhibited two distinct values: 1, indicating the occurrence of the event (leaving from university), observed 80 times, and 0, representing non-occurrence (staying in university), observed 561 times, resulting in a total of 641 observations. The regression equation derived from the analysis provides insights into the relationship between the predictor variables and the probability of the event occurring.

Equation 13,

$$P(1) = \frac{1+exp(\mathrm{Y}')}{exp(\mathrm{Y}')} \tag{13}$$

and Equation 14,

$$(\mathrm{Y}')=2.00+0.2049X_1+0.955X_2+0.300X_3+0.0120X_4+3.54X_5-2.302X_6-0.0542X_7+0.2476X_8-0.974X_9-0.670X_{10}-0.001X_{11}-0.93X_{12} \tag{14}$$

delineates the relationship between the predictor variables and the probability of the event. The results of LR are given in the Appendix. The following variables were considered in this research: Age (X_1), Gender (X_2), Region (X_3), Order of Preference

(X_4), Preparatory Class Information (X_5), Grade Point Average (X_6), Total Credits Received (X_7), Total Credits Failed (X_8), Number of FF/FD Courses (X_9), Number of DD/DC Courses (X_{10}), Scholarship Status (X_{11}), and Previous Year Enrollment Status (X_{12}) and Y (prediction of retention or leaving).

Analysis of the coefficients associated with each predictor variable reveals varying magnitudes of influence. For instance, variables such as X_5 exhibit a notable coefficient of 3.54, suggesting a substantial impact on the event's probability. Conversely, variables like X_6 demonstrate a significantly negative coefficient of -2.302, indicating an inverse relationship with the variable of leaving university.

Model summary statistics indicate that the regression model accounts for approximately 76.29% of the deviance observed in the data. Goodness-of-fit tests, including the Deviance and Pearson tests, yielded p-values close to 1.000, suggesting that the model adequately fits the data. Analysis of variance indicates significant results ($p < 0.05$) for several predictor variables, including X_1, X_3, X_5, X_6, X_7, X_8, X_9, and X_{10}, suggesting that these variables significantly impact the event's probability. Fits and diagnostics for unusual observations highlight instances where observed probabilities deviate significantly from fitted values, indicating potential outliers warranting further investigation.

Table 3 compares the predicted values from the logistic regression model with the actual outcomes. Here's the interpretation of the results:

Table 3. Confusion Matrix for Logistic Regression Model

Row Labels	0	1	Grand Total
0	553	13	566
1	8	67	75
Grand Total	**561**	**80**	**641**

True Negatives (0,0): The model correctly predicted 553 out of 561 instances where students stayed in university (TN).

True Positives (1,1): It correctly predicted 67 out of 80 instances where students left the university (TP).

False Negatives (1,0): The model incorrectly predicted that 8 students would stay when they left (FN).

False Positives (0,1): It incorrectly predicted that 13 students would leave when they stayed (FP).

The model has high accuracy (0.9672), effectively distinguishing between students who stay and those who leave the university. Precision is 0.997 and Recall value is 0.9857.

4.2. Random Forest

The analysis was conducted using the WEKA software, employing the Random Forest algorithm with specific variables. The dataset, included 641 instances with 13 variables. The evaluation was performed on the training data itself. The Random Forest model, utilizing bagging with 100 iterations - was utilized. Subsequently, predictions were made on the training set. These predictions, comparing actual values with predicted ones, were analyzed for each instance. The analysis revealed both accurately predicted values and instances where predictions varied from the actual values. The confusion matrix created for the random forest method is shown in Table 4.

Table 4. Confusion Matrix for Random Forest Method

Row Labels	0	1	Grand Total
0	561	0	561
1	31	49	80
Grand Total	**592**	**49**	**641**

The model correctly predicted 49 instances where students left the university (TP), which represents 61.25% of the actual cases of students leaving. It erroneously predicted 31 instances (FP), accounting for 5.24%, and correctly identified 561 instances (TN), representing 87.47%, with no missed predictions (FN), resulting in a 0% rate. The Random Forest model achieved an accuracy of 0.9516, a precision of 1and recall of 0.9476.

4.3. Support Vector Machine

The Support Vector Machine (SVM) was utilized to analyze the dataset using Weka, which comprises 641 instances with 13 variables.

The performance analysis of the SVM algorithm was conducted using different kernel functions (Linear, Polynomial, Radial Basis Function, and Sigmoid) and evaluated across various metrics. The Sigmoid kernel demonstrated the highest True Positive Rate (0.878) and Precision (0.878), making it noteworthy; however, the absence of False Positive Rate and Recall values limits a comprehensive assessment of its overall performance. On the other hand, the Radial Basis Function (RBF) kernel exhibited a balanced and consistent performance with high Precision (0.984) and Recall (0.984) values. These results highlight the critical importance of kernel selection in SVM models for achieving accuracy and balanced classification. The performance metrics of the SVM, evaluated using different kernel functions (Linear, Polynomial, Radial Basis Function, and Sigmoid), are summarized in Table 5, highlighting metrics such as True Positive Rate, False Positive Rate, Precision, and Recall.

Table 5. Performance Metrics of Support Vector Machine (SVM) with Different Kernel Functions

Kernel	True Positive Rate	False Positive Rate	Precision	Recall
Linear	0,948	0,032	0,968	0,948
Polynomial	0,949	0.051	0,94	0,949
Radial Basis Function	0,9563	0	1	0.9525
Sigmoid	0,878	N/A	0,878	N/A

Based on the results presented in Table 5, the radial basis transformation demonstrated the highest precision (1) and recall (0.984), indicating its superior ability to accurately predict and identify students likely to stay or leave. This suggests that the radial basis function is the most suitable option for this study, as it achieves a balanced and reliable performance across all metrics.

The model presents an array of weights, indicating the influence of each variable on the prediction outcome. Variable X_{12} exhibits the highest positive impact with a weight of 0.9751, followed by X_5 and X_{11} with weights of 0.0021 and 0.0013 respectively. On the contrary variables X_6, X_7, X_8, X_9, and X_{10} depict negative impacts on the prediction outcome.

During the training phase, the model required 205761 kernel evaluations, 97.028% of which were cached, indicating computational efficiency. The model was built in 0.31 seconds.

Upon evaluation of the training set, the model demonstrates a moderate correlation coefficient of 0.4433, signifying a discernible but not strong linear relationship between the predicted and actual values. Additionally, the MAE is reported at 0.1, with a root mean squared error of 0.31, indicating the average and overall discrepancies between predicted and actual values, respectively. Relative absolute error and root relative squared error are calculated at 45.79% and 94.28%, respectively, offering insights into the relative performance of the model across the dataset.

In conclusion, the model trained on the dataset exhibits a fair predictive performance, characterized by moderate correlation coefficients and errors within acceptable ranges. Further refinement or exploration of alternative algorithms may be warranted to enhance prediction accuracy, particularly in scenarios demanding higher precision. The confusion matrix created for the SVM the radial basis function is shown in Table 6.

Table 6. Confusion Matrix for Support Vector Machine with Radial Basis Function Kernel

Row Labels	0	1	Grand Total
0	561	0	**561**
1	28	52	**80**
Grand Total	**589**	**52**	**641**

4.4. Artificial Neural Networks (ANNs)

In order to determine the minimum RMSE, the number of neurons is varied between 7 and 15. The x-axis shows the values of different hidden neurons (7, 8, 9, ..., 15), and the y-axis shows the Root Mean Square Error (RMSE) values calculated for these values.

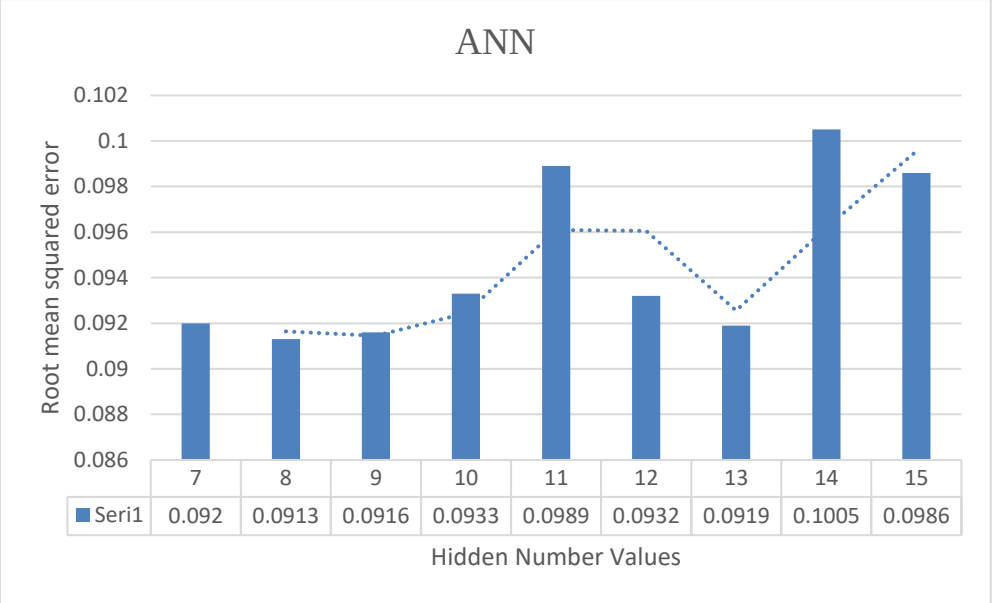

Figure 6. The number of neurons versus Root Mean Square Error

When looking at the results in Figure 6, 8 hidden neurons have the lowest RMSE value, meaning the best performance of the model was observed in these two cases. Lower or higher hidden layer numbers caused an increase in the error rate. This shows that increasing the optimum hidden layer number does not always give better results. The number of hidden layers is accepted as 8.

In the analysis performed with the Artificial Neural Network (ANN) model, 99.06% accuracy, 99.64% precision and 99.29% recall rates were achieved. As seen in Table 7, the total number of individuals attending school is 561, 559 of these individuals were correctly classified, and only 2 individuals were incorrectly predicted as having dropped out. For individuals who dropped out of school, 76 out of a total of 80 individuals were correctly predicted, and 4 individuals were incorrectly classified as continuing school. These results show that the ANN model can predict school attendance and school dropout status with high accuracy.

Table 7. Confusion Matrix for Artificial Neural Network (ANN) Model

Row Labels	0	1	Grand Total
0	559	2	561
1	4	76	80
Grand Total	**603**	**78**	**641**

The performance results obtained show that the model can distinguish both classes with a high accuracy rate and generally makes correct predictions. The number of incorrect classifications in the confusion matrix is limited, indicating that the overall success of the model is high. This analysis performed on a total of 641 data confirms that the ANN model offers an effective method in predicting student behavior.

4.5. K-Nearest Neighbors (K-NN) Algorithm

In the K-Nearest Neighbors (KNN) algorithm, the k value determines how many neighbors will be considered when making a prediction and directly affects the performance of the model.

In Figure 7, the x-axis shows different k values (from 2 to 13), and the y-axis shows the Root Mean Square Error (RMSE) values calculated for these k values. According to the results, the lowest RMSE value was obtained for k = 2, meaning that the best performance of the model was observed at this point. Therefore, other stages and comparisons were made by taking k = 2 as a reference.

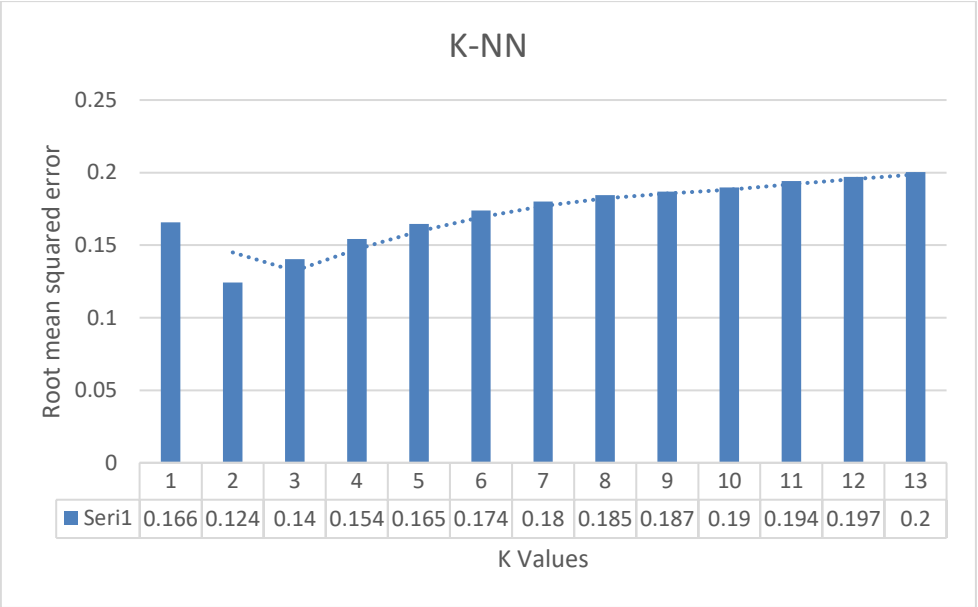

Figure 7. K Values versus Root Mean Square Error

The K-Nearest Neighbors (K-NN) classifier, implemented with k=2, achieved a classification accuracy of 0.9641 over a dataset consisting of 641 instances. In this method, all values between 1 and 10 were given for k, and the best result was observed at k=2. This outcome highlights the algorithm's capability to provide reliable predictions. The confusion matrix presented in Table 8 indicates that 564 instances were correctly classified as TP, and 54 instances were accurately classified as TN.

Table 8. Confusion Matrix for K-Nearest Neighbors (K-NN) Algorithm

Row Labels	0	1	Grand Total
0	564	0	564
1	23	54	77
Grand Total	**587**	**54**	**641**

The classifier also yielded an accuracy of 0.9641, a precision of 1, and a recall of 0.9608. These metrics indicate that the model performs well in reducing Type I errors FP while maintaining a relatively high capacity to detect TP. Despite these strengths, the presence of 23 FN suggests a limitation in correctly identifying all positive instances within the dataset.

CHAPTER 5: CONCLUSION AND FURTHER RESEARCH

This study underscores the critical importance of understanding and addressing student attrition in higher education. By employing logistic regression analysis and exploring the relationships between various demographic, academic, and enrollment-related variables, significant predictors of student attrition have been identified. The findings highlight the necessity of early identification of at-risk students and the implementation of targeted interventions to enhance retention rates.

The comparative analysis of logistic regression, the application of Random Forest, Artificial Neural Networks (ANNs), and K-NN using WEKA, illustrates the potential of machine learning techniques in predicting student attrition with high accuracy. These predictive models provide educational institutions with powerful tools to support students and reduce attrition rates proactively.

This study has delved into the complex phenomenon of student attrition within higher education institutions, employing a comprehensive approach that integrates demographic, academic, and enrollment-related variables. Through the analysis of data spanning from 2017 to 2023 and utilizing logistic regression techniques, we have uncovered significant insights into the determinants of student retention across various faculties.

The findings underscore the importance of considering socio-economic background, academic performance, and engagement metrics in understanding student attrition patterns. By identifying at-risk students and implementing targeted interventions, institutions can proactively address factors contributing to attrition, thus fostering a supportive and conducive academic environment.

Moreover, the predictive capabilities developed through this research offer a valuable tool for educational institutions to anticipate retention rates before each semester, enabling them to allocate resources more effectively and tailor interventions to individual student needs.

Moving forward, further research could explore longitudinal trends in student attrition, delve deeper into the effectiveness of specific intervention strategies, and examine the impact of external factors such as economic conditions or changes in educational policy. Additionally, qualitative studies could provide richer insights into the lived experiences of students at risk of attrition, shedding light on the underlying reasons for their departure.

Ultimately, by fostering a deeper understanding of the drivers behind student departure and equipping institutions with actionable insights, this study contributes to the ongoing efforts to enhance student retention and promote the holistic development and success of all students within higher education.

This study evaluated five machine learning methods using four key performance metrics: Accuracy, Precision, and Recall.

The models under comparison include Logistic Regression, Random Forest, Artificial Neural Network (ANN), Support Vector Machine (SVM), and K-Nearest Neighbors (K-NN). The results of the analysis are summarized in Table 9.

Table9. Performance Comparison of Classification Models Across Key Metrics

	Accuracy	Precision	Recall	F1 Score
LR	0.967	0.977	0.986	0.981
RF	0.952	1	0.948	0.973
SVM	0.927	1	0.923	0.959
ANN	0.991	0.996	0.993	0.994
K-NN	0.964	1	0.961	0.980

The performances of the classification models evaluated in this study were compared, especially through the F1 score. The F1 score is an important metric that reflects the overall classification success of the model by balancing both precision and recall values. Especially in unbalanced data sets such as the results obtained in this study, it allows the evaluation of model performance by establishing a balance between false positive and false negative predictions and is considered a critical measure in

classification problems.

According to the results obtained, the model that exhibited the highest performance with an F1 score of 0.994 was the ANN. This value shows that the model minimizes misclassifications and makes balanced and reliable predictions. LR (0.981) and K-NN (0.980) models also had values close to ANN, but ANN provided the highest success. RF (0.973) and SVM (0.959) models have relatively lower F1 scores.

In Table 9, Logistic Regression (LR) has a wide range of applications in the literature as a simple and effective classification algorithm. In our study, the accuracy value obtained with this method was calculated as 0.9672, the precision value as 0.977 and the recall value as 0.9857. These results show that Logistic Regression provides high accuracy and recall. Especially the high recall value reveals that this method exhibits a strong performance in correctly predicting the dropout status of students. In addition, it is seen that the model keeps false negative classifications to a minimum and is effective in identifying risky students. Logistic Regression has become an attractive option for researchers and practitioners due to the explainability and simplicity of the model.

Random Forest (RF) is a method that consists of multiple decision trees and makes predictions by combining the outputs of these trees. The accuracy value obtained with the Random Forest method was found as 0.9516, the precision value as 1 and the recall value as 0.9476. The perfect precision value shows that the model almost completely eliminates false positive predictions. However, the relatively lower recall value indicates that the model cannot predict some dropout cases correctly. This shows that the model, although providing high accuracy, falls behind Logistic Regression and ANN in terms of recall.

Support Vector Machine (SVM) is known as a powerful algorithm in classification problems and shows effective performance in data sets with non-linear relationships. In the study, the accuracy value of 0.9267, the precision value of 1, and the recall value of 0.9227 were obtained with the SVM method. The high precision value shows that the model makes positive classifications quite accurately. However, the relatively lower recall value indicates that the model misses some dropout cases. This shows that

SVM is less effective compared to other methods. However, it should be kept in mind that SVM is a method sensitive to variable optimization and its performance can be improved with appropriate hyperparameter settings.

Artificial Neural Network (ANN) is known for its superior performance in complex data sets and relationships. In our study, the accuracy value was calculated as 0.9906, the precision value as 0.9964, and the recall value as 0.9929 with the ANN method. Reaching the highest values in all metrics, ANN showed superior performance in predicting students' school dropout status. The strong results of ANN are an indicator of the algorithm's multi-layered structure and its capacity to learn complex relationships. It is predicted that this superiority of ANN will become more evident especially in large data sets and in analyses with more variables.

K-NN is a classification algorithm based on neighborhood relationships and generally stands out with its simple structure. In the study, the accuracy value was determined as 0.9641, the precision value as 1, and the recall value as 0.9608 with the K-NN method. K-NN, which showed a very successful performance in terms of precision, fell behind ANN and Logistic Regression in recall and accuracy metrics. However, the performance of K-NN depends on the size of the data set and the number of neighbors selected. It is known that K-NN can perform better, especially in small or homogeneous data sets.

These results show that the methods used to predict students' dropout status have different strengths and weaknesses. Artificial Neural Network (ANN) has achieved the highest accuracy, precision, and recall values, outperforming other methods. The success of ANN requires it to be preferred especially in cases where complex data structures are worked on. Logistic Regression and K-NN have provided results close to ANN and have stood out as effective alternatives. Random Forest and SVM have shown strong performance in terms of precision, but have reached relatively lower values in recall and accuracy metrics.

According to the study results, the best-performing model was ANN. It reached the highest values in accuracy (0.9906), precision (0.9964), and recall (0.9929) metrics. These results reveal that ANN is a superior model in predicting students' dropout status

and provides balanced success in all performance metrics. ANN ranked first with high accuracy in both positive and negative classifications thanks to its ability to model complex data structures. score is the criterion that best reflects the classification success and ANN stands out as the most reliable model. Providing both high precision and recall, it shows that the model exhibits strong performance even in imbalanced data sets.

Following ANN, the second-best model was evaluated as Logistic Regression (LR). LR provided very strong results in accuracy (0.9672), precision (0.977), and recall (0.9857) metrics and showed a performance close to ANN. Logistic Regression stands out due to the simple structure of the model and its low computational cost. However, it is seen that LR falls slightly behind compared to ANN's advantage in learning complex data relationships.

The third place is taken by the K-NN model. K-NN performed well in terms of accuracy (0.9641), precision (1) and recall (0.9608), but was slightly lower in the recall metric than ANN and Logistic Regression. Although the performance of K-NN is not as strong as ANN and LR, it provides excellent results in the precision metric, showing that it is a reliable model in positive classifications.

These analyses guide educational institutions and researchers in determining which method is more suitable by comparing the prediction performance of different machine learning algorithms. While the strong performance of ANN emphasizes that it is an ideal method for modeling complex data structures, methods such as Logistic Regression and K-NN can offer simpler and more effective solutions. It should also be noted that Random Forest and SVM offer high values in terms of precision and therefore can be effective in certain scenarios. The information obtained from these methods will contribute to the development of proactive policies to improve educational processes and prevent students from dropping out of school.

This study focused on identifying the risk factors contributing to student attrition in higher education institutions, emphasizing academic and demographic variables. However, to provide a more comprehensive understanding, the following future research directions are suggested:

1) Time Series Analysis

   Although this study analyzed data from 2017 to 2023, a time series analysis could delve deeper into the periodic or annual changes in student attrition rates. Such an analysis would help uncover the impact of economic, social, or political events on student performance and retention. Additionally, time series forecasting can equip higher education institutions with the tools to better anticipate and mitigate future risks.

2) Enhancement of Machine Learning Models

   The supervised and unsupervised methods used in this research could be improved by exploring alternative machine learning algorithms and optimizing hyperparameters. Deep learning techniques, in particular, could reveal complex relational networks affecting student attrition. Moreover, incorporating explainable AI techniques could enhance the interpretability of model outputs, aiding decision-makers in understanding and applying the findings effectively.

3) Integration of Psychosocial Factors

   While the current study considered demographic and academic variables, integrating psychosocial factors could provide a more holistic perspective. Variables such as stress levels, access to social support mechanisms, and family circumstances could enrich the analysis and improve predictive accuracy. This approach would enable universities to design more effective student support programs tailored to individual needs.

4) Analysis of Long-Term Effects

   Exploring the long-term consequences of student attrition on individuals and society could expand the scope of existing research. For instance, examining the unemployment rates, income levels, and life satisfaction of students who leave university would provide insights into the broader economic and social implications of education. Such an analysis would inform the strategic planning of education policies for sustainable development.

   These suggestions aim to foster a more comprehensive approach to addressing student attrition, contributing to both the literature and the efforts of higher education institutions to enhance student engagement and retention.

REFERENCES

Delua, J. (2021) *Supervised versus unsupervised learning: What's the difference?* IBM. [Online] Available at: https://www.ibm.com/think/topics/supervised-vs-unsupervised-learning. (Accessed: 20 May 2024).

Bagabir, A. M., Zaino, M. R., Abutaleb, A., Fagehi, A. (2021) *Predictive analysis of higher-education graduation and retention in Saudi Arabia using multinomial logistic regression.* International Journal of Basic and Applied Sciences, Vol. 3(6).

Qvortrup, A., and Lykkegaard, E. (2022) *Study environment factors associated with retention in higher education*. Higher Education Pedagogies, Vol. 7(1), pp. 37–64.

Herodotou, C., Naydenova, G., Boroowa, A., Gilmour, A., and Rienties, B. (2020) *How can predictive learning analytics and motivational interventions increase student retention and enhance administrative support in distance education?* Journal of Learning Analytics, Vol. 2, pp. 72-83.

Aypay, A., Çekiç, O., and Boyacı, A. (2012*) Student retention in higher education in Turkey: A qualitative study.* Journal of College Student Retention: Research, Theory and Practice, Vol. 14.

Dissanayake, H. U. (2016) *Predicting Student Retention: A Comparative Study of Predictive Models for Predicting Student Retention at St. Cloud State University.* St. Cloud State Universit, Vol. 10.

Shafiq, D. A., Marjani, M., Habeeb, R. A. A., and Asirvatham, D. (2022) *Student retention using educational data mining and predictive analytics: A systematic literature review.* IEEE Access, Vol. 10.

Ram, S., Wang, Y., Currim, F., and Currim, S. (2015) *Using Big Data for Predicting Freshman Retention.* Proceedings of the Thirty Sixth International Conference on Information Systems, Fort Worth 2015 University of Arizona, United States.

Matz, S. C., Bukow, C. S., Peters, H., Deacons, C., Dinu, A., and Stachl, C. (2023*) Using machine learning to predict student retention from socio-demographic characteristics and app-based engagement metrics*. Scientific Reports, Vol. *13*.

Peng, P., Chao-Ying Joanne, L., Kuk Lida Lee, and Ingersoll, G. M. (2002) *An introduction to logistic regression analysis and reporting.* The Journal of Educational Research, Vol. 96.

Larner, A. J. (2002) *The 2x2 matrix: Contingency, confusion and the metrics of binary classification.* 2nd edition. University College London, London, UK.

Bugbee, Erin, and Jared Wilber. (2022) *Logistic Regression.* [Online] Available at: https://mlu-explain.github.io/logistic-regression/ (Accessed: 25 May 2024).

Tavoosi, S. (2019) *A beginner's guide to machine learning with R.* Kaggle. [Online] Available at: https://www.kaggle.com/learn/beginners-guide-to-machine-learning-with-r. (Accessed:16 May 2024).

Patel, S. (2017) *Machine Learning 101,* Medium. [Online] Available at: https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72. (Accessed: 20 May 2024).

Nieuwoudt, J. E., and Kelly, M. L. (2021) *Student Retention in Higher Education: Why Students Choose to Remain at University*. Journal of College Student Retention.

Villegas-Ch, W., Govea, J., and Revelo-Tapia, S. (2023) *Improving student retention in institutions of higher education through machine learning: A sustainable approach.* Sustainability, Vol. 15(19).

Nieuwoudt, J. E., and Kelly, M. L. (2021) *Student retention in higher education: Why students choose to remain at university*. Journal of College Student Retention.Vol. 25.

Lai, A. H. Y., Wong, E. L. Y., Lau, W. S. Y., Tsui, E. Y. L., and Leung, C. T. C. (2024) *Life-World Design: A career counseling program for future orientations of school students.* Children and Youth Services Review, Vol. 161.

Bhandari, A. (2024) *Confusion Matrix in Machine Learning* Analytics Vidhya. [Online] Available at: https://www.analyticsvidhya.com/articles/confusion-matrix-in-machine-learning/. (Accessed: 16 November 2024).

Peng, J., Jury, E. C., Dönnes, P., and Ciurtin, C. (2021) *Machine learning techniques for personalised medicine approaches in immune-mediated chronic inflammatory diseases: Applications and challenges.* Frontiers in Pharmacology, Vol. 12. [Online] Available at: https://doi.org/10.3389/fphar.2021.720694. (Accessed: 15 April 2024)

Çelik, Ö., and Altunaydın, S. S. (2018) *A Research on Machine Learning Methods and Its Applications.* Journal of Educational Technology and Online Learning, Vol. 1(3).

Jiang, T., Gradus, J. L., and Rosellini, A. J. (2020) *Supervised machine learning: A brief primer.* Behavior Therapy. Vol. *51*(5).

Sperandei, S. (2014) *Understanding Logistic Regression Analysis.* Biochemia Medica, Vol. 24, pp. 12–18.

Brijith, A. (2023) *ANN (Artificial Neural Networks): A basic guide.* Asia University, Taichung, Taiwan.

Farooq, U., Mohd Rahim, M. S., Sabir, N., and Abid, A. (2021). *Advances in machine translation for sign language: Approaches, limitations, and challenges.* Neural Computing and Applications. Vol.33(7).

APPENDICES

Appendix A: Ethics Committee Decision

SAYI: B.30.2.İEÜFMB.0.05.05-20-056 24.06.2024

KONU: Etik Kurul Kararı hk.

Sayın Aslıhan Bal,

24.06.2024 tarih ve 47 numaralı Etik Kurul toplantısında yöneticisi olduğunuz "Üniversiteyi Bırakma Oranı Tahminlemesi" başlıklı projeniz görüşülmüştür. Başvurunun etik açıdan uygun olduğu gerekçesiyle onaylanmasına, toplantıya katılan üyelerin oy birliği ile karar verilmiştir.

Gereği için bilginize sunarım.
Saygılarımla,

Prof. Dr. İsmihan Bayramoğlu
Fen ve Mühendislik Bilimleri
Etik Kurul Başkanı

Appendix B: Results of Logistic Regression

Regression Equation

P(1) = exp(Y')/(1 + exp(Y'))

Y' = 2.00 + 0.2049 c1 + 0.955 c2 + 0.300 c3 + 0.0120 c4 + 3.54 c5 - 2.302 c6 - 0.0542 c7+ 0.2476 c8 - 0.974 c9 - 0.670 c10 - 0.001 c11 - 0.93 c12

Model Summary

Deviance R-Sq	Deviance R-Sq(adj)	AIC	AICc	BIC	Area Under ROC Curve
76,29%	73,81%	140,39	140,97	198,41	0,9819

Odds Ratios for Continuous Predictors

	Odds Ratio	95% CI
c1	1,2274	(1,0696, 1,4084)
c2	2,5998	(0,7489, 9,0245)
c3	1,3500	(1,0281, 1,7726)
c4	1,0121	(0,9094, 1,1263)
c5	34,3661	(2,7852, 424,0340)
c6	0,1000	(0,0258, 0,3877)
c7	0,9473	(0,9193, 0,9761)
c8	1,2810	(1,0913, 1,5037)
c9	0,3776	(0,2514, 0,5672)
c10	0,5117	(0,3594, 0,7283)
c11	0,9987	(0,2856, 3,4926)
c12	0,3953	(0,0058, 26,9890)

Coefficients

Term	Coef	SE Coef	Z-Value	P-Value	VIF
Constant	2,00	2,80	0,71	0,476	
c1	0,2049	0,0702	2,92	0,004	1,35
c2	0,955	0,635	1,50	0,132	1,43
c3	0,300	0,139	2,16	0,031	1,21
c4	0,0120	0,0546	0,22	0,826	1,14
c5	3,54	1,28	2,76	0,006	2,26
c6	-2,302	0,691	-3,33	0,001	7,38
c7	-0,0542	0,0153	-3,54	0,000	4,22
c8	0,2476	0,0818	3,03	0,002	12,24
c9	-0,974	0,208	-4,69	0,000	9,63
c10	-0,670	0,180	-3,72	0,000	1,56
c11	-0,001	0,639	-0,00	0,998	1,31
c12	-0,93	2,15	-0,43	0,667	2,53