GRADUATE PROGRAMS INSTITUTE

ENHANCING STUDENT RETENTION: A PREDICTIVE STUDY ON DROPOUT RATES AND REMEDIAL STRATEGIES

Raghdah RAGHDAH

MASTER'S THESIS

Istanbul, August 2025

ENHANCING STUDENT RETENTION: A PREDICTIVE STUDY ON DROPOUT RATES AND REMEDIAL STRATEGIES

Raghdah RAGHDAH

MASTER'S THESIS

COMPUTER ENGINEERING DEPARTMENT

COMPUTER ENGINEERING MASTER'S PROGRAM WITH THESIS

MASTER'S THESIS ADVISOR

Asst. Prof. Dr. Gizem TEMELCAN ERGENECOSAR

MASTER'S THESIS JURY MEMBERS

Asst. Prof. Dr. Inal BEGUM TURNA DEMIREL

Asst. Prof. Dr. Engin HENGIRMEN

ACKNOWLEDGMENT

I would like to express my gratitude and respect to all the individuals who have guided, mentored and supported me during the process to help me complete my thesis. First of all, I would like to thank Almighty Allah for showering me with strength, intelligence and courage to focus and study for the study.

I extend my sincere gratitude to my supervisor Asst. Prof. Dr. Gizem TEMELCAN ERGENECOSAR whose knowledge, insights, attention to details and constructive feedback for not just theoretical but practical work has played crucial work in the final shape of my work.

I would like to express deep appreciation for my parents whose sacrifices, love and support has always given me strength to always learn and thrive in different environments. Their trust in me has always been pillars of my success. To my extra ordinary siblings whose achievements have been source of motivation for me to get the inspiration and deliver quality work.

To my fellow friends, whose encouragement has been source of positivity throughout my work. Finally, I acknowledge the University for providing us the platform to learn and grow our knowledge span that can be helpful to shape our future.

CONTENT

ABSTRACT

ENHANCING STUDENT RETENTION: A PREDICTIVE STUDY ON DROPOUT RATES AND REMEDIAL STRATEGIES

Education is important for economic and social well-being where dropout is a concern for students and institutions worldwide. This thesis seeks to improve students' retention by predicting the students' dropouts and remedial strategies. Prediction algorithms can identify students with poor academic achievement, demographics, and social conduct who are most likely to drop out. In this thesis, the machine learning techniques are used to build a predictive model to analyze dropout or retention of the students at early stages. Key features like grades, age of enrollment, courses taken, courses approved, gender, tuition fees status, class duration etc. are evaluated and analyzed using Random Forest classifier and correlation analysis.

The findings of the thesis show that institutional issues, socio-economic factors, academic challenges, personal and psychological factors have a significant impact on dropout rates. The output shows the students who are likely to be retained or dropped out from the institution with a data-driven predictive model that can be helpful to structure and increase strategies for retention. This thesis not only helps in shaping retention strategies but also can provide detailed analysis for advisors, administrations and policy makers to create a support mechanism by applying such educational structure that can be shaped for students having different economical, social, educational and psychological backgrounds.

Keywords: Retention, Dropout, Remedial Strategies, Data-Driven Predictive Model, Prediction Algorithm.

Date: August 2025.

ÖZET

ÖĞRENCİ DEVAMLILIĞININ GELIŞTIRILMESI: TERK ORANLARI VE TELAFİ STRATEJİLERİ ÜZERİNE BİR ÇALIŞMA

Eğitim, öğrencilerin ve kurumların dünya çapında okulu bırakmasının bir endişe kaynağı olduğu ekonomik ve sosyal refah için önemlidir. Bu çalışma, öğrencilerin okulu bırakmalarını ve telafi stratejilerini tahmin ederek öğrencilerin tutulmasını iyileştirmeyi amaçlamaktadır. Tahmin algoritmaları, okulu bırakma olasılığı en yüksek olan zayıf akademik başarıya, demografiye ve sosyal davranışa sahip öğrencileri belirleyebilir. Bu çalışmada, makine öğrenimi teknikleri, öğrencilerin erken aşamalarda okulu bırakmasını veya tutmasını analiz etmek için tahmini bir model oluşturmak için kullanılmıştır. Notlar, kayıt yaşı, alınan küfürler, onaylanan dersler, cinsiyet, öğrenim ücreti durumu, sınıf süresi vb. gibi temel özellikler, Random Forest sınıflandırıcısı ve korelasyon analizi kullanılarak değerlendirilir ve analiz edilir.

Çalışmanın bulguları kurumsal sorunların, sosyoekonomik faktörlerin, akademik zorlukların, kişisel ve psikolojik faktörlerin terk oranları üzerinde önemli bir etkiye sahip olduğunu göstermektedir. Çıktı, kurumda kalma veya kurumdan ayrılma olasılığı yüksek öğrencileri, tutma stratejilerini yapılandırmak ve artırmak için yardımcı olabilecek veri odaklı bir tahmin modeliyle göstermektedir. Bu tez yalnızca tutma stratejilerini şekillendirmeye yardımcı olmakla kalmaz, aynı zamanda danışmanlar, idareciler ve politika yapıcılar için farklı ekonomik, sosyal, eğitimsel ve psikolojik geçmişlere sahip öğrenciler için şekillendirilebilecek böyle bir eğitim yapısını uygulayarak bir destek mekanizması oluşturmak için ayrıntılı analiz sağlayabilir.

Anahtar Kelimeler: Devamlılık, Terk Oranı, Telafi Stratejileri, Tahmin Algoritmaları.

Tarih: Ağustos, 2025.

ABBREVIATIONS

AI	: Artificial Intelligence
AUC-ROC	: Area under the Curve - Receiver Operating Characteristic
DT	: Decision Tree
EDM	: Educational Data Mining
FN	: False Negative
FP	: False Positive
FPR	: False Positive Rate
GDP	: Gross Domestic Product
GPA	: Grade Point Average
GUI	: Graphical User Interface
HE	: Higher Education
KDD	: Knowledge Discovery in Databases
LA	: Learning Analytics
LMS	: Learning Management System
ML	: Machine Learning
RF	: Random Forest
RFC	: Random Forest Classifier
SVM	: Support Vector Machine
TN	: True Negative
TP	: True Positive
TPB	: Theory of Planned Behavior
TPR	: True Positive Rate
UK	: United Kingdom
US	: United States

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

This thesis focuses on the predictive analysis performed on higher education by using machine learning techniques to help understand and analyze the patterns linked with students’ various attributes, including socioeconomic, financial, academic, and institutional factors. These attributes can be used to develop the predictive model for accurately identifying whether students will be dropping out or staying enrolled. The aim is to enable institutes to make decisions based on historical data that can give them useful insights for taking timely interventions. These decisions can lead to the general success ratio of the educational institutes in having higher number of student retentions.

Students’ retention is a key metric to measure an institutions’ success as good retention rate contributes to attract new students, faculty and staff, improving their reputation and refining policies to improve academic quality (Burke, 2019); (Crosling, 2017). Araque et al. (2009) have highlighted that students’ dropout rate is the percentage of students who leave school before finishing their education and it provides basis for understanding key factors driving dropouts. According to Roberts (2020); Tagliaferro (2023), 6.3% students in UK drop out from university out of which 33% of students leave within first year in UK. Wild and Heuling (2020) signified that students’ retention is important to reduce dropouts and with the rise of big data and analytics, academics have used analytics to predict dropouts, hence this thesis emphasizes enhancing students’ retention by using predictive models of dropout rates and remedial strategies.

Countries have very different academic standards and benchmarks, which makes it difficult for international students to adapt to new learning contexts. The grading scheme is one obvious point of difference. For instance, grades in the United States are frequently given on a 4.0 GPA scale, whereas in the UK, grades are usually presented as percentages, with 70% being regarded well (Tagliaferro, 2023). Students from other countries may find this standard confusing, as 90% is expected for top results. In a similar vein, systems in nations like India or Pakistan, where larger numerical scores signify greater results, contrast with those in Germany, where a lower numerical grade suggests higher performance. Students may become confused about their performance and how to interpret comments as a result of these variations in grading practices (Tagliaferro, 2023). Expectations in the classroom and teaching methods also differ greatly. Classroom environments in many Western nations, like Canada and Australia, frequently place an emphasis on critical thinking, active involvement, and candid communication between teachers and students. Students are encouraged to participate in group discussions, pose questions, and challenge preconceived notions (Roberts, 2020). However, the classroom may be more teacher-centered in places like China, Japan, or many Middle Eastern countries, where pupils are expected to pay attention in class and refrain from interrupting. When moving from a more hierarchical educational system to a participative one, students may find it difficult to adjust and may not know how or when to express their ideas (Roberts, 2020).

Methods of assessment are still another important distinction. Continuous assessments, including as group projects, essays, and presentations, are typical in several nations, like the UK and Finland. On the other hand, end-of-term exams frequently account for the majority of a student's mark in nations like South Korea and India. It could be difficult for a foreign student accustomed to exam-based assessments to adjust to assignments that call for independent study, critical analysis, and copious referencing (Roberts, 2020). For example, if academic integrity standards, like plagiarism regulations, are more tightly enforced in the host nation, a student from a rote-learning background may find it intimidating to write persuasive essays or present research findings. These challenges may be made worse by language limitations, especially in non-native English-speaking nations. For instance, foreign students studying in France or Germany may run into technical terms in French or German that make it more difficult for them to understand the course material (Song et al., 2023). Students may find it difficult to communicate when faced with foreign academic jargon or cultural nuances, even in English-speaking nations. This can make it challenging to understand assignment requirements in their entirety or to contribute productively to class discussions.

Cultural perspectives on workload and time management might also differ. Students in nations like the US frequently balance their studies with internships and part-time jobs, which are regarded as crucial for advancing their careers. On the other hand, education may be seen as a full-time undertaking with little focus on extracurricular activities in places like Saudi Arabia or many African countries (Matz et al., 2023). Managing time in a new system may be too much for a foreign student who is not used to juggling academic and extracurricular obligations. For international students, these distinctions make the transfer difficult, especially when combined with social and cultural adaptations. By identifying and resolving these issues through academic seminars, peer support, and orientation programs, the gap can be closed and international students can succeed in a variety of educational environments.

1.1 Background

Matz et al. (2023) have highlighted that dropping out from institutions without a degree negatively impacts students' finances and mental health. Drop-out rates refer to the percentage of students who leave schools before finishing their study or completing the requirements of graduation (Araque et al., 2009). Hanushek and Woessmann (2010) highlighted that education is important for economic and social well-being where dropout is a concern for students and institutions. Furthermore, Song et al. (2023) highlighted that many educational institutions worldwide struggle with student attrition. According to Unesco.org. (2023) 24 million pupils drop out due to the COVID-19 epidemic. The Unesco's research anticipates a 3.5% reduction in enrolment, 7.9 million fewer students, and the highest higher education dropout rate (Song et al., 2023). The students' dropouts from schools are also important as it become social pressure once they enter society due to their academic failure and early entry. This could lead to poverty, joblessness, and crime (Song et al., 2023). While analyzing the factors of students' dropouts, Del Savio et al. (2022); Contreras et al. (2022) highlighted that family, community, psychology, and school contribute to students' dropouts. The key factors of students' dropouts are lack of motivation as academic factors, financial factors, socio-economic factors and personal factors such as low satisfaction with educational experiences. Through understanding dropout rates, educational institutions improve educational and learning experiences.

Seo et al. (2024) highlighted that due to the speed and complexity of modern life, lifelong education has become more important. In this age of rapid technological advancements, global economic shifts, and shifting employment marketplaces, everyone must continually adapt and gain new skills to stay competitive and relevant (Laal and Laal, 2012). In relation to lifelong and continuing education, digital platforms and online institutions can make education more accessible where students' dropout rates are important to study. In relation to dropouts, students' risk having their career and social growth halted and suffering psychological effects including low self-esteem and a lack of direction. Due to high dropout rates, universities face lower retention rates, financial cuts, and reputation damage (Simpson, 2013). These difficulties may impair their academic ranking, recruitment, and retention; hence this study considers the significance of dropout rates.

With the rise of AI and data analytics, educational institutions have used the predictive analytics models such as educational data mining (EDM) and learning analytics to identify at-risk students (Agrusti et al., 2019). Student retention is crucial for institutions and students. From the student's perspective, retention is important for academic success, job prospects, and personal growth. In relation to institutions, students' retention contributes to their financial sustainability, academic reputation, and institutional rankings (Burke, 2019). Webster and Showers (2011) highlighted that increasing retention rates saves money since obtaining new students costs more than retaining current ones. For enhancing students' retentions, Seo et al. (2024) highlighted students' engagement, sense of community, academic performance through learning management system and efficient communication contributes to students' retention. Furthermore, Seo et al. (2024) signified that academic tutoring, campus environment and counselling help students to keep them retain and contribute to students' engagement.

Considering the significance of students' retention, the recent advancements of technology have helped in improving students' retention where authors have used machine learning and predictive learning models (Matz et al., 2023). Hamdane et al. (2023) have signified that LA technology contributes to analyzing the database and summarizing the experiences to predict at risk factors of dropping outs. LA measures, collects, analyses, and reports student data to improve learning environments (Siemens and Baker, 2012). Learning analytics also provides real-time student learning insights where using LA, teachers can monitor student behavior on learning management systems to identify key issues of dropouts. This allows students to get support when they need it, preventing them from falling behind or quitting out. Gašević et al. (2015) have signified that LA technologies, along with customized learning plans and continual feedback, can boost higher education retention by 15% – 20%, which underlined the significance of learning models for students' retention.

In addition to this, Agrusti et al. (2019) have signified the role of educational data mining techniques to predict university dropouts and highlighted that research is focusing on using data mining techniques for analyzing dropouts. Romero and Ventura (2020) defined EDM as employing computational methods to examine massive educational datasets to predict student performance and behaviour in class. Both methodologies can reveal student involvement, academic performance, and dropout risk factors. Dekker et al. (2009) found that using EDM to predict university attrition rates improved student retention. Their findings showed that test scores, academic progress, and previous academic performance can be used to predict at-risk students. Tempelaar et al. (2025) found that traditional academic indicators and learning management system data improved student dropout forecasts.

1.2 Problem Statement

To understand and predict students' dropouts are important for enhancing students' retention which improves institutions' reputation and success. Gonzalez-Nucamendi et al. (2023) highlighted that both secondary and university institutions struggle with student attrition and increased dropouts. The previous studies on students' dropouts have studied the key factors such as poor socioeconomic status, lack of academic aid, poor social and academic skills have been linked to student attrition and dropouts (Nurmalitasari et al., 2023). Cameron and Heckman (2001) highlighted a lack of desire leading to sporadic class attendance as a factor in academic failure. Barbera et al. (2020); Pyle and Wexler (2012) discovered that emotionally or structurally disturbed homes predisposed adolescents to school misbehavior. Some families have high expectations because they do not provide their kids with the resources they need. Lack of parental and family academic support may contribute (Balfanz et al., 2007). First-generation students may need a supportive school environment due to their lack of formal education at home (Gonzalez-Nucamendi et al., 2013).

Recent research studies have used machine learning and data analytics to predict students' dropouts. However, studies like Aulck et al. (2016) have shown dropout causes and solutions but these studies are less effective for highlighting a single explanation for student dropouts due to their complexity. Also, large datasets are studied utilizing data analysis methods to understand variable correlations. According to current research, institutional retention programs should target students at risk of dropping out yet most likely to be retained by adequate selection algorithms.

Many research studies have examined academic contexts, student activities data, personal and societal variables as dropout predictors (Nurmalitasari et al., 2023). In online distance education, digital data from various learning activities can predict outcomes (Mubarak et al., 2022); (Seo et al., 2024). However, too many variables might cause overfitting and analysis complexity. Moreover, Shiao et al. (2023) research has emphasized the significance of input variables which require larger dataset for reliable sources. As the model becomes more complicated, educational practitioners struggle to evaluate and apply their implications. Hence, it has been assessed that most studies are focused towards students' retention and predicting students' dropouts. Considering the complexities of current models, this study seeks to predict students' dropouts and remedial strategies for enhancing students' retention.

Giving students sufficient opportunities to pursue their interests and realize their full potential is a major problem for the educational industry. One significant problem is that kids are not exposed enough, which restricts their comprehension of the variety of job options and paths. This is made worse by the propensity to downplay the significance of subjects like the arts and commerce while elevating others like engineering and medical. Students with diverse interests are frequently discouraged by such cultural biases, which force them to pursue disciplines in which they lack enthusiasm (Shiao et al., 2023). This can result in indifference, subpar academic performance, and even dropout. Furthermore, pupils are treated differently depending on their financial circumstances. Systemic hurdles impede the advancement of disadvantaged students due to unfair regulations and restricted access to high-quality education. The lack of inspiration and support from mentors to promote creativity, innovation, and critical thinking exacerbates these problems. Students find it difficult to explore original ideas or entrepreneurial endeavors without the right direction and assistance (Shiao et al., 2023). A more inclusive and creative learning environment can be fostered by addressing these issues through structural improvements such as providing diversified career recommendations, allowing students to pursue their passions, and establishing equal opportunities.

1.3 Research Gap

However, despite the fact that research on student retention has focused on identifying critical characteristics that influence dropout rates, there are still gaps in our understanding of these difficulties and our ability to address them with predictive and preventative techniques. Numerous studies have concentrated on demographic and academic characteristics, frequently ignoring behavioral and emotional factors that may have a substantial impact on the degree to which students continue their education (Oliveira and Medeiros, 2024).

Furthermore, although prediction models for dropouts are making progress, they frequently lack the capability to intervene in real time, which makes it impossible to offer timely assistance to students who are at risk of dropping out altogether. Furthermore, there is a dearth of data on the effectiveness of remedial measures that are specifically suited to specific dropout predictors, such as mental health or financial hardship, which differ greatly from student population to student population (Oliveira and Medeiros, 2024). It is necessary to conduct additional research in order to build solutions that are both comprehensive and individualized, taking into account both predictive data and holistic support systems. By bridging this gap, it may be possible to increase the development of preventative measures and tailored support programs, which will ultimately lead to more successful tactics for increasing student retention rates and boosting student success.

Giving students sufficient opportunities to pursue their interests and realize their full potential is a major problem for the educational industry.

1.4 Research Aims and Objectives

This thesis seeks to improve students’ retention by predicting the students’ dropouts and remedial strategies. It seeks to accomplish following research objectives:

- To present key factors contributing to students’ dropout rates in educational institutions.
- To develop a predictive model for student’s dropouts in educational institutions.
- To evaluate the effectiveness of remedial strategies for improving students’ retention.
- To recommend strategies for enhancing students’ retention based on predictive dropout rates of educational institutions.

1.5 Significance of Study

This thesis is presented to answer the question: What are key predictors of students’ dropouts and how can remedial strategies enhance students’ retention in educational institutions? Educational institutions struggle with student dropouts and enhancing students’ retention. Student dropouts affect more than simply academic paths, including unemployment and economic inequalities (Hanushek and Woessmann, 2010). Understanding the causes of student dropout and developing appropriate remedial strategies can enhance educational outcomes.

This thesis seeks to study student dropout characteristics and build predictive models to help schools with early intervention and retention. This study examines the complex interaction between personal, academic, and institutional factors to improve student retention. To improve students’ retention, educational institutions identify student dropout factors of at-risk kids where predictive algorithms can help them receive early intervention and support. Academic performance, graduation rates, and educational outcomes can improve with prolonged schooling. The study's findings can influence academic advising, financial aid, and student support policies. This thesis examines how dropout factors interact to better understand them and develop solutions to retain more students.

1.6 Theoretical Framework

As the number of students who drop out of school continues to be a problem all over the world, educational institutions are placing a significant emphasis on improving student retention. Fincham et al. (2021) Tinto's Model of Student Integration and the Theory of Planned Behavior are two significant theories that are frequently utilized in the process of comprehending and enhancing the retention of students. These frameworks provide useful insights into the factors that lead to student dropout as well as effective solutions to reduce the consequences of this phenomenon.

Fincham et al. (2021) Tinto's Model of Student Integration emphasizes the significance of social and intellectual integration within educational institutions as critical elements in the process of retaining students. The research conducted in suggests that students who have a strong feeling of connection to their academic environment are more likely to persevere and achieve success in their academic endeavors (Fincham et al., 2021). It is the contention of this paradigm that a student's integration is influenced by both the formal (classroom interactions) and the informal (peer relationships) parts of their educational experience. Students have a lower likelihood of dropping out of school if they have established robust academic goals and are actively involved in the academic community. Specifically, the model places an emphasis on the necessity for institutions to cultivate environments that are supportive and that encourage both academic and social activity (Fincham et al. (2021). The sense of belonging among students can be strengthened by institutions through the implementation of methods such as peer mentoring, teacher assistance, and inclusive community events. This will result in a reduction in the number of students who drop out of school. Tinto's idea has been helpful in driving programs that aim to promote student retention by establishing a cohesive campus community and increasing meaningful interactions between staff and students (Müller and Klein 2023).

Another helpful framework for comprehending the concept of student retention is Ajzen's Theory of Planned Behavior (TPB), which was designed specifically for this purpose. According to this theory, the intention of a student to continue their education is determined by three primary factors: attitudes toward persistence, subjective norms, and perceived behavioral control as well as perceived behavioral control. The term "attitude" refers to the degree to which a student has a favorable outlook on continuing their education, which is influenced by their scholastic experiences and the consequences they anticipate (Bosnjak et al., 2020) A person's social pressure to continue their education, whether it comes from their family, their classmates, or cultural standards, is an example of a subjective norm. As a last point of discussion, the student's perceived behavioral control represents their confidence in their capacity to overcome challenges that could result in them dropping out of school. Although the TPB emphasizes the significance that both internal and external variables play in the decisions that students make, it suggests that remedial tactics may be successful if they take into account both the students' personal motivations and their perceptions of support (Ajzen, 2020). By way of illustration, educational establishments can provide students with counseling services, academic support, and financial aid in order to assist them in feeling more in charge of their academic pathways.

Fincham et al. (2021) Tinto's Model of Student Integration and the Theory of Planned Behavior both offer valuable insights into the process of increasing the number of students who remain in school. Through the promotion of academic and social engagement as well as addressing students' objectives, educational institutions are able to gain a deeper understanding of the myriads of factors that contribute to student dropout and so conduct tailored interventions that increase retention rates.

1.7 Ethical Considerations

During the process of data gathering and analysis a number of factors were kept under consideration including data privacy, consent and bias mitigation. Realinho et al. (2021) dataset was captured from an open source and is publicly available for academic use, individual consent is not required and it is fully anonymized before use. Feature importance is interpreted with caution to avoid reinforcing existing inequalities.

1.8 Structure of Study

The study consists of six chapters where the first chapter elaborates the rationale and significance of research by highlighting problems. Chapter 2 reviews available literature on factors of students’ dropouts, students’ retention, current models such as learning analytics and educational data mining are reviewed. Also, this chapter indicates the literature gap which this study seeks to fulfil. Chapter 3 presents the methodology of data collection and analysis to investigate international students' perceptions. Chapter 4 analyses the collected data and presents the findings which were discussed in Chapter 5 in relation to literature. Finally, Chapter 6 concludes and summarizes the findings with limitations and scope of future research.

2. LITERATURE REVIEW

This chapter discusses related work for student dropouts and retention predictions accomplished by various researchers. It focuses on the implementation of data mining and machine learning algorithms used in different stages of education. These past studies help us to identify key factors that influence student dropout rates and gaps that are further classified below in following sub sections.

2.1 Student Retention

One of the most important issues facing higher education is student retention. In education, retention rate is a crucial assessment parameter. High retention rates are achieved as students re-enroll from one academic year to the next. In order to accomplish this, educational institutions must offer suitable guidance and instructional strategies in addition to other activities that improve student performance and keep them from putting off finishing their education (Caruth, 2018). Over the years, this topic has received a great deal of attention in the education sector due to the curiosity of several scholars and organizations regarding what variables lead to students not finishing their degrees and what aspects do not. Because of low student retention, institutions may find it difficult to maintain a high graduation rate. According to general estimates, 40% of undergraduate college students drop out of school altogether, with 30% of first-year students quitting before their second year (Haverila et al., 2020).

Even while there is a wide range of reasons why students choose to leave university, depending on how different institutions evaluate student retention and what data is used, the problem is still very much a worry, particularly for higher education institutions searching for more effective retention tactics. Numerous theoretical and empirical research projects aim to investigate the factors that may contribute to student dropout rates. A primary element found in recent research is financial anxiety stemming from the high expense of school or insufficient support for education (Sarrico, 2018).

Students' financial situation may be impacted by low student retention at universities. A low graduation rate in higher education institutions is caused by the requirement that students repay their education loans even if they are unable to complete their studies. Low-paying jobs may arise from not living up to students' expectations as the need for highly qualified graduates grows (Alsharari and Alshurideh, 2020). Moreover, low retention rates may be a sign that an institution is not fulfilling its objective. Universities can create fewer professionals and practical skills, which can result in a significant loss of human capital if they fail to prepare students' mindsets for the corporate world. Parents, teachers, and children alike all benefit greatly from a high retention rate (Kučak et al., 2018). For instance, if a student is not participating in the assigned readings or coursework, an examination of their development might yield useful information that teachers can use to improve their lesson plans and instructional resources to increase student interest and learning. Students who demonstrate inadequate academic progress or are at risk of discontinuing their studies owing to various factors are classified as "at-risk" students. Early identification of these individuals can yield a number of advantages and beneficial resources for both educational institutions and learners, including the creation of an early warning system designed to address problems that could cause students to postpone or stop their studies (Ortiz-Lozano et al., 2018).

Predicting the successful students can also assist teachers in identifying the qualities of an exceptional student, which can then be codified into a set of guidelines that new university students can be given. Predictive analytics, or more precisely learning analytics, is the term for sophisticated, high-tech solutions focused on data and analysis that are being adopted by numerous organizations (Kučak et al., 2018). Using vast volumes of historical data, analytics are used in education to identify trends and patterns that can be used to forecast future student success. Numerous empirical investigations have demonstrated the need of leveraging learning analytics to improve retention rates. By analyzing data to forecast students' overall performance as early as feasible before their second semester, Georgia State University and an estimated 1,400 other schools and universities are employing learning analytics to reroute students toward a better and more successful career (Caruth, 2018). Students receive recommendations based on these forecasts, such as advice to enroll in tutoring sessions, pursue particular degrees, and so forth. While it is true that "student's degree completion" is generally correlated with student retention, this review gathered and examined empirical studies about student retention and attrition in three distinct classroom environments—online, traditional, and blended learning.

2.2 Educational Data Mining

A new area of study being used in the educational setting is called educational data mining, or EDM. It makes use of the many types of data that are acquired from the institutions, such as student information, administrative records, etc. In order to enhance the teaching and learning environment, it is essential to comprehend how students behave in educational settings and to distinguish successful and unsuccessful students (Hashim et al., 2020). Hidden patterns that suggest correlations between numerous qualities observed in sizable datasets derived from university records can be detected with the use of EDM. EDM is widely used in various educational applications, including performance prediction, warning creation, course maintenance and improvement, recommendation generation, and detection of undesired behavior. The word comes from "data mining," which is the practice of looking for important information in huge datasets by searching through them (Bakshinategh et al., 2018). In order to investigate hidden patterns, it extracts and analyzes data. Numerous fields, including healthcare, higher education, business and financial research, and intrusion detection, can benefit from the use of data mining. Through the six primary steps of the KDD process, we can extract knowledge for Data Mining (Shrestha and Pokharel, 2020). KDD uses data mining techniques to find knowledge by identifying target data sources and cleaning the data to remove inconsistent or incomplete information in order to produce a standardized dataset (Sultan et al., 2021). Additionally, the data is transformed, and data mining techniques like statistics and algorithms are used to find patterns in the data that were previously unknown. The results are then analyzed and assessed to convey the knowledge that has been found.

The use of data mining, machine learning, and statistics to information derived from educational settings, such as colleges and intelligent tutoring systems, is the focus of the study field known as EDM. The field's overarching goal is to create and enhance techniques for examining this data, which frequently includes several levels of relevant hierarchy, in order to gain fresh perspectives on how individuals learn in these kinds of environments (Mohamad and Tasir, 2013). By doing this, EDM has added to the theories of learning that scholars in the learning sciences and educational psychology have studied. The field and learning analytics are closely related, and they have been contrasted and compared. The term "educational data mining" describes methods, resources, and studies created to automatically glean insights from vast databases of information produced by or associated with individuals' educational experiences. This information is frequently detailed, accurate, and comprehensive. A number of LMS, for instance, keep track of data like the time each student accessed each learning object, the number of times they did so, and the duration of time the learning object was visible on the user's computer screen (Scheuer and McLaren, 2012). As an additional illustration, intelligent tutoring systems log information each time a student turns in a solution to a problem. The time since the last submission, if the solution matches the expected solution, the sequence in which the solution components were typed into the interface, and other details may be gathered (Scheuer and McLaren, 2012). Even a very brief session with a computer-based learning environment, such as 30 minutes, may yield a significant amount of process data for analysis due to the high precision of this data. The data is less detailed in other situations. For instance, a student's university transcript might include a chronologically arranged record of the courses they have completed, their grades for each course, and the dates they chose or altered their major. EDM uses both kinds of data to find useful information on the structure of domain knowledge, the impact of instructional strategies incorporated into diverse learning environments, and the sorts of learners and how they learn. It would be challenging to identify new information from the raw data without these techniques. For instance, examining data from an LMS can show a connection between a student's final course grade and the learning resources they used during the course (Peña-Ayala, 2014). Analyzing student transcript data may also show a connection between a student's choice to switch their major and their grade in a certain course. Students, teachers, school administrators, and educational policy makers can use this information to gain insight into how learning environments are designed and make well-informed decisions about how to use, provide, and oversee educational resources.

2.3 Learning Analytics

Through a variety of interactions, including portal logins, assignment uploads, attendance in class, and other activities, students engage with universities (Lemay et al., 2021). Frequently, these interactions create digital trails that can be used to analyze how the university is running its operations and/or helping its students. These trailways can be kept as potentially useful data for Learning Analytics (LA). LA is characterized as the procedure for evaluating student and teacher progress by analyzing, assessing, gathering, and reporting data (West et al., 2016). Most universities are currently facing a problem with student retention. A student's education may be impacted by a number of issues, making it difficult for teachers to keep them until the end of the school year. In order to effectively analyze student learning that is influenced by variables like demographics, clickstream behavior, and so forth, LA has been established (Lemay et al., 2021). In order to create a better learning environment and guarantee that students can complete their degrees on schedule, LA approaches can be used to uncover the elements that contribute to student dropout. Enhancing the standard of student management and the learning environment is the common objective shared by LA and EDM. According to a recent review by Mubarak et al. (2022), there is a growing overlap between these two phrases due to their interchangeable usage, which causes continuous confusion in the field of student retention.

2.4 Factors Contributing to Students' Dropout Rates in Educational Institutions

Understanding the factors that contribute to student dropouts is the first step in preventing student dropouts. The reasons why students leave their studies in higher education are extremely complicated and are influenced by a number of different factors. According to the findings of Mouton et al. (2020) there are a variety of factors that contribute to student dropouts in Germany. In many cases, the cause is a confluence of a number of different elements. Mouton et al. (2020) identified kids who had dropped out of school by using latent class analysis. Students drop out of school for a variety of reasons, including their relationships with study programs or colleges, socioeconomic issues, student performance, academic self-concept, and their intention to drop out of school. Ortiz-Lozano et al. (2018) conducted an investigation on the factors that influence student dropouts in Spain, taking into account socio-demographic and academic information. Although the rationale behind selecting this particular variable is not made entirely clear, the findings of the research indicate that this particular variable does have a major impact.

The prediction study of pupils who choose to drop out of school in Colombia was examined by (Pérez et al., 2018) Students' demographics and their transcript records are the factors that influence the number of students that drop out of school in Colombia. Students who drop out of school are predicted using these characteristics, and the variables that are produced as a result have a considerable impact on students who drop out. Another area of research that Chen et al. (2018) investigated was the forecasting of dropouts in the United States. For the purpose of Chen's research, the variables that were utilized to predict dropout were information regarding high school, demographics, college enrollment, and information regarding each semester. The rationale behind using these factors for prediction is not articulated in a clear and concise manner. Nevertheless, according to the findings of the analysis, the variables that were chosen are able to predict student dropout rates in a significant way. In their study, Troelsen and Laursen (2014) investigated the factors that influence children in Denmark who drop out of school. There are two assumptions that they believe influence kids who drop out of school. Students who drop out of school are influenced by their parents' financial status and educational level, according to the first hypothesis. According to the second hypothesis, students drop out of school as a result of policies implemented by the Danish government regarding education. These regulations cause students to switch to different study programs, switch universities, or not continue their education.

The selection of variables is one of the most important stages in the process of forecasting kids who will drop out of school. This is due to the fact that the variable is the primary construct in a study. According to the findings of the Troelsen and Laursen (2014) research that was discussed earlier, the reasons that influence children to drop out of school vary from country to country. Another piece of evidence that lends credence to this assertion is the research conducted by Troelsen and Laursen (2014), which reveals that nations with different cultures have different perspectives on the importance of education. In light of this, the elements that determine whether or not pupils are successful in their academic endeavors are diverse. Consequently, the factors that influence student dropout are suited to the conditions that exist in the country (Nurmalitasari et al., 2023). Additionally, the factor that effects dropout students does not originate from direct information from dropout students in the research that has been done so far; therefore, the correctness of the factors still needs to be determined.

2.5 Predictive Model for Students' Dropouts in Educational Institutions

An investigation was carried out by Lottering et al. (2020) at a South African University of Technology to identify students who were considered to be at danger. The investigation utilized EDM methodologies and ML. Additionally, they enriched the dataset with 19 variables that were extracted from the database of the school by using the final grades of the students. They carried out under sampling in order to correct the imbalance in the data, which resulted in a dataset consisting of 1.156 records. After that, they utilized a number of different supervised classification methods, with the SVM exhibiting the most impressive performance, attaining an F-Measure score of 99.32%. Providing insights into models with balanced data and highlighting the superior performance of SVM are two ways in which this methodology adds to the ongoing research that is being conducted (Santos et al., 2021). The research that was carried out at a Brazilian institutional of higher education utilized data from seven different qualities in order to develop 10 different data models that corresponded to different semesters. The accuracy of the DT method ranged from 79.31% to 98.25%, with some models earning a perfect score of 100% when it came to predicting dropouts. The accuracy of dropout forecasts, on the other hand, varied from semester to semester, with the model for the third semester demonstrating the lowest accuracy. Some models were able to achieve a recall rate of one hundred percent for graduation predictions, while others achieved a precision ranging from 75.68% to 97.22% for the graduation class (Oliveira and Medeiros, 2024). As the semesters went, the models acquired more properties, which resulted in an increase in their ability to accurately forecast future outcomes. In addition to shedding insight on the relationship between academic achievement and dropout or graduation probability across successive semesters, these findings contribute to the development of dropout prediction models that are based on undergraduate course data. Within the context of a Brazilian higher education institution, Manrique et al. (2019) study analyzes three different techniques to the prediction of school dropouts. The first strategy makes use of general variables that can be applied to any class, whereas the second strategy makes use of data that is specific to the class. Both methods make use of Machine Learning techniques, including Gradient Boosting Tree, Support Vector Machine, Random Forest, and Naive Bayes, among others. Through the utilization of the K-Nearest Neighbors technique, the third method examines the progression of student data throughout the whole of the course duration. The results demonstrate that the model that makes use of global features in conjunction with the RF algorithm obtains the highest levels of accuracy, recall, precision, and F-Measure metrics. This brings to light the exceptional performance of this technique when dealing with generic variables that are not reliant on the course. When compared to other studies in the same field, this work is more advanced, particularly with regard to the data context (Oliveira and Medeiros, 2024). By giving a more in-depth understanding of the variables and patterns that contribute to dropout, predictive research on school dropout that is based on data from a one-of-a-kind undergraduate course self-assessment instrument has the potential to deliver significant scientific gains. This research has the potential to uncover previously unknown connections and risk factors by employing data mining techniques and prediction models. As a result, it can provide useful insights that can be used to build intervention tactics that are more effective (Oliveira and Medeiros, 2024). These scientific discoveries make a contribution to the field of education, perhaps enhancing the retention of students and assisting in the development of ways that are more individualized and evidence-based in order to solve the dilemma of school dropouts.

3. RESEARCH METHODOLOGY

This chapter covers research approaches used to create a prediction model for student retention and dropout. It includes the preprocessing procedures, feature selection, machine learning methods used, and data gathering procedure. Numerous academic, demographic, and socioeconomic characteristics are included in the dataset, including GPA, gender, type of attendance, tuition fee status, and macroeconomic variables like unemployment and inflation.

In this thesis, RF classifiers are used in a supervised learning technique to accurately classify students for being retained to continue, graduate, or drop out. The justification for choosing particular features and models, as well as the validation methods implemented to guarantee the predicted robustness and generalizability of results, are also covered in this chapter.

3.1 Research Design

In this thesis, quantitative and predictive design approaches are applied by using historical data that can help in predicting the percentages of student dropout and early proposed plan for retention. The reason for using these approaches is to utilize data driven decision making and to propose insights that can be actionable through machine learning models.

The selected approaches discussed below focus on the classification algorithm implementations that predict binary values (0 and 1) for dropout and retention. To ensure accuracy and dependability, these dropout and retention models are tested on unseen samples after being trained on labeled data. In order to improve model performance and enable evidence-based educational interventions, the methodology also highlights the significance of feature selection, data preparation, and validation techniques.

3.2 Data Collection

3.2.1 Data Sources

The dataset used in this thesis predicts students' dropout and academic success from **UCI Machine Learning Repository**. The records present in the dataset belong to the students enrolled in higher education institution based in Portugal. The parameters used in the dataset are as follows:

- Academic factors: Grades, Curriculum Status, GPA, Courses, Class Durations.
- Demographic factors: Age, Nationality, Gender, Marital Status, Day or Evening time.
- Socioeconomic factors: Family occupation, Personal Occupation, Tuition, Fees.
- Academic status: Qualification; Undergraduate, Postgraduate, Drop-out, Enrolled.

The selected dataset is available on open-source platform and downloaded from the UCI Machine Learning Repository. The dataset is provided and approved for analysis, research and study purposes due to which it is not required to take consent for processing or usage.

Given that the dataset is anonymized, ethical principles are still applied by making sure the data is used, managed, analyzed and prepared.

3.3 Data Variables and Features

3.3.1 Feature Selection

Since the dataset used in the thesis includes 36 different attributes. There is a probability of having repeated or unnecessary data and to avoid such inconsistencies, feature selection is used. It helps in training model efficiently by selecting the features having strong relevancy with the target and also key predictors are identified using Random Forest feature importance. Furthermore, user input is either averaged or omitted for less significant variables (such as GDP) to streamline the dataset.

3.3.2 Input Features (Predictors)

Numerous academic, demographic, and socioeconomic factors that affect student retention and dropout are included in the dataset utilized in this thesis. Course information, number of enrolled courses, parental education and occupation, application method, admission grade, tuition fee status, debtor status, marital status, prior credentials, age at enrollment, attendance and GPA by semester, and so on are some of the parameters that help us predict student drop-out or retention probability. Each attribute offers important background information for comprehending and forecasting student results. The details of the attributes included in the dataset are represented further into Figure 3.1.

Variable Name	Description
Marital Status	1 – single 2 – married 3 – widower 4 – divorced 5 – facto union 6 – legally separated
Application mode	1 - 1st phase - general contingent 2 - Ordinance No. 612/93 5 - 1st phase - special contingent (Azores Island) 7 - Holders of other higher courses 10 - Ordinance No. 854-B/99 15 - International student (bachelor) 16 - 1st phase - special contingent (Madeira Island) 17 - 2nd phase - general contingent 18 - 3rd phase - general contingent 26 - Ordinance No. 533-A/99, item b2) (Different Plan) 27 - Ordinance No. 533-A/99, item b3 (Other Institution) 39 - Over 23 years old 42 - Transfer 43 - Change of course 44 - Technological specialization diploma holders 51 - Change of institution/course 53 - Short cycle diploma holders 57 - Change of institution/course (International)
Application order	Application order (between 0 - first choice; and 9 last choice)
Course	33 - Biofuel Production Technologies 171 - Animation and Multimedia Design 8014 - Social Service (evening attendance) 9003 - Agronomy 9070 - Communication Design 9085 - Veterinary Nursing 9119 - Informatics Engineering 9130 - Equinculture 9147 - Management 9238 - Social Service 9254 - Tourism 9500 - Nursing 9556 - Oral Hygiene 9670 - Advertising and Marketing Management 9773 - Journalism and Communication 9853 - Basic Education 9991 - Management (evening attendance)
Daytime/evening attendance	1 – daytime 0 - evening
Previous qualification	1 - Secondary education 2 - Higher education - bachelor's degree 3 - Higher education - degree 4 - Higher education - master's 5 - Higher education - doctorate 6 - Frequency of higher education 9 - 12th year of schooling - not completed 10 - 11th year of schooling - not completed 12 - Other - 11th year of schooling 14 - 10th year of schooling 15 - 10th year of schooling - not completed 19 - Basic education 3rd cycle (9th/10th/11th year) or equiv. 38 - Basic education 2nd cycle (6th/7th/8th year) or equiv. 39 - Technological specialization course 40 - Higher education - degree (1st cycle) 42 - Professional higher technical course 43 - Higher education - master (2nd cycle)
Previous qualification (grade)	Grade of previous qualification (between 0 and 200)

Nacionality	1 - Portuguese; 2 - German; 6 - Spanish; 11 - Italian; 13 - Dutch; 14 - English; 17 - Lithuanian; 21 - Angolan; 22 - Cape Verdean; 24 - Guinean; 25 - Mozambican; 26 - Santomean; 32 - Turkish; 41 - Brazilian; 62 - Romanian; 100 - Moldova (Republic of); 101 - Mexican; 103 - Ukrainian; 105 - Russian; 108 - Cuban; 109 - Colombian
Mother's qualification	1 - Secondary Education - 12th Year of Schooling or Eq. 2 - Higher Education - Bachelor's Degree 3 - Higher Education - Degree 4 - Higher Education - Master's 5 - Higher Education - Doctorate 6 - Frequency of Higher Education 9 - 12th Year of Schooling - Not Completed 10 - 11th Year of Schooling - Not Completed 11 - 7th Year (Old) 12 - Other - 11th Year of Schooling 14 - 10th Year of Schooling 18 - General commerce course 19 - Basic Education 3rd Cycle (9th/10th/11th Year) or Equiv. 22 - Technical-professional course 26 - 7th year of schooling 27 - 2nd cycle of the general high school course 29 - 9th Year of Schooling - Not Completed 30 - 8th year of schooling 34 - Unknown 35 - Can't read or write 36 - Can read without having a 4th year of schooling 37 - Basic education 1st cycle (4th/5th year) or equiv. 38 - Basic Education 2nd Cycle (6th/7th/8th Year) or Equiv. 39 - Technological specialization course 40 - Higher education - degree (1st cycle) 41 - Specialized higher studies course 42 - Professional higher technical course 43 - Higher Education - Master (2nd cycle) 44 - Higher Education - Doctorate (3rd cycle)

Father's qualification	1 - Secondary Education - 12th Year of Schooling or Eq. 2 - Higher Education - Bachelor's Degree 3 - Higher Education - Degree 4 - Higher Education - Master's 5 - Higher Education - Doctorate 6 - Frequency of Higher Education 9 - 12th Year of Schooling - Not Completed 10 - 11th Year of Schooling - Not Completed 11 - 7th Year (Old) 12 - Other - 11th Year of Schooling 13 - 2nd year complementary high school course 14 - 10th Year of Schooling 18 - General commerce course 19 - Basic Education 3rd Cycle (9th/10th/11th Year) or Equiv. 20 - Complementary High School Course 22 - Technical-professional course 25 - Complementary High School Course - not concluded 26 - 7th year of schooling 27 - 2nd cycle of the general high school course 29 - 9th Year of Schooling - Not Completed 30 - 8th year of schooling 31 - General Course of Administration and Commerce 33 - Supplementary Accounting and Administration 34 - Unknown 35 - Can't read or write 36 - Can read without having a 4th year of schooling 37 - Basic education 1st cycle (4th/5th year) or equiv. 38 - Basic Education 2nd Cycle (6th/7th/8th Year) or Equiv. 39 - Technological specialization course 40 - Higher education - degree (1st cycle) 41 - Specialized higher studies course 42 - Professional higher technical course 43 - Higher Education - Master (2nd cycle) 44 - Higher Education - Doctorate (3rd cycle)

Mother's occupation	0 - Student 1 - Representatives of the Legislative Power and Executive Bodies, Directors, Directors and Executive Managers 2 - Specialists in Intellectual and Scientific Activities 3 - Intermediate Level Technicians and Professions 4 - Administrative staff 5 - Personal Services, Security and Safety Workers and Sellers 6 - Farmers and Skilled Workers in Agriculture, Fisheries and Forestry 7 - Skilled Workers in Industry, Construction and Craftsmen 8 - Installation and Machine Operators and Assembly Workers 9 - Unskilled Workers 10 - Armed Forces Professions 90 - Other Situation 99 - (blank) 122 - Health professionals 123 - teachers 125 - Specialists in information and communication technologies (ICT) 131 - Intermediate level science and engineering technicians and professions 132 - Technicians and professionals, of intermediate level of health 134 - Intermediate level technicians from legal, social, sports, cultural and similar services 141 - Office workers, secretaries in general and data processing operators 143 - Data, accounting, statistical, financial services and registry-related operators 144 - Other administrative support staff 151 - personal service workers 152 - sellers 153 - Personal care workers and the like 171 - Skilled construction workers and the like, except electricians 173 - Skilled workers in printing, precision instrument manufacturing, jewelers, artisans and the like 175 - Workers in food processing, woodworking, clothing and other industries and crafts 191 - cleaning workers 192 - Unskilled workers in agriculture, animal production, fisheries and forestry 193 - Unskilled workers in extractive industry, construction, manufacturing and transport 194 - Meal preparation assistants

Father's occupation (Part 1)	0 - Student 1 - Representatives of the Legislative Power and Executive Bodies, Directors, Directors and Executive Managers 2 - Specialists in Intellectual and Scientific Activities 3 - Intermediate Level Technicians and Professions 4 - Administrative staff 5 - Personal Services, Security and Safety Workers and Sellers 6 - Farmers and Skilled Workers in Agriculture, Fisheries and Forestry 7 - Skilled Workers in Industry, Construction and Craftsmen 8 - Installation and Machine Operators and Assembly Workers 9 - Unskilled Workers 10 - Armed Forces Professions 90 - Other Situation 99 - (blank) 101 - Armed Forces Officers 102 - Armed Forces Sergeants 103 - Other Armed Forces personnel 112 - Directors of administrative and commercial services 114 - Hotel, catering, trade and other services directors 121 - Specialists in the physical sciences, mathematics, engineering and related techniques 122 - Health professionals 123 - teachers 124 - Specialists in finance, accounting, administrative organization, public and commercial relations 131 - Intermediate level science and engineering technicians and professions 132 - Technicians and professionals, of intermediate level of health 134 - Intermediate level technicians from legal, social, sports, cultural and similar services

Father's occupation (Part 2)	135 - Information and communication technology technicians 141 - Office workers, secretaries in general and data processing operators 143 - Data, accounting, statistical, financial services and registry-related operators 144 - Other administrative support staff 151 - personal service workers 152 - sellers 153 - Personal care workers and the like 154 - Protection and security services personnel 161 - Market-oriented farmers and skilled agricultural and animal production workers 163 - Farmers, livestock keepers, fishermen, hunters and gatherers, subsistence 171 - Skilled construction workers and the like, except electricians 172 - Skilled workers in metallurgy, metalworking and similar 174 - Skilled workers in electricity and electronics 175 - Workers in food processing, woodworking, clothing and other industries and crafts 181 - Fixed plant and machine operators 182 - assembly workers 183 - Vehicle drivers and mobile equipment operators 192 - Unskilled workers in agriculture, animal production, fisheries and forestry 193 - Unskilled workers in extractive industry, construction, manufacturing and transport 194 - Meal preparation assistants 195 - Street vendors (except food) and street service providers
Admission grade	Admission grade (between 0 and 200)
Displaced	1 – yes 0 – no
Educational special	1 – yes 0 – no
Debtor	1 – yes 0 – no

Tuition fees up to date	1 – yes 0 – no
Gender	1 – male 0 – female
Scholarship holder	1 – yes 0 – no
Age at enrollment	Age of studend at enrollment
International	1 – yes 0 – no
Curricular units 1st sem	Number of curricular units credited in the 1st semester
Curricular units 1st sem	Number of curricular units enrolled in the 1st semester
Curricular units 1st sem	Number of evaluations to curricular units in the 1st semester
Curricular units 1st sem	Number of curricular units approved in the 1st semester

Curricular units 1st sem (grade)	Grade average in the 1st semester (between 0 and 20)
Curricular units 1st sem (without evaluations)	Number of curricular units without evalutions in the 1st semester
Curricular units 2nd sem (credited)	Number of curricular units credited in the 2nd semester
Curricular units 2nd sem (enrolled)	Number of curricular units enrolled in the 2nd semester
Curricular units 2nd sem (evaluations)	Number of evaluations to curricular units in the 2nd semester
Curricular units 2nd sem (approved)	Number of curricular units approved in the 2nd semester
Curricular units 2nd sem (grade)	Grade average in the 2nd semester (between 0 and 20)
Curricular units 2nd sem (without evaluations)	Number of curricular units without evalutions in the 1st semester

Unemployment rate	Unemployment rate (%)
Inflation rate	Inflation rate (%)
GDP	GDP
Target	Target. The problem is formulated as a three category classification task (dropout, enrolled, and graduate) at the end of the normal duration of the course

Figure 3.1 Dataset Attributes with Description

3.3.2.1 Gender

This attribute indicates the student's gender, which can be classified as either male or female as displayed in Figure 3.1. Because of observed behavioral, academic, or sociocultural variations in educational results, gender is thought to be a potential factor impacting retention and dropout trends. The dataset has records including both Male (1556) and Female (2869) candidates and their division is represented in Figure 3.2.

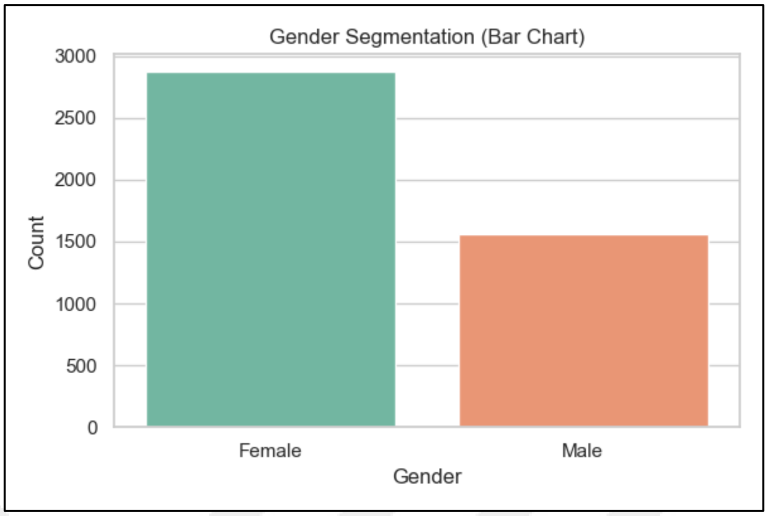

Figure 3.2 Female against Male segregation

3.3.2.2 Class Timing (Evening/Day Attendance)

This attribute indicates if a student attends classes in the evening or during the day. Students' external obligations (such as employment or family) may be reflected in the schedule of their classes, which can affect their degree of participation and likelihood of dropping out. The purpose of its inclusion is to determine whether scheduling flexibility and student retention are related. Figure 3.3 displays the number of students taking evening and day time classes.

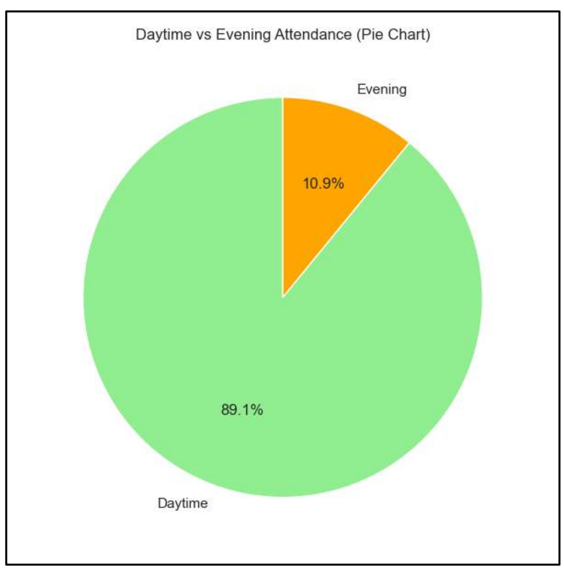

Figure 3.3 Percentage of students attending Day time vs Evening Classes

3.3.2.3 Grade: Admission Grade (GPA)

This attribute represents the student's entry-level academic performance, or GPA, usually from high school or before attending college. It helps in forecasting academic readiness of the student. Figure 3.4 represents the distribution of students based on the grade points achieved by them.

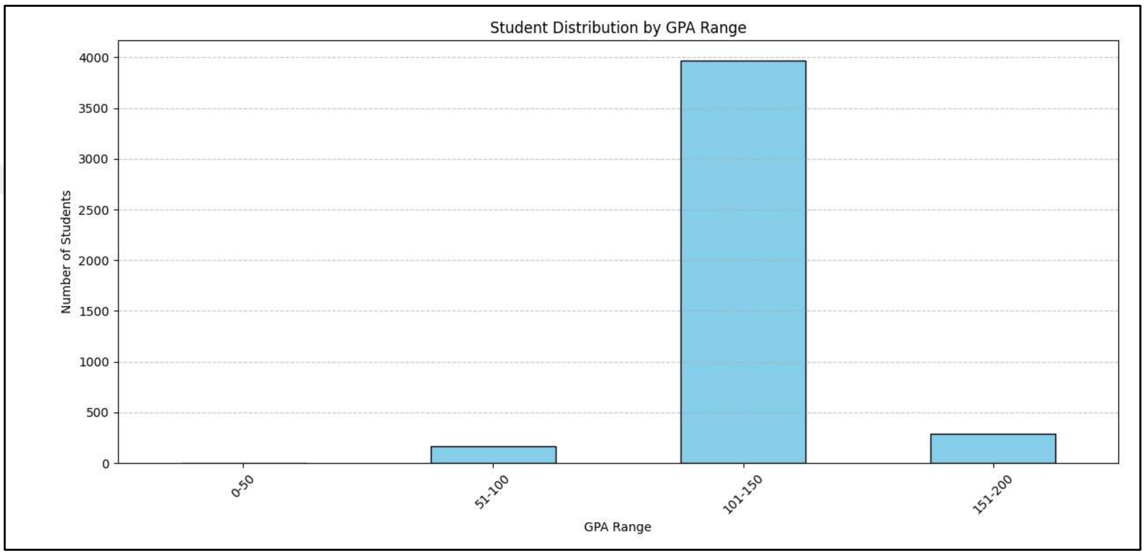

Figure 3.4 Student Distribution by GPA Score Range

3.3.2.4 Paid Tuition Fees

Paid tuition fees represented in binary value (0 or 1) indicates if the student has made timely and complete tuition payments, in Figure 3.5 it can be seen that most of the students from the dataset have been successfully paying the tuition fees. Payment patterns can be an indication of course commitment and financial stability.

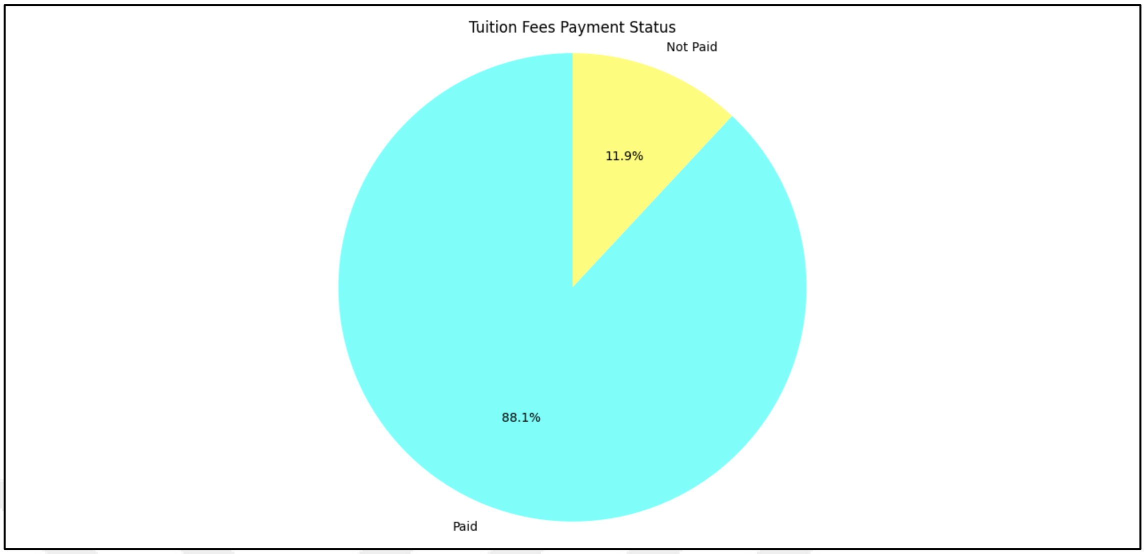

Figure 3.5 Tuition Fees Payment Status: Paid or Unpaid

3.3.2.5 Unemployment Rate

The percentage of the labor force that is unemployed and actively looking for work during a specific time period is represented by the unemployment rate, a macroeconomic attribute displayed in Figure 3.6. Examining how outside economic forces might affect a student's choice to continue or stop their education plays an important role in this thesis. Due to financial burden, high unemployment may prevent students from continuing, or on the other hand, it may promote retention as they postpone entering a weak labor market.

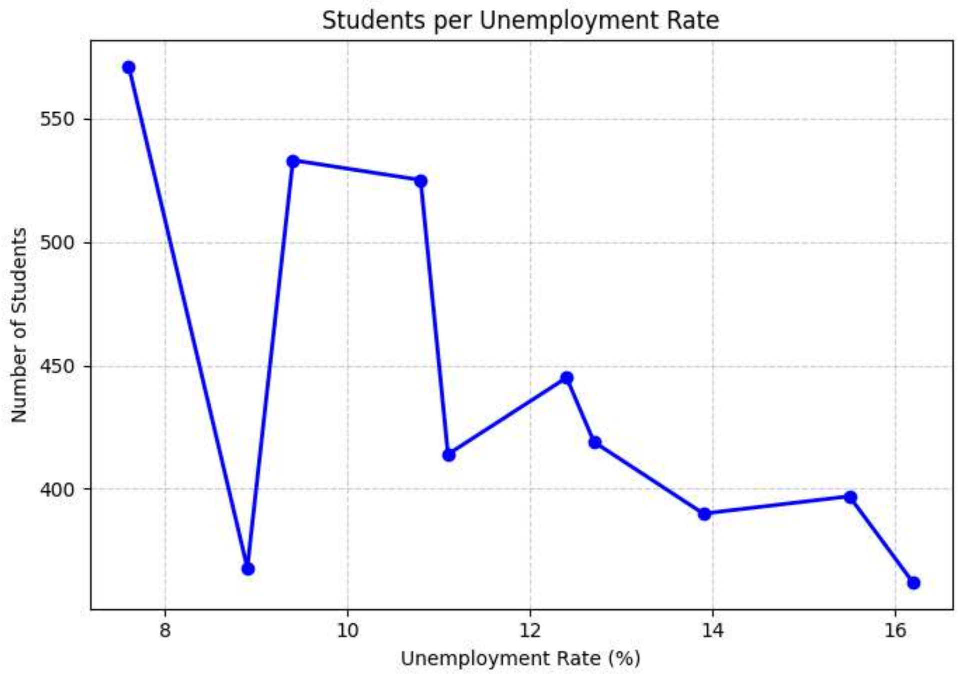

Figure 3.6 Unemployment Rate Trend

3.3.2.6 Inflation Rate

The overall increase in prices over time that lowers buying power is measured by the inflation rate and is represented in Figure 3.7. As an independent variable in this thesis, it supports in evaluating the wider economic strain on households that can have an impact on a student's capacity to pay for living expenses, tuition, or supplies, which may lead to dropout decisions.

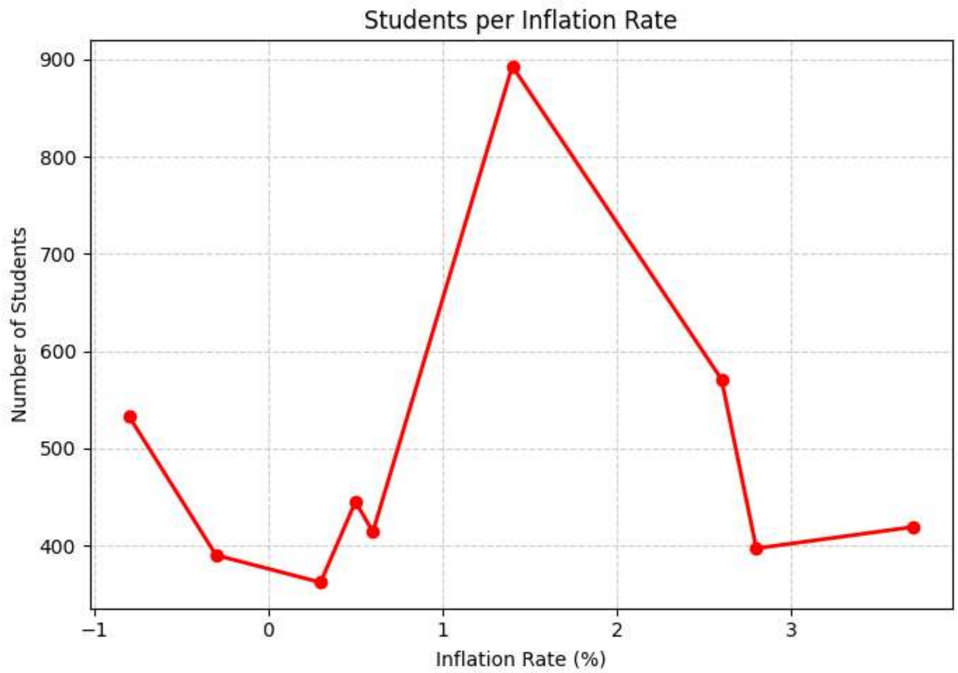

Figure 3.7 Inflation Rate against Students

3.3.2.7 Age at the Time of Enrollment

Age of the student at the start of the present course of study plays an important role in the prediction as it indicates maturity, outside obligations, or non-traditional student. The students' age mentioned in the dataset is further classified into age group segmentation in Figure 3.8 and a line graph showing the age distribution of students in Figure 3.9.

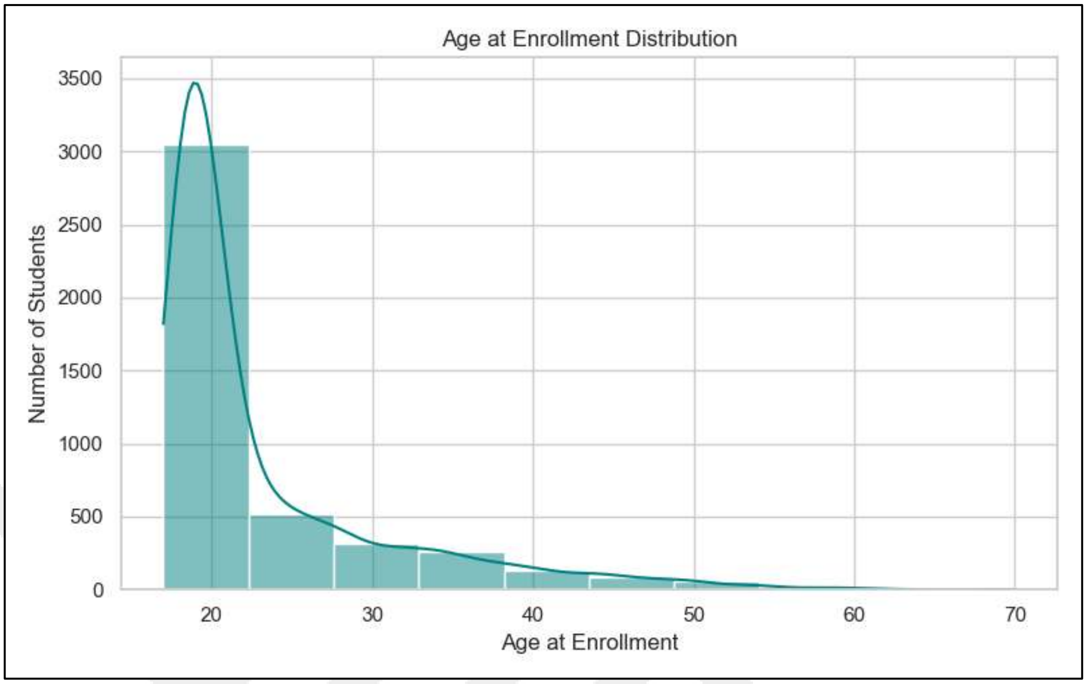

Figure 3.8 Age of Enrollment Distribution

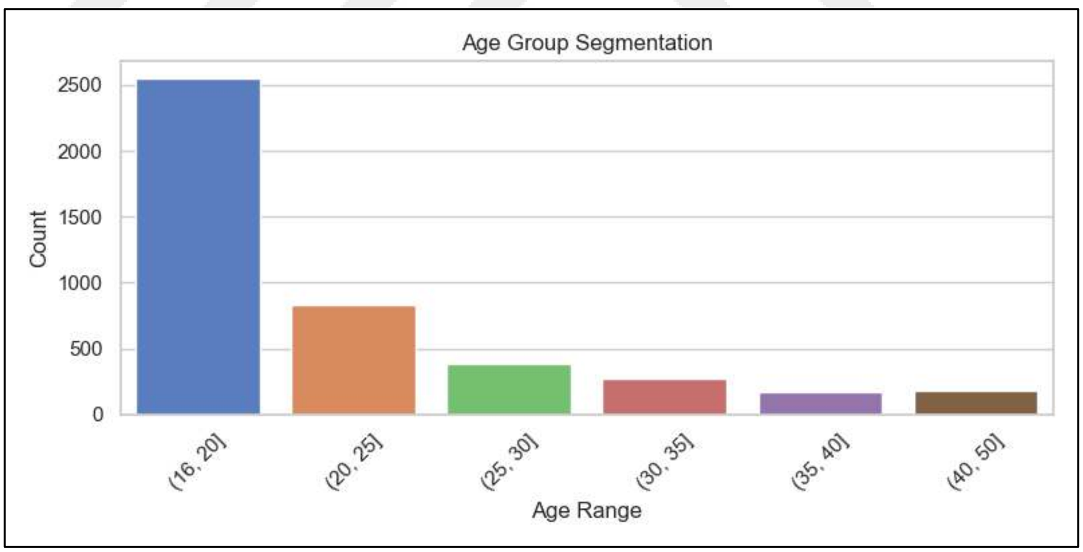

Figure 3.9 Age Group Visualization and Segmentation

3.3.2.8 Attendance and GPA (per Semester Records)

GPA and other semester-specific measures (e.g., grades, evaluations, and passed credits) show continuous academic success and are reliable predictors of retention or dropout risk represented in Figure 3.10.

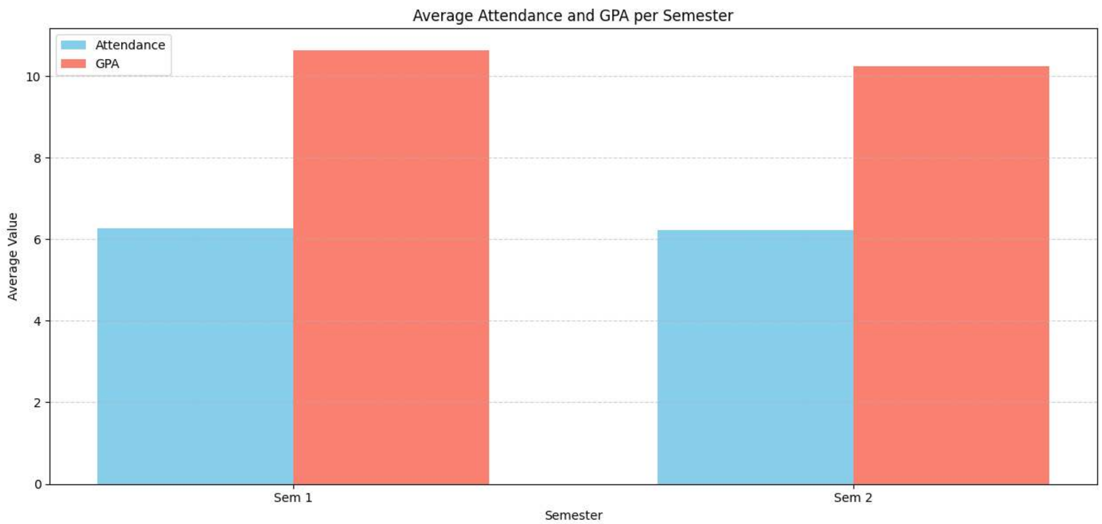

Figure 3.10 Attendance and GPA Average Ratio for 2 Semesters

3.3.2.9 Course Details and Number of Enrolled Courses

This attribute represents the number of classes the student is registered for each semester as well as the academic program or major they are enrolled in (such as engineering, law, or economics).

3.3.3 Weightage of Input Features

The input features are ranked on the graph according to how well they predict student dropout. Figure 3.11 represents the contribution of each attribute to the decision making of drop out prediction model. The influence of attributes on the model is important and critical to understand as it helps in identifying the gaps to fill and finding areas where strategies can be applied to intervene and reduce drop out ratio.

Features like "Curricular units 2nd sem (approved)", "Curricular units 2nd sem (grade)", and "Curricular units 1st sem (approved)" are of the most relevance, suggesting that dropout is significantly predicted by students' academic progress and performance.

Age upon enrollment, acceptance grade, and tuition fee payment status are other significant factors that emphasize the importance of both academic and financial readiness. Furthermore, background and demographic factors including gender, nationality, and special education needs are comparatively less significant in the prediction. Which qualities are most useful for early detection and intervention measures can be determined with the use of this image.

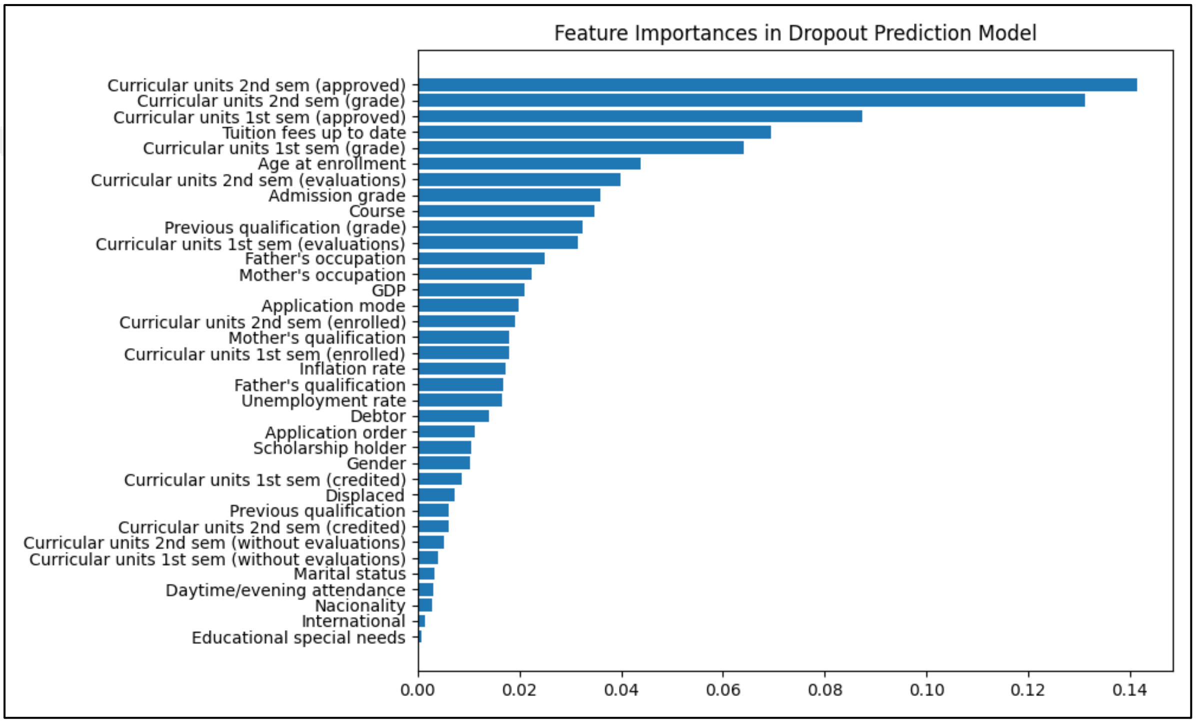

Figure 3.11 Weightage of Input Features

3.3.4 Data Preprocessing

Before processing data, a number of steps are carried out for optimization, compatibility, analysis and performance of the chosen model. These steps are given as follows:

3.3.5 Missing Data – Handling

In the given dataset, records have missing numerous values that cause the algorithm to give errors. Missing, null, mismatched and empty values are handled for independent and dependent variables. The average (mean) value of the corresponding columns is used to refer numerical features like GPA, admission grade, and unemployment/inflation rates while maintaining the data distribution.

In order to maintain statistic consistency throughout dataset the column's average is added for missing values, for example if a row in GPA column has missing data it is replaced with the average of the overall GPA column value (i.e. 150.4). Furthermore, if unemployment rate has a missing value, the mean of employment rate (i.e. 20.7%) is added to avoid high values in the respective column, that may create biasness and effect the results.

In order for the model to interpret categorical or binary-type attributes like tuition fee status, debtor status, and scholarship holding effectively and without bias or distortion, missing variables are substituted with default binary values (e.g., 0 for No, 1 for Yes). For example, Debtor column's default value is set as 0 (No) therefore if there is any value missing in the column, the value is replaced with 0 whereas for Tuition fees it is set as 1 (Yes), replacing all the missing values in Tuition fees column as 1. This process helps the model to prevent misinterpretation of data and ensures encoding process to be consistent.

3.3.6 Feature Scaling

StandardScaler is used to scale numerical features. By eliminating the mean and scaling to unit variance, the StandardScaler standardizes numerical features, essentially converting the data to have a mean of 0 and a standard deviation of 1. By doing this, bias toward features with higher values is avoided and features are guaranteed to contribute evenly to the model.

3.3.7 Target Engineering

The process to establish and transform the output variables (target variables) of a dataset to make it capable of training a machine learning model is referred to as target engineering. The aim in this thesis is to identify two different but related outcomes (retention and dropout). To increase the probability of getting targeted and accurate predictions, classification models are used, and target variables are allocated in two different binary columns (0 or 1).

The prediction is made simpler using binary encoding by training two different machine learning models: target retention and target dropout. Target retention is added to highlight whether students have been successfully retained or dropped out of the institute, where ‘1’ (true) refers to a student who is enrolled or graduated and ‘0’ (false) refers to dropout. On the other hand, the target dropout column refers to the column that highlights whether a student has dropped out; here, 1 (true) refers to the student who has dropped out, and 0 (false) refers to the student who is successfully enrolled or graduated.

The approach used in this thesis for target engineering encourages and helps to analyze different cases of retention and dropouts which enable decision makers or institutes to plan preventive strategies and provide better interventions.

3.3.8 Feature Selection

One of the critical steps in the machine learning model training process is feature selection, which identifies the selection of the most relevant variables from the dataset. The use of feature selection in this thesis is to improve performance of the model, reduce complexity, and to guarantee that model explicitly concentrates on unique attributes. Given the aim of predicting dropout and retention of students, features were selected from the dataset based on their effect on students’s educational process and journey. These features are categorized under 2 factors: academic and socioeconomic background attributes.
Academic attributes help in identifying how well students are handling their education, including admission grade, curricular units enrolled, curricular units approved, and GPA. Whereas, socioeconomic attributes such as gender, marital status, scholarship holder, parental occupation, age at enrollment, tuition fees up to date, debtor, and parental qualification give understanding into the financial and social status of the student. Inflation and unemployment rates are also used as external factors to identify how other economic factors affect the decision of students in continuing or dropping the education.

3.3.9 Categorical Encoding

Features such as gender, attendance, etc. are encoded using label encoding and one-hot encoding. One-hot encoding generates binary columns for each category to indicate their existence, and on the other hand label encoding gives each category in a feature a distinct number.

These encoding approaches are used in the thesis to transform categorical variables, such as attendance type, marital status, and gender, into numerical representations that are used to train machine learning models that predict student retention and dropout rates.

3.3.10 Data Splitting

A reliable machine learning model is implemented by dividing the dataset into training, validation, and testing. This process guarantees that the model is trained, validated, and evaluated accurately. 70% of the portion of data is used for training the model, which helps in learning patterns and connections between input features and the target outcomes. 15% is used during training to optimize hyperparameters and standardize parameters of the model. This phase prevents overfitting, making sure the performance of the model is effective on both trained and unseen data.

After training and validation of the model, testing takes place where the remaining 15% is used. The testing subset helps in identifying the final performance of the model. This method of dataset splitting ensures the accuracy of model development and helps to prevent data leakage. It also enhances the reliability of results, enhancing the robustness to use the model in future predictions and applications.

3.4 Predictive Modeling Techniques

RFC model is implemented to train the predictive model it is an ensemble learning system which creates several decision trees during training and outputs the mode, or most common outcome, of each tree's predictions. It is renowned for its excellent precision, robustness, and capacity to manage both categorical and numerical data, as well as missing values to a certain degree.

RFC has been widely employed in comparable studies pertaining to dropout prediction and student retention because of its efficacy in categorization challenges (Agrusti et al., 2019); (Mubarak et al., 2022). Using characteristics including grades, attendance, socioeconomic condition, and demographic data, it is used to find the student who is at risk in educational datasets. For instance, RFC have been used in previous research to determine the best predictors of dropout (such as GPA or tuition payment status) and rank feature importance (Dekker et al., 2009); (Aulck et al., 2016); ((Gonzalez-Nucamendi, et al., 2023).

In comparison with SVM and LR, RFC has consistently had higher performance in EDM tasks due to its readability via feature scoring (Hashim et al., 2020); (Pérez, 2018). Developing comprehensible models is essential as it helps educational establishments to make data-driven choices about early interventions. By identifying trends in past student data, Random Forest aids in the prediction of binary outcomes (retention or dropout) in the thesis, which makes it appropriate for the developed predictive framework.

3.4.1 Retention Model

The Retention Model is constructed using the RFC to forecast the probability of students remaining enrolled. As an ensemble approach, RFC builds several decision trees and combines their results to increase the accuracy and stability of predictions. This technique fits appropriately with datasets related to educational backgrounds having various attributes since it reduces the probability of overfitting and ensures numerical and categorical attributes are managed efficiently.

The target label is established so that '1' denoted a retained student and '0' denoted otherwise, and the model is trained using selective attributes from the dataset. As a result, the model is able to identify trends and traits that are frequently linked to students who are likely to pursue further education. The RFC model helps in identifying influential attributes that possess most effect on the student retention. These analyses are critically important for the institutions to not just implement early predictions and warning systems but also prepare intervention strategies that can allow them to have visibility of student's who are at high risk of dropping out. In general retention model based on RFC provides understandable predictive framework which helps in data driven decisions in educational institutes.

3.4.2 Dropout Model

To identify students who are at danger of dropping out, a distinct Dropout Model is created using the RFC. This model focuses on the prediction of negative academic outcomes based on previous student data which makes it critical and significant part of the thesis. However, the aim of thesis is to develop model that can help in identifying and enhancing student retention but developing a dropout model ensures an approach that help in understanding the socio economic, personal or educational factors that might encourage students to dropout or discontinue their study.

The dropout model is trained using similar preprocessed dataset as retention model which includes academic, parental occupation and education, socioeconomic, demographic and performance attributes. For dropout model the target label has been changed in a way that ‘1’ indicates dropped out students and ‘0’ indicates enrolled students. Due to the label inversion, RFC refines the learning process to predict dropouts instead of retentions. The model becomes specialized in identifying risk factors and patterns that usually result in discontinuation by being trained independently using dropout-focused labeling which makes it a valuable tool for early intervention planning.

3.4.3 Tools Used

Data preparation, binary target development, data splitting, and reliable model training with interpretable machine learning approaches are all incorporated into the methodology. The models' generalizability and dependability are guaranteed by validation using the following tools and libraries.

Due to the adaptability of Python language and framework, this language has been used in the model implementation. The framework is used in this thesis, since it provides wide range of data science libraries, data preprocessing methods and GUI implementation which allows ease of use and readability for the readers.

Several libraries are used for different stages to design, create and evaluate the model. For data analysis and manipulation Pandas and NumPy are used and to create the machine learning model and evaluation tools, scikit-learn is utilized. On the other hand, Joblib is used for model serialization and to develop GUI for prediction of retain and drop-out, Tkinter is used.

The implementation is done in a virtual environment with Visual Studio Code to effectively handle dependencies. This platform ensures consistency, and easier debugging throughout the development of the thesis project.

3.5 Evaluation Metrics

The performance of the predictive model is determined using the evaluation metrics. Accuracy is given less weight than precision, recall, and F1-score due to the class imbalance common in dropout prediction. Model performances in differentiating between dropout and non-dropout instances are compared using AUC-ROC. These metrics are used to evaluate the effectiveness and assessment of the model, especially in the datasets where dropout ratio is higher than retention ratio or vice versa.

These uneven datasets can cause accuracy to be incorrect because if the dataset has higher number of retained students the model may show higher accuracy by predicting most of the students as retained while it may fail to correctly predict students who are high risk of dropping out students. Due to these factors recall, precision and f1 score are used to get in dept insights of the model's prediction.

All these metrics are used individually on both retention and dropout model, which are trained by RFC. To ensure efficiency and scalability of these metrics a small and reserved test data set is evaluated. Using these targeted metrics, it can be guaranteed that models not just showcase statistical and figurative data but also, they can be practically used in institutions for making educational decisions.

3.5.1 Accuracy

The predictive model's accuracy is determined by the proportion of accurate predictions it makes about student outcomes. Based on this model it can be determined whether a student will remain in school or drop out. One essential evaluation statistic that shows the classification model's overall efficiency is accuracy.

To get the percentage of accurate predictions:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

is used, where TP (True Positives) corresponds to the number of students who are accurately predicted to be retained, TN (True Negatives) corresponds to the number of students who are accurately predicted to drop out, FP (False Positives) corresponds to the number of students who are inaccurately predicted as drop-out who are retained (incorrectly predicted retention as drop-out) and FN (False Negatives) corresponds to the number of students who are inaccurately predicted as retained who are actually drop outs (incorrectly predicted drop-out as retention).

In the context of educational analytics, where decisions may have a direct influence on students' academic paths, a higher accuracy means the model is making more accurate predictions and fewer mistakes.

3.5.2 Precision

Precision is utilized in this thesis to measure the proportion of cases that are actually positive, or forecasted as positive (dropped out or retained), in order to evaluate the accuracy of positive predictions. Finding students who are either retained or at risk of dropping out helps assess the model's reliability. When the cost of FP is high—for example, incorrect identification of a dropout risk, precision becomes critically important because it can prevent the correct people from receiving the help or action they need.

Precision can be calculated for both retention and dropout. The precision for the retention class determines the percentage of students that are retained against the students who are expected to be retained. It is calculated and it assesses how well the model predicts which students will remain enrolled.

In Equation (2), TP represents the number of students who are accurately predicted and are retained. Whereas FP represents the number of students who dropped out but are mistakenly anticipated to be retained.

Equation (2) and assesses how well the model predicts which students will remain enrolled.

$$Precision(Retention) = \frac{TP}{TP + FP} \tag{2}$$

3.5.2.1 Precision for Drop-out:

Likewise, Precision for Drop-out assesses the percentage of anticipated dropouts that actually did drop out. This metric is crucial for determining how effectively the model separates students who are actually in danger or at risk.

To calculate precision for drop-out equation (3) is used where TN represents the number of students accurately predicted as drop-outs and FN represents the number of students who are retained but predicted as drop-outs.

$$Precision(Drop-out) = \frac{TN}{TN + FN} \tag{3}$$

This thesis guarantees that the model not only detects student outcomes but also does so with a level of certainty that makes the predictions actionable. It also ensures reliable predictions by analyzing precision across both classes.

3.5.3 Sensitivity (Recall)

Recall or sensitivity is used to assess the ability of model to identify all true dropout students. The process of identifying all at risk students or most likely to continue is a critical stage in our thesis as it specifies the filtration of true positives (students who are predicted as dropouts) out of actual positives (students who actually dropped out).

High recall guarantees that at-risk students are not missed, which is essential for institutional support programs given the significance of early intervention in education. It is especially helpful in situations where failing to identify a student who would drop out, for example, when missing a positive case (false negative) is more harmful than incorrectly flagging a case (false positive).

The ability of the model to accurately identify students who are retained out of all those who are actually retained is known as recall for retention. To calculate the recall for retention Equation (4) is used and it guarantees that the majority of students who will remain enrolled are identified.

$$Recall(Retention) = \frac{TP}{TP + FN} \tag{4}$$

In Equation (4) TP corresponds to the number of students predicted as retained and are retained and FN corresponds to the number of students who are predicted to drop out but are instead retained.

On the other hand, recall for drop-out determines how well the model predicts which students will really drop out. It enables organizations to actively interact with students who are most in need of assistance. It can be calculated with:

$$Recall(Drop - out) = \frac{TN}{TN + FP} \tag{5}$$

where TN corresponds to students accurately identified as dropouts, and FP corresponds to students who are predicted as retained but in reality, are dropouts.

A higher recall value in both categories suggests that the model is successful in reducing missed cases, which is a crucial step for putting preventative measures into practice for student retention.

3.5.4 F1-Score

The harmonic mean of precision and recall is used to get the F1-score. This penalizes significant differences between the two indicators and guarantees that they are given equal weight. A high F1-score shows that the model does an accurate job of correct identification and prediction. The F1-score is used in this thesis as a balanced statistic that sums recall and precision into a single value.

It is a crucial metric for assessing the model's overall performance because it works especially well in situations where there is a class imbalance, such as when there are more retained students than dropouts. By taking into account both the accuracy of positive predictions, precision and the capacity to detect every positive instance, the F1-score offers a more thorough perspective - recall.

By balancing FP and FN, the retention F1-score measures effectively the model predicts students who are likely to stay enrolled. To calculate the retention F1-score:

$$F1 - Score(Retention) = 2 \times \frac{Precision(Retention) \times Recall(Retention)}{Precision(Retention) + Recall(Retention)} \tag{6}$$

is used where precision (retention) refers to how many of the predicted retained students are actually retained and recall (retention): It refers to How many of the students who are actually retained are accurately predicted.

The model's ability to identify students who are likely to drop-out is indicated by the drop-out F1-score. This score is particularly useful for guaranteeing prompt intervention which is calculated by:

$$F1 - Score(\mathrm{Drop-out}) = 2 \times \frac{Precision(Drop-out) \times Recall(Drop-out)}{Precision(Drop-out) + Recall(Drop-out)} \quad (7)$$

where precision (drop-out) is the percentage of actual dropouts that are predicted and recall (drop-out) measures the percentage of real dropouts that are accurately identified.

In this thesis, the use of F1-score guarantees an equitable assessment of model performance by taking into consideration both over- and under-prediction, which is essential for projecting student outcomes.

3.5.5 AUC-ROC

The area under the Receiver Operating Characteristic (ROC) curve is used for binary classification of models. It helps in assessing how effectively the model differentiates classes given various thresholds by plotting true positives against false positives. Datasets often include details for retained students more than dropped out students leads to inaccurate outcomes, but ROC curve gives balanced view of model’s performance for all thresholds.

This technique enables educational institutes to select priorities; for example, institute may prioritize identification of all possible dropouts even if it inaccurately includes false positives (students who will be retained but are predicted as possible dropouts).

The model's capacity to differentiate between students who are likely to drop out and those who are retained is demonstrated by the AUC-ROC Curve (Figure 3.12). The curve displays the trade-off between specificity and sensitivity (recall).

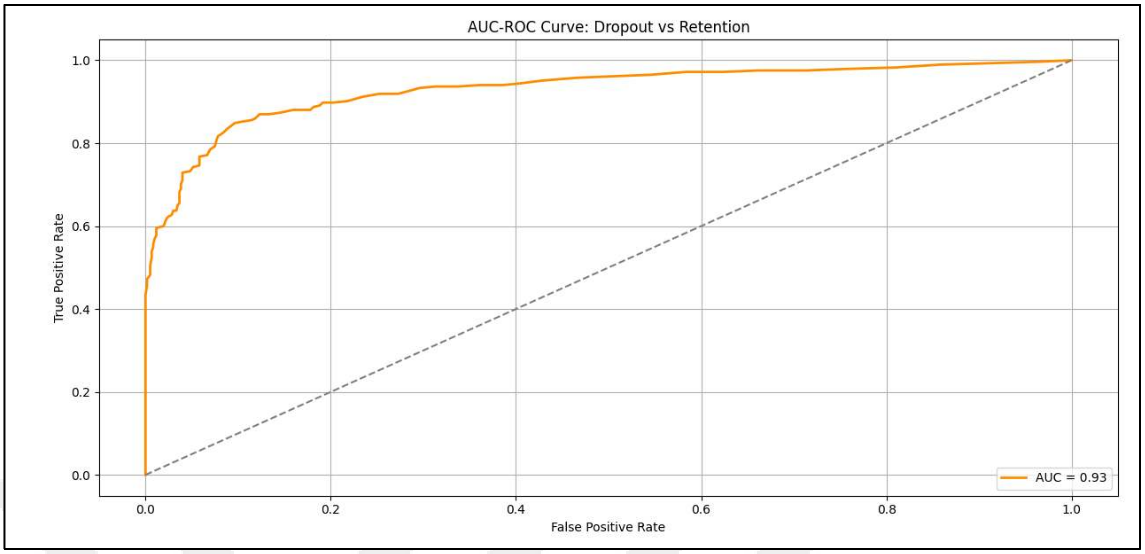

Figure 3.12 AUC-ROC Curve: Dropout against Retention

Plots of the Receiver Operating Characteristic (ROC) curve: In Figure 3.9 False Positive Rate (FPR) is the proportion of retained students who were mistakenly predicted as drop out is the X-axis. While at Y-axis, True Positive Rate (TPR) is the proportion of real dropouts who were accurately classified as dropouts. A distinct threshold value is used to categorize dropout probability which is represented by each point on the curve.

Curve Interpretation: The curve begins at (0,0) and climbs sharply in the direction of the upper-left corner, which is the ideal point at (0,1). In the majority of thresholds, the model maintains low FPR while achieving high TPR, which means many dropouts are appropriately identified by it and retained students are rarely mislabeled as dropouts.

The area under the curve is ‘0.93’ which is extremely high. This indicates that the model has a 93% chance of accurately differentiating between a retained student and a randomly selected dropout. The scale of interpretation is presented in Table 3.1.

Table 3.1 AUC Score Implementation

AUC Score	Interpretation
0.90–1.0	Outstanding model
0.80–0.90	Good Model
0.70–0.80	Moderate Model
0.60–0.70	Inadequate Model
0.50–0.0	Below Standard

The dropout prediction model has outstanding discriminative power (AUC = 0.93) in differentiating between students who are expected to drop out and those who are kept, according to the AUC-ROC curve. Table 3.1 demonstrates how well the approach works to identify at-risk students early on, which can help with proactive intervention techniques.

3.6 Hypotheses

The hypotheses are as follows:

H1: There is an impact of academic challenges on students' dropout rates in educational institutions.

H2: There is an impact of socio-economic factors on students' dropout rates in educational institutions.

H3: There is an impact of institutional issues on students' dropout rates in educational institutions.

H4: There is an impact of personal and psychological factors on students' dropout rates in educational institutions.

4. IMPLEMENTATION AND FINDINGS

This chapter covers predictive models created to determine which students are likely to be retained and which are at risk of dropping out. The performance results and a demonstration of the GUI-based prediction system are also covered to provide practical insights on the implementation.

4.1 Model Implementation

A set of sequential steps are taken to design and execute the student retention and drop-out prediction model which is explained in the Workflow Diagram and Data flow Diagram as illustrates in Figure 4.1 and Figure 4.2, respectively.

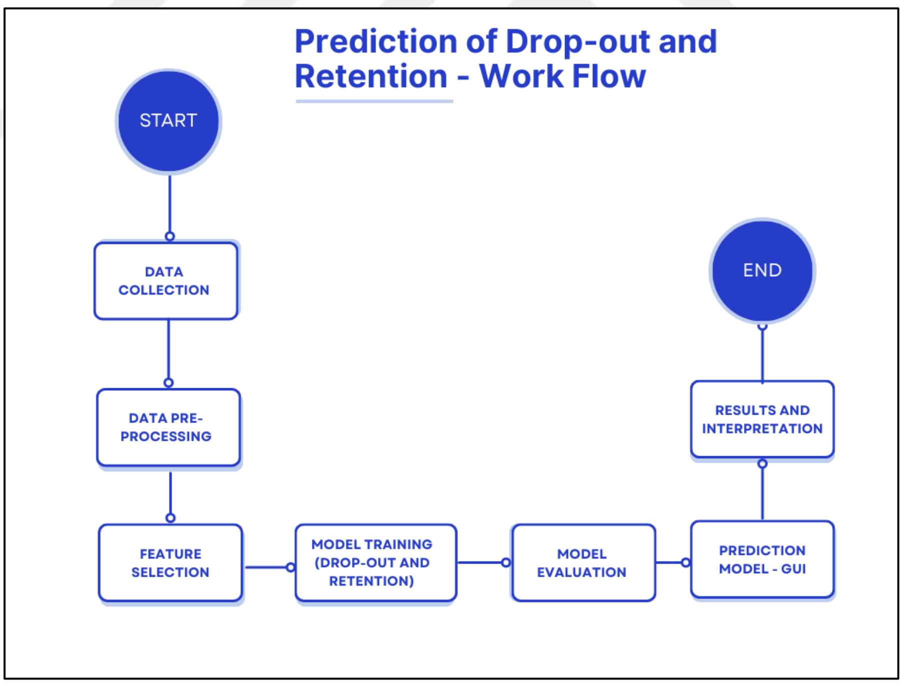

Figure 4.1 Prediction of Drop-out and Retention - Work Flow

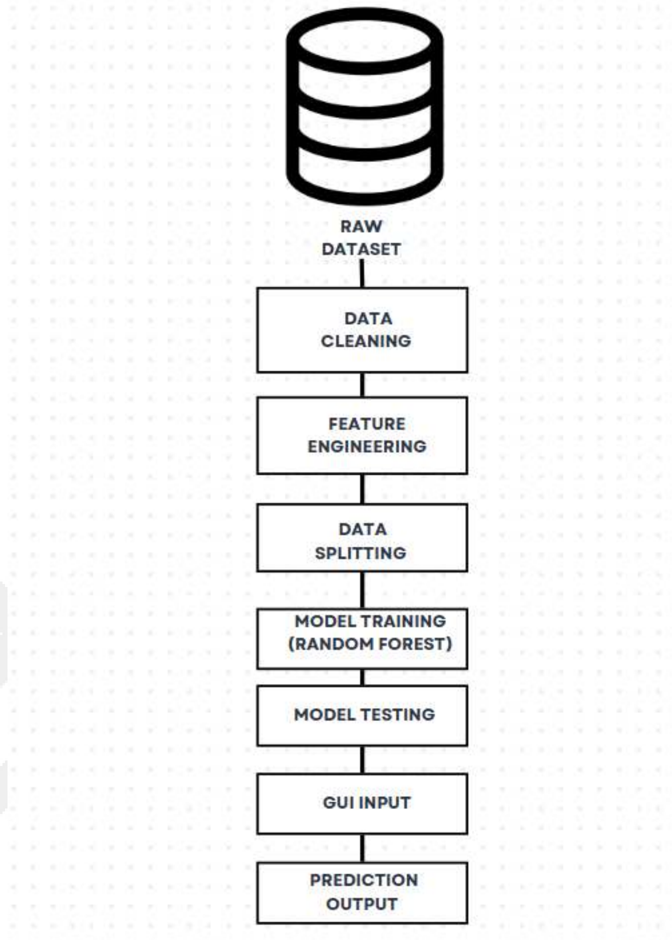

Figure 4.2 Data Flow Diagram for Prediction Model

Several attributes are gathered from an open-source dataset including academic, personal, demographic and financial parameters as a first step. Once the data is collected a process is followed which helps in making the data consistent, encoded (categorical attributes), normalized (numerical attributes) and ensure handling of missing values and illustrated in Figure 4.1 and Figure 4.2 respectively.

The next steps taken after data pre-processing are feature selection, this step helps the model to identify the attributes which are having higher impact on retention and drop-out rates. Once feature selection is completed, model is moved for training where two Random Forest classifiers are trained, one helps in predicting the percentage or probability of a student to drop out while the other helps in predicting the probability of student to be retained. To ensure the model is performing as per the standard, accuracy, precision, recall and confusion matrix are used. Once the model is verified and evaluated using the required parameters the model can finally be used to predict results.

To make the system have more accessibility and user friendliness a GUI predictive model is developed in which user inputs student information, system analyzes and applies the trained model which then predicts whether the student is at risk of dropping out or education can be continued without any difficulty.

The workflow diagram given in Figure 4.1 convers all steps taken from data gathering till prediction model to reflect end to end process while on the other hand, Data Flow diagram focuses on how the data is moving from user input to prediction output. It gives visibility for the internal processes of the developed model and emphasizes the dependency of data variables that affect system’s function.

4.2 Model Validation and Evaluation

To evaluate Model Validation and Evaluation a reserved validation set is used. A reserved validation set is a portion of the original dataset kept aside especially for evaluating the model's performance after training. Since this part of dataset is not used, it helps to assess the ability of models to generalize new data.

This restricted validation set helps to assess the accuracy of dropout and retention projections in the thesis before testing on the final test set.

4.2.1 Retention Model Results

Several classification criteria, including as accuracy, precision, recall, and F1-score, are used to assess the retention prediction model's performance. Figure 4.3 illustrates the result, which shows good prediction ability and class balance.

```
PS C:\Users\hp\Desktop\Raghdah Thesis> python validatemodel.py
📊 Retention Model Evaluation
[[144  58]
 [ 19 443]]
              precision    recall  f1-score   support

           0       0.88      0.71      0.79       202
           1       0.88      0.96      0.92       462

    accuracy                           0.88       664
   macro avg       0.88      0.84      0.85       664
weighted avg       0.88      0.88      0.88       664
```

Figure 4.3 Retention Model Evaluation

Figure 4.3 shows the model's accuracy is 0.88 for both the retained (1) and dropped out (0) student classes. This indicates that the model is quite accurate at predicting positive outcomes; that is, it is 88% accurate when predicting that a student would be retained.

Additionally, the model is very good at identifying students who are actually retained, as evidenced by the class 1 (retention) recall of 0.96. The recall for class 0 (dropouts) is 0.71, which is still acceptable but raises the possibility that some real dropouts is incorrectly labeled. In Figure 4.3 by F1 score the model is better at predicting retention than dropout, as evidenced by the F1-score of 0.92 for retained students and 0.79 for dropouts, which strikes a balance between precision and recall.

The confusion matrix in Figure 4.3 displays TP is 443 which means the retained students are accurately predicted, TN is 144 which corresponds to the number of accurately predicted dropouts, FP is 58 which refers to the number of dropouts that are mistakenly predicted to be retained while FN is 19 that reflects the retained students who are mistakenly predicted as dropout. With a high accuracy of 88%, the model is able to accurately predict the results for 88% of the 664 students.

Macro Average and unweighted average for both classes is found 0.85 for F1, 0.88 for Precision and 0.84 for Recall. All metrics equal 0.88 for the weighted average (which takes class imbalance into account), indicating that the model manages imbalance reasonably well.

4.2.2 Dropout Model Results

The standard classification criteria of accuracy, precision, recall, and F1-score are used to evaluate the dropout prediction model. These measures reveal how well the model detects students who are at danger of dropping out. The evaluation's findings are shown in Figure 4.4.

```
Dropout Model Evaluation
[[443  19]
 [ 60 142]]
              precision    recall  f1-score   support

           0       0.88      0.96      0.92       462
           1       0.88      0.70      0.78       202

    accuracy                           0.88       664
   macro avg       0.88      0.83      0.85       664
weighted avg       0.88      0.88      0.88       664

PS C:\Users\hp\Desktop\Raghdah Thesis>
```

Figure 4.4 Dropout Model Evaluation

From Figure 4.4 the Results after running drop-out model present effective identification of students at high risk and performance evaluated using classification report metrics. The model consistently achieves accurate predictions for both courses, as evidenced by its 0.88 precision (dropout = 0, retention = 1). The model is efficiently identifying students who will actually drop out, as evidenced by the recall score of 0.96 for class 0 (dropouts). This supports the thesis's goal of early identification of high-risk students.

When dropout identification is prioritized, there is a trade-off because the recall for class 1 (retained students) is 0.70, indicating that some retained students are mistakenly categorized as dropouts. The Dropouts (class 0) F1-score is 0.92 whereas for Class 1 - Retained Students, the F1-score is 0.78. This suggests that although the model performs rather well in predicting retained students, it is more successful at predicting dropouts.

The confusion matrix displayed in Figure 4.4 show TP is 443 which corresponds to number of students who are correctly predicted as dropouts and TP is 142 which means number of students predicted as retained. While FP is 60 which refers to incorrectly predicted retained students as dropouts and FN is 19 which means the number of dropout students incorrectly predicted as retained. The model has reached an accuracy of 88%, confirming its dependability on the 664-student test dataset.

4.2.3 Correlation Matrix

Since the dataset includes several demographics, social, financial and academic attributes, it is critically important to understand the relationship among them. This can be achieved by correlation matrix that can help to determine the chances of students' drop out or retention ratio. The correlation matrix presented in Figure 4.5 determines the linear correlation between numeric attributes and target attribute (drop-out/retention).

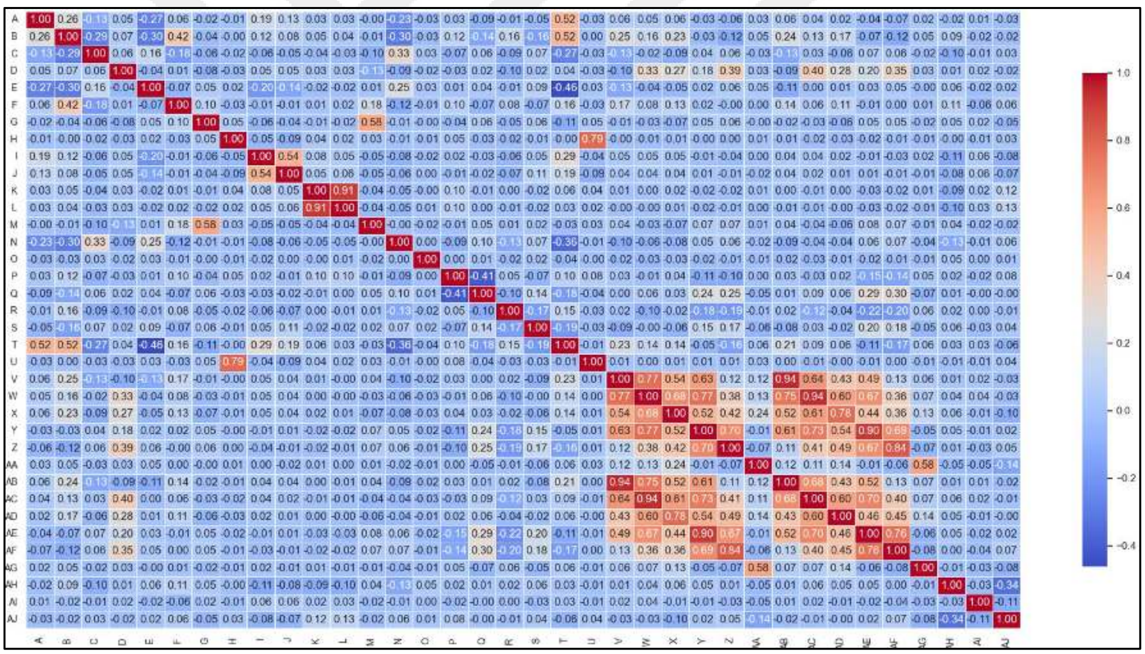

Figure 4.5 Correlation Matrix

The attributes included in dataset are presented in A-Z alphabetical order in Table 4.1.

Table 4.1 Data Attributes

Attribute	Meaning	Attribute	Meaning
A	Marital status	S	Scholarship holder
B	Application mode	T	Age at enrollment
C	Application order	U	International
D	Course	V	Curricular units 1st sem (credited)
E	Daytime/evening attendance	W	Curricular units 1st sem (enrolled)
F	Previous qualification	X	Curricular units 1st sem (evaluations)
G	Previous qualification (grade)	Y	Curricular units 1st sem (approved)
H	Nationality	Z	Curricular units 1st sem (grade)
I	Mother's qualification	AA	Curricular units 1st sem (without evaluations)
J	Father's qualification	AB	Curricular units 2nd sem (credited)
K	Mother's occupation	AC	Curricular units 2nd sem (enrolled)
L	Father's occupation	AD	Curricular units 2nd sem (evaluations)
M	Admission grade	AE	Curricular units 2nd sem (approved)
N	Displaced	AF	Curricular units 2nd sem (grade)
O	Educational special needs	AG	Curricular units 2nd sem (without evaluations)
P	Debtor	AH	Unemployment rate
Q	Tuition fees up to date	AI	Inflation rate
R	Gender	AJ	GDP

Correlation matrix heatmap represented in Figure 4.5 is generated by pre-processing dataset to make sure the target attribute is encoded in numerical data and further records are consistent throughout the file. The degree to which all the attributes are affecting student’s path is analyzed and higher negative or positive values represent the closest association with the outcome. It can be determined from the developed matrix that several attributes such as tuition payment, curriculum performance, GPA or grades have higher impact on the final results. The attributes can further be used to train and enhance the efficiency of the model in terms of predictions.

The attributes labelled from A-Z order which represent the correlation with one another where each cell is displaying a value varying between 1.00 (strongest) to -0.4 (weakest). The red highlighted cell is representing strong positive correlation among the attributes indicating their strong relation, while blue highlighted cells are representing strong negative correlation. On the other hand, the white or pale color-coded cells are indicating no or weak correlation. The attributes indicating strong correlation, greater than or equal to 0.6, are shown in Table 4.2 with pairs.

Table 4.2 Attributes highlighting strong correlation

Attribute 1	Attribute 2	Correlation
Nacionality	International	0.79
Mother's occupation	Father's occupation	0.91
Curricular units 1st sem (credited)	Curricular units 1st sem (enrolled)	0.77
Curricular units 1st sem (credited)	Curricular units 1st sem (approved)	0.63
Curricular units 1st sem (credited)	Curricular units 2nd sem (credited)	0.94
Curricular units 1st sem (credited)	Curricular units 2nd sem (enrolled)	0.64
Curricular units 1st sem (enrolled)	Curricular units 1st sem (evaluations)	0.68
Curricular units 1st sem (enrolled)	Curricular units 1st sem (approved)	0.77
Curricular units 1st sem (enrolled)	Curricular units 2nd sem (credited)	0.75
Curricular units 1st sem (enrolled)	Curricular units 2nd sem (enrolled)	0.94
Curricular units 1st sem (enrolled)	Curricular units 2nd sem (approved)	0.67
Curricular units 1st sem (evaluations)	Curricular units 2nd sem (enrolled)	0.61
Curricular units 1st sem (evaluations)	Curricular units 2nd sem (evaluations)	0.78
Curricular units 1st sem (approved)	Curricular units 1st sem (grade)	0.7
Curricular units 1st sem (approved)	Curricular units 2nd sem (credited)	0.61
Curricular units 1st sem (approved)	Curricular units 2nd sem (enrolled)	0.73
Curricular units 1st sem (approved)	Curricular units 2nd sem (approved)	0.9
Curricular units 1st sem (approved)	Curricular units 2nd sem (grade)	0.69
Curricular units 1st sem (grade)	Curricular units 2nd sem (approved)	0.67
Curricular units 1st sem (grade)	Curricular units 2nd sem (grade)	0.84
Curricular units 2nd sem (credited)	Curricular units 2nd sem (enrolled)	0.68
Curricular units 2nd sem (enrolled)	Curricular units 2nd sem (evaluations)	0.6
Curricular units 2nd sem (enrolled)	Curricular units 2nd sem (approved)	0.7
Curricular units 2nd sem (approved)	Curricular units 2nd sem (grade)	0.76

Table 4.2 represents attributes having strong relationship in the dataset which represents the pair of attributes that are highly associated with each other varying between 1.0 to 0.7. This shows that if one attribute from the pair changes it directly impacts the influence of the other attribute which can have high impact on the final results (in increasing or decreasing way). This analysis helps to determine which attributes need to be considered and utilized for development of prediction models making sure that the relevant attributes are picked to evaluate the results and help in making decisions.

4.3 Prediction System Based on GUI

To further evaluate and determine if the prepared model is working accurately, a graphical user interface is developed using Tkinter that allows the user to get different results depending on the values given for each parameter.

The parameters and attributes of the predicted model are GPA, gender, age, admission grade, tuition status, and academic performance are important characteristics. All other fields have pre-defined numerical ranges, while dropdowns were utilized for category inputs. To confirm to model expectations, GPA input is translated into semester grades and for GDP handling a default average value is given since it is not user facing. Once the input values are entered a message box is displayed with the predicted retention and dropout probabilities showing up as outcome.

To test the model, an instance of input list is given in Table 4.3.

Table 4.3 Inputs added into Attributes for Prediction

Features	**Values**
Gender	Male
Daytime/evening attendance	Daytime
Tuition fees up to date	No
Student GPA	0.1
Admission Grade (0-200)	80
Age at Enrollment	18
Approved Subjects (1st Sem)	2
Average Grade (1st Sem)	6.1
Subjects Enrolled (1st Sem)	10
Evaluations Taken (1st Sem)	10
Approved Subjects (2nd Sem)	2
Average Grade (2nd Sem)	7.5
Subjects Enrolled (2nd Sem)	4
Evaluations Taken (2nd Sem)	4
Unemployment Rate (%)	41.1
Inflation Rate (%)	50

The output of the model is illustrated in Figure 4.6 with the help of GUI.

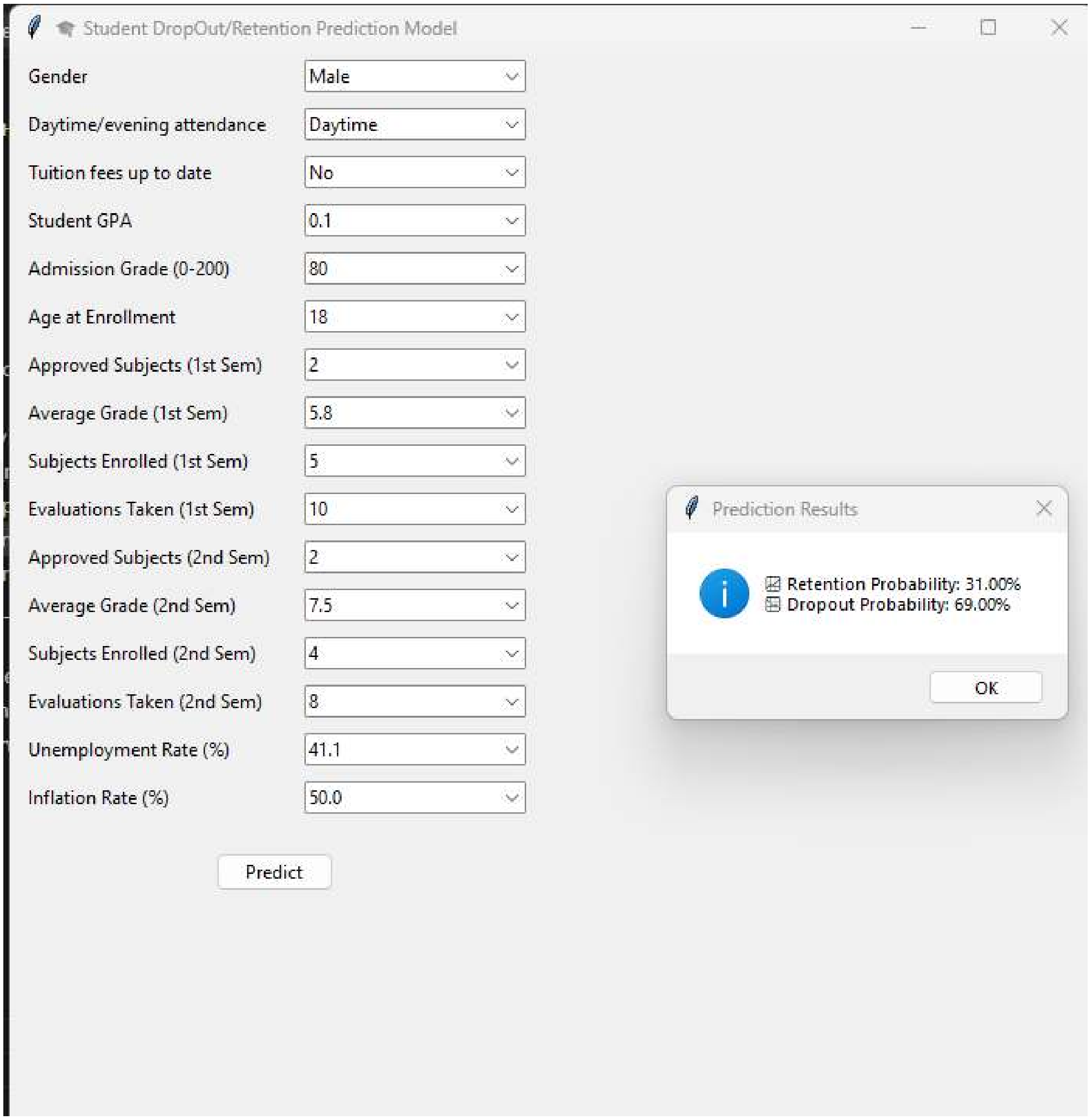

Figure 4.6 GUI Based Prediction System - Retention and Dropout percentage

In Figure 4.6, the probability of retention and dropout are shaped in percentages. These percentages are highly dependable on the values given in GPA, class schedule (Daytime/Night time), and Semester Grades. The input values represented in Table 4.3 are inserted in the form to get the prediction output. However, these values can be varied and users can add them differently to get the probability of student being retained or dropping out.

GPA, admission grade, fee payment status, and academic achievement during semesters are identified as key contributing factors and the algorithm accurately forecasts the likelihood of both dropout and retention. Institutions can proactively identify and assist students who are at risk thanks to the GUI tool. The thesis shows how different attributes play critical role in predicting the possibilities of students to continue their studies or to drop out from the institute.

5. DISCUSSION AND EVALUATION

This chapter offers a critical evaluation of the study's results, emphasizing important discoveries, the effectiveness of the model, and the importance of the implemented GUI application. The review discusses the predictive models' advantages and disadvantages while evaluating them in light of the study's goals.

5.1 Analysis of the Results

The project's main objective is to increase student retention by using machine learning models based on academic, personal, and financial characteristics to forecast dropout risks. A dual viewpoint for institutional decision-making is provided by the training of two distinct models, one for dropout prediction and the other for the chance of student retention. Key findings include:

High-impact characteristics: Age at enrollment, entrance grades, tuition payment status, GPA (curricular unit grades), and type of course attendance (evening or day) is found to be significant predictors of both dropout and retention.

Model performance: Random Forest classifiers performed well in terms of prediction for both objectives. Metrics including F1-score, recall, accuracy, and precision shows that the models are dependable for real-world applications.

GUI Application: To make forecasts understandable to non-technical users, such as academic advisers or administrative personnel, a user-friendly GUI is successfully designed. Numerical fields are simplified for simple data entry, and dropdown menus are used to select categorical inputs.

5.2 Model Effectiveness Assessment

According to performance criteria, the models are especially effective at identifying students who are at a high risk of dropping out, which is essential for preventative intervention. Among the notable evaluation results are:

Retention Prediction Model: To guarantee that the majority of students who are likely to stay are accurately recognized, high recall was given priority.

Dropout Prediction Model: To make sure that identified at-risk students actually needed intervention, high precision is marginally preferred.

The overall resilience is enhanced by feature scaling and data pretreatment procedures such label encoding and handling of non-critical missing data. By using Random Forests, feature interactions can be handled efficiently without requiring a lot of hyperparameter tweaking.

Even though the unemployment and inflation rates are averaged for GUI simplicity, the models are still quite sensitive to changes in these socioeconomic variables. This offers a chance to make it better in subsequent versions.

5.3 Limitation of Study

A number of restrictions are identified:

Dataset Balance: Despite efforts to balance classes, there were marginally less dropout cases in the original dataset than retention instances, which could have an impact on model learning.

External Factors: Personal situations and emotional aspects that are not included in the dataset have an impact on students' decisions in the real world.

GUI simplification: A few non-essential features, such as GDP, were averaged out for GUI purposes, which might have resulted in slight reductions in forecast granularity.

6. FUTURE WORK AND CONCLUSION

This chapter covers the possible future work in terms of expanding the areas covered in the thesis. Prediction performance can be improved with vast range of datasets from different demographics and can be helpful in future studies. It explains different applications and systems where predictive models for student drop out and retention can be utilized in efficient manner.

6.1 Future Work

Even though the goal of thesis has been met, a number of areas for further study and system enhancements have been noticed:

More Real-Time Data: Including attendance records and real-time academic performance (such as semester-by-semester GPA) could improve model accuracy.

Deep Learning Models: Investigating neural networks may improve predicting skills, particularly when dealing with sequential student data (such as year-by-year performance).

Explainable AI (XAI): Advisors may be better able to comprehend the logic behind certain forecasts if model interpretability tools like SHAP or LIME are added.

Alert and Warning System: Integrating the GUI into a comprehensive early warning system that automatically identifies at-risk students and suggests focused interventions to help institutes guide their students for better results.

Wider Dataset Collection: The generalizability of the model would be enhanced by extending the dataset across several colleges, institutes and a wider range of demographics.

6.2 Conclusion

The study can make a substantial contribution to educational institutions' capacity to raise student success rates by effectively creating a machine learning-based method for forecasting student dropout and retention. The models offer practical insights for early interventions by highlighting important impacting elements such as enrollment characteristics, financial status, and GPA.

By bridging the gap between sophisticated prediction models and real-world applications, the GUI application enables academic staff to employ data-driven tactics without the need for technical knowledge.

All things considered; the experiment shows that machine learning provides strong instruments to tackle important issues in education. Such predictive systems have the potential to revolutionize student support services and promote improved academic performance in the future with ongoing development and wider data integration.

REFERENCES

Agrusti, F., Bonavolontà, G., and Mezzini, M. (2019). University dropout prediction through educational data mining techniques: A systematic review. Journal of e-learning and knowledge society. 15(3), 161-182.

Ajzen, I. (2020). The theory of planned behavior: Frequently asked questions. Human behavior and emerging technologies. 2(4), 314-324.

Alsharari N. M., and Alshurideh M. T. (2020). Student retention in higher education: the role of creativity, emotional intelligence and learner autonomy. International Journal of Educational Management. 35(1), 233-247.

Araque, F., Roldán, C., and Salguero, A. (2009). Factors influencing university drop out rates. Computers and Education, 53(3), 563-574.

Aulck, L, Velagapudi, N., Blumenstock, J., and West J. (2016). Predicting student dropout in higher education. arXiv preprint arXiv:1606.06364.

Balfanz, R., Herzog, L., and Mac Iver, D. J. . (2007). Preventing student disengagement and keeping students on the graduation path in urban middle-grades schools: Early identification and effective interventions. Educational Psychologist, 42(4), 223-235.

Barbera, S. A., Berkshire, S. D., Boronat, C. B., and Kennedy, M. H. (2020). Review of undergraduate student retention and graduation since 2010: Patterns, predictions, and recommendations for 2020. Journal of College Student Retention: Research, Theory and Practice, 22(2), 227-250.

Bosnjak, M., Ajzen, I., and Schmidt, P. . (2020). The theory of planned behavior: Selected recent advances and applications. Europe's journal of psychology, 16(3), 352.

Burke, A. (2019). Student retention models in higher education: A literature review. College and University. 94(2), 12-21.

Cameron, S. V., and Heckman, J. J. (2001). The dynamics of educational attainment for black, hispanic, and white males. Journal of political Economy, 109(3), 455-499.

Caruth, G. D. (2018). Student engagement, retention, and motivation: Assessing academic success in today’s college students. Participatory Educational Research. 5(1), 17-30.

Chen, Y., Johri, A., and Rangwala, H. (2018). Running out of stem: a comparative study across stem majors of college students at-risk of dropping out early. In Proceedings of the 8th international conference on learning analytics and knowledge (pp. 270-279).

Contreras, D., González, L., Láscar, S., and López, V. (2022). Negative teacher–student and student–student relationships are associated with school dropout: Evidence from a large-scale longitudinal study in Chile. International Journal of Educational Development, 91, 102576.

Crosling, G. (2017). Student retention in higher education, a shared issue. Encyclopedia of international higher education systems and institutions. 1-6.

Dekker, G. W., Pechenizkiy, M., and Vleeshouwers, J. M. (2009). Predicting students drop out: A case study. In Proceedings of the 2nd International Conference on Educational Data Mining, EDM 2009, July 1-3, 2009. Cordoba, Spain (pp. 41-50).

Del Savio, A. A., Galantini, K., and Pachas, A. (2022). Exploring the relationship between mental health-related problems and undergraduate student dropout: A case study within a civil engineering program. Heliyon, 8(5).

Fincham, E., Rozemberczki, B., Kovanović, V., Joksimović, S., Jovanović, J., and Gašević, D. (2021). Persistence and Performance in Co-Enrollment Network Embeddings: An Empirical Validation of Tinto's Student Integration Model. IEEE Transactions on Learning Technologies, 14(1), 106-121.

Gašević, D., Dawson, S., and Siemens, G. (2015). Let’s not forget: Learning analytics are about learning. TechTrends, 59, 64-71.

Gonzalez-Nucamendi, A., Noguez, J., Neri, L., Robledo-Rella, V., and García-Castelán, R. M. G. (2023). Predictive analytics study to determine undergraduate students at risk of dropout. In Frontiers in Education (Vol. 8, p. 1244686).

Hamdane, K., El Mhouti, A., Massar, M., and Chihab, L. (2023). Potentialities of learning analytics to overcome students dropout in distance higher education. In International Conference on Digital Technologies and Applications (pp. 397-404).

Hanushek, E. A., and Woessmann, L. (2010). Education and economic growth. Economics of education, 60(67), 1.

Hashim, A. S., Awadh, W. A., and Hamoud, A. K. . (2020). Student performance prediction model based on supervised machine learning algorithms. In IOP conference series: materials science and engineering (Vol. 928, No. 3, p. 032019). IOP Publishing.

Haverila, M. J., Haverila, K., and McLaughlin, C. . (2020). Variables affecting the retention intentions of students in higher education institutions: A comparison between international and domestic students. Journal of International Students, 10(2), 358-382.

Kučak, D., Juričić, V., and Đambić, G. (2018). Machine Learning In Education-A Survey Of Current Research Trends. Annals of DAAAM and Proceedings, 29.

Laal, M., and Laal, A. (2012). Challenges for lifelong learning. Procedia-social and behavioral sciences, 47, 1539-1544.

Lemay, D. J., Baek, C., and Doleck, T. (2021). Comparison of learning analytics and educational data mining: A topic modeling approach. Computers and Education: Artificial Intelligence, 2, 100016.

Lottering, R., Hans, R., and Lall, M. (2020). A model for the identification of students at risk of dropout at a university of technology. In 2020 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD) (pp. 1-8). IEEE.

Realinho, V., Vieira Martins, M., Machado, J., & Baptista, L. (2021). Predict Students' Dropout and Academic Success [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5MC89.

Manrique, R., Nunes, B. P., Marino, O., Casanova, M. A., and Nurmikko-Fuller, T. (2019). An analysis of student representation, representative features and classification algorithms to predict degree dropout. In Proceedings of the 9th International Conference on Learning Analytics and knowledge (pp. 401-410).

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, 13(1), 5705.

Mohamad, S. K., and Tasir, Z. (2013). Educational data mining: A review. Procedia-Social and Behavioral Sciences, 97, 320-324.

Mouton, D., Zhang, H., and Ertl, B. (2020). German university student's reasons for dropout. Identifying latent classes. Journal for educational research online, 12(2), 190-224.

Mubarak, A. A., Cao, H., and Zhang, W. (2022). Prediction of students' early dropout based on their interaction logs in online learning environment. Interactive Learning Environments, 30(8), 1414-1433.

Mubarak, A. A., Cao, H., and Hezam, I. M. (2022). Deep analytic model for student dropout prediction in massive open online courses. Computers and Electrical Engineering, 93, 107271.

Müller, L., and Klein, D. (2023). Social inequality in dropout from higher education in Germany. Towards combining the student integration model and rational choice theory. Research in Higher Education, 64(2), 300-330.

Nurmalitasari, Awang Long, Z., and Faizuddin Mohd Noor, M. (2023). Factors influencing dropout students in higher education. Education Research International, 2023(1), 7704142.

Oliveira, R. and. (2024). Predictive Model of School Dropout Based on Undergraduate Course Self-assessment Data. 195-207.

Oliveira, R., and Medeiros, F. (2024). Predictive Model of School Dropout Based on Undergraduate Course Self-assessment Data. In International Conference on Artificial Intelligence in Education (pp. 195-207). Cham: Springer Nature Switzerland.

Ortiz-Lozano, J. M., Rua-Vieites, A., Bilbao-Calabuig, P., and Casadesús-Fa, M. (2018). University student retention: Best time and data to identify undergraduate students at risk of dropout. Innovations in education and teaching international.

Peña-Ayala, A. (2014). Educational data mining. 524.

Pérez, B., Castellanos, C., and Correal, D. (2018). Predicting student drop-out rates using data mining techniques: A case study. In IEEE Colombian Conference on Applications in Computational Intelligence (pp. 111-125).

Pyle, N., and Wexler, J. (2012). Preventing students with disabilities from dropping out. Intervention in School and Clinic, 47(5), 283-289.

Roberts, M. (2020). Universities With Highest and Lowest Dropout Rates.

Romero, C., and Ventura, S. . (2020). Educational data mining and learning analytics: An updated survey. Wiley interdisciplinary reviews: Data mining and knowledge discovery, 10(3), e1355.

Santos, C. H. D., de Lima Martins, S., and Plastino, A. (2021). É Possível Prever Evasão com Base Apenas no Desempenho Acadêmico?. In Anais do XXXII Simpósio Brasileiro de Informática na Educação (pp. 792-802). SBC.

Sarrico, C. S. (2018). Completion and retention in higher education. Encyclopedia of international higher education systems and institutions.

Scheuer, O., and McLaren, B. M. (2012). Educational data mining. Encyclopedia of the Sciences of Learning.

Seo, E. Y., Yang, J., Lee, J. E., and So, G. (2024). Predictive modelling of student dropout risk: Practical insights from a South Korean distance university. Heliyon, 10(11).

Shrestha, S., and Pokharel, M. (2020). Data mining applications used in education sector. Journal of Education and Research, 10(2), 27-51.

Siemens, G., and Baker, R. S. D. (2012). Learning analytics and educational data mining: towards communication and collaboration. In Proceedings of the 2nd international conference on learning analytics and knowledge (pp. 252-254).

Song, Z., Sung, S. H., Park, D. M., and Park, B. K. (2023). All-year dropout prediction modeling and analysis for university students.

Sultan, E., Norshahriel, M., Rani, A., Filzah, N., Radzuan, M., and Huay, L. (2021). Predictive analytics on university student dropouts from online learning due to MCO. In Proc. Knowl. Manage. Int. Conf.(KMICe) (pp. 117-123).

Tagliaferro, L. (2023). Graduation statistics and fun facts in 2024 - Future Fit.

Tempelaar, D. T, Rienties, B., and Giesbers, B. . (2025). In search for the most informative data for feedback generation: Learning analytics in a data-rich context. Computers in Human Behavior, 47, 157-167.

Troelsen, R., and Laursen, P. F. (2014). Is drop-out from university dependent on national culture and policy? The case of D enmark. European Journal of Education, 49(4), 484-496.

Unesco.org. (2023). 250 million children out-of-school: What you need to know about UNESCO's latest education data.

Webster, A. L., and Showers, V. E. (2011). Measuring predictors of student retention rates. American Journal of Economics and Business Administration, 3(2), 301-311.

West, D., Heath, D., and Huijser, H. (2016). Let's talk learning analytics: A framework for implementation in relation to student retention. Online Learning Journal, 20(2), 1-21.

Wild, S., and Heuling, L. S. (2020). Student dropout and retention: An event history analysis among students in cooperative higher education. International Journal of Educational Research, 104, 101687.