

Rochester Institute of Technology

RIT Digital Institutional Repository

Theses

12-2022

Effective Academic Advising Strategy in Higher Education using Machine Learning

Reem Abuzayeda
rma8393@rit.edu

Follow this and additional works at: <https://repository.rit.edu/theses>

Recommended Citation

Abuzayeda, Reem, "Effective Academic Advising Strategy in Higher Education using Machine Learning" (2022). Thesis. Rochester Institute of Technology. Accessed from

This Master's Project is brought to you for free and open access by the RIT Libraries. For more information, please contact repository@rit.edu.

**Effective Academic Advising Strategy
in Higher Education
using Machine Learning**

by

Reem Abuzayeda

**A Capstone Submitted in Partial Fulfilment of the Requirements for
the Degree of Master of Science in Professional Studies: Data
Analytics**

Department of Graduate Programs & Research

Rochester Institute of Technology

RIT Dubai

December 2022

RIT

Master of Science in Professional Studies: Data Analytics

Graduate Capstone Approval

Student Name: **Reem Abuzayeda**

Graduate Capstone Title: **Effective Academic Advising in Higher Education using Machine Learning**

Graduate Capstone Committee:

Name: **Dr. Sanjay Modak**

Date:

Chair of the Committee

Name: **Dr. Ioannis Karamitsos**

Date:

Member of committee

Acknowledgments

I would like to express my gratitude and deep appreciation to Dr. Ioannis Karamitsos, my capstone project mentor. Without his supervision, support, encouragement, and guidance; I would not have been able to complete this project.

I would also like to thank all professors in this Master's program for the knowledge I gained through their classes.

Finally, I would like to thank my family and friends for their support during my studies.

Abstract

Covid-19 pandemic has caused many obstacles to higher education students, especially students with academic, financial, and family disadvantages. Moreover, Covid-19 has caused negative impacts on enrollment of junior students and students' academic goals, which has led to a higher rate of dropouts worldwide.

University enrollment is a complex process for students and families, and the decision to drop out is overwhelming to both. Many factors might cause the dropout, but the most important factors that might affect this decision are financial factors and academic standing. Higher education institutions can boost their academic advising plans, through the use of their strategic resources of data and Machine Learning techniques.

This study investigated the important factors that influence students' dropout and also studied the factors that indicate that a particular student needs extra academic advising.

This study used two datasets, and different machine learning algorithms. For the dropout prediction, the K-Nearest Neighbor model outperformed the Random Forest and the Decision tree models. While for the extra advising prediction, the Random Forest model outperformed the Decision tree and the Artificial Neural Network models.

The study found that Tuition fees status and age at enrollment seriously affect student decision of drop out. Also, the study found that academic standing, year of study, and special needs cases are the most factors that indicate if a student needs extra academic advising.

Keywords: Higher Education, Academic Advising, Student Dropout, Student Retention, Supervised Machine Learning.

Table of Contents

ACKNOWLEDGMENTS.....	II
ABSTRACT.....	III
LIST OF FIGURES.....	1
LIST OF TABLES.....	2
CHAPTER 1 - INTRODUCTION	3
1.1 INTRODUCTION TO THE PROBLEM.....	3
1.2 STATEMENT OF THE PROBLEM.....	4
1.3 PROJECT GOALS	4
1.4 AIMS AND OBJECTIVES.....	5
1.5 RESEARCH QUESTIONS.....	5
1.6 LIMITATIONS OF STUDY	6
CHAPTER 2 - LITERATURE REVIEW	7
2.1 INTRODUCTION	7
2.2 COVID-19 EFFECTS ON HIGHER EDUCATION.....	7
2.3 ACADEMIC ADVISING.....	11
2.4 ACADEMIC ADVISING CHALLENGES.....	15
2.5 MACHINE LEARNING MODELS FOR ACADEMIC ADVISING.....	18
2.6 TAKEAWAYS FROM LITERATURE REVIEW.....	22
CHAPTER 3- RESEARCH METHODOLOGY	23
CHAPTER 4 - PROJECT DESCRIPTION	26
4.1 ACADEMIC DROPOUT DATASET	27
4.2 ACADEMIC ADVISING DATASET	32
CHAPTER 5- DATA ANALYSIS.....	38
5.1 ACADEMIC DROPOUT DATA PROCESSING	38
5.2 ACADEMIC ADVISING DATA PROCESSING	41
5.3 ACADEMIC DROPOUT MACHINE LEARNING MODELS	44
5.4 ACADEMIC ADVISING MACHINE LEARNING MODELS	48
5.5 RESULTS.....	53
CHAPTER 6- CONCLUSION	58
6.1 CONCLUSION	58
6.2 RECOMMENDATIONS	59
6.3 FUTURE WORK.....	62
BIBLIOGRAPHY.....	63

List of Figures

Figure 1- Replies - 4 Regions	9
Figure 2 - Do you believe COVID-19 will affect enrollment numbers for the new academic year?	10
Figure 3 - Learning Analytics Dashboard.....	19
Figure 4 - J48 Decision Tree	20
Figure 5 - Data set Fields and Types	21
Figure 6- CRISP_DM.....	23
Figure 7 - Dataset 1 Dimension.....	27
Figure 8 - Dataset 1 Structure.....	28
Figure 9 - Dataset 1 Summary	29
Figure 10 - Dataset 2 Dimensions.....	32
Figure 11 - Dataset 2 Structure.....	32
Figure 12 - Dataset 2 Summary	33
Figure 13 - CGPAAttribute	34
Figure 14 - Academic Standing	35
Figure 15 - Level (Year & Semester) Plot.....	35
Figure 16 - Employed vs Non-Employed Plot.....	36
Figure 17 - College Counselling Cases Plot.....	36
Figure 18 - Extra Advising Cases.....	37
Figure 19 - Target Processing.....	38
Figure 20 - Academic Dropout Training Set.....	39
Figure 21 - Academic Dropout Testing Set	40
Figure 22 - Academic Advising Attributes Correlation.....	41
Figure 23 - Academic Advising Data Selection	42
Figure 24 - Academic Advising Training Set.....	43
Figure 25 - Academic Advising Testing Set.....	43
Figure 26 - Minimal Depth Academic Dropout Set.....	45
Figure 27 - Result of Random Forest Model for Academic Dropout	45
Figure 28 - Decision Tree - Academic Dropout	46
Figure 29 - Decision Tree Model Results for Academic Dropout.....	47
Figure 30 - KNN Model Results for Academic Dropout	47
Figure 31 - Random Forest Minimal Depth Plot - Dataset 2.....	49
Figure 32 - Random Forest Model Optimal Number of Trees	49
Figure 33 - Number of Trees and Nodes.....	49
Figure 34 - Random Forest Results for Academic Advising	50
Figure 35 - Decision Tree for Academic Advising.....	51
Figure 36 - Decision Tree Results for Academic Advising	51
Figure 37 - ANN Model for Academic Advising.....	52
Figure 38 - ANN Results for Academic Advising.....	52

List of Tables

Table 1 - Dataset 1 Visualization	31
Table 2 - Academic Dropout Correlation.....	39
Table 3 - Cost Matrix Illustration	53
Table 4 – Part 1 Cost Matrix	54
Table 5 - Part 2 Cost Matrix	54
Table 6 - Precision True Positive and False Positive	55
Table 7 - Academic Dropout Models Comparison	56
Table 8 - Academic Advising Models Comparison	57

Chapter 1 - Introduction

1.1 Introduction to the Problem

According to United Nations Educational (2020), due to Covid-19, about 24 million students worldwide are at risk of not returning to education. They recommended that more efforts have to be done for the re-enrolment process and to take into consideration different factors like gender, geographical location, social, and economic factors.

Higher education institutions are competing in a battle to improve student retention and therefore graduate more students. The traditional role of an academic advisor which is mainly focused on helping students meet their degree graduation requirements via one or maximum two meetings per semester is not enough to retain and satisfy students in the current era.

With different levels of academic standing, different financial status, unemployment rates, marital and family status, and students' motivation factors; academic advisors are encouraged to deploy innovative communication strategies and creative and more friendly yet effective and more personalized advising techniques. While prescriptive advising allows the traditional communication strategy between the student and the advisor, intrusive advising initiates extra communication when needed in some difficult situations like year one students, graduation, and in-risk cases for more effective results. On the other hand, developmental advising helps students take responsibility for their personal and vocational goals, and tries to improve the student's behavior, problem-solving skills, and decision-making skills.

The power of academic advising is enormous, in students' lives, success, and satisfaction, this is why it is very important to have continuous interactions and personalized communication according to the need and situations of each student.

Higher education institutions have experienced growth in their data in the last few years but most data are not discovered or used properly; applying machine learning models can help them to

discover how to improve academic advising process and personalize it, which will increase students' retention percentage and reduce students' dropout cases.

1.2 Statement of the Problem

Higher Education institutions are facing a challenge which is the **increase in the number of students dropouts** worldwide due to the impacts of Covid 19, which negatively affects students' enrollment, students' retention, and students' satisfaction which will affect their future employment chances. Moreover, losing students will affect the institution's revenue and reputation, and also enrolling new students will cost institutions more from marketing, communication, and negotiation perspectives. Higher education student retention is also of high importance to students who invest their time and resources in the hope of earning a degree.

Many higher education students depend on part-time jobs which are most of the time not secured or with low and unstable incomes. In other cases, many students depend on their family's financial resources which have been affected during the last 2 years since Covid-19. Also, some students face other difficulties rather than financial issues which prevent their graduation, like family issues, health issues, and poor academic standing.

Covid-19 has caused many complications for students, especially students with academic, financial, and family disadvantages. The effects of Covid-19 on the economy have caused a significant dropout in higher education.

Academic advisors in higher education institutions try to improve graduation rates and decrease the loss of tuition revenues caused by students who either drop out or transfer to another university. In this study, Machine learning algorithms were used to gain analytical insights to help academic advisors to achieve their objectives.

1.3 Project goals

This study aimed to explore the challenges of Covid-19 on Higher Education, academic advising strategies, and academic advising current challenges. The study explored Machine learning models in the academic advising field. Then the study applied machine learning models on two datasets,

one dataset is an online dataset about student dropout and the other dataset was collected from one Higher Education institution in the UAE about academic advising need. The main purpose is to develop **an effective academic advising recommendation strategy** for Higher Education institutions which will be recommended to the UAE Ministry of Higher Education.

1.4 Aims and Objectives

The main objectives of this study are:

- Investigate the effects of Covid-19 on Higher Education.
- Investigate current academic advising challenges.
- Investigate Machine learning models in the academic advising field.
- Identify the most significant variables to be used in dropout prediction and academic advising machine learning models.
- Develop effective dropout prediction models.
- Develop effective academic advising models for students who need extra academic advising.
- Recommend the models to the Ministry of Higher of Education.

1.5 Research Questions

This capstone project aimed to address the following research questions:

1. What are the most important factors that indicate that a student is more likely to drop out?
2. What are the most important factors that indicate that a student needs extra academic advising?
3. How accurately can the proposed model predict the dropout?
4. How accurately can the proposed model predict the extra academic advising need?
5. What can higher education institutions do to decrease dropout rates?

1.6 Limitations of Study

During working on this project, I experienced some limitations which I am listing below:

The secondary dataset:

- The quality of the dataset was collected from Kaggle.
- Also, the dataset has many predictors which I had to try to find the most significant to my models.

Literature Review:

- No enough research was found about machine learning models for academic advising and student retention.

Chapter 2 - Literature Review

2.1. Introduction

A well-designed academic advising strategy that is easily accessible by students might fix many challenges in higher education including the increase in dropout rates. Applying effective academic advising strategies allows students to achieve their personal and academic goals, which will lead to higher student satisfaction. Gaining students' and parents' satisfaction is a strategic objective of higher education institutions as it leads to higher level of student retention which is one of the key success indicators of an academic institution.

Moreover, putting students at risk, pulling scholarships increasing tuition costs of higher education are great concerns for students and their families.

This chapter explored the effects of Covid-19 on higher education, academic advising strategies, and academic advising current challenges. The chapter also explored some past applied machine learning models in the academic field.

2.2. Covid-19 Effects on Higher Education

Covid-19 has negatively affected the higher education enrolment process and students' satisfaction, and since 2020 there is an increase in students dropout cases. Marinoni et al., (2020) mentioned that according to UNESCO, on April 2020, schools and higher education institutions were closed in 185 countries, affecting around 1,542,412,000 students, which is about 89.4% of total students globally.

Naughton (2021) mentioned that COVID-19 pandemic has caused significant disruption to students across the globe, especially students who were at risk of falling behind because of academic, financial, and racial reasons. These students faced some additional challenges in perusing their graduation goals. The author also mentioned that the path to university for students

with low income, first-generation, and racially minoritized has been plagued by COVID-19 which hurts college enrollment, especially for students from marginalized backgrounds.

“Although not being able to contact students at all was rare, advisers struggled to provide proactive, timely assistance to students they would have normally seen in person” Naughton (2021, p. 5).

Getting financial aid is an issue for low-income students, especially with the increase in the of higher education tuition in the last few years. Also, many families lost their main source of income, and with Covid-19 impacts this challenge has doubled. Moreover, some families lack the essentials of the requirements of study from home, like computers, laptops, and Internet connection.

Naughton (2021) mentioned that in the United States the overall first-year enrollment was down in the fall of 2020 by 16%, and community colleges had a 23% drop in enrollment for first-time students. This means the challenges the students face to access higher education institutions are larger than ever before.

McFarlane and Wallder (2021) said that Covid-19 affected academic advising globally which negatively affected university students' experience, satisfaction, retention, and success. This is proved by exploring the effectiveness of academic advising since the start of Covid-19 and how it affected student enrollment, engagement and course completion.

They surveyed 196 participants, and they found that most participants agreed on how communication is crucial between advisors and advisees but it was affected by Covid-19 because in many cases it was shifted from face-to-face to online forms with fewer advising sessions.

Aucejo et al., (2020) surveyed 1500 students at Arizona State University to find out more about the effects of COVID-19 on higher education; and found that 13% of students have delayed graduation due to COVID-19. They also found that lower-income students are 55% more likely to delay graduation compared to their colleagues of higher-income families.

Marinoni et al., (2020) conducted a survey that covered the impact of COVID-19 on higher education around the world. While the majority of questions in the survey were closed questions; which means the respondent had to choose between some answers, there were two optional open

questions, to allow the respondent to report any challenges, 320 of the respondents (around 75% of the sample) provided contributions to these questions. Many respondents from all countries said that the most important challenge that will affect their higher education institution is the financial implications. Below are more of Marinoni and his team's findings:

Enrollment and teaching implications findings:

- Figure 1 shows that they received 576 replies from 424 unique Higher Education institutions in 109 countries. According to Figure 1, 46% of replies were from Europe, which is almost half the participants. 21% of replies were from Africa, which is a fifth of the participants. And 17% of replies were from Asia and the Pacific and 15% from the Americas, which is low compared to the number of higher education institutions in both Asia and the Americas.

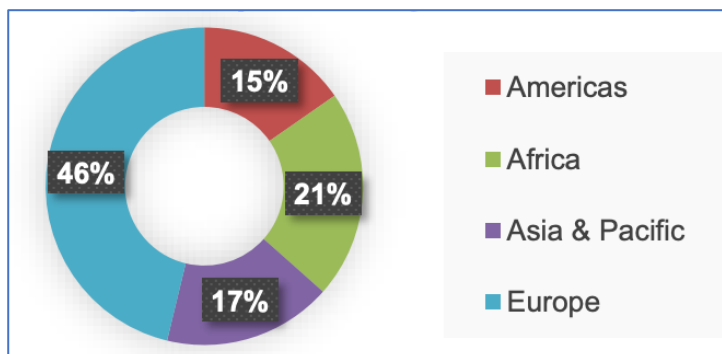


Figure 1- Replies - 4 Regions

- According to Figure 2, they found that 78% of responders think that the pandemic has an impact on the new academic year enrollment, and 46% think that Covid-19 will affect local and international students' enrollment.

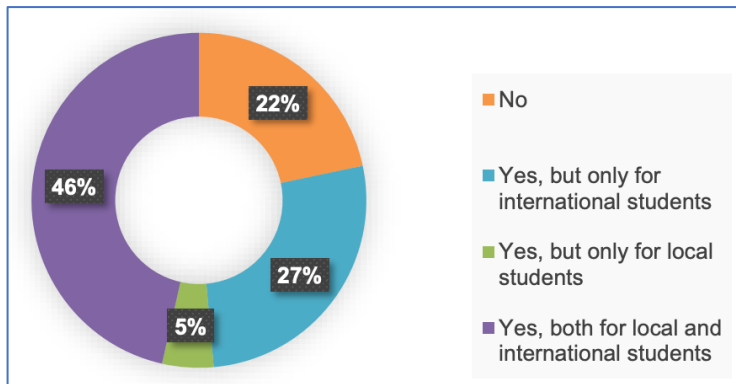


Figure 2 - Do you believe COVID-19 will affect enrollment numbers for the new academic year?

- 59% of the respondents mentioned that all campus activities have stopped, and the institution is completely closed.
- Two-thirds of the respondents reported that classroom teaching has been replaced by distance learning, with many challenges, like access to technical infrastructure, competencies & pedagogies for distance learning, and the requirements of some academic programs that were not fulfilled by distance learning.

Financial implications:

- Many respondents from all countries said that the most important challenge that will affect their higher education institution is the financial implications.
- All institutions; but especially private higher education institutions, explained their concerns regarding the tuition payments and the increased risk of students being dropped.
- Most respondents stressed that a decrease in student enrolment will have negative implications on the institution's financial future.
- Many respondents mentioned the financial impact on students who will not be able to work or get support from their families.
- Some respondents predicted long-term financial implications which will lead to a decline in student enrolment, and this will cause institutions to experience a shortfall in revenue through a lack of tuition fees.

2.3. Academic Advising

Kuhan (2008) defined an academic advisor as an institutional representative who gives direction to students about academic, social, or/and personal matters. The direction might be done through suggestion, counseling, coaching, or advising to achieve several goals.

2.3.1 - History of Academic Advising

Academic advising developed through the years as higher education, curriculums, and academic programs developed. Folsom et al., (2015) mentioned that in the first era of academic advising which was between 1636 and 1870, students had limited academic curriculum and programs, and students used to refer to their professors if they have concerns or out-curriculum questions. The second era was between 1871 and 1971 when many higher education institutions offered different undergraduate programs and elective courses to make sure students contributed to their societies. At this time faculty members used to advise students and support them in any academic, social, and personal matters but still the focus was on graduating.

The authors also mentioned that the third era started at 1972, with the foundation of developmental advising which is not only concerned with academic advising; but supporting skills like problem-solving, decision-making, and evaluation skills. Advisor's role shifted from prescribing some sort of steps and actions, to finding the causes of the student's concerns and helping him/her develop and improve the skills needed to overcome challenges in their academic and personal life as well.

2.3.2 - Academic Advisors' Role

Oertel (2020) described good advisors as ones who will support their advisees once they feel there is a need for the support. Many types of research show that students who undergo continuous advising sessions are more likely to persist and graduate. The author also mentioned that there is a strong correlation between the outside personalized interaction with students and their success.

McFarlane and Wallder (2021) found that academic advisors have an important role in providing advisees with both academic and career help. Their role is more than just connecting students with learning outcomes, but is to shift student focus and goals toward life achievements and long-run targets. They see academic advising as a system that must develop students' academic plans to

match their values, career goals, and life goals. This shall prepare students to act as lifelong learners and positive individuals in their society. Students have to learn how to be initiative and take responsibility for their own choices and actions. The academic advising system must be built with robust objectives that match the university's vision and society's culture and needs, to provide students with more opportunities for their academic and career goals.

It is important that the advisors are friendly and they try to build a relationship and trust with their advisees. A smart way for the advisor is to get to know the advisee outside the academic setting. Advisors are advised to not only seek out their advisees but also to anticipate their needs for help and support in specific situations. They also have to be available and easy to reach.

2.3.3- Importance of Academic Advising

Hagen and Jordan (2008) described advising as a methodology that helps students access past experiences to find what keeps them from reaching their goals and find solutions to their challenges, also it helps students use their education to create their new stories.

One research about student retention stated that “Effective measures for student retention must be implemented to increase the retention of qualified students at institutions of higher learning” (Lau, 2003, p. 1). One of the most effective measures that have to be applied to increase student retention is Academic advising, “Academic advising should be treated as an ongoing process, to be complemented with periodical follow-up sessions throughout the semester” (Lau, 2003, p. 8).

Student success and satisfaction in higher education is a challenge, especially after Covid-19. It is the role of everyone to improve the student journey in academic institutions, faculty, administrators, advisors, and students. Faculty can encourage their students through innovative teaching and learning techniques, administrators can provide the best scholarships plans and the innovative facilities, and advisors have a vital role in changing the direction and allowing students to be aware of themselves and their community not only helping them achieve their graduation requirements but their lives goals, and students have to be motivated and willing to learn, change and improve.

Drake (2011) stressed on the power of advising, communicating, and mentoring in student success. He encourages advisors to build strong relationships with their advisees, locate when and where

the students feel uncomfortable or disconnected, and build a strategy to support and help them get reconnected.

Folsom et al., (2015) mentioned the effect of the transition phase a student faces; which applies to higher education students who suffer from major challenges like a home departure, changing family roles, and returning to university as an adult. This transition theory helps advisors to understand the challenges and new experiences the students are facing and try to help them. On the other hand, they mentioned that the most popular advising issues apart from the personal ones are the demand for detailed information about specific courses, programs, individual faculty expectations which are very hard to predict, and the difficulties the students face with academic policies. Simply but deeply, Academic advising requires effort, patience, understanding, and compassion.

Hunter and White (2004) discussed if academic advising can fix the higher education system in their paper (Academic Advising Fix Higher Education). They mentioned that academic advising itself can't change the curriculum, programs, and system, but it can create a very important connection between students and their education environment factors, which hopefully will help students to be more strategic about their academic and life choices.

A well-structured academic advising system; is the only system that allows students to access a healthy interaction with a caring and concerned specialists to help them with their choices. But the challenge for any higher education institution is to create a trustable and accessible academic advising system which students trust and view as an essential tool for a successful academic life, not as peripheral! The base of such a system starts from a group of well-trained, caring, and professional academic advisors, besides high-quality data, technology and facilities.

Hunter and White (2004) said that around 50 years of academic advising experience in higher education proves that advising can serve as a connection to curricular and non-curricular aspects. With the help and support of an academic adviser, students can identify their objectives and achieve personal and academic goals.

Effective advising and counseling for higher education students make a huge positive impact especially for disadvantaged students to overcome different barriers and challenges.

2.3.4 - Developmental and Intrusive Academic Advising Techniques

Crookston (1994) mentioned that **developmental** advising aims to help students become more aware of their changing selves and behavior. The author also added that developmental advising is concerned with personal and vocational decisions and is also focusing to help the student's environmental and interpersonal interactions, problem-solving, decision-making, and evaluation skills.

Developmental advising is considered as a close relationship and is built on trust between the advisor and the advisees; because it focuses on personal skills not only academic goals. Advisee has to trust the academic advising system, and only in this case, he/she will be open to listening and discussing. Here the advisee is encouraged to initiate and ask for meetings if needed.

On the other hand, Wilder (2016) mentioned that **intrusive** advising is a model that may not be initiated by the student, it is initially initiated by the advisor in some critical situations like at-risk cases, year one students, or graduation.

Intrusive advising is intentional communication with students due to some academic or personal reasons that in some cases might lead to negative outcomes or in other cases needs direct help like the case of graduation, with the advisor's goal to develop a relationship that hopefully will lead to positive outcomes.

Finally, in all cases academic advisors must be available to their advisees even the ones who don't seek help; they have to demonstrate their knowledge of the university policies, curriculum, and resources, but most important their compassion and caring for their advisees.

2.3.5 - Academic Advising Summary

Higher education enrollment and financing is a challenge for students who face academic, financial, social, health, and psychological issues. This is why the rates of dropout and institution enrollment decreased more after Covid-19.

This is why effective advising can help students find financial aid, overcome, adapt & face their academic, social, health, and psychological issues. Effective Academic Advising techniques like

intrusive and Developmental Advising are the main solutions to the students' dropout challenge in Higher Education, instead of the most common passive advising strategy.

2.4. Academic Advising Challenges

2.4.1 - Passive Academic Advising

Tudor (2018) mentioned that most academic advising programs are **passive** where advisors just help students to register for the correct courses to complete their degree. He also mentioned that academic advising in higher education worldwide needs improvement to improve students' success, self-esteem, and satisfaction. He highly recommends integrating academic advising with proactive career advising to help students achieve graduation goals, this means that students will be asked what they want to achieve when they graduate, besides their academic goals.

Drake (2011) mentioned that one important challenge in academic advising is that a student seems like a normal student with no major issues till he/she decides to withdraw. Drake said that this happened to him with one student and he tried the best strategy with this shy and afraid student to compete with his peers. After extensive advising sessions, the student finally found his confidence!

Many experts view student retention as an important performance indicator that measures the effectiveness of the way the higher institution is satisfying students' needs. This is why low student retention is a red flag that the institution is failing to meet students' needs which increases the dropout numbers. And here comes the role of effective academic advising and why it is very core to have a successful interaction with every student individually, this must be done by analyzing every student's case which surely will help to improve students' retention and satisfaction. The means of academic advising, how frequent and easy students have access to academic advising, and what are the needs of students are the three factors that any academic advising strategy must be built upon.

2.4.2 - Academic Advising Quality

Gutiérrez et al., (2020) mentioned that the Global Community for Academic Advising defines four dimensions of academic advising at universities as below:

1. Who advises?
2. How the advising responsibilities are divided?
3. Is the advising service divided according to topics or challenges?
4. Where the advising service takes place: on-campus or off-campus?

These four dimensions summarize the whole process of academic advising: the advisor, responsibilities, approach, and service location. Poor quality of any of these dimensions will negatively affect the whole process of academic advising.

Richard (2008) said that a student might claim that the loss of his/her study is due to poor advising or negligence from the advisor. Also, some students may claim that there were ethical issues in the advising sessions, especially if the students revealed personal information.

Gudep (2007) administrated a questionnaire with 47 Questions to 50 respondents from Skyline College, Sharjah, and revealed that some students find that their academic advisor is not serious, and the advisor avoided meeting his/her advisees with a non-professional attitude, and this is one of the major factors in students' dropout!

Listening to students is just a crucial factor and observing & tracking them are major responsibilities of academic advisors. Passive advising is the worst an advisor can offer to students.

2.4.3 - Virtual Academic Advising

There are many implications if virtual advising is applied in higher education. The challenges of hard and rare communication and the barriers of face-to-face meetings made it harder for students to persist and graduate especially students with the most severe cases and issues. “Although not being able to contact students at all was rare, advisers struggled to provide proactive, timely assistance to students they would have normally seen in person” (Naughton, 2021, p. 5).

Argüello (2014) mentioned that with the changes to higher education in the last few years due to Covid-19 and some other economic reasons, Academic advising has mostly shifted to remote process to adapt to the social distance needs. She added that the virtual advising technique had added to the challenges of engaging, satisfying, and retaining students.

Arhin et al., (2017) said that student retention in distance learning has some concerns, for example, in the Open University in the United Kingdom, the graduation percentage is 22 percent which is low compared to any full-time institution's graduation average percentage which is around 82 percent.

Although many higher education institutions are planning and applying different strategies to enhance student retention, the dropout percentages are still considered higher in distance learning programs compared to full-time programs. Arhin and his team investigated the effect of virtual academic advising on student retention at the University of Cape Coast, in a distance learning environment. From a selected sample of 727 students, 625 students (around 86%) participated in the survey. One of the questions was if the student has been assigned an academic advisor, but most of the students (around 87.5 %) had not been assigned an academic advisor. Also, 10.7% of the students were not sure if they have been assigned an academic advisor. Only 1.8% of the students agreed that they have been assigned an academic advisor. Another question was about the interaction with the academic advisor, around 70% of the students mentioned that they had no interaction with their academic advisor last year. And the shocking news was that only 17.4% of the students had only one to two meetings with their academic advisors last year. Only 44.5% of the students which is less than half of the sample, agreed that academic advisors helped them to set their academic goals and to work toward achieving their academic and life goals.

Increasing students' access to academic advising meetings can help in identifying students who are at risk of planning to drop out. This can be done by allocating a face-to-face meeting schedule for academic advising in higher education institutions.

Naughton (2021) mentioned that few researchers suggested that virtual advising may address important informational issues at higher education institutions, however, virtual advising has not successfully replaced engaged advising provided through interpersonal relationships and in-person meetings, even with the use of innovative technology in virtual advising like; online meetings,

emails, and reminders. For example, a recent study with 80,000 students, found that one-way text-based messages about financial aid made no impact on college enrollment.

Some studies found that there is a little to no impact of virtual advising on college enrollment with disadvantaged students because these students require more intensive advising like in-person advising meetings because their issues are too overwhelming to be solved using virtual tools. These students need direct help and guidance. Also, due to Covid-19, the lack of full orientation which was replaced by virtual orientation has affected the support that used to be delivered to first-year students.

2.5. Machine Learning Models for Academic Advising

2.5.1 - Introduction

Every advisor would ask himself/ herself how can I be a better advisor. How can the academic advising experience be improved? How can the advisor help in satisfying the student's academic and personal goals? What kind of techniques or models can reduce the gap between the advisor and the advisees? How can Machine Learning improve academic Advising and thus improve student retention? And many more questions!

Machine learning can be used to help academic advisors to take immediate and fast actions; for example, when a student is academically at risk, has a psychological problem, has financial problems, and/or has family issues. Machine learning can detect such cases and give intelligent and faster decisions once the Machine learning model recognizes that a student is more likely to drop out and needs extra support via personalized academic advising. Advisors would immediately take action to support the student before the case reach a dropout, which can improve student retention and satisfaction, and of course, will positively affect the institution's revenue and reputation.

2.5.2 - Learning Analytics Dashboard

Gutiérrez et al., (2020) mentioned that as much research studied the power of analyzing the data of higher education institutions, results indicate that the majority of these studies focus on suggestions for learning materials and the prediction of student performance. But little research has been done to support the academic advising process. They believe that most higher education institutions only offer basic support and statistics to academic advisers. In their work, they presented the implementation of a Learning Analytics Dashboard for Advisers (Figure 3), which they called (LADA), to help academic advisors in the decision-making process using analysis techniques. Their Prediction “clustering” model is used to predict what they call “chance of success”, for the **prediction of academic risk**.



Figure 3 - Learning Analytics Dashboard

The model predicts the academic risk of each student using historical data, this is done by clustering previous students using some fields like grades and the number of courses the student did each semester. So, for a new student, the model clusters the student to the most similar student category according to his/her grades and other features. An outlier student, who is considered a unique case, will be advised better to avoid a dropout case. They concluded that the results of trying their model are encouraging; because the LADA offered support to explore different cases and improved academic advisors’ decision-making. They compared the LADA with some traditional tools and found that academic advisors found the LADA more helpful and appealing.

2.5.3 - A classification Academic Advisor Model

Hegazy and Waguih (2018) studied how advising students in selecting their academic major early plays an important role in student life and career. They generated a decision tree that used the major attribute as the root node, with 16 rules that **help students to choose the right track**. They applied three classification models on a real dataset from a higher institute in Egypt to advise students to select their academic majors in the first academic year.

Their study was done on a higher education system, they applied three decision tree algorithms, J48, random tree, and REP using some fields like the student's results. The data set contains 8080 records and 7 fields like personal information and student academic results (genders, place of birth, high school score, CGPA, College major). They used WEKA to implement the three decision tree algorithms and their collected data was prepared and converted to an (arff) file.

They found that the J48 algorithm produces a decision tree with 25 nodes and 16 leaves (Figure 4). They recommend the use of this model in advising students for selecting their major, they see this model as a guide to identify the suitable track to improve students' performance and decrease the dropout percentage.

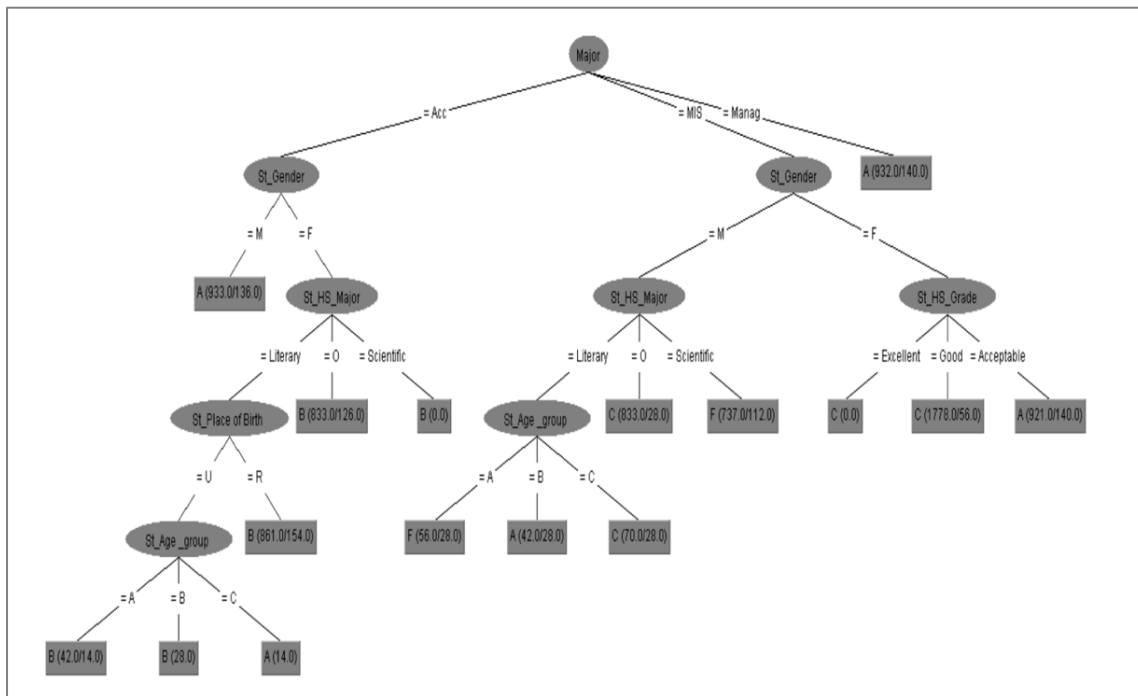


Figure 4 - J48 Decision Tree

2.5.4 - An Educational Data Mining System for Advising Higher Education Students

Nagy et al., (2013) used a classification algorithm to **recommend a suitable major** for the student. They also used a clustering algorithm to segment students into several clusters based on grades. They combined the results from both models to predict more accurate results to improve students' success.

Figure 5 shows that the dataset includes some fields like student scores and student first-year results.

DATA SET META DATA		
ATTRIBUTE	DATA TYPE	RANGE
Secondary Stage Type	Discrete	9 values(SSA1,SSA2,...SSA9)
Total Marks SS	Continues	0-420
English Mark	Continues	0-50
Math Marks	Continues	0-100
Physics Marks	Discrete	0-50
First Year Grade	Discrete	8 values (A,B+,B-,C+,C,D+,D,F)
Department	Discrete	4 values (AE,CE,CS,MIS)

Figure 5 - Data set Fields and Types

In the classification stage, the output is the recommended major. But they prepared the data first by removing all failed students in year one. While in the Clustering Phase they divided the students' records into several clusters based on their academic results. They also created a user-friendly interface that is mapped to the models to make it easier for academic advisors to use their algorithms.

2.6. Takeaways from Literature Review

Higher education institutions are facing many challenges, especially in the last decade. The most concerning challenge is the **increase in the number of student dropouts** which threatens student retention, and at the same time impacts the reputation and the income of the institution.

One important solution for this challenge is to move from **passive** academic advising method to **intrusive** and **Developmental** Advising methods.

Higher Education Institutions have experienced growth in the **data** in the past years, and these institutions can improve student retention and decrease the number of dropouts by applying **machine learning models** to provide effective **personalized academic advising strategies** to each student according to his/her situation. Models can make much more intelligent and faster recommendations to advisors to take the best and most accurate actions.

Chapter 3- Research Methodology

Schröerabet al., (2021) mentioned that the **Cross Industry Standard Process for Data Mining (CRISP-DM)** is the worldwide used **standard** for applying **data mining** projects. The authors considered CRISP-DM as an independent and not restricted to any technology process model for data mining projects.

According to HOTZ (2022), the Cross Industry Standard Process for Data Mining (Figure 6 - CRISP-DM) has **six phases** that describe the data science process cycle. These phases helped in this capstone project as a road map to understand, analyze, mine data, find results, and recommend solutions. Figure 6 shows the phases of CRISP-DM:

1. Business understanding or the needs of the business.
2. Data understanding.
3. Data preparation for modeling.
4. Modeling or the machine learning algorithms that suites the data we have.
5. Evaluate and compare results to select the best model for the business needs and available data.
6. Deployment to allow stakeholders to start using the selected model.

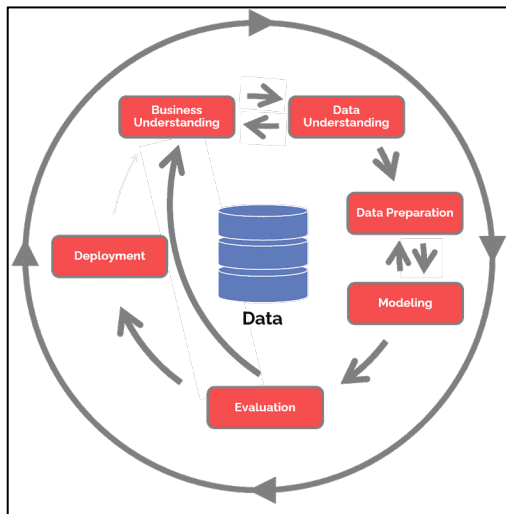


Figure 6- CRISP_DM

Source: <https://www.datascience-pm.com/crisp-dm-2/>

For this project, the details of the six CRISP-DM phases are listed below:

Phase 1 - Business Understanding:

- Understand the business objectives: what are the goals of higher education institutions, what they are trying to achieve & improve, and what are they trying to prevent and avoid?
- Determine the objectives of the data mining model: what can the solution achieve with the available datasets, is it applicable, and will it produce an accurate result that matches the objectives the stakeholder is looking for?

Phase 2 - Data Understanding

- Datasets to be collected
- Datasets are to be initially examined to make sure it is suitable for the planned model.
- Datasets are to be described: how many records and variables, and variables types.
- Datasets are to be deeply explored, by examining which variables will be used in the model, and also exploring the relationship between variables.
- Datasets have to be assessed: assess the quality of data, is the data accurate, complete, valid, unique, consistent, and represents reality from the required point of time.

Phase 3 - Data Preparation

- Select variables: Determine which variables will be used in the model.
- Clean data: the most important step to correct and impute data.
- Transform data: for example, use discretization to transform a continuous attribute into a discrete attribute

Phase 4 - Modeling

- Select machine learning methods: depending on the datasets, (like classification for a discrete label or regression for the estimation of a continuous value) if the dataset is labeled, or unsupervised method (association or clustering) if the dataset is not labeled.
- Divide the datasets into at least training and testing parts for some models like classification.
- Build the machine learning model(s) and fine tune the models.
- Test the model(s) according to some measures, interpret, and then test again.

Phase 5 - Evaluation

- Evaluate if the models meet the business objectives.
- If there is more than one model, evaluate which one best meets the business objectives using different measurements.
- Initially present to stakeholders and get feedback.
- Correct any part if needed.
- Summarize results and present them to stakeholders.
- Revisit any of the above phases.

Phase 6 - Deployment

- Produce recommendation.
- Finally present to stakeholders.
- Develop a maintenance strategy.
- Final project review for future improvements.

Chapter 4 - Project Description

Academic Advising has been always an impactful tool to help students overcome most barriers they face in their academic study. The more innovative advising strategies and technologies higher education institutes apply, the higher percentages of student success and student satisfaction they might achieve. The use of high-quality data and Machine learning models to improve the academic advising process can lead to improvement in students' success, retention, and satisfaction.

It is considered a challenge but at the same time a solution to build an effective academic advising system which is now an essential component of any higher education system. Recently, machine learning has had great input to many fields in our lives and education is not an exception.

Two datasets were used in this project as below:

- 1- **Academic Dropout** Dataset to predict if a student will continue to be enrolled or if there is a chance, he/she will be a dropout case and what are the parameters that affect dropouts, and if there are any factors that can be avoided.
- 2- **Academic Advising** Dataset to predict who are the students who need extra advising sessions and what factors increase the likelihood of this need.

For both datasets, several machine learning algorithms were used. The Decision Tree algorithm was used for both datasets. Sharda et al., (2020) described decision trees as the most common classifier! Also, James et al., (2021) considered decision trees as an easy classifier to implement and interpret. K – nearest neighbor algorithm was applied to the first dataset with all numeric predictors, where some odd k values were tried to build the model.

On the other hand, the Random Forest algorithm was applied to both datasets, as Han et al., (2012) random forest is a collection of trees or classifiers, each split or tree is formed using a random set of features, and this is the reason it is called random forest.

Finally, Neural Network was applied to the second dataset. Shadra et al., (2020) said that Neural Networks are also one of the common classifiers. They accept inputs, and via some weights and calculations, they send an output to the next layer. All codes were done using R.

4.1. Academic Dropout Dataset

4.1.1 - Source of the Data:

The dataset was collected for a study about how to reduce academic dropout in higher education to identify at-risk students for developing strategies to improve student retention.

The dataset includes information about student enrollment, demographics, and the country's economic factors. The origins of the dataset (University and country) were not mentioned on Kaggle.

Link to dataset:

<https://www.kaggle.com/datasets/ankanhore545/dropout-or-academic-success?resource=download>

4.1.2 - Read the Dataset:

The dataset was read and the spaces were removed from the attribute names.

4.1.3 - Data Dimension:

Figure 7 shows that the academic dropout dataset contains 4427 records and 37 attributes.

```
[1] 4427
[1] 37
```

Figure 7 - Dataset 1 Dimension

4.1.4 - Data Structure:

Figure 8 shows the structure of the dropout dataset. The attributes data types are integer, float, and one character feature which is the response variable “Target”:

```
'data.frame': 4427 obs. of 37 variables:
 $ Marital.status      : int  1 1 1 1 2 2 1 1 1 1 ...
 $ Application.mode    : int  17 15 1 17 39 39 1 18 1 1 ...
 $ Application.order   : int  5 1 5 2 1 1 1 4 3 1 ...
 $ Course              : int  171 9254 9070 9773 8014 9991 9500 9254 9238 9238 ...
 $ Daytime.evening.attendance : int  1 1 1 1 0 0 1 1 1 1 ...
 $ Previous.qualification : int  1 1 1 1 1 19 1 1 1 1 ...
 $ Previous.qualification.grade : num  122 160 122 122 100 ...
 $ Nationality        : int  1 1 1 1 1 1 1 1 62 1 ...
 $ Mother.s.qualification : int  19 1 37 38 37 37 19 37 1 1 ...
 $ Father.s.qualification : int  12 3 37 37 38 37 38 37 1 19 ...
 $ Mother.s.occupation  : int  5 3 9 5 9 9 7 9 9 4 ...
 $ Father.s.occupation  : int  9 3 9 3 9 7 10 9 9 7 ...
 $ Admission.grade     : num  127 142 125 120 142 ...
 $ Displaced           : int  1 1 1 1 0 0 1 1 0 1 ...
 $ Educational.special.needs : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Debtor              : int  0 0 0 0 0 1 0 0 0 1 ...
 $ Tuition.fees.up.to.date : int  1 0 0 1 1 1 1 0 1 0 ...
 $ Gender              : int  1 1 1 0 0 1 0 1 0 0 ...
 $ Scholarship.holder  : int  0 0 0 0 0 0 1 0 1 0 ...
 $ Age.at.enrollment  : int  20 19 19 20 45 50 18 22 21 18 ...
 $ International       : int  0 0 0 0 0 0 0 0 1 0 ...
 $ Curricular.units.1st.sem..credited. : int  0 0 0 0 0 0 0 0 0 ...
 $ Curricular.units.1st.sem..enrolled. : int  0 6 6 6 6 5 7 5 6 6 ...
 $ Curricular.units.1st.sem..evaluations. : int  0 6 0 8 9 10 9 5 8 9 ...
 $ Curricular.units.1st.sem..approved. : int  0 6 0 6 5 5 7 0 6 5 ...
 $ Curricular.units.1st.sem..grade. : num  0 14 0 13.4 12.3 ...
 $ Curricular.units.1st.sem..without.evaluations. : int  0 0 0 0 0 0 0 0 0 ...
 $ Curricular.units.2nd.sem..credited. : int  0 0 0 0 0 0 0 0 0 ...
 $ Curricular.units.2nd.sem..enrolled. : int  0 6 6 6 6 5 8 5 6 6 ...
 $ Curricular.units.2nd.sem..evaluations. : int  0 6 0 10 6 17 8 5 7 14 ...
 $ Curricular.units.2nd.sem..approved. : int  0 6 0 5 6 5 8 0 6 2 ...
 $ Curricular.units.2nd.sem..grade. : num  0 13.7 0 12.4 13 ...
 $ Curricular.units.2nd.sem..without.evaluations. : int  0 0 0 0 0 5 0 0 0 ...
 $ Unemployment.rate   : num  10.8 13.9 10.8 9.4 13.9 16.2 15.5 15.5 16.2 8.9 ...
 $ Inflation.rate      : num  1.4 -0.3 1.4 -0.8 -0.3 0.3 2.8 2.8 0.3 1.4 ...
 $ GDP                 : num  1.74 0.79 1.74 -3.12 0.79 -0.92 -4.06 -4.06 -0.92 3.51 ...
 $ Target              : chr  "Dropout" "Graduate" "Dropout" "Graduate" ...
```

Figure 8 - Dataset 1 Structure

4.1.5 - Data Summary:

Figure 9 shows the dropout dataset summary, some NAs values were found, which were dealt with in the next parts.

Marital.status	Application.mode	Application.order	Course	Daytime.evening.attendanc	Curricular.units.1st.sem..approved.	Curricular.units.1st.sem..grade.	
Min. :1.000	Min. :1.00	Min. :0.000	Min. :33	Min. :0.0000	Min. :0.000	Min. :0.00	
1st Qu.:1.000	1st Qu.:1.00	1st Qu.:1.000	1st Qu.:9085	1st Qu.:1.0000	1st Qu.:3.000	1st Qu.:11.00	
Median :1.000	Median :17.00	Median :1.000	Median :9238	Median :1.0000	Median :5.000	Median :12.29	
Mean :1.179	Mean :18.67	Mean :1.728	Mean :8857	Mean :0.8908	Mean :4.707	Mean :10.64	
3rd Qu.:1.000	3rd Qu.:39.00	3rd Qu.:2.000	3rd Qu.:9556	3rd Qu.:1.0000	3rd Qu.:6.000	3rd Qu.:13.40	
Max. :6.000	Max. :57.00	Max. :9.000	Max. :9991	Max. :1.0000	Max. :26.000	Max. :18.88	
NA's :3	NA's :3	NA's :3	NA's :3	NA's :3	NA's :3	NA's :3	
Previous.qualification	Previous.qualification..grade.	Nationality	Mother.s.qualificat	Curricular.units.1st.sem..without.evaluations.	Curricular.units.2nd.sem..credited.		
Min. :1.000	Min. :95.0	Min. :1.000	Min. :1.00	Min. :0.0000	Min. :0.0000		
1st Qu.:1.000	1st Qu.:125.0	1st Qu.:1.000	1st Qu.:2.00	1st Qu.:0.0000	1st Qu.:0.0000		
Median :1.000	Median :133.1	Median :1.000	Median :19.00	Median :0.1377	Median :0.0000		
Mean :4.578	Mean :132.6	Mean :1.873	Mean :19.56	3rd Qu.:0.0000	Mean :0.5418		
3rd Qu.:1.000	3rd Qu.:140.0	3rd Qu.:1.000	3rd Qu.:37.00	Max. :12.0000	Max. :19.0000		
Max. :43.000	Max. :190.0	Max. :109.000	Max. :44.00	NA's :3	NA's :3		
NA's :3	NA's :3	NA's :3	NA's :3	Curricular.units.2nd.sem..enrolled.	Curricular.units.2nd.sem..evaluations.		
Father.s.qualification	Mother.s.occupation	Father.s.occupation	Admission.grade	Displaced	Min. :0.000		
Min. :1.00	Min. :0.00	Min. :0.00	Min. :95.0	Min. :0.00	1st Qu.:5.000		
1st Qu.:3.00	1st Qu.:4.00	1st Qu.:4.00	1st Qu.:117.9	1st Qu.:0.00	Median :6.000		
Median :19.00	Median :5.00	Median :7.00	Median :126.1	Median :1.00	Mean :6.232		
Mean :22.28	Mean :10.96	Mean :11.03	Mean :127.0	Mean :0.54	3rd Qu.:7.000		
3rd Qu.:37.00	3rd Qu.:9.00	3rd Qu.:9.00	3rd Qu.:134.8	3rd Qu.:1.00	Max. :23.000		
Max. :44.00	Max. :194.00	Max. :195.00	Max. :190.0	Max. :1.00	NA's :3		
NA's :3	NA's :3	NA's :3	NA's :3	NA's :3	Curricular.units.2nd.sem..approved.	Curricular.units.2nd.sem..grade.	
Educational.special.needs	Debtor	Tuition.fees.up.to.date	Gender	Min. :0.0000	Min. :0.00		
Min. :0.00000	Min. :0.0000	Min. :0.0000	Min. :0.000	1st Qu.:0.00000	1st Qu.:10.75		
1st Qu.:0.00000	1st Qu.:0.0000	1st Qu.:1.0000	1st Qu.:0.000	Median :0.00000	Median :12.20		
Median :0.00000	Median :0.0000	Median :1.0000	Median :0.000	Mean :0.01153	Mean :10.23		
Mean :0.01153	Mean :0.1137	Mean :0.8807	Mean :1.351	3rd Qu.:0.00000	3rd Qu.:13.33		
3rd Qu.:0.00000	3rd Qu.:0.0000	3rd Qu.:1.0000	3rd Qu.:1.000	Max. :1.00000	Max. :18.57		
Max. :1.00000	Max. :1.0000	Max. :1.0000	Max. :2868.000	NA's :3	NA's :3		
NA's :3	NA's :3	NA's :3	NA's :1	Curricular.units.2nd.sem..without.evaluations.	Unemployment.rate	Inflation.rate	
Scholarship.holder	Age.at.enrollment	International	Curricular.units.1st.sem..credited.	Min. :0.0000	Min. :7.60	Min. :0.800	Min. :-4.068
Min. :0.0000	Min. :17.00	Min. :0.00000	Min. :0.00	1st Qu.:0.00000	1st Qu.:9.40	1st Qu.:0.300	1st Qu.:-1.706
1st Qu.:0.0000	1st Qu.:19.00	1st Qu.:0.00000	1st Qu.:0.00	Median :0.00000	Median :11.10	Median :1.400	Median :0.520
Median :0.0000	Median :20.00	Median :0.00000	Median :0.00	Mean :0.2484	Mean :11.57	Mean :1.228	Mean :0.001
Mean :0.2484	Mean :23.27	Mean :0.02486	Mean :0.71	3rd Qu.:0.00000	3rd Qu.:13.90	3rd Qu.:2.600	3rd Qu.:1.790
3rd Qu.:0.00000	3rd Qu.:25.00	3rd Qu.:0.00000	3rd Qu.:0.00	Max. :1.00000	Max. :16.20	Max. :3.700	Max. :3.510
Max. :1.0000	Max. :70.00	Max. :1.00000	Max. :20.00	NA's :3	NA's :3	NA's :3	NA's :3
NA's :3	NA's :3	NA's :3	NA's :3	Target	Length:4427	Class:character	Mode:character

Figure 9 - Dataset 1 Summary

4.1.6 - Missing Values:

From the summary in Figure 9, some missing values could have affected the planned models, these missing values needed to be removed.

4.1.7 - Outliers and Leverage Points:

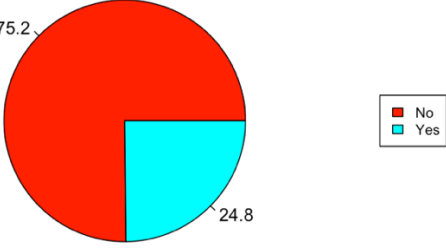
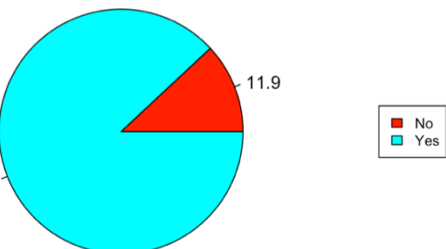
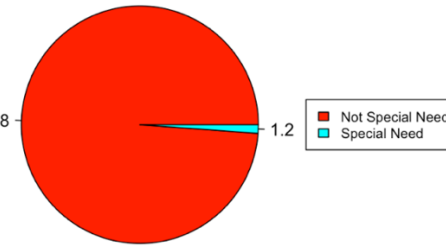
Outliers are data points with **response** values that are extreme compared to the typical observations. Running the below code to show only the values of the response “Target” feature, no strange or extreme values were found so no observations were removed.

table(dropFile\$Target)

On the other hand, leverage points are data points that have extreme **predictors** values. From the data summary (Figure 9), it is shown that the predictors seem not to have extreme values.

4.1.8 - Data Visualization:

Table 1 shows some plots of the predictors and response features.

<p>Percentage of Students by Scholarship</p>  <table border="1"><thead><tr><th>Scholarship Status</th><th>Percentage</th></tr></thead><tbody><tr><td>No</td><td>75.2%</td></tr><tr><td>Yes</td><td>24.8%</td></tr></tbody></table>	Scholarship Status	Percentage	No	75.2%	Yes	24.8%	<p>This plot shows that 24.8% of students have a scholarship and 75.2% don't have a scholarship, which is a high percentage and might affect the dropout rate.</p>
Scholarship Status	Percentage						
No	75.2%						
Yes	24.8%						
<p>Percentage of Students by Tuition Fees Up to Date</p>  <table border="1"><thead><tr><th>Tuition Fees Status</th><th>Percentage</th></tr></thead><tbody><tr><td>No</td><td>11.9%</td></tr><tr><td>Yes</td><td>88.1%</td></tr></tbody></table>	Tuition Fees Status	Percentage	No	11.9%	Yes	88.1%	<p>This plot shows that 11.9% of the students haven't paid the tuition fees, which might affect the dropout rate.</p>
Tuition Fees Status	Percentage						
No	11.9%						
Yes	88.1%						
<p>Percentage of Students by Educational Special Needs</p>  <table border="1"><thead><tr><th>Special Needs Status</th><th>Percentage</th></tr></thead><tbody><tr><td>Not Special Need</td><td>98.8%</td></tr><tr><td>Special Need</td><td>1.2%</td></tr></tbody></table>	Special Needs Status	Percentage	Not Special Need	98.8%	Special Need	1.2%	<p>This plot shows only 1.2% of the students are registered as special needs, this might have or might not have an effect on the dropout rate.</p>
Special Needs Status	Percentage						
Not Special Need	98.8%						
Special Need	1.2%						

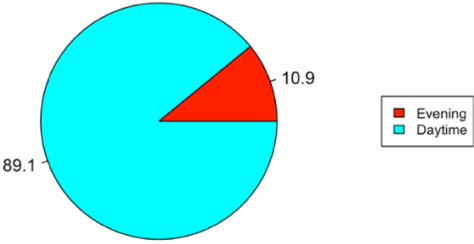
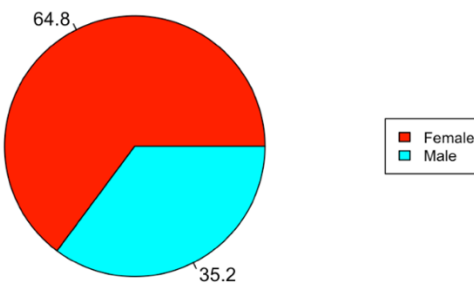
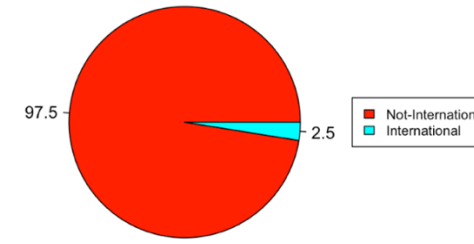
<p>Percentage of Students by Type of Attending</p>  <p>A pie chart titled 'Percentage of Students by Type of Attending'. The chart is divided into two segments: a large cyan segment representing 'Daytime' at 89.1% and a smaller red segment representing 'Evening' at 10.9%. A legend to the right shows a red square for 'Evening' and a cyan square for 'Daytime'.</p>	<p>The plot shows that 10.9% of the students are attending evening classes maybe due to employment or family reasons. This was studied to check if it has an effect on the dropout rate.</p>
<p>Percentage of Students by Gender</p>  <p>A pie chart titled 'Percentage of Students by Gender'. The chart is divided into two segments: a red segment representing 'Female' at 64.8% and a cyan segment representing 'Male' at 35.2%. A legend to the right shows a red square for 'Female' and a cyan square for 'Male'.</p>	<p>The plot shows that 64.8% are female students and 35.2% are male students.</p>
<p>Percentage of International Students</p>  <p>A pie chart titled 'Percentage of International Students'. The chart is almost entirely red, representing 'Not-International' students at 97.5%. A very small cyan slice represents 'International' students at 2.5%. A legend to the right shows a red square for 'Not-International' and a cyan square for 'International'.</p>	<p>The plot shows that only 2.5% are international students.</p>
<p>Percentage of Students by Target</p>  <p>A pie chart titled 'Percentage of Students by Target'. The chart is divided into three segments: a blue segment representing 'Graduate' at 49.9%, a red segment representing 'Dropout' at 32.1%, and a green segment representing 'Enrolled' at 17.9%. A legend to the right shows a red square for 'Dropout', a green square for 'Enrolled', and a blue square for 'Graduate'.</p>	<p>This plot is for the label attribute "Target". 49.9% are graduating students. While 17.9% are enrolled. But, 32.1% are dropout cases, which is a high percentage of around third the sample.</p>

Table 1 - Dataset 1 Visualization

4.2. Academic Advising Dataset

4.2.1 - Source of the Data:

This dataset was collected from a higher education institution in the UAE.

4.2.2 - Read the Dataset:

The dataset was read and the spaces were removed from the attribute names.

4.2.3 - Data Dimension:

Figure 10 shows the dimensions of the academic advising dataset which contains 428 records and 9 attributes.

```
[1] 428
[1] 9
```

Figure 10 - Dataset 2 Dimensions

4.2.4 - Data Structure:

Figure 11 shows the structure of the academic advising dataset. Some features are numbers while some are characters. The response variable “Extra Advising” is character:

```

$ Cgpa          : num  2.6 2.79 1.94 1.74 2.83 1.94 3.82 2.15 2.16 NA ...
$ Academic.Standing : chr  "Good Standing" "Good Standing" "Academic Probation" "Academic Probation" ...
$ Level         : int   4 4 6 8 7 2 5 4 8 1 ...
$ College.Counselling.Case: chr  "" "" "" "" ...
$ Employed      : chr  "" "" "" "" ...
$ Extra.Advising : chr  "" "" "yes" "yes" ...
```

Figure 11 - Dataset 2 Structure

4.2.5 - Data Dictionary:

Continuous attribute: 1 Attribute

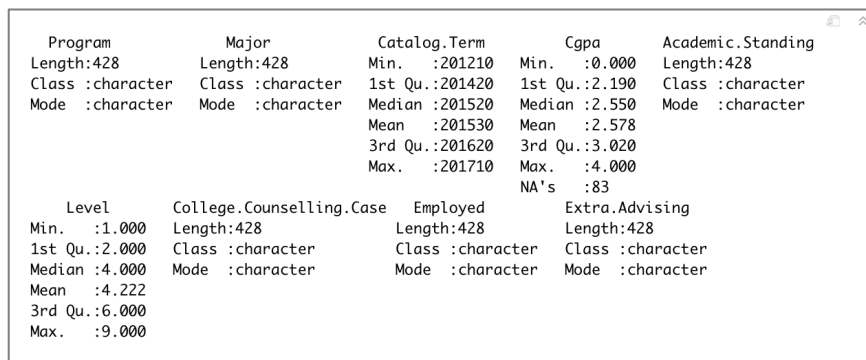
- **CGPA** a float number variable that holds the student's cumulative grade point average, ranges from 0 to 4.

Discrete attribute: 8 Attributes

- **Level:** Discrete integer variable to save the student level (semester, for example, year one is level 1 and 2, year 2 is level 3 and 4, etc.).
- **CollegeCounsellingCase:** Discrete character variable saves a binary value (yes for counselling cases).
- **Employed:** Discrete character variable saves a binary value (yes for employed student).
- **ExtraAdvising:** Discrete character variable saves a binary value (yes for extra advising needed).

4.2.6 - Data Summary:

Figure 12 shows the academic advising dataset summary, showing the minimum, 1st quartile, Mean, 3rd quartile, and maximum for numeric variables. Also showing the length and class of character variables.



Program	Major	Catalog.Term	Cgpa	Academic.Standing
Length:428	Length:428	Min. :201210	Min. :0.000	Length:428
Class :character	Class :character	1st Qu.:201420	1st Qu.:2.190	Class :character
Mode :character	Mode :character	Median :201520	Median :2.550	Mode :character
		Mean :201530	Mean :2.578	
		3rd Qu.:201620	3rd Qu.:3.020	
		Max. :201710	Max. :4.000	
			NA's :83	
Level	College.Counselling.Case	Employed	Extra.Advising	
Min. :1.000	Length:428	Length:428	Length:428	
1st Qu.:2.000	Class :character	Class :character	Class :character	
Median :4.000	Mode :character	Mode :character	Mode :character	
Mean :4.222				
3rd Qu.:6.000				
Max. :9.000				

Figure 12 - Dataset 2 Summary

4.2.7 - Missing Values:

Figure 13 shows that there are 83 missing values in the CGP Attribute, and these are for level 1 students who are new to the program so they don't have CGPA.

These NAs are level 1 students and should not be removed, but CGPA was not used in building the planned models.

Cgpa	
Min.	:0.000
1st Qu.	:2.190
Median	:2.550
Mean	:2.578
3rd Qu.	:3.020
Max.	:4.000
NA's	:83

Figure 13 – CGPA Attribute

4.2.8 - Outliers and Leverage Points:

Outliers are data points with **response** values that are extreme compared to the typical observations. Running the below code to show only the values of the response “Extra Advising” feature, no strange or extreme values were found so observations were removed.

```
table(ac$Extra.Advising)
```

On the other hand, leverage points are data points that have extreme **predictors** values. From the summary in (Figure 12), it is shown that the predictors seem not to have extreme values.

4.2.9 - Data Visualization:

Figure 14 shows that most students (52.5%) are in the “Good Standing” category, 19.3% are in the “New Students” category, 14.9% of the students are in the “Academic Warning” category, and 13% of the students are in the “Academic Probation” category.

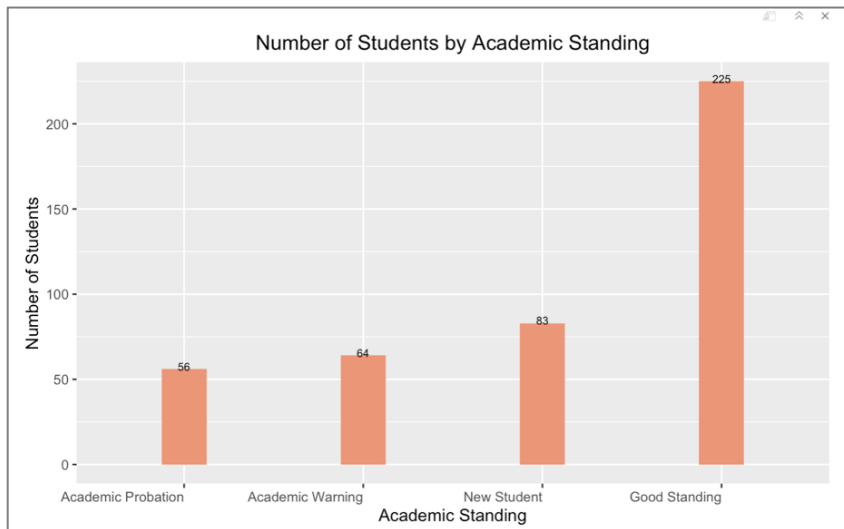


Figure 14 - Academic Standing

In Figure 15, the columns are sorted by the number of students and academic Level (year and semester of study) in the x-axis. The plot shows that around fifth of the students are in level 1 or new students (19.3%), and around 1.2% only are in level 9 (Graduating students).

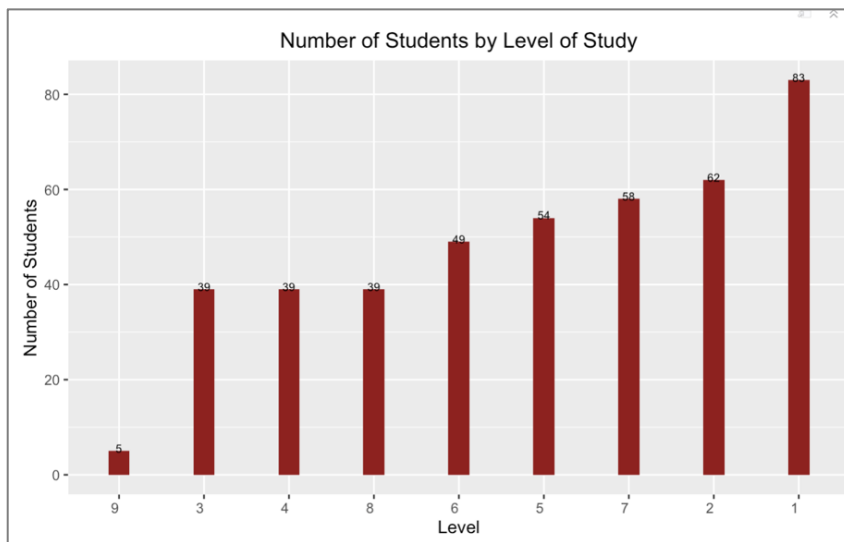


Figure 15 - Level (Year & Semester) Plot

Figure 16 shows that 95.8% of the students are not employed and only 4.2% are employed while studying.

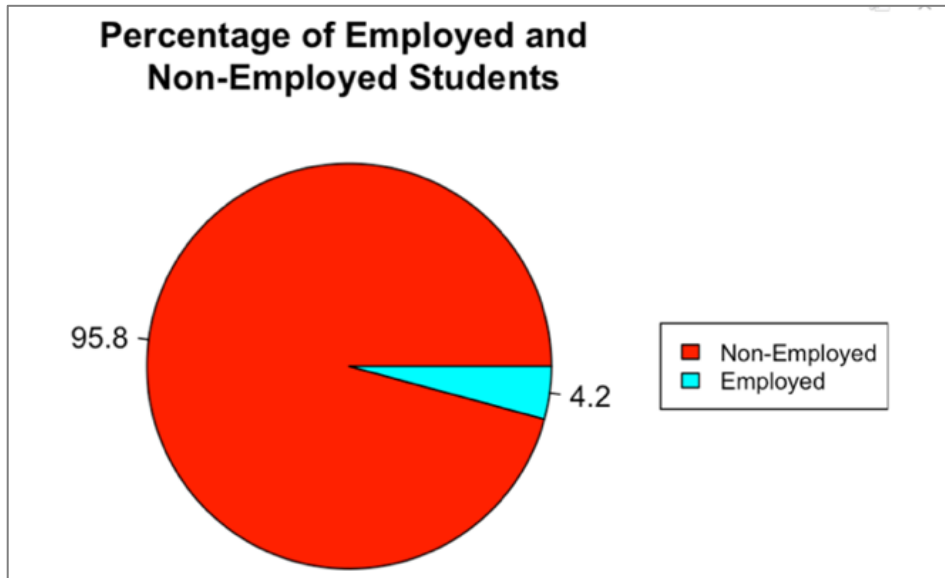


Figure 16 - Employed vs Non-Employed Plot

Figure 17 shows that only 1.9% of the students are registered as college counselling cases.

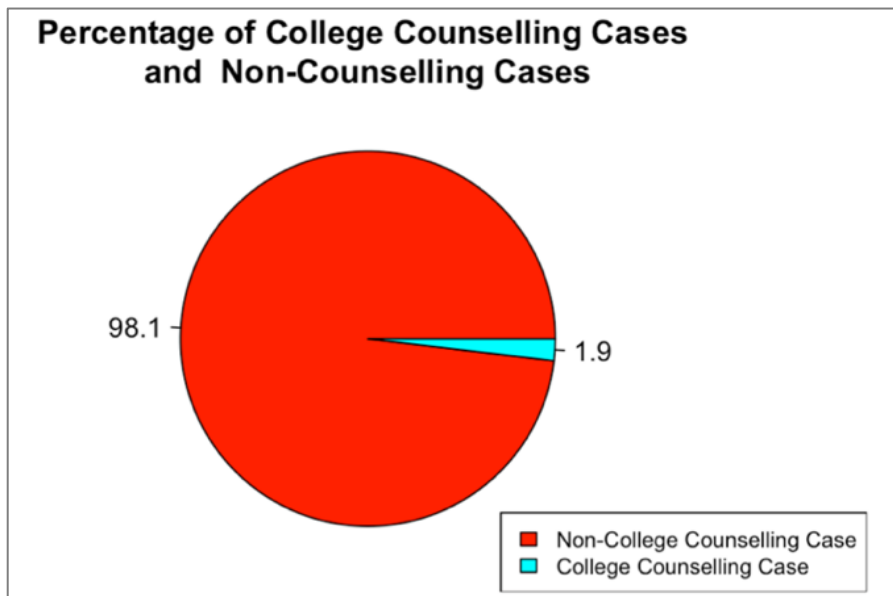


Figure 17 - College Counselling Cases Plot

Figure 18 shows that 57.9% of the students need extra advising sessions, which is more than half of the total number of students. The report studied the factors that increased this need.

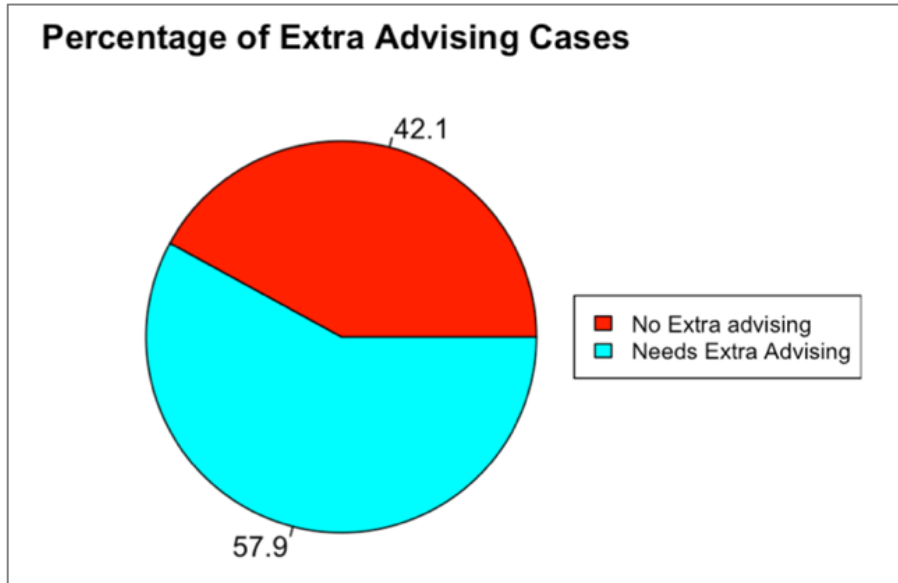


Figure 18 - Extra Advising Cases

Chapter 5- Data Analysis

5.1. Academic Dropout Data Processing

5.1.1 - Response Processing:

As can be seen from the Figure 19, the “**Graduate**” value was removed from the response attribute “**Target**” and the two values “**Dropout**” and “**Enrolled**” were maintained, to be able to predict whether a student will remain **Enrolled** or will **Dropout**.

Then, the value “**Dropout**” was replaced by 1 and the value of “**Enrolled**” was replaced by 2 for the response “**Target**”.

Dropout	Enrolled
1421	794

Figure 19 - Target Processing

5.1.2 - Variables Correlation:

Table 2 shows the correlation of the dataset attributes.

The highest **positive** correlation with the response “**Target**” is with “**Tuition fees up to date**”, which means if its value is high (**1 = paid**) then “**Target**” will be also high (**2**) which means the student is **enrolled** not drop out. Next comes “**Scholarship holder**”, “**Daytime evening attendance**”, “**Mother’s occupation**” and “**Father’s occupation**”.

The highest **negative** correlation with the response “**Target**” is with “**Age at enrollment**”, which means if its value is **low** then “**Target**” will be **high (2 Enrolled)**, which means that the student is likely to dropout at older age.

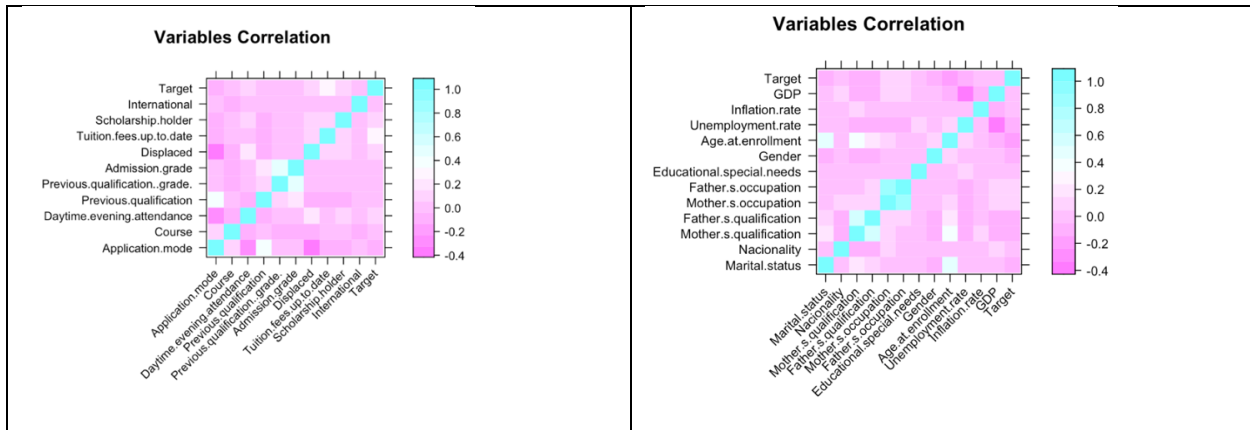


Table 2 - Academic Dropout Correlation

5.1.3 - Training and Testing Sets:

Two sets were created from the original dataset. Training set “80% of the data” was used to build the planned models while Testing set “20% of the data” was used to validate these models.

Figure 20 shows the structure of the “Training” set. It is 80% of the whole dataset with 1791 records.

```
'data.frame': 1791 obs. of 37 variables:
 $ Marital.status           : int  1 1 1 1 1 1 1 1 1 1 ...
 $ Application.mode         : int  1 18 1 18 1 18 1 1 17 44 ...
 $ Application.order        : int  5 4 1 1 1 4 1 1 2 1 ...
 $ Course                  : int  9070 9254 9238 9238 9853 9556 9238 9085 9500 9003 ...
 $ Daytime.evening.attendance : int  1 1 1 1 1 1 1 1 1 1 ...
 $ Previous.qualification   : int  1 1 1 1 1 1 1 1 1 39 ...
 $ Previous.qualification.grade : num 122 119 138 137 140 127 151 138 127 150 ...
 $ Nationality             : int  1 1 1 1 1 1 1 1 1 1 ...
 $ Mother.s.qualification   : int  37 37 1 19 19 1 19 19 3 4 ...
 $ Father.s.qualification   : int  37 37 19 38 19 38 38 19 3 5 ...
 $ Mother.s.occupation      : int  9 9 4 5 7 4 9 3 1 4 ...
 $ Father.s.occupation      : int  9 9 7 8 7 7 9 3 10 2 ...
 $ Admission.grade         : num 125 113 123 137 125 ...
 $ Displaced               : int  1 1 1 1 1 1 1 1 1 0 ...
 $ Educational.special.needs : int  0 0 0 0 0 0 0 0 0 ...
 $ Debtor                  : int  0 0 1 0 0 0 1 0 0 0 ...
 $ Tuition.fees.up.to.date : int  0 0 0 1 1 1 1 1 1 1 ...
 $ Gender                  : int  1 1 0 0 0 0 0 0 1 1 ...
 $ Scholarship.holder      : int  0 0 0 0 0 0 1 0 0 0 ...
 $ Age.at.enrollment       : int  19 22 18 18 18 20 18 18 19 21 ...
 $ International           : int  0 0 0 0 0 0 0 0 0 ...
 $ Curricular.units.1st.sem..credited : int  0 0 0 0 0 0 0 0 0 ...
 $ Curricular.units.1st.sem..enrolled : int  6 5 6 6 7 7 6 5 7 6 ...
 $ Curricular.units.1st.sem..evaluations : int  0 5 9 10 7 14 8 9 9 15 ...
 $ Curricular.units.1st.sem..approved : int  0 0 5 1 6 7 5 5 6 5 ...
 $ Curricular.units.1st.sem..grade : num 0 0 11.4 12 11.7 ...
 $ Curricular.units.1st.sem..without.evaluations : int  0 0 0 0 0 0 2 0 0 ...
 $ Curricular.units.2nd.sem..credited : int  0 0 0 0 0 0 0 0 0 ...
 $ Curricular.units.2nd.sem..enrolled : int  6 5 6 6 7 8 6 5 7 6 ...
 $ Curricular.units.2nd.sem..evaluations : int  0 5 14 14 8 9 12 7 7 17 ...
 $ Curricular.units.2nd.sem..approved : int  0 0 2 2 6 8 4 4 6 5 ...
 $ Curricular.units.2nd.sem..grade : num 0 0 13.5 11 13.5 ...
 $ Curricular.units.2nd.sem..without.evaluations : int  0 0 0 0 0 0 0 0 0 ...
 $ Unemployment.rate       : num 10.8 15.5 8.9 10.8 16.2 12.7 7.6 9.4 16.2 16.2 ...
 $ Inflation.rate          : num 1.4 2.8 1.4 1.4 0.3 3.7 2.6 -0.8 0.3 0.3 ...
 $ GDP                     : num 1.74 -4.06 3.51 1.74 -0.92 -1.7 0.32 -3.12 -0.92 -0.92 ...
 $ Target                  : num 1 1 1 2 2 2 2 2 2 ...
```

Figure 20 - Academic Dropout Training Set

Figure 21 shows the structure of the “Testing” set. It is 20% of the whole dataset with 424 records.

```
'data.frame': 424 obs. of 37 variables:
 $ Marital.status           : int 1 1 1 1 2 1 1 1 1 ...
 $ Application.mode         : int 17 1 1 39 39 7 17 1 39 15 ...
 $ Application.order        : int 5 2 1 1 1 1 1 1 1 ...
 $ Course                   : int 171 9853 9773 33 9991 9254 9070 171 9003 9130 ...
 $ Daytime.evening.attendance : int 1 1 1 1 0 1 1 1 1 ...
 $ Previous.qualification    : int 1 1 1 1 3 1 1 4 1 ...
 $ Previous.qualification.grade : num 122 133 127 130 133 ...
 $ Nacionality              : int 1 1 1 1 1 1 1 1 41 ...
 $ Mother.s.qualification   : int 19 19 19 38 37 1 1 19 3 1 ...
 $ Father.s.qualification   : int 12 37 37 37 37 3 1 19 3 1 ...
 $ Mother.s.occupation       : int 5 4 9 9 9 4 3 5 4 4 ...
 $ Father.s.occupation       : int 9 9 3 6 9 3 4 10 2 4 ...
 $ Admission.grade          : num 127 130 121 102 110 ...
 $ Displaced                : int 1 1 1 0 0 0 1 1 0 0 ...
 $ Educational.special.needs : int 0 0 0 0 0 0 0 0 0 ...
 $ Debtor                   : int 0 0 0 1 0 0 0 0 1 0 ...
 $ Tuition.fees.up.to.date  : int 1 1 1 0 1 1 1 1 1 ...
 $ Gender                   : int 1 0 0 1 0 0 1 1 1 0 ...
 $ Scholarship.holder       : int 0 0 0 0 0 0 0 0 0 ...
 $ Age.at.enrollment        : int 20 19 20 37 29 39 19 19 24 26 ...
 $ International            : int 0 0 0 0 0 0 0 0 1 ...
 $ Curricular.units.1st.sem..credited. : int 0 0 0 0 0 0 0 6 0 ...
 $ Curricular.units.1st.sem..enrolled. : int 0 6 6 7 5 6 6 0 12 6 ...
 $ Curricular.units.1st.sem..evaluations. : int 0 6 6 7 0 10 6 0 17 7 ...
 $ Curricular.units.1st.sem..approved. : int 0 0 5 0 0 0 5 0 6 2 ...
 $ Curricular.units.1st.sem..grade. : num 0 0 13.2 0 0 ...
 $ Curricular.units.1st.sem..without.evaluations. : int 0 0 0 0 0 0 0 0 1 ...
 $ Curricular.units.2nd.sem..credited. : int 0 0 0 0 0 0 0 7 0 ...
 $ Curricular.units.2nd.sem..enrolled. : int 0 6 6 7 5 6 6 0 12 6 ...
 $ Curricular.units.2nd.sem..evaluations. : int 0 0 7 7 0 6 8 0 14 6 ...
 $ Curricular.units.2nd.sem..approved. : int 0 0 0 1 0 0 1 0 10 0 ...
 $ Curricular.units.2nd.sem..grade. : num 0 0 0 10 0 0 10 0 13.3 0 ...
 $ Curricular.units.2nd.sem..without.evaluations. : int 0 0 0 0 0 0 0 0 0 ...
 $ Unemployment.rate        : num 10.8 12.7 15.5 8.9 16.2 12.4 10.8 15.5 12.4 8.9 ...
 $ Inflation.rate           : num 1.4 3.7 2.8 1.4 0.3 0.5 1.4 2.8 0.5 1.4 ...
 $ GDP                      : num 1.74 -1.7 -4.06 3.51 -0.92 1.79 1.74 -4.06 1.79 3.51 ...
 $ Target                   : num 1 1 1 1 1 2 1 1 1 ...
```

Figure 21 - Academic Dropout Testing Set

5.2. Academic Advising Data Processing

5.2.1 - Variables Correlation:

Figure 22 shows the correlation of the dataset attributes.

CGPA is **not selected** because new students don't have CGPA, and this field is empty for new students.

The highest **positive** correlation with the response “**Extra Advising**” is with “**Employed**”, which means if “**Employed**” value is **high (1)** then “**Extra Advising**” will be also **high (1)** which means **the employed student needs extra advising**.

The highest **negative** correlation with the response “**Extra Advising**” is with “**Academic Standing**”, which means if the “**Academic Standing**” value is **low (1 for new student) (2 for Probation)** then “**Extra Advising**” will be **high (1)** which means **the student needs extra advising**.

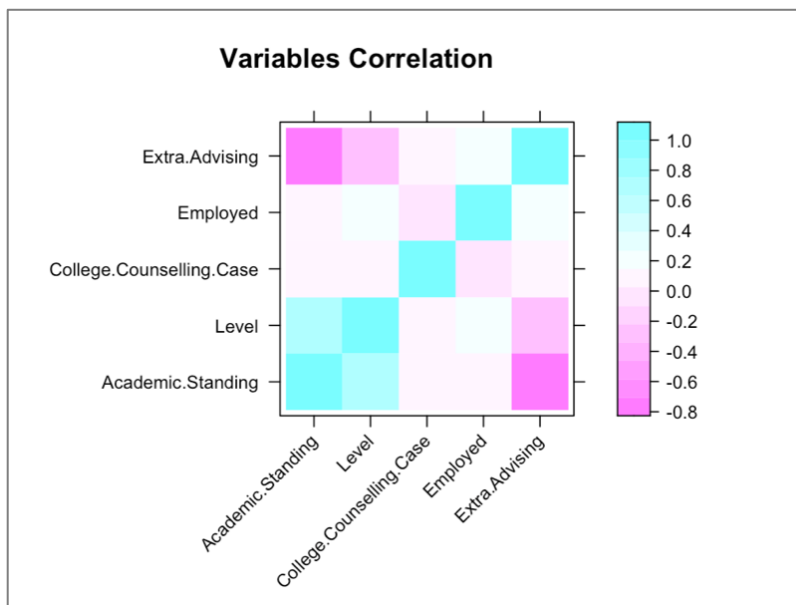


Figure 22 - Academic Advising Attributes Correlation

5.2.2 - Data Selection:

Figure 23 shows the attributes which were used in building the models from the previous part (Correlation).

Extra Advising is the label, and the selected predictors are academic standing, level, college counseling case, and employment.

```
'data.frame':  428 obs. of  5 variables:
 $ Academic.Standing      : num  4 4 3 3 4 3 4 2 2 1 ...
 $ Level                  : int  4 4 6 8 7 2 5 4 8 1 ...
 $ College.Counseling.Case: num  0 0 0 0 0 0 0 1 0 0 ...
 $ Employed               : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Extra.Advising         : num  0 0 1 1 0 1 0 1 1 1 ...
```

Figure 23 - Academic Advising Data Selection

5.2.3 –Values Replacement:

Four attribute values were replaced to build the planned models. The attributes are:

- **College counseling case:** the **empty** was replaced by **0** and “yes” was replaced by **1**.
- **Employed:** : the **empty** was replaced by **0** and “yes” was replaced by **1**.
- **Extra advising (the label):** : the **empty** was replaced by **0** and “yes” was replaced by **1**.
- **Academic standing:** The **empty** value was replaced by **1** (for **new students**), “**Academic Probation**” by **2**, “**Academic Warning**” by **3** and “**Good Standing**” by **4**.

5.2.4 - Normalization:

Normalization is done to make sure that all attributes are on the same scale. Minimum-Maximum normalization was done for Academic Standing and Level attributes.

Aksu et al., (2014), mentioned that normalization has an important role when applying artificial neural network algorithms. This is because data will be transformed into smaller intervals. The input nodes and hidden layers of the artificial neural network have weights, and if the data is left non-normalized then those weights will not be calculated accurately, especially with large scales of data values, and the model will be slower or/and will not predict accurately.

5.2.5 - Training and Testing Sets:

Two sets were created from the original dataset. Training set “80% of the data” was used to build the planned models while Testing set “20% of the data” was used to validate these models.

Figure 24 shows the structure of the “Training” set. It is 80% of the whole dataset with 354 records

```
'data.frame': 354 obs. of 5 variables:
 $ Academic.Standing : num 4 2 2 4 3 3 1 4 4 1 ...
 $ Level : int 4 6 8 5 4 8 1 6 8 1 ...
 $ College.Counselling.Case: num 0 0 0 0 1 0 0 0 0 0 ...
 $ Employed : num 0 0 0 0 0 0 0 0 0 0 ...
 $ Extra.Advising : num 0 1 1 0 1 1 1 0 1 1 ...
```

Figure 24 - Academic Advising Training Set

Figure 25 shows the structure of the “Testing” set. It is 20% of the whole dataset with 74 records.

```
'data.frame': 74 obs. of 5 variables:
 $ Academic.Standing : num 4 4 2 4 1 2 1 4 3 4 ...
 $ Level : int 4 7 2 3 1 3 1 3 6 2 ...
 $ College.Counselling.Case: num 0 0 0 0 0 0 0 0 0 0 ...
 $ Employed : num 0 0 0 0 0 0 0 0 0 0 ...
 $ Extra.Advising : num 0 0 1 0 1 1 1 0 1 0 ...
```

Figure 25 - Academic Advising Testing Set

5.3. Academic Dropout Machine Learning Models

5.3.1 - Chosen Algorithms:

The academic dropout dataset is labeled and the response is “**Target**”. This is why classification-supervised machine learning algorithms were used.

For the dropout dataset, three classifiers were applied: Random Forest, Decision Tree, and K Nearest Neighbor. Then the results were compared using the prediction accuracy, test MSE (Mean Squared Error), cost of the models, and precision to decide which one is the best model for this dataset.

5.3.2 - Random Forest Model:

A random forest algorithm was applied to the training set. After trying some predictors, the ones that gave the best results are 14 predictors, which are: Scholarship holder, Tuition fees up to date, Previous qualification grade, Age at enrollment, Mother’s occupation, Father’s occupation, Daytime evening attendance, Educational special needs, Gender, Age at enrollment, Unemployment rate, Inflation rate, GDP, and Displaced.

The below hyperparameters were used which gave together the highest accuracy:

1. Boehmke and Greenwell (2020) mentioned that in random forest models the number of trees can start with the number of predictors time 10, and according to the results it can be increased till the optimal results are reached. The model has 14 predictors, so it started with 140 trees, and then with increasing the number of trees the testing accuracy increased till the number of trees reached **1000**. More than 1000 trees reduced the accuracy while all hyperparameters are fixed.
2. Mtry (the random number of parameters used in any split) = **5**
3. Min Node size which is the minimum number of data points used in each node = **40**
4. Maximum number of Nodes (terminal nodes) = **200**

Below (Figure 26) is the plot of the minimal depth, the lowest the depth the more important the variable to the random forest model. So, **Tuition fees up to date** is the most important variable in building the model followed by **Age at enrollment**.

These two variables were found in section 5.1.2 of this report to have the highest positive and negative correlation with the response “Target” attribute.

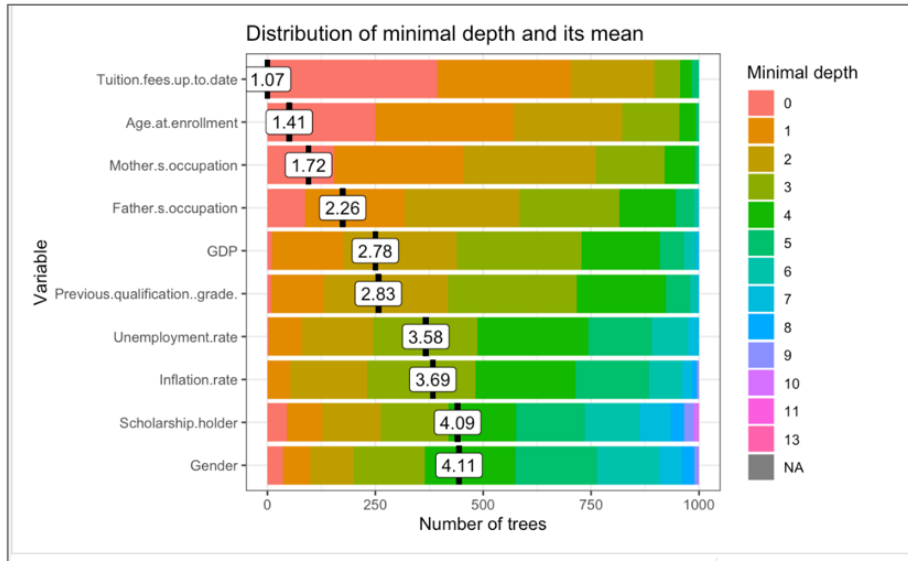


Figure 26 - Minimal Depth Academic Dropout Set

Later, the model was validated using the testing set. Figure 27 shows the confusion matrix, prediction accuracy, and test MSE of the model. Prediction Accuracy is 70.28% and Test MSE is 0.297, more classifiers were applied next.

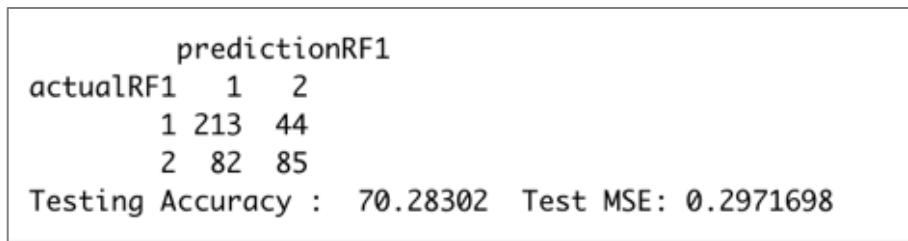


Figure 27 - Result of Random Forest Model for Academic Dropout

5.3.3 - Decision Tree Model:

The decision tree algorithm was applied to the training set of the Dropout dataset. **Predictors** used are Scholarship holder, Tuition fees up to date, Previous qualification grade, Age at enrollment, Mother’s occupation, Father’s occupation, Daytime evening attendance, Educational special needs, Gender, Age at enrollment, Unemployment rate, Inflation rate, and GDP.

Figure 28 shows the decision tree plot, the model selected “Tuition fees up to date”, “Mother’s occupation”, “Age at enrollment”, “Father’s occupation”, and “Gender” from the provided above predictors in the model.

If the tuition fee up to date is 0 (not paid), then the student is more likely to drop out. While if the “tuition fees up to date” is 1 (paid), then the model checks the mother's occupation and then the student's age to predict if he/she is more likely to drop out.

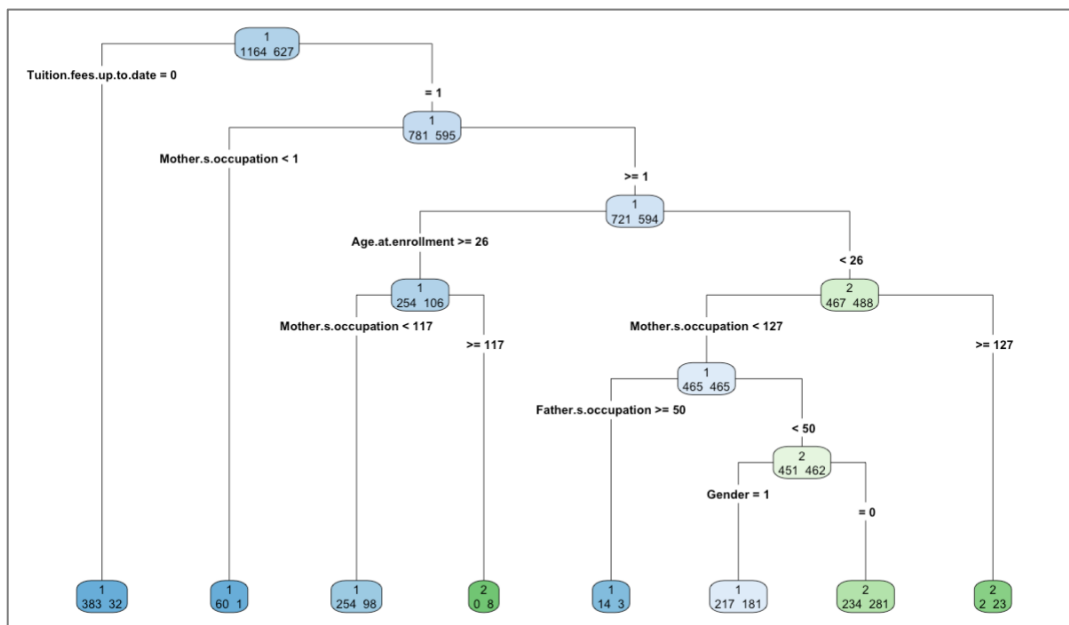


Figure 28 - Decision Tree - Academic Dropout

Later, the model was validated using the testing set. Figure 29 shows the confusion matrix, prediction accuracy, and test MSE of the model. Prediction Accuracy is 65.8% which is lower than the random forest model and Test MSE is 0.341 which is higher than the random forest model.

```

DToutput
  1  2
1 191 66
2  79 88
Testing Accuracy : 65.80189 Test MSE: 0.3419811

```

Figure 29 - Decision Tree Model Results for Academic Dropout

5.3.4 - K-Nearest Neighbor Model:

KNN algorithm was applied, and after some tries of the k value (the number of nearest neighbors to include), k=15 gave the best prediction accuracy for this model.

Figure 30 shows the confusion matrix, prediction accuracy, and test MSE of the model. Prediction Accuracy is 70.754% which is the highest compared to the random forest and decision tree models and Test MSE is 0.292 which is the lowest compared to the random forest and decision tree models.

Results were finalized and compared to decide the selected model for the academic dropout dataset in section 5.5.3 of this report.

```

tar1
knn1  1  2
  1 213 80
  2  44 87
Testing Accuracy : 70.75472 Test MSE: 0.2924528

```

Figure 30 - KNN Model Results for Academic Dropout

5.4. Academic Advising Machine Learning Models

5.4.1 - Chosen Algorithms:

The academic dropout dataset is labeled and the response is “**Extra Advising**”. This is why supervised machine learning classification algorithms were used.

For the academic advising dataset, three classifiers were applied: Random Forest, Decision Tree, and Artificial Neural Network. Then the results were compared using the prediction accuracy, test MSE (Mean Squared Error), cost of the models, and precision to decide which one is the best model for this dataset.

CGPA was not used in building the models, because level 1 students have no CGPA. Extra Advising was used as the label, while Employed, College Counselling Case, Level, and Academic Standing were used as predictors.

5.4.2 - Random Forest Model:

A random forest algorithm was applied to the training set. Predictors used are Level, College Counselling Case, and Academic Standing.

The below hyperparameters were used which gave together the highest accuracy:

1. Number of trees = 100
2. Mtry (the random number of parameters used in any split) = 3
3. Min Node size which is the minimum number of data points used in each node = 15
4. Maximum number of Nodes (terminal nodes) =50

Figure 31 shows the minimal depth plot, the lowest the depth the more important the variable to building the model. So, Academic standing is the most important variable in building the random forest model followed by Level, then the College Counselling case.

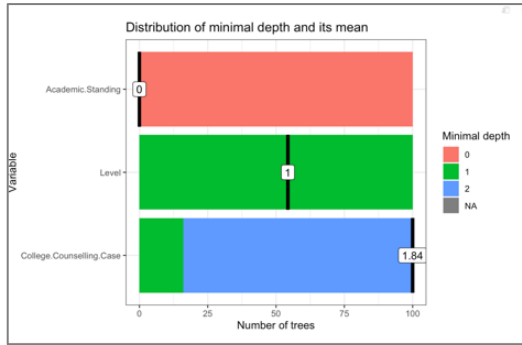


Figure 31 - Random Forest Minimal Depth Plot - Dataset 2

Figure 32 shows that the optimal number of trees for this data is 100 which gave the lowest error rate.

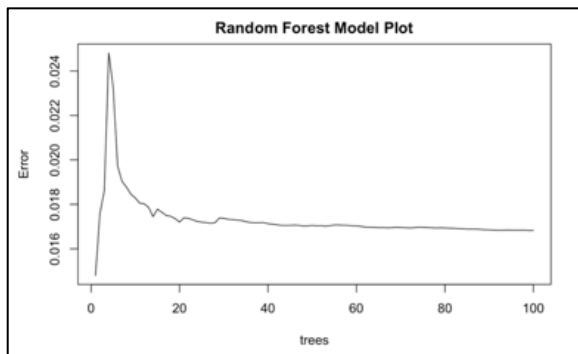


Figure 32 - Random Forest Model Optimal Number of Trees

Figure 33 shows that more than 40 trees have between 12 and 15 nodes, around 12 trees have less than 10 nodes and around 2 trees have 25 nodes

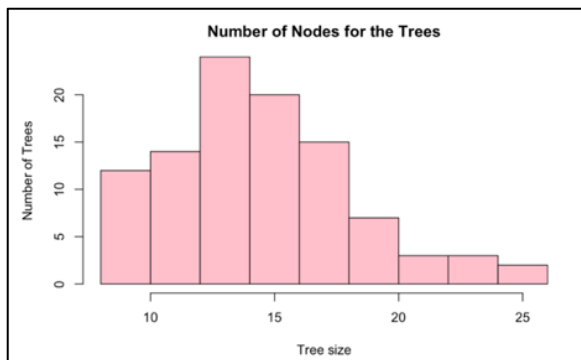


Figure 33 - Number of Trees and Nodes

Later, the model was validated using the testing set. Figure 34 shows the confusion matrix, prediction accuracy, and test MSE of the model. Prediction Accuracy is 97.297% and Test MSE is 0.027, results are very good but were compared to the next classifiers in section 5.5.4.

		predictionRF	
actualRF	0	1	
0	31	0	
1	2	41	
Testing Accuracy :		97.2973	Test MSE: 0.02702703

Figure 34 - Random Forest Results for Academic Advising

5.4.3 - Decision Tree Model:

A decision tree algorithm was applied to the training set. **Predictors** used are Employed, College Counselling Case, and Academic Standing.

Figure 35 shows the decision tree plot, if academic standing (is less than 4): 1 (new student), 2 (academic probation), or 3 (academic warning) then there is a need for extra academic advising.

If academic standing is 4, then the model checks whether the student is Employed or not, extra advising is needed for Employed students.

If the academic standing is 4 and the student is not employed, then the model checks if the student is registered as a college counseling case, if yes then he/she will need extra advising.

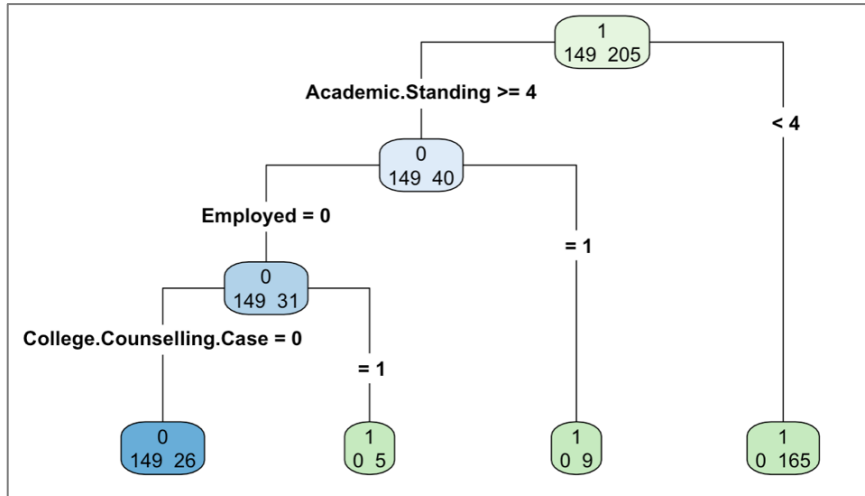


Figure 35 - Decision Tree for Academic Advising

Later, the model was validated using the testing set. Figure 36 shows the confusion matrix, prediction accuracy, and test MSE of the model. Prediction Accuracy is 95.946% which is lower than the random forest model and Test MSE is 0.041 which is higher than the random forest model. One more model (neural network) was applied to this dataset in the next page of this study.

DTtest	
actual\DT	0 1
0	31 0
1	3 40
Testing Accuracy : 95.94595 Test MSE: 0.04054054	

Figure 36 - Decision Tree Results for Academic Advising

5.4.4 - Artificial Neural Network Model:

An Artificial Neural Network algorithm was applied to the training set, using the **neuralnet** library. The **predictors** used are Employed, College Counselling Case, and Academic Standing. Four hidden layers were used.

Below (Figure 37) is the plot of the model, which shows 4 hidden layers. The black lines show the connections between layers with the weights, while the blue lines show the bias which is a kind of intercept of the model.

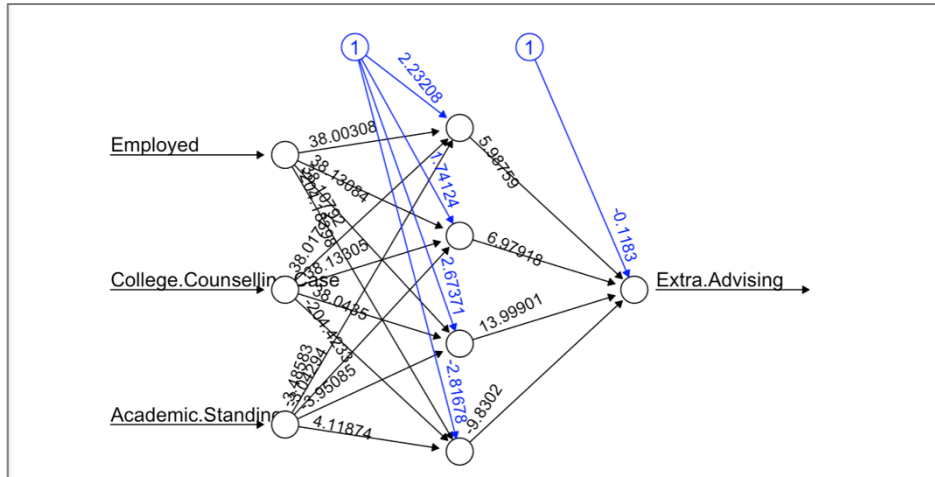


Figure 37 - ANN Model for Academic Advising

Later, the model was validated using the testing set. Figure 38 shows the confusion matrix, prediction accuracy, and test MSE of the model. Prediction Accuracy is 90.2439% which is the lowest compared to the random forest and decision tree models and Test MSE is 0.09756 which is the highest compared to the random forest and decision tree models.

Results were finalized and compared to decide the selected model for the academic advising dataset in section 5.5.4

predictionANN		
actualANN	0	1
0	29	0
1	8	45
Testing Accuracy : 90.2439 Test MSE: 0.09756098		

Figure 38 - ANN Results for Academic Advising

5.5. Results

5.5.1 - Cost of Models:

Gorunescu (2011) mentioned that comparing classifiers is not an easy task! Each classifier can work better compared to other classifiers depending on the problem and the dataset the model is trying to solve.

The author mentioned that classification models are better compared not only using the accuracy but also the cost which is associated with the correct or incorrect classifications. A cost matrix can be created to minimize the model cost or maximize benefit. Table 3 shows an illustration of a cost matrix.

COST MATRIX	Predicted Class		
		Class= No	Class = Yes
Actual Class	Class = No	A	B
	Class= Yes	C	D

Table 3 - Cost Matrix Illustration

The model cost can be then calculated as:

cost = Number of True Negatives * **A** + Number of False Positives * **B** + Number of False Negatives * **C** + Number of True Positives * **D**

In summary, the cost purpose is to **reward** the correct classifications and **penalize** misclassifications. The lower the cost the better the performance of a model.

Gyanchandani et al., (2014) mentioned that the cost matrix depends on the field of the problem. For example, in a loan problem, the cost of rejecting a customer who will not pay back is lower than the cost of accepting a customer who will not pay back. So, it is like a balance between gains and losses.

As mentioned by Tenini (2019), one way to calculate the cost matrix weights is shown below:

$$0 \quad N(0)/N(1)$$

$$1 \quad 0$$

For part 1 of this study, the dataset has the below response values:

Dropout (1) Enrolled (0)

1421 794

$$N(0)/N(1) = 0.558761436$$

Table 4 shows the cost matrix that was used to calculate and compare the cost of the classifiers in parts 1 of this study.

WEIGHTS	Predicted Class		
Actual Class		No	Yes
	No	0	0.558761436
	Yes	1	0

Table 4 – Part 1 Cost Matrix

For part two of this project, the dataset has the below response values:

No(0) Yes(1)

180 248

$$N(0)/N(1) = 0.725806452$$

Table 5 shows the cost matrix that was used to calculate and compare the cost of the classifiers in parts 2 of this study.

WEIGHTS	Predicted Class		
Actual Class		No	Yes
	No	0	0.725806452
	Yes	1	0

Table 5 - Part 2 Cost Matrix

5.5.2 - The precision of Models:

Another measure called **Precision** was used to compare the performance of the classifiers. Sharda et al., (2020) described precision as an evaluation measure of model exactness. In other words, it is the proportion of correctly predicted positive among the positive points. The **higher the precision the more accurate the model.**

Precision = True Positive / (True Positive + False Positive) (Table 6)

CLASSIFICATION	Predicted Class		
Actual Class		No	Yes
	No	True Negative	False Positive
	Yes	False Negative	True Positive

Table 6 - Precision True Positive and False Positive

5.5.3 - Academic Dropout Models Comparison:

Random Forest Model

```

predictionRF1
actualRF1  1  2
          1 213 44
          2  82 85
Testing Accuracy : 70.28302 Test MSE: 0.2971698
    
```

Random Forest Model Cost = $(213 * 0) + (44 * 0.558761436) + (82 * 1) + (85 * -0) = 106.586$

Random Forest Model Precision = $85 / (85 + 44) = 65.891473\%$

Decision Tree Model

```

DToutput
         1  2
1 191 66
2  79 88
Testing Accuracy : 65.80189 Test MSE: 0.3419811
    
```

Decision Tree Model Cost = $(191 * 0) + (66 * 0.558761436) + (79 * 1) + (88 * 0) = 115.878$

Decision Tree Model Precision = $88 / (88 + 66) = 57.142857\%$

K-Nearest Neighbor Model

```

tar1
knn1  1  2
      1 213 80
      2  44 87
Testing Accuracy : 70.75472 Test MSE: 0.2924528
    
```

K-Nearest Neighbor Model Cost = $(213 * 0) + (80 * 0.558761436) + (44 * 1) + (87 * 0) = 88.701$

K-Nearest Neighbor Model Precision = $87 / (87 + 80) = 52.095808\%$

According to table 7, random forest and KNN models compete, their prediction accuracy and test MSE are almost similar but the KNN model cost is lower and the random forest model precision is higher!

The **K-Nearest Neighbor model** was selected as it has the highest prediction accuracy, the lowest test MSE, and the lowest cost, although the random forest model has the highest precision.

Model	Prediction Accuracy	Test MSE	Precision	Cost
Random Forest	70.283%	0.29716	65.891473%	106.586
Decision Tree	65.802%	0.34198	57.142857%	115.878
KNN	70.754%	0.29245	52.095808%	88.701

Table 7 - Academic Dropout Models Comparison

5.5.4 - Academic Advising Models Comparison

Random Forest

```

predictionRF
actualRF 0 1
0 31 0
1 2 41
Testing Accuracy : 97.2973 Test MSE: 0.02702703
    
```

$$\text{Random Forest Model Cost} = (31 * 0) + (0 * 0.725806452) + (2 * 1) + (41 * 0) = 2$$

$$\text{Random Forest Model Precision} = 41 / (41 + 0) = 100\%$$

Decision Tree

```

DTtest
actualDT 0 1
0 31 0
1 3 40
Testing Accuracy : 95.94595 Test MSE: 0.04054054
    
```

$$\text{Decision Tree Model Cost} = (31 * 0) + (0 * 0.725806452) + (3 * 1) + (40 * 0) = 3$$

$$\text{Decision Tree Model Precision} = 40 / (40 + 0) = 100\%$$

Artificial Neural Network

```

predictionANN
actualANN 0 1
0 29 0
1 8 45
Testing Accuracy : 90.2439 Test MSE: 0.09756098
    
```

$$\text{Artificial Neural Network Model Cost} = (29 * 0) + (0 * 0.725806452) + (8 * 1) + (45 * 0) = 8$$

$$\text{Artificial Neural Network Model Precision} = 45 / (45 + 0) = 100\%$$

According to table 8, the **Random Forest model** was selected as it has the highest prediction accuracy, the lowest test MSE, 100% precision (the three models have 100% precision) and the lowest cost.

Model	Prediction Accuracy	Test MSE	Precision	Cost
Random Forest	97.2973%	0.0270	100%	2
Decision Tree	95.9459%	0.040	100%	3
ANN	90.243%	0.097	100%	8

Table 8 - Academic Advising Models Comparison

Chapter 6- Conclusion

6.1 Conclusion

This study's focus was to explore one of the serious challenges that higher education institutions are facing which is the increase in student dropout rates worldwide. Losing students affect institutions' revenue and reputation. Higher education student retention is of high importance to students who invest their time and resources in the hope of earning a degree.

This study used Machine Learning algorithms to help academic advisors to achieve their objectives to improve graduation rates, decrease student dropout rates, and decrease the loss of tuition revenues caused by students who either drop out or transfer to another school.

This study explored the challenges of Covid-19 on Higher Education, academic advising strategies, and academic advising challenges. The study also explored Machine learning models for academic advising.

The study applied machine learning models on two datasets, one dataset is an online dataset about student dropout and the other dataset is from one Higher Education institution in the UAE. The main purpose was to develop an effective academic advising recommendation strategy for Higher Education institutions which will be recommended to the UAE Ministry of Higher Education.

The K nearest neighbor model outperformed the random forest and the decision tree models for the academic dropout dataset. While the random forest model outperformed the decision tree and the neural network models for the academic advising dataset.

In earlier parts of this research, the below questions were proposed to be answered by the proposed models:

1. What are the most important factors that indicate that a student is more likely to drop out?

The study found that the most important factors that indicate that a student is more likely to drop out are Tuition fees up to date and Age at enrollment.

2. What are the most important factors that indicate that a student needs extra academic advising?

The study found that the most important factors that indicate that a student needs extra academic advising are Academic standing, Level, and College Counselling case.

3. How accurately can the proposed model predict the dropout?

For the dropout prediction, the K Nearest Neighbor model has the best results with 70.754% prediction accuracy.

4. How accurately can the proposed model predict the extra academic advising need?

For academic advising prediction, the random forest model has the best results with 97.297% prediction accuracy.

5. What can higher education institutions do to decrease dropout rates?

The answer to this question is explained in detail in the next section (6.2 Recommendations).

6.2 Recommendations

To be able to advise students properly and create a personal advising plan for each student, advisors have to be aware of the below outputs that affect the percentage of student dropouts, and the need for extra academic advising sessions. Academic advising must be changed to a data-driven decision-making system using updated data which has to be continuously provided to advisors. This must be applied instead of advising based on old techniques which deal with all students the same, by meeting them only once per semester or even having no meetings, and just sending them reminders about some important dates!

Each part of my study allowed the researcher to extract some outputs and conclude the below outcomes and recommendations:

6.2.1 - Academic Dropout Management:

Output.1.1: If student tuition is paid on time, then the dropout percentage is low.

Recommendation.1.1: Advisors to communicate with accountants to get a list of students with non-paid tuition and follow-up, check the best scholarships and university financial plans.

Outcome.1.1: A decreased number of late or non-paid tuitions.

Output.1.2: Older age students with more family and work commitment have a high percentage of dropouts.

Recommendation.1.2: Advisors to get a list of students' ages and CGPA and assign extra advising sessions if needed and find the best plans for these students.

Outcome.1.2: A decreased number of dropouts in the older age students category.

Output.1.3: Male students have a higher percentage of dropouts due to military and work commitments. On the other hand, female students have a lower percentage of dropouts but it is still significant due to family commitments.

Recommendation.1.3: Advisors to take into consideration the variables that affect the academic results and choices of each gender. Different plans for each gender might work. Also, it is a good idea to provide different graduation pathways for example dual enrollment and night or weekend classes.

Outcome.1.3: A decreased number of dropouts based on the student's gender.

6.2.2 - Advising Extra Sessions:

Output.2.1: The lowest the academic standing of a student, the higher the percentage he/she will need extra classes.

Recommendation.2.1: Advisors to frequently check their advisees' academic results, and set more sessions for whoever has a problem or decreased results, try to find the reasons, and offer help.

Outcome.2.1: A decreased number of dropouts due to low academic standing.

Output.2.2: College counseling students and students with determination need extra classes.

Recommendation.2.2: Advisors to check frequently with campus counselors and set extra sessions, and meetings with counselors, parents, and teachers to discuss how these cases are progressing. Find what extra resources these students need and if the campus can offer them.

Outcome.2.2: A decreased number of dropouts caused by college counseling students and students with determination.

Output.2.3: Employed students need extra advising sessions.

Recommendation.2.3: Advisors to follow up with all employed students to check how they are balancing work and study, and offer them the help and support they need, for example getting excuse letters in the period of the final exams.

Outcome.2.3: A decreased number of dropouts due to student employment.

Output.2.4: New students need extra advising sessions.

Recommendation.2.4: Advisors to follow up with all new students to check how they are adjusting to the new academic life. Extra sessions, group meetings, and workshops are needed to help them to adapt.

Outcome.2.4: A decreased number of dropouts in the new student category.

6.3 Future Work

The researcher has an intention to apply the built models to more academic datasets, which might contain more observations and at the same time different types of predictors than the used predictors in this study.

The researcher is also motivated to continue seeking knowledge in the area of machine learning to find more algorithms that can help in the field of academic advising and student retention.

Finally, it would be interesting if the researcher can collaborate with more higher education institutions to be able to investigate and study the current factors that affect students' dropout rates using machine learning algorithms.

Bibliography

1. United Nations Educational, S. a. (30-July-2020). UNESCO COVID-19 Education Response . *Advocacy paper* .
2. Naughton, M. (January-December 2021,). College Advising During COVID-19. <https://journals.sagepub.com/home/ero-Vol. 7, No. 1, pp. 1-12>
DOI:<https://doi.org/10.1177/23328584211018715>. <https://journals.sagepub.com/home/ero>.
3. Marinoni, G., Land, H., & Jense, T. (May 2020). The impact of Covid-19 on Higher Education around the World - IAU Global Survey Report. *The International Association of Universities (IAU)-ISBN: 978-92-9002-212-1* .
4. Aucejo , E., Jacob, F., Araya, M., & Zafa, B. (2020). The impact of COVID-19 on student experiences and expectations: Evidence from a survey. *Journal of Public Economics* 191 (2020) 104271.
5. Tudor, T. (March 8, 2018). Fully integrating academic advising with career coaching to increase student retention, graduation rates and future job satisfaction: An industry approach. <https://doi.org/10.1177/0950422218>.
6. Gutiérrez, F., Seipp, K., Ochoa, X., Chiluiza, K., De Laet, T., & Verbert, K. (2020). A learning analytics dashboard for academic advising. *Computers in Human Behavior* 107 (2020) 105826.
7. Drake, J. (July 1, 2011). The Role of Academic Advising in Student Retention and Persistence. <https://doi.org/10.1002/abc.20062>.
8. Kuhan , T. (2008). Historical Foundation Of Academic Advising. In *Academic Advising: A Comprehensive Handbook* (pp. 3-16). san francisco: Jossey-Bass.
9. Hagen, P., & Jordan , P. (2008). Theoretical Foundation of Academic Advising. In *Academic Advising: A Comprehensive Handbook* (pp. 17-36). san francisco: Jossey-Bass.
10. Richard, M. (2008). Legal Foundation of Academic Advising. In *Academic Advising* (pp. 50-67). san francisco: Jossey-Bass.
11. Folsom, P., Yoder, F., & Joslin, J. (2015). *The New Advisor Guidebook: Mastering the Art of Academic Advising*. San Francisco: Jossey-Bass.
12. Argüello, G. (2020). In *The Impact of Covid19 on the International Education System*. Nova Southeastern University, United States: DOI: 10.51432/978-1-8381524-0-6_14.
13. Gudep, V. (2007). *Issues And Challenges In Academic Advising: A Multivariate Study Of Students' Attitudes Towards Academic Advising In United Arab Emirates (UAE)*. Sharjah, UAE.
14. Hegazy , M., & Waguih, H. (2018). *A proposed academic advisor model based on data mining classification techniques*. Cairo: DOI:10.19101/IJACR.2018.836003.
15. Nagy, H., Aly, W., & Hegazy, O. (2013). An Educational Data Mining System for Advising Higher Education Students. *International Journal of Computer, Control, Quantum and Information Engineering* Vol:7, No:10.
16. Lau, L. (2003). *Institutional factors affecting student retention*. Alabama: Project Innovation, Vol. 124, Issue 1.
17. LEWIS , P., THORNHILL , A., & SAUNDE, M. (8th ed, 2019). *RESEARCH METHODS FOR BUSINESS STUDENTS*. Harlow CM17 9NA, UK: Pearson Education Limited.
18. Diertrich, D., Heller, B., & Yang, B. (2015). *Data Science & Big Data Analytics, Discovering, Analyzing, Visualizing and Presenting Data*. Indianapolis: John Willey & Sons.
19. Crookston, B. (1994). A Developmental View of Academic Advising As Teaching. *NACADA Journal*, doi.org/10.12930/0271-9517-14.2.5.

20. Wilder, M. B. (2016). *A Qualitative Case Study on Intrusive Academic Advising*. Semantic Scholar.
21. Oertel , B. (2020). *CREATING THE CASE FOR A NEW ACADEMIC ADVISING MODEL at Winona State University*. Minnesota.
22. Naughton, M. (January-December 2021). College Advising During COVID-19. <https://journals.sagepub.com/home/ero-Vol. 7, No. 1, pp. 1-12>
DOI:<https://doi.org/10.1177/23328584211018715>. <https://journals.sagepub.com/home/ero>.
23. McFarlane, R., & R.Wallder, S. (2021). *Engaging Students for Success during a Pandemic: The Impact of Academic Advisement upon Course of Study Completion Rates amongst Business & Computer Studies and Industrial Technology Students at the University of Technology, Jamaica*.
24. Arhin , V., Wang'eri , T., & Kigen , E. (Vol 7 No 3 Social Research September 2017). Academic Advising and Student Retention in Distance Learning: The Case of University of Cape Coast, Ghana. *Journal of Educational* .
25. Hunter, M., & White, E. (2014). Could Fixing: Academic Advising: Fix Higher Education. Could Fixing: Academic Advising: Fix Higher Education. <https://doi.org/10.1177/108648220400900103>.
26. James , G., Witten , D., Hastie, T., & Tibshirani , R. (2021). *An Introduction to Statistical Learning: with Applications in R*. New York: Springer Science & Business Media.
27. Sharda, R., Delen , D., & Turban, E. (2020). *ANALYTICS, DATA SCIENCE, & ARTIFICIAL INTELLIGENCE SYSTEMS FOR DECISION SUPPORT*. Hoboken, NJ: Pearson Education.
28. Han, J., Kamber , M., & Pei, J. (2012). *Data Mining*. Waltham, USA: Morgan Kaufmann Publishers.
29. Brotby, K. (2006). *Information Security Governance, Guidance for Boards of Directors and Executive Management*. IT Governance Institute.
30. Schröderab, C., Kruseb, F., & Gómezb, o. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model. *ScienceDirect*.
31. HOTZ, N. (2022, August 8). *What is CRISP DM?* Retrieved from Data Science: <https://www.datascience-pm.com/crisp-dm-2/>
32. Boehmke, B., & Greenwell , B. (2020). *Hands-On Machine Learning with R*. Oxfordshire, UK: CRC Press. Taylor & Francis Group.
33. Gorunescu, F. (2011). *Data Mining*. Berlin: Springer.
34. Gyanchandani, M., Rana, J., & Yadav, R. (2014). Search for Optimized Cost matrix for Performance Enhancement of Anomaly Based Intrusion Detection System using cost sensitive classifier. *International Journal of Computaional Science*.
35. Aksu, G., Güzeller, C., & Eser, M. (2014). The Effect of the Normalization Method Used in Different Sample Sizes on the Success of Artificial Neural Network Model. *INTERNATIONAL JOURNAL OF ASSESSMENT TOOLS IN EDUCATION*.
36. Tenini, J. (2019, Feb 25). *How to do Cost-Sensitive Learning: Be right in classification modeling when it matters most*. Retrieved from medium.com: <https://medium.com/rv-data/how-to-do-cost-sensitive-learning-61848bf4f5e7>
37. HORE, A. (2022, May). *Predict Dropout or Academic Success*. Retrieved from Kaggle: <https://www.kaggle.com/datasets/ankanhore545/dropout-or-academic-success?resource=download>