Predict Student Success and Performance factors by analyzing educational data using data mining techniques

توقع عوامل أداء ونجاح الطلاب من خلال تحليل بيانتهم التعليمية باستخدام تقنيه التنقيب عن البيانات

by

MUHAMMAD ATIF

Dissertation submitted in fulfilment

of the requirements for the degree of

MSc INFORMATION TECHNOLOGY MANAGEMENT

at

The British University in Dubai

March 2022

DECLARATION

I warrant that the content of this research is the direct result of my own work and that any use made in it of published or unpublished copyright material falls within the limits permitted by international copyright conventions.

I understand that a copy of my research will be deposited in the University Library for permanent retention.

I hereby agree that the material mentioned above for which I am author and copyright holder may be copied and distributed by The British University in Dubai for the purposes of research, private study or education and that The British University in Dubai may recover from purchasers the costs incurred in such copying and distribution, where appropriate.

I understand that The British University in Dubai may make a digital copy available in the institutional repository.

I understand that I may apply to the University to retain the right to withhold or to restrict access to my thesis for a period which shall not normally exceed four calendar years from the congregation at which the degree is conferred, the length of the period to be specified in the application, together with the precise reasons for making that application.

Signature of the student

COPYRIGHT AND INFORMATION TO USERS

Abstract

Academic institutions around the globe strive to become highly reputable and make continuous efforts to improve their students' ability to gain and apply knowledge concepts in the field. The primary outcome of the academic institutions is their student's quality of education. The academic institutions are known for their outcome product that are their students work in the practical field. The educational institutions desire to have beneficial insights to ensure the success of students and to enable them to acquire knowledge and improve their abilities. This enables the institutions to retain students, graduate students on time, make students' workplace ready and improve the institution's reputation. The primary aim of the study is to identify key attributes that contribute to the performance of the student. Past research has mainly focused on data related to student academic assessments grades, GPA, and student demographics. The research study includes more aspects like the number of students in class, attendance of the student in class, and due to the fact that the United Arab Emirates is a diversified multicultural country, English Language Proficiency, nationality and age of students and the instructor contributes towards student performance. The research study is performed as experimental analysis and develop models from nine machine learning algorithms including KNN, Naïve Bayes, SVM, Logistic regression, Decision Tree, Random forest, Adaboost, Bagging Classifier, and voting Classifier. The model is then applied to data collected from a reputable university that included 126,698 records with twenty-six (26) initial data attributes. The results show that the Random forest model performed better in terms of accuracy of 90.12% as compared to other models. The attendance in class attribute showed positive correlation while the number of students in class attribute showed negative correlation with the grades. The Future enhancement of the research study is to include more attributes from various aspects and also to further the study to provide recommendations for the students, instructor, and the educational institution.

Keywords: Education Data mining, EDM, Machine Learning, Student Performance Prediction, KNN, Naïve Bayes, Random Forest, Support Vector Machines, Logistic Regression, Decision Tree, Feature Importance.

الملخص

تسعى المؤسسات الأكاديمية في جميع أنحاء العالم إلى أن يذيع صيتها وتبذل جهودًا متواصلة في سبيل تحسين قدرة طلابها على اكتساب وتطبيق مفاهيم المعرفة في هذا المجال. إن جودة تعليم طلاب المؤسسات الأكاديمية هي النتيجة الأولية لجهود تلك المؤسسات وما يزيد من شهرة المؤسسات الأكاديمية، دخول طلابها في مجال العمل وتطبيق ما تمت دراسته. كما وترغب المؤسسات التعليمية في الحصول على رؤى مفيدة لضمان نجاح الطلاب وتمكينهم من اكتساب المعرفة وتحسين قدراتهم مما يتيح ذلك للمؤسسات الحصول على ولاء الطلاب واستمرارهم في مواصلة تلقي تعليمهم في تلك المؤسسات وتساهم في تخريج الطلاب في الوقت المحدد وتهيئة مكان عمل للطلاب الخريجين مما يساعد في تحسين سمعة المؤسسة. إن الهدف الأساسي من الدراسة هو تحديد السمات الرئيسية التي تحدد أداء الطالب. لقد ركزت الأبحاث السابقة بشكل محوري على البيانات المتعلقة بدرجات التقييمات الأكاديمية للطلاب والمعدل التراكمي والتركيبة السكانية للطلاب. بينما تتضمن هذه الدراسة البحثية جوانب أكثر مثل عدد الطلاب في الفصل، ونسبة حضور الطالب في الفصل وبما أن الإمارات العربية المتحدة تعد بلد متعدد الثقافات فإن إتقان اللغة الإنجليزية، وجنسية وعمر الطلاب والمعلمين كذلك يساهم بشكل أو بأخر في أداء الطلاب. يتم إجراء الدراسة البحثية كتحليل تجريبي وتطوير نماذج من تسعة خوارزميات للتعلم الآلي بما في ذلك KNN و Naïve Bayes و SVM و Logistic regression, Decision Tree و Random Forest و Adaboost و Bagging Classifier و voting Classifier. ثم يتم تطبيق النموذج على البيانات التي تم جمعها من إحدى الجامعات ذات السمعة المرموقة والتي تضمنت 126698 سجلًا مع ستة وعشرين (26) سمة بيانات أولية. أظهرت النتائج أن نموذج Random Forest كان أداؤه أفضل من حيث الدقة بنسبة 90.12٪ مقارنة بالنماذج الأخرى. أظهرت سمة الحضور في الفصل ارتباطًا إيجابيًا بينما أظهر عدد الطلاب في الصف ارتباطًا سلبيًا بالدرجات. يتمثل التعزيز المستقبلي للدراسة البحثية في تضمين المزيد من السمات والجوانب المختلفة في العملية التعليمية التي قد تؤثر على أداء الطلاب وأن تقوم الدراسة بتقديم توصيات للطلاب والمدرسين والمؤسسات التعليمية تساهم في تحسين الأداء.

Acknowledgement

All praise is for Allah that He provided me the ability and opportunity to carry out the research. Without the will of Allah this would have not be possible. I am greatly thankful to Allah for His blessings and enabling me to complete the research work.

I am sincerely grateful for the guidance and support shown by my supervisor and mentor Professor Abdallah Sherief throughout my journey of the masters degree. I am also grateful to all my teachers especially Prof. Khaled Shalan who have dedicatedly taught us well. I have gained good knowledge and learned from them.

I would like to especially thank my dearest friend and colleague Hani Abdel Hadi for the support and encouragement throughout my masters program.

Lastly but not the least, I would like to thank my family who have shown a lot of patience and made it comfortable for me to pursue my studies.

Table of Contents

List of figures

List of Tables

1 Introduction

The introductory chapter of the research provides an overview of the topic, the motivation behind conducting the experimental research, and the main objectives. The chapter also includes a brief methodology plan explained in chapter 3 in detail later on. The last section of the chapter provides the organization of the research study.

1.1 Overview

Educational institutions worldwide have recognized the advantage of using data mining techniques to predict student performance. The use of machine learning algorithms enables the institutions to take timely measures and improvements for the students to succeed [13]. The prediction of student academic performance is of significant interest for educational institutes. Education data mining (EDM) is a way educational institutes use to discover the information and perform the prediction enabling them to get early detection of student performance [9]. Artificial intelligence and data mining techniques have improved the user experience by allowing the machines to perform rigorous computations make wise decisions [26]. The nine dimensions are identified and are used to explain the models for student performance. The dimensions include Educational and performance levels, problem type, predictors and predictor type, methods, stage, scope, and explainable output type [27].

The academicians are focused on improving student achievements and exploring the factors and critical attributes that play a vital role in student progress. The large volume of data and the variety of features makes it complex to determine standard vital factors. The context of the university is essential in such explorations. Data mining techniques have shown impactful analysis that has benefited the students and educational institutions. [22]. The rate is increasing rapidly for the academic data generated. Data mining techniques are applied to reveal more helpful information for better outcomes [7][12]. The combination of data mining techniques with data analytics can provide beneficial outcomes. These learning analytics have the objective of enhancing the knowledge transfer and delivery for the academic institutions [8]

The students' results are the product of various aspects, including the learning material selection, choosing the activities, and the student's ability or potential to make progress and achieve high performance [1]. Academic institutions strive to improve retention rates, and many research analytics have been conducted in this scope. The institutions want to reduce the drop-out rates of students, and the machine learning model can assist in early detection, and institutions can take preventative remedies [38]. Graduating on time assurance can be achieved effectively by discovering the student progress factors and predicting the student performance using the in-process academic factors and assessments result. The goal can be attained by tweaking the ways of teaching and studying for the student to succeed and graduate on time [3][5].

1.2 Research Motivation

The advantage of exploring the insights using data mining techniques in predicting student performances provides the opportunity of personalized support to the students. Thus, resulting in an engaging, adaptive learning environment [24] and enabling educational institutions to implement intervention strategies [25]. The education data has a lot of hidden knowledge. Data mining techniques are applied to extract these hidden patterns and knowledge that educational institutions can use for predicting student academic performance and taking in-time measures [14]. The academicians considered the use of advanced data mining techniques to classify the students and predict their performance to be important for better advisement on the program progression and also utilized the model for selection of universities [2]. One of the key investigation areas in academics carried out is student performance prediction that allows the entities to develop strategies for effective student recruitment and retainment of the students. The factors identification for analysis and prediction remains a complicated section that is varied from the institutions and the variety and volume of data that is available [21]. The diversity of the student having different qualification backgrounds and culture, the nature and categories of the courses, and the student maturity level growth makes the prediction tasks complex and challenging [3].

The educational challenges and problems can be explored using data mining techniques and analyzing the student performance prediction [11]. The student behavior and the presence in the class have an impact on the student performance. The data mining technique can be applied to explore the hidden information and predict the performance as early detection. This enables the educational institutions to improve the quality of their education [15].

The primary and preschool institutions are not yet explored to a satisfactory extent in the context of educational data mining. The scope for primary and preschool can be analyzed to predict whether the student will continue for further higher studies [32].

The impact of the success for students reflects on the employability in the field. Unemployment can cause students to go into depression and take harmful actions. With the help of machine learning algorithms, the pattern and insights can be predicted, and proactive actions can provide successful results [29]. Educational institutions around the globe make efforts to retain students. Higher drop-out rates destroy the reputation of the educational institution. Data mining techniques can aid in identifying potential drop-out students, and necessary actions can be carried out by the educational institution in time to improve retention [35].

The use of data mining techniques provides useful information to recommend the students to select the courses depending on the impactful attributes. The recommender system can predict the successful completion of courses and allow the students for on-time graduation [10]. The hybrid classification and ensemble techniques can increase the performance of the classification models to provide better predictions [28].

1.3 Research Objectives

Academic institutions around the globe strive to make sincere efforts to improve the quality of education and, most importantly, aim to prepare their student to acquire the knowledge and wisdom of the field of study. During the past few decades, with the increase in data storage, educational data mining has become desirous and lucrative for many academicians to explore the hidden secrets and patterns for improving the student abilities and potential, ultimately extending their success achievement targets. There are great researches conducted in the past give clearly depict that the various academic and non-academic factors have an impact on student performance.

The aim of the research is

- To thoroughly study the past research in the field of education data mining
- Identify the attributes that have an impact on the student performance
- Identify and use the most used machine learning algorithms and evaluation techniques
- And provide a model of classification technique to predict the performance of the students studying in a private university in the United Arab Emirates.

1.4 Research Methodology Plan

The research is divided into the steps of Planning in order to study the past literature and research in the field of education data mining, followed by the step of data collection and preparation that includes the data wrangling process. Further, the selected algorithms will be applied, and analysis will be carried out in the Analysis phase. The final step of the research will be to deduce the conclusion of the research from the selected evaluation methods.

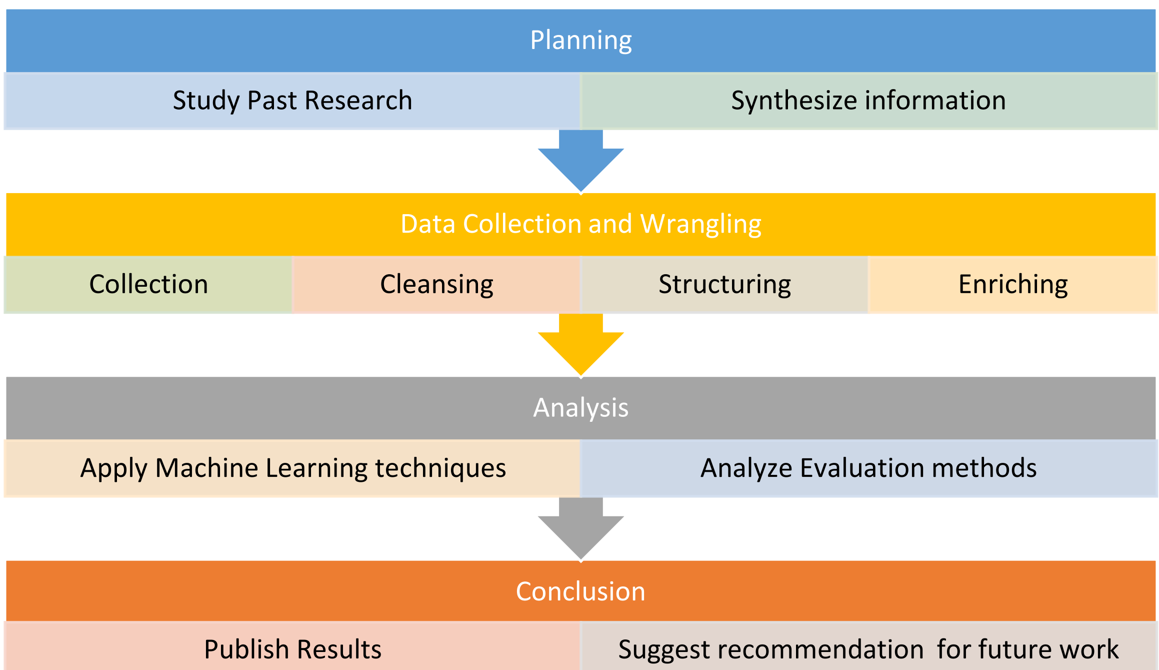


Figure 1.4-1 Research Methodology Steps

The tools that are utilized include SQL Server for data extraction, MS Excel to perform initial data exploration and filtration of records, Power BI for visualizations and data analysis, Anaconda Jupyter notebook for the execution of python code to perform the preprocessing steps and implement the machine learning algorithms and deduce the results for analysis.

1.5 Research Study Organization

The organization of the research study is planned into five chapters.

- Chapter 1 provides the introductory information and sets up the background knowledge that will be beneficial for understanding the context for the study of research.
- In Chapter 2, the review of the past research is briefly discussed. This chapter includes the distribution analysis based on the year of publications, conferences, and journals. The utilization of machine learning algorithms and evaluation techniques are depicted with the aid of illustrations. A synthesized table summary is also provided for the research papers.
- Chapter 3 explains the methodology of the research conducted. It includes the understanding of the business, discovering the Data to have more understanding, preparation of data, followed by Modeling, Evaluation and Deployment phases.
- Chapter 4 provides the analysis of the results and discussions.
- The final chapter 5 provides the concluding discussion and recommendation for future work.

2 Literature Review

This chapter discusses the theoretical perspective from the studied research papers, visual representation of the analysis of studied research papers that include year-wise research papers, conference or publisher-wise distribution, the machine learning algorithms used, and the evaluation metrics analyzed to determine the performance of the algorithms. The student attributes that are studied in the research papers are also shown in the form of a chart. Finally, the chapter provides a synthesis matrix table for the research papers.

2.1 Theoretical Perspective

The education institutions want to succeed, and they consider the success of the students is their success, and the failure of students is their failure. For improving the success rate, data mining techniques can effectively discover the hidden knowledge, solutions, and patterns that can benefit the educational institute to reduce the chance of failure and increase the opportunity of success for the students [6]. The attributes that had an impact on predicting the student potential abilities and progress were proposed in the form of a student attribute matrix [1]. The behavior features of students are important to be analyzed for the student performance prediction [19]. The EDM and Deep learning, in conjunction, can identify the weak students and provide a recommendation to enhance their academic performance [20]. The use of the tensor flow algorithm on the test data resulted in up to 91% accuracy and can be comprehensively improved with the inclusion of non-academic attributes [2]. The majority of the prediction methods rely on the achieved scores and historical information of the students. The assessment text is not primarily included in the training of the classification models. This area requires deeper exploration [4]. Education Data mining is gaining popularity in researching the useful insights and patterns that could be beneficial for educational institutes. The algorithms popularly utilized are regression, classification, association analysis, clustering, and outlier analysis [17].

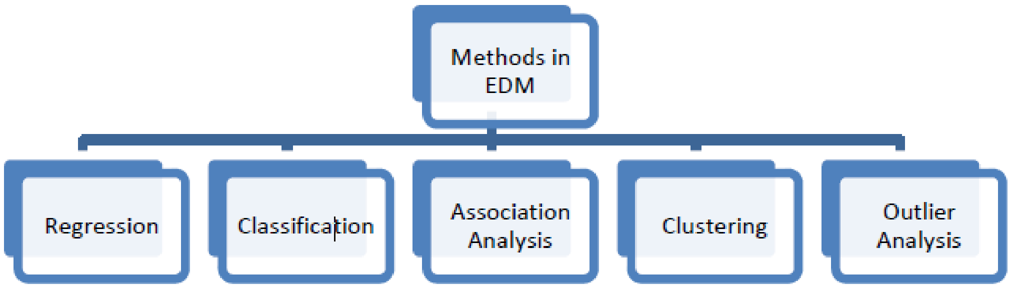


Figure 2.1-1 Educational Data Mining popular algorithms [17]

Most researches in the past have used the academic attributes collected for the prediction. The research uses additional data extracted from detailed log files related to internet activities and time used to determine the impact on the performance prediction [9]. The student performance prediction enables the educational institutes to detect the failure and perform preventive measures [16][18]. The neural network techniques can provide beneficial insights on the cognitive level of understanding for specific knowledge concepts. Academic institutions can take targets measures in improving the specific concepts of knowledge [23].

The educational institution strives to enroll quality students in order to improve its reputation in the world. Determining the student performance early at the time of admission provides ample time for the university to set a progressive path for students' success [33]. The prediction of student performance through literature review can be broadly categorized into twelve domains that include medical, Engineering, Computer Science, Chemistry, Physics, Marketing, Business administration, Sociology, Industrial Design, Landscape, Business Administration, English, and General Domains [34].

2.2 Research Papers Analysis

The research papers analysis section provides visual representation with the aid of charts and brief explanations about the chart. The sub section includes the distribution of studied research papers by publishing year and by conferences and Journals. The next sub-sections highlight the machine learning algorithms evaluation metrics being selected in the research papers. The last sub-section mentions the student attributes that were analyzed by the researchers.

2.2.1 Distribution of Research Paper

A total of forty scrutinized research papers were thoroughly studied. The below chart provides a visual illustration of the research papers in contract to their published years.

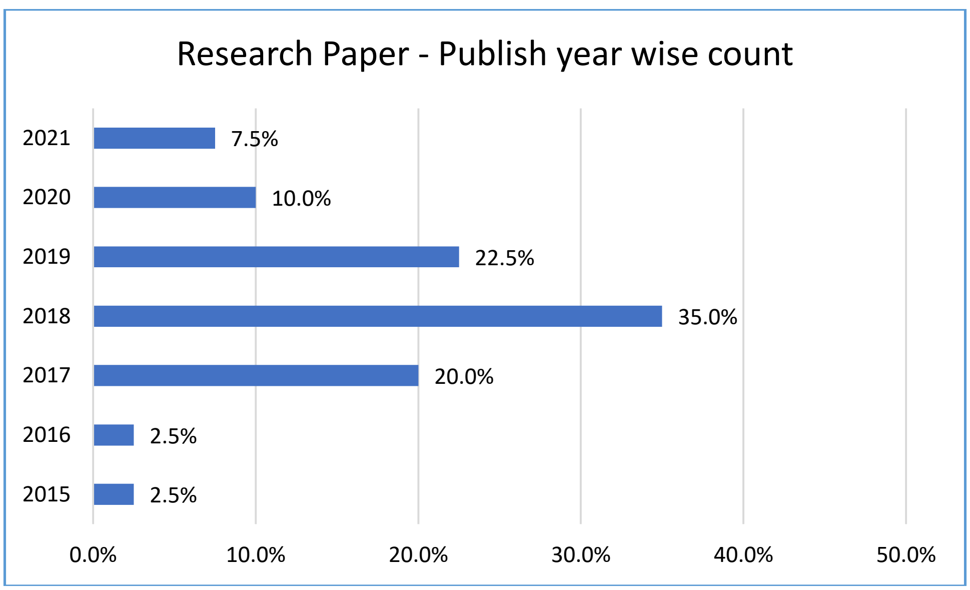


Figure 2.2-1 Research Paper - Year-wise distribution

The research papers selected for the study majorly belong to the year 2018 (35%), followed by 2019 (22.5%) and 2017 (20%).

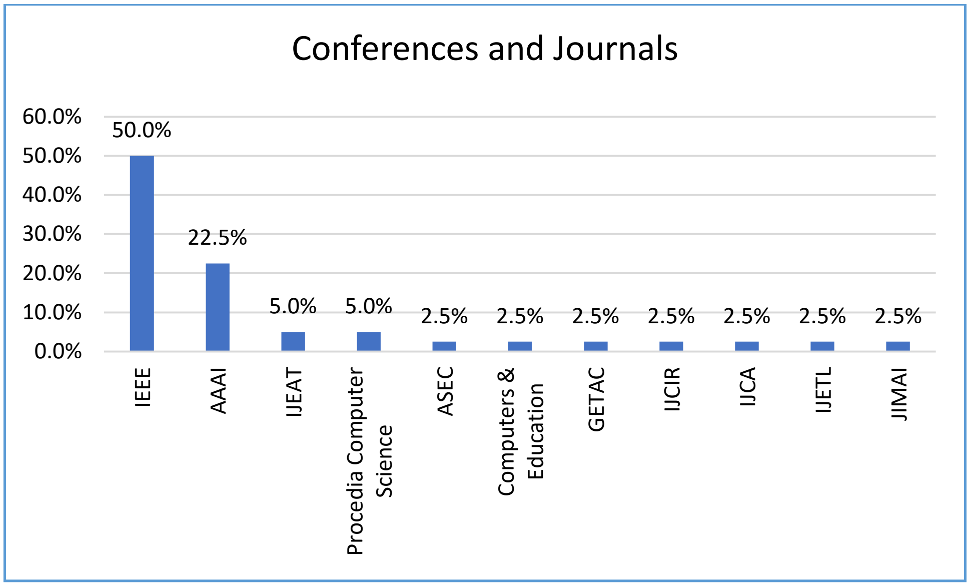


Figure 2.2-2 Conferences and Journals

The research papers selected were majorly 50% from IEEE 22.5% from AAAI. The rest of the papers were selected for a similar topic of research study purpose.

2.2.2 Machine Learning Algorithms and Techniques

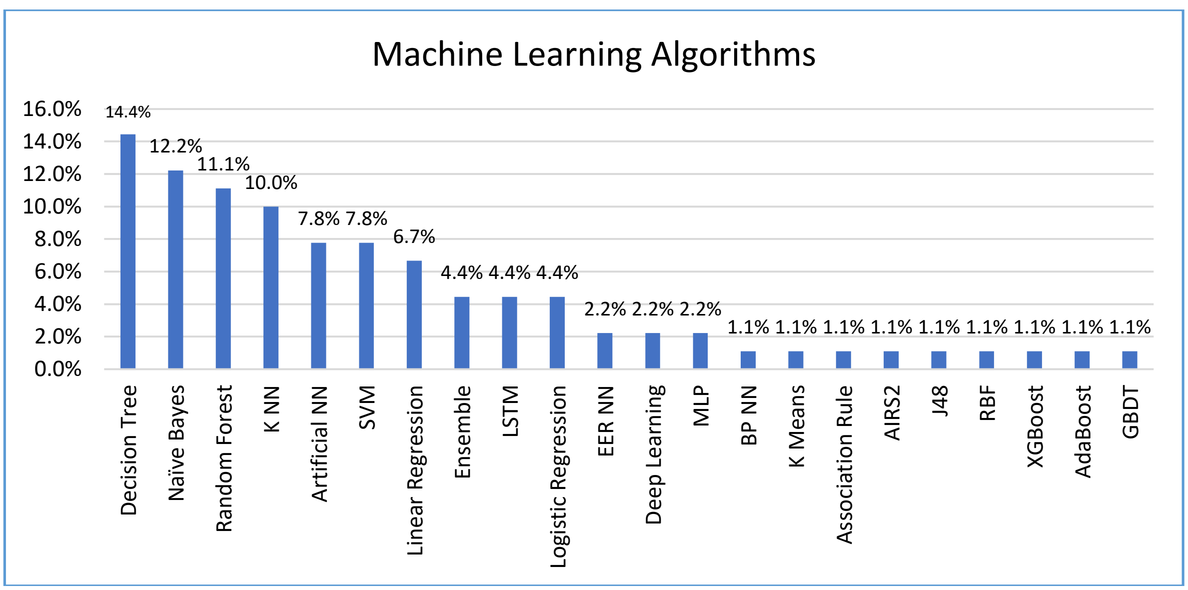


Figure 2.2-3 Machine Learning Algorithms

The studied research papers have used several machine learning algorithms and various evaluation methods. The selected research papers have majorly utilized decision trees, Naïve Bayes, Random forest, and K-NN (K-Nearest Neighbor) algorithms in their studies.

2.2.3 Evaluation Techniques

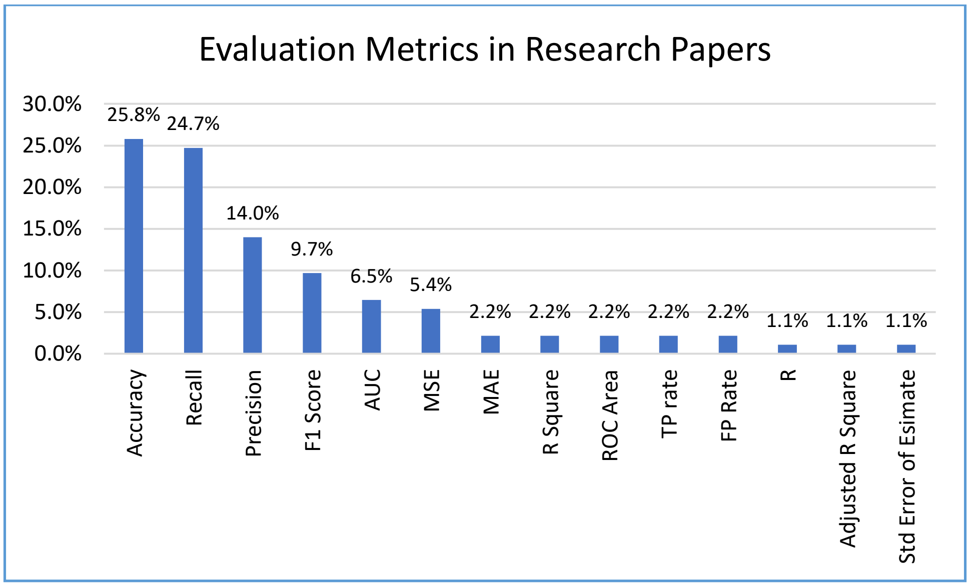


Figure 2.2-4 Evaluation Techniques

Accuracy is most popularly used as the evaluation technique, followed by the utilization of precision, recall, and F measure. The research studies that analyze classification and regression machine learning algorithms usually use these top four evaluation metrics.

2.2.4 Student Data

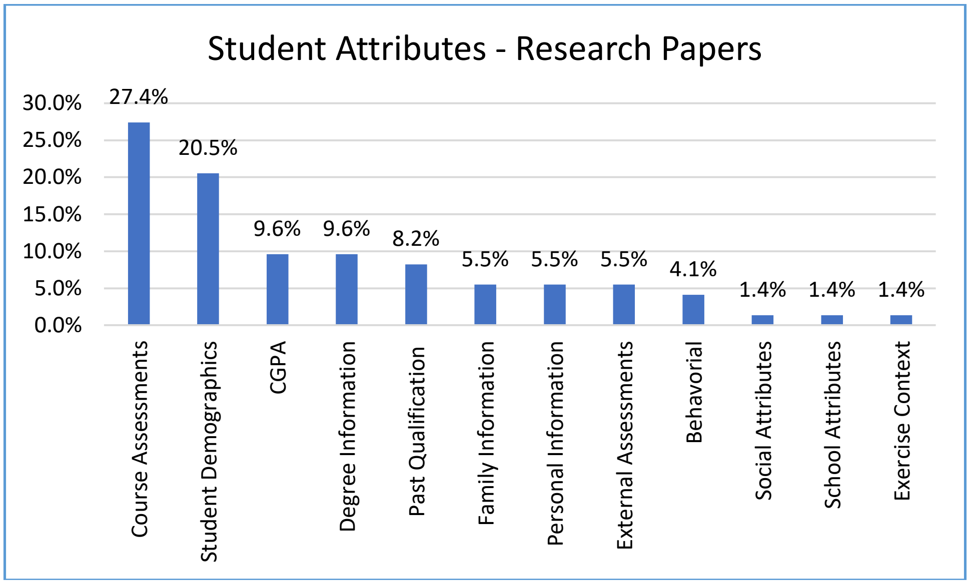


Figure 2.2-5 Student Data Attributes

Most of the Research papers have used Course Assessments as their primary input for the prediction tasks. Student Demographics followed by CGPA and Degree Information were among the majorly used attributes. The Demographics data included Gender, Date of Birth, Place of birth, city, and Marital Status. Few of the research papers used grades, grade point average, and cumulative grade point average as well. The other attributes include Secondary school name, specialization, Degree name, weight, interest, employment information, siblings, father income, accommodation, class size, and more.

2.3 Summary

RP No	Objective	Factors	Dataset Size	Algorithms	Evaluation
1	Predict Student Performance Estimation and Student Potential	Performance and Non-Performance Attributes	62 records of students	BP-NN classification	MSE (Mean Squared Error) **Test**: One-Way ANOVA and F-Test
2	Improve Advisement and University selection process by classification of student and prediction of student performance	Academic Performance	2000 Records (75% Training and 25% Testing Data)	Deep Learning Tensor Flow	Accuracy
3	To evaluate and predict the performance of the student to improve the advisement process and on-time graduation.	The academic grades of students and the course map of prerequisites.	1169 Undergraduate Data	Linear Regression, Logistic Regression, Random Forest and KNN, Ensemble-based Progressive Prediction (EPP)	MSE (Mean Square Error)
4	Predict performance based on historical data and the text of the assessment.	Academic Achievement Records, Historical data, and text of exercise.	Data was collected from http://www.zhixue.com	Exercise-Enhanced Recurrent Neural Network (EERNN), LSTM	Accuracy (ACC) and AUC (Area under the curve)
5	To predict performance progressively to satisfy on-time graduation.	Evolving GPA, Credit Hours records of students	367 Student Records.	SVM KNN and EPP	Accuracy

6	Predict Student Progress by using decision trees	Academic and Non-Academic attributes.	161 records	Decision Trees (J48, Random Tree and REPTree), Weka 3.8 tool was used.	Precision and Recall
7	Discover the hidden information from the raw data using KNN, Naïve Bayes, and Decision tree using comparative analysis.	Student Attributes, financial information, Employment information, GPA	230 Students	KNN, Naïve Bayes, and Decision Tree	Accuracy
8	Predict performance of the student by using multiple linear regression	Student Academic information	-	Multiple Linear Regression	R, R Square, Adjusted R Square, and Std. Error of Estimate **Test**: Anova
9	Analyze and predict that internet usage activities have an impact on student performance.	CGPA, Demographics, Internal & External Assessments, Psychometric test result.	360,000 Records per day. Filtered to target for 294 Student	Naïve Bayes, Logistic Regression, Neural Network, Decision Tree, Random Forest	Accuracy, Precision, Recall, F-measure, and ROC Area
10	Recommend selection of courses on time for on-time graduation.	GPA records	-	K Means, Association Rule, Naïve Bayes, SVM and KNN	Accuracy

11	Determine Student Performance using Decision Trees	Question Types Student Scores	-	Decision Trees	Accuracy
12	Predict Early Student Performance to assist Ministry of Education to improve student performance	Demographics, Degree Information and Past Qualification	2000 Records with 8 Attributes	KNN and Naïve Bayes	Accuracy
13	Reviewing and comparing prediction techniques	Student Attributes, Academic Attributes, Personal Attributes, Family & Social Attributes, School Attributes	-	Decision Trees, Neural Network, Naïve Bayes, KNN, SVM	Accuracy
15	Impact of Student behavior and presence on the performance prediction	Behavioral and student absences	460 instances with 16 Attributes	Naïve Bayes; K-Nearest Neighbor; Decision Tree; Artificial Neural Network; Ensemble Techniques	Accuracy, Precision, and Recall
16	To provide a solution for predicting early detection of student failures	-	-	Neural Network and Linear Regression	Accuracy
18	Detect the prediction of performance by tracking the	Demographics, Degree Information, Past	68 thousand records	EERNN and LSTM	Accuracy and AUC, MAE and RMSE

	student knowledge acquisition	Qualification, and Assessment Information			
19	To improve the student performance by early prediction	Demographics, Academic and Behavioral	500 Student Records with 16 Attributes	KNN, SVM, Decision tree, and Ensemble techniques.	Accuracy, Precision, Recall, and F-Measure
20	Detection of weak students and provide recommendations for performance improvement	Internal Assessment Marks	10,000 Records with 10 Attributes	Deep Learning, Recurrent Neural Networks	Accuracy, Precision, Recall, and F-Measure
23	Determine the proficiency level of students on concept levels	Exercise Details and Assessment marks	ASSISTments 2009-2010" skill builder"; Open Dataset	Artificial Neural Network	Accuracy, AUC and RMSE
24	Predict student performance for adaptive learning	Assessment Marks	181 Student records	LSTM	MSE, MAE, R^2
25	Predicting Student Performance for K 12 Education	Demographics, Past Qualification, Assessment Marks, and Class Information	403 Records with 27 attributes	Linear Regression, Decision Tree and Naïve Bayes	Accuracy
26	Predict Student Performance using Classification Algorithms	Student Grades and Streams information	72010 records of students enrolled between 2000 till 2015	Decision Tree, Random Forest, and Linear Regression	Accuracy
28	To enhance the accuracy of the classification model for predicting student	Student Demographics and course grades	480 Samples	Radial Basis Function (RBF), J48, Multilayer	TP Rate, FP Rate, Precision, Recall, F-Measure, ROC Area, Accuracy

	performance with the help of ensemble techniques			perceptron (MLP) and Random Forest, Voting Classifier	
29	To predict student employability based on their academic performance	Demographics, academic performance, and employment Data	2,133 student records	LSTM, SVM, Random Forest, GBDT and XGBoost	Accuracy, Precision, Recall, and F1 Score
30	To predict on-time graduation	Academic information, graduation on-time information	3.6M course enrollment records	Logistic Regression and Neural Network	Accuracy
31	Predict student scores for an international exam (PISA)	Academic information	4,400 participants data	Linear regression	AUC
33	To forecast student performance at the time of admission	School Grades, Admission test score and Aptitude test scores	2039 Students	Decision Tree, Support Vector Machines and Naïve Bayes	Accuracy, Recall, Precision, and F1 Score
35	Predict the drop-out students	Academic, Demographical, psychological, health, and student behavior	Dataset initial attributes of 54 that were further reduced to 24 main attributes	Decision Tree and Naïve Bayes	TP rate, FP rate, Precision, recall, F1 score, ROC Area, PRC Area, and Accuracy
36	Student Achievement Prediction in Smart Campus.	Course GPA and Internet usage information	3733 Student records	MLP , Naïve Bayes, SVM and Logistic Regression	Precision, Recall, and F1 Score

37	Develop a framework to determine the early prediction of student achievements and their effectiveness on teamwork	Basic Demographics and Team activity logs	350 Students with over 30,000 data point	Random Forest	Accuracy, precision, and recall.
38	Early detection of students at risk	Grades data, Attendance information, and portal usage logs	202 Student records	Random Forest	Accuracy, Precision, Recall, and F1 score
39	Analyze comparatively the prediction of classification algorithms	Basic Demographics and subject grades	394 students	KNN, Decision Trees, naïve Bayes, adaboost, extratree, bernaoulli naïve bayes and Random Forests	Precision, Recall, F1 Score, and AUC
40	Using Random Forest Classification from determining student performance	Demographics, Assessment grades, the Absence of information	73 Student records	Random Forest	Accuracy, Precision, Recall, and F1 Score

Table 2.3-1 Research Synthesis Matrix

3 Methodology

This chapter provides the guidelines that are followed during the research. The research methodology steps are based upon the cross-industry standard process for data mining (CRISP-DM)[1]. The research followed the six phases that are listed below

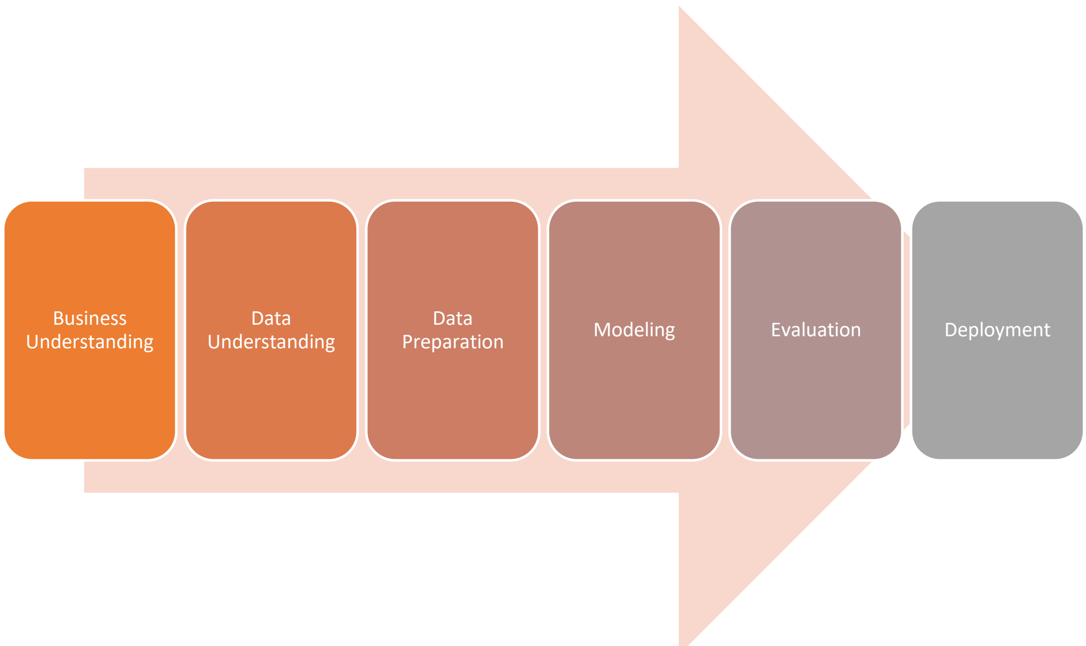


Figure 2.3-1 Research Methodology Phases

3.1 Business Understanding

This section enlightens that the data collected is from a private university operating in United Arab Emirates (U.A.E) for more than twelve years. The university comprises seven colleges and offers thirty-six (36) specialized programs. The university enrolls students throughout the year. The university operates through the semester system. The semesters are divided into two main categories regular and optional semesters. Fall Semester and Spring Semester are the official

[1] CRISP-DM - Data Science Process Alliance (datascience-pm.com)

regular semesters, while summer semester courses are offered as an optional semester for the students that would like to speed track the program duration.

The objective of this research is to design a framework model to identify whether the student is good at courses or is on the verge of failing and marginally passing in courses. This will proactively alarm the academic institution to take extra measures to improve the student capabilities and understanding of the courses. Eventually, student success contributes toward building a good reputation of the academic institution in the world, also enabling the student to perform well in the practical field.

3.2 Data understanding

3.2.1 Data Collection

The data is collected from the information technology department of a private university. The university has developed an in-house Enterprise Resource Planning level campus solution. There are various modules that are integrated with each other. The university is using SQL server as their database server. SQL queries are being used to extract the attributes of the students, course information, and the grades achieved. The data includes all bachelor's and master's degree students. A total of twenty-six (26) initial attributes and 134,171 records are part of the initial raw data.

3.2.2 Initial Description of Attributes

The below table illustrates the factor type, attribute name, a brief description, and the last column shows the possible values for the attribute.

Factor type	**Attribute Name**	**Description**	**Possible Values**
Student General	StudentRefNo	Anonymous Identifier for Records	
	AdmissionDescriptionEnglish	The Admission Type of the student	4 Possible values (Degree, Visiting, External Transfer, and Re-admitted).
	class rank	The level of the student according to the number of completed credit hours	4 Values for Bachelor degree (Freshman, Sophomore, Junior and Senior) While master is assigned 1 Value of "Graduate."
Student Demographic	DateOfBirth	Date of Birth of the student	
	Gender	Gender of the student	Male or Female
	Nationality	Nationality of the student	91 Unique Nationality Values
Degree Information	DegreeName	Enrolled degree for the student	Two Values (Bachelors or Masters)
	ProgramName	Enrolled Program Name for the student	36 Possible values (Specializations offered by the university)
Past Qualification	SchoolName	The last institution attended by the student	1207 School and University Names
	SchoolCountry	The country of the last attended institutions	59 Unique values
	SchoolGraduationRate	Grades achieved in the last Attended Institution	The score is out of 100 high schools, and the score is a CGPA out of 4 for university passed students.
English Proficiency	English exam	The English proficiency level is determined by the International English test, and their scores	IELTS, TOEFL, EmSAT, City Guilds, and Their Scores
	ELIScore		
Course Information	course code	Course Code for the enrolled Course	673 Course Codes (Text Value)
	CourseName	Course Name for the course enrolled by the student	Text Value
	CourseCategory	The course category according to the student study plan	five possible values (Preparatory, General Education, Core, Specialization and Elective)
	academic year	Academic Year of the course enrolled	YYYY
	SemesterName	Semester name of the course enrolled	Fall, Spring or Summer followed by year description
	Presence	Student presence percentage.	Value is out of 100
	StudentCount	The class size, i.e., the number of students enrolled in the same class.	
	AcademicStart	Date of the start of the classes for the semester	
Grades Achieved	Grade	Grade Symbol as per the university grade standards	Letter symbols like A, A+, B, B+
	GradeTotal	The score out of 100 in the course	
Instructor Demographics	FacultyGender	Gender of the instructor for the course	Male or Female
	FacultyDOB	Date of Birth of the instructor for the course.	
	FacultyNationality	Nationality of the instructor	61 Unique Nationality Values

Table 3.2-1 Raw Data Attribute and Description

3.3 Data Preparation

3.3.1 Feature Selection

The feature selection process is one of the core processes in the data preparation phase. The attributes have an influential impact on the machine learning algorithms training. The selection of features immensely impacts the accuracy of the results. There are multiple techniques for feature selection, such as univariate Selection, feature importance, and Correlation Matrix with Heatmap. In this research, we will use the correlation matrix with heatmap as the feature selection technique. The matrix provides a demonstration of positively and negatively related attributes.

3.3.2 Cleaning Data

3.3.2.1 Missing Data

Data can have missing information attributes that can be due to data entry or the non-availability of information. There are a few ways to deal with the missing data. These techniques include filling the missing information, replacing the information, and dropping the entire data record containing missing information. The research utilizes the python library of pandas to deal with the missing information.

3.3.2.2 Outliers Identification

The data was loaded into Jupyter Notebook, and with the use of python language, diagrams for scatter and boxplot was generated for the last Institution grades. For plotting the graphs, the data frames were separately analyzed for bachelor's degree students and master's degree students.

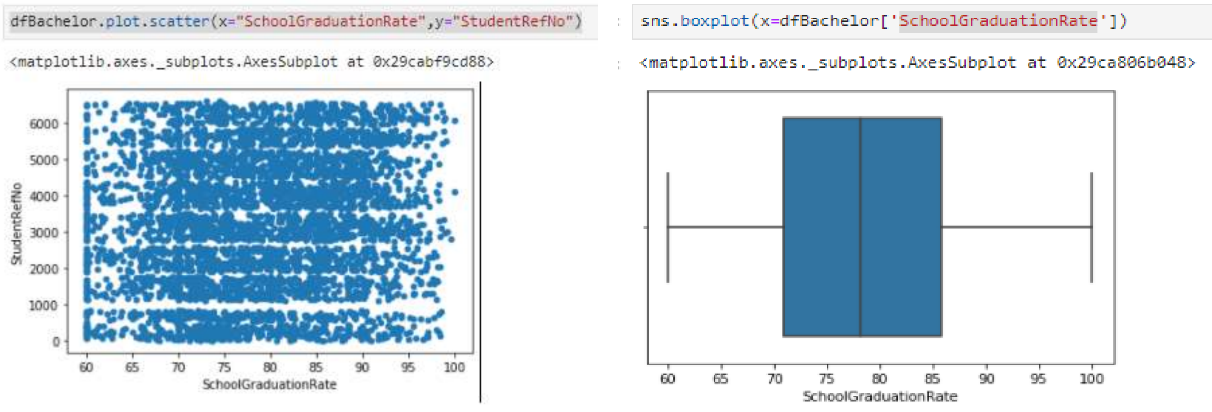


Figure 3.3-1 Outlier Diagram

The scatter diagram and box plot diagram generated for bachelor degree-seeking students displayed no outliers.

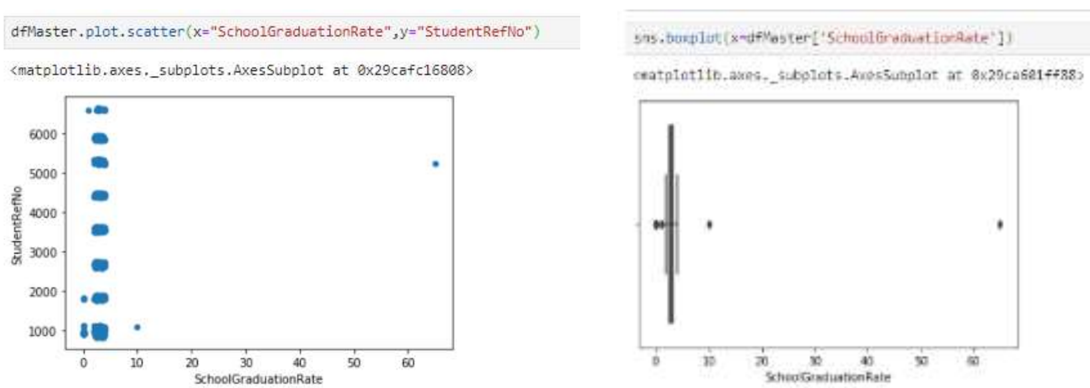


Figure 3.3-2 Scatter Plot Diagram

The scatter diagram and box plat diagram generated for master degree-seeking students identified two (2) data points that were identified as outliers and were removed from the dataset for further processing.

3.3.2.3 Irrelevant Data Identification

The Student records included the course with TR grade. The TR Grade represents the transfer courses. These records are eliminated as they are irrelevant in the context of predicting student performance. By Applying this step, the unique student count was reduced from 5660 to 5627 Records.

3.3.3 Data Transformation

3.3.3.1 Feature Extraction

The classification models for Machine learning algorithms primarily function on numerical data. Hence it is essential to covert the categorical information into numerical data. There are mainly two ways for conversion of categorical data into numerical: label Encoding and One Hot Encoding.

Label Encoding

Food Name	Categorical #	Calories
Apple	1	95
Chicken	2	231
Broccoli	3	50

→

One Hot Encoding

Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50

Figure 3.3-3 One Hot Encoding Example[2]

Label Encoding is a technique that uses alphabetical ordering for providing the numerical values for the conversion, while one-hot encoding is a popular technique in which each of the categorical values is converted into a column, and a value of 1 or 0 is assigned to it. In this research, one hot encoding is used.

3.3.3.2 Feature Scaling

Feature scaling is one of the essential preprocessing steps that remove the skewness and biasness of the classification model from the dominant value groups. The feature scaling can significantly improve the performance of the machine learning model from a weaker model to a better one[3]. There are several feature scaling techniques such as normalization, standardization, absolute maximum scaling, and Min-max scaling.

[2] https://medium.com/@michaeldelsole/what-is-one-hot-encoding-and-how-to-do-it-f0ae272f1179
[3] https://towardsdatascience.com/all-about-feature-scaling-bcc0ad75cb35

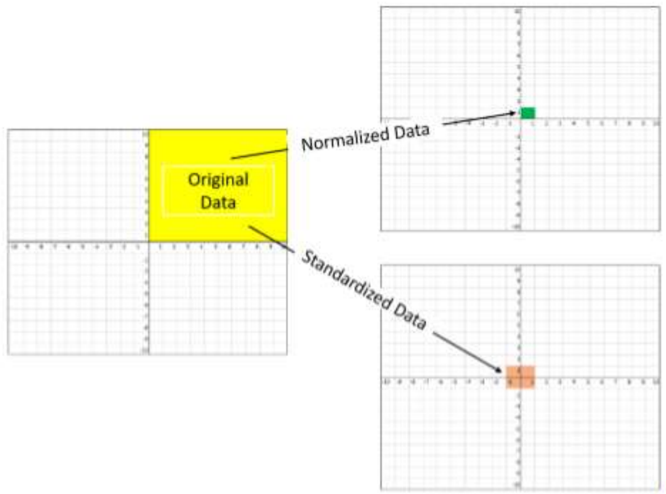


Figure 3.3-4 Scaling Techniques Example

In this research, the standardization scaling is used for the age columns, and the min-max scaling is used for grades, last attended school grades, and the presence percentage.

3.3.3.2.1 Standardization

In this technique, the features come in close proximity by the use of mean value and standard deviation.

$$X_{new} = \frac{X - X_{mean}}{\sigma}$$

3.3.3.2.2 Min-Max Scaling

In this technique, the maximum and minimum values are taken into consideration. The data values result in a range between 0 and 1.

$$X_{new} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

3.3.3.3 Feature Creation

The two attributes age for Student and Instructor at the time of course engagement was derived from the semester start date and the date of birth of students and faculty members. The columns are numerical in nature.

3.3.3.4 Class Label "BorderLineORFailure"

The Class label identified is labeled as "BorderLineORFailure." The ranges of the class labels are different for bachelor's and master's degrees. The following table shows the class label information.

Degree Level	BorderLineORFailure	
	YES (1)	NO (1)
Bachelor	Below 70	Above and equal to 70
Master	Below 74	Above and equal to 75

Table 3.3-1 Class Label Information

3.4 Modeling

3.4.1 K-nearest Neighbors (KNN)

KNN is one of the supervised classification techniques. KNN algorithm works on the principle of similarity and dissimilarity. In order to evaluate the similarity of the data points, KNN computes a distance matrix. KNN can use various distance measuring mechanisms such as the most popular Euclidean distance, Manhattan distance, or Minkowski distance. The research uses the Euclidean distance that uses the below formula to compute the distance

$$d(p,q) = \sqrt{\sum_{n=1}^{n}(q_i - p_i)^2}$$

Determining the value of K is essential for the model to provide effective results and better accuracy. The research has incrementally predicted the values using values of K from 1 till 40 and plotted the value of K versus the error rate. This plot will identify the best value of K to be used for the model.

3.4.2 Naïve Bayes

Naïve Bayes is a supervised machine learning algorithm. It is one of the algorithms that use the least computation power. The algorithm is based on Bayes Theorem. The algorithms tend to have high accuracy measures when the models are implemented for large data sets. The primary consideration of the algorithms is assuming that each attribute is independent of the other. The assumption is known as conditional independence and can be a demonstration by the following formula

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

3.4.3 Support Vector Machines

Support Vector machines (SVM) is one of the supervised machine learning algorithms. SVM algorithm can be utilized for regression purposes as well. The principle of the SVM evolves around the concept of constructing a hyperplane and separating the data points according to the class labels. SVM invests more time in training the model compared to the other classification models where as SVM performs faster to predict the class labels and provides good accuracy. The algorithm also uses less memory.

3.4.4 Logistic Regression

Logistic regression is one of the most popular classification techniques used in machine learning. It uses a sigmoid function

$$f(x) = \frac{1}{1 + e^{-(x)}}$$

3.4.5 Decision Tree

The decision tree is one of the simplest and fast classification machine learning algorithms. The algorithms, as a result, produce a hierarchal flow chart. The leave nodes that are the last nodes on the tree diagram represent the class labels. The internal branches form the features conjunction that is based on the classification rules represented as the path leading to the class labels.

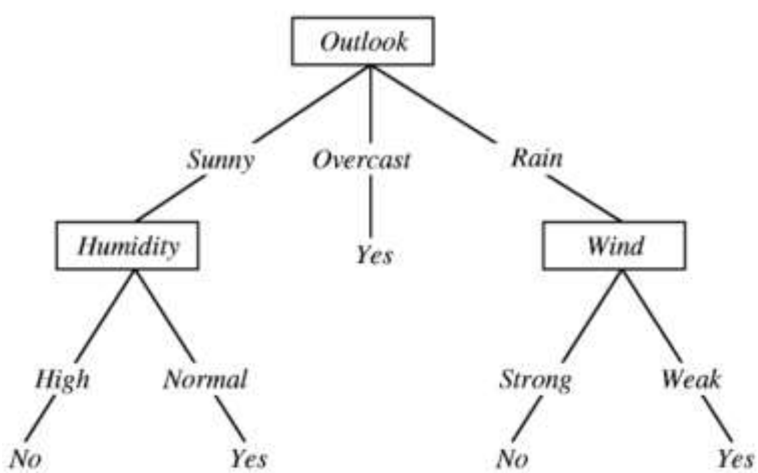


Figure 3.4-1 Decision Tree Example[4]

3.4.6 Random Forest

Random forest is one of the ensemble machine learning algorithms that use decision trees as the base mode algorithm. The enhancement to the random forest classification is that the trees are generated from multiple subsets within the training dataset. The class labels are determined by computing the average or highest ranking. The decision trees are prone to overfitting situations which are handled in the random forest classification model. However, the random forest classification algorithm is slower than the decision tree but provides more optimal results.

3.4.7 ADA Boost

Adaboost is the short name for the adaptive boost classification model. It is also an ensemble classifier that is used to boost or increase the accuracy of other standard classification models. The research uses a decision tree to be boosted by using the Adaboost classification algorithm.

3.4.8 Bagging Classifier

Bagging classifier is another ensemble algorithm. The research uses the decision tree as a base algorithm. The bagging classifier estimates the prediction on random data sets and, as a result, provide the output based on aggregation by a voting or averaging mechanic

[4] https://towardsdatascience.com/

3.4.9 Voting Classifier

The Voting classification is also an ensemble technique that provides the analysts to combine various standard machine learning algorithms as based models and perform the prediction. The output prediction of class labels is determined on the average of the predicted values from the different models used. In this research, decision tree and Logistic regression are combined to provide the voting classification predictions.

3.5 Evaluation

The research has identified evaluation metrics that were used in the past research studies and will be using seven metrics that are Accuracy, Precision, Recall, F1 Score, Mean Squared Error, ROC AUC SCORE, and AUC. The evaluation measures primarily depend on the confusion matrix that is used to describe the performance of the models.

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

Table 3.5-1 Confusion Matrix Explanation

The matrix forms the basis for computing the other evaluation measures. The four entries in the confusion matrix are the counts generated by the prediction model.

3.5.1 Accuracy

Accuracy is the most widely used performance measure in research studies. Accuracy shows how accurately the model has predicted the actual class labels.

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{True Negative} + \text{False Positive} + \text{False Negative}}$$

3.5.2 Precision

The precision measure indicates the performance of the model based on the actual positives versus the predicted positives.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

3.5.3 Recall

The recall is, also sometimes referred to as sensitivity, is a measure that bases the calculation on the ratio of the predicted positives to the actual positives

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive + False Negatives}}$$

3.5.4 F1 Score

F1 score is also one of the popular evaluation metrics that is the combination of precision and recall.

$$\text{Recall} = 2 \text{ x } \frac{\text{Precision x Recall}}{\text{Precision + Recall}}$$

3.5.5 Mean Squared Error

Mean Squared Error (MSE) is one of the simplest evaluation metrics to provide the common loss function value.

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

3.5.6 ROC AUC SCORE

The ROC is dependent on the true positive rate and false-positive rates.

3.5.7 AUC

The area under the curve (AUC) measure depicts how likely data points will be predicted highly towards positives or negatives. The value of AUC ranges between 0 and 1. Zero is the value of the model where all the predictions are wrong. In contrast, the AUC value will be one if all the predictions are correct.

3.6 Deployment

The research has used Python as a tool to implement the models and evaluate the performance of the algorithms. The research has used 70 percent of the data set in training the models, while the remaining 30 percent of the data set is used for testing and predictions of the models.

4 Results and Analysis

This research chapter discusses the analysis of the data that includes visual representations of the data and the results from applying the machine learning algorithms using python code. The evaluation metrics summary will conclude this chapter.

4.1 Data analysis

The sub-section of Data analysis is divided into more subsections that provide information of the initial data exploration and correlation of the attributes with the class label.

4.1.1 Initial Data Exploration

In this section, the Power BI tool is utilized for visual illustrations. The below chart demonstrates the distribution of student records by bachelor's degree and master's degree.

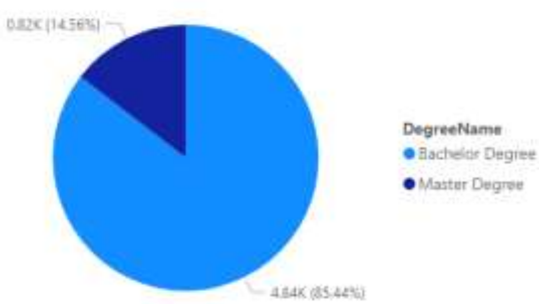


Table 4.1-1 Degree wise Student Distribution

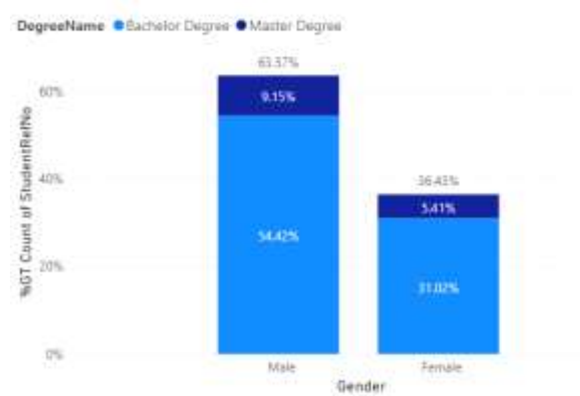


Figure 4.1-1 Gender wise distribution

The initial raw data consisted of 85.4% (4836) Bachelor and 14.6% (824) Master degree-seeking students. The raw data consisted of 63.57% male and 36.43% female student records. The above figure depicts that out of 63.57% male student records, 54.42% students are bachelor while 9.15% students are master degree-seeking students. Further, the 36.43% female student records are comprised of 31.02% bachelor and 5.41% master degree-seeking student records.

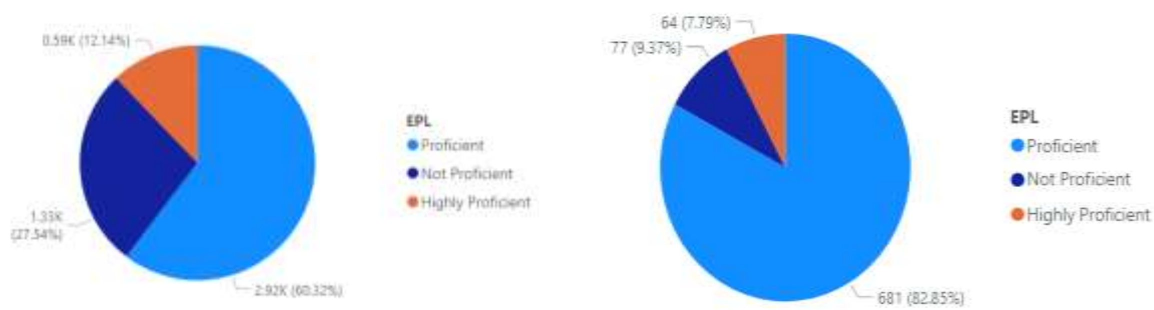


Figure 4.1-2 English Proficiency Degree wise distribution

The above-left side figure depicts the bachelor degree-seeking students that shows that the majority of the students, 87.8% are either proficient or highly proficient in the English language. The above right-side figure shows that the master degree-seeking students are 92% proficient and highly proficient in English skills.

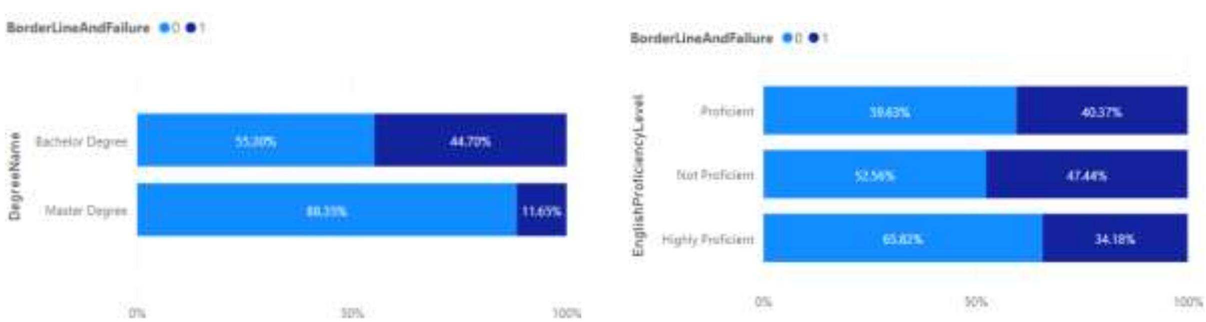


Figure 4.1-3 Class Label statistics with Degree and English Proficiency

The above figures provide a one hundred percent stack view in perspective of the class label with the student degree and English Proficiency. The bachelor students are at the riskier situation of being categorized as borderline and failure as compared to the master degree-seeking students.

The figure to the right depicts that lower English proficiency level leads to higher chances of being labeled as borderline and failure student. There is an inverse relation.

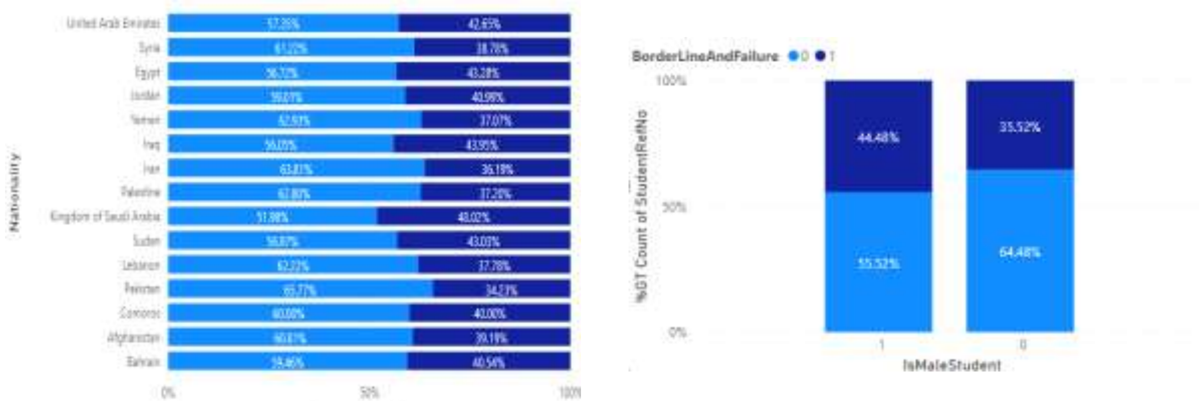


Figure 4.1-4 Class Label Statistics with Nationality and Gender

The above figure provides a relation of student nationalities and gender in comparison to the class label of Border line and failure. The above figure on the left shows the top 15 nationalities' views, and each of the nationalities is almost having slight variances. The above figure on the right depicts that female students are at slightly lesser risk to be labeled as Border line and failure class.

4.1.2 Correlation of Attributes

The research at the feature selection stage generated the below correlation matrix and heat map

	StudentRefNo	DegreeName	ProgramName	IsMaleStudent	Nationality	SchoolName	SchoolCountry	SchoolGraduationRate	AdmissionDescriptionEnglish	EnglishExam	...	GradeTotal	ClassRank	Presence	StudentCount
StudentRefNo	1.000000	0.024482	0.021795	0.006001	0.069810	0.028026	0.044410	-0.000591	0.228692	-0.199794	...	-0.016655	-0.088722	0.180908	-0.014123
DegreeName	0.024482	1.000000	0.681416	-0.001210	0.106830	0.061734	-0.005524	-0.946340	-0.042794	-0.196855	...	0.207483	-0.269388	0.308739	-0.280666
ProgramName	0.021795	0.681416	1.000000	-0.003183	0.021799	0.034667	-0.004847	-0.640891	-0.063988	-0.139575	...	0.195354	-0.185874	0.332176	-0.268414
IsMaleStudent	0.006001	-0.001210	-0.003183	1.000000	0.137025	-0.056169	0.049231	-0.039516	0.037134	0.076182	...	-0.207577	-0.048694	-0.043954	0.033656
Nationality	0.069810	0.106830	0.021799	0.137025	1.000000	-0.037150	0.362833	-0.130711	0.061060	-0.073195	...	-0.075485	-0.064042	0.029640	-0.005679
SchoolName	0.028026	0.061734	0.034667	-0.056169	-0.037150	1.000000	-0.124018	-0.072857	0.049922	0.020032	...	0.039947	-0.022609	0.025589	-0.018970
SchoolCountry	0.044410	-0.005524	-0.004847	0.049231	0.362833	-0.124016	1.000000	0.033841	0.023041	-0.020674	...	-0.021146	-0.016631	0.010513	0.018263
SchoolGraduationRate	-0.000591	-0.946340	-0.640891	-0.039516	-0.130711	-0.072857	0.033841	1.000000	0.002955	0.158243	...	-0.146131	0.267155	-0.279455	0.263209
AdmissionDescriptionEnglish	0.228692	-0.042794	-0.063988	0.037134	0.061060	0.049922	0.023041	0.002955	1.000000	0.098625	...	-0.017723	0.073548	0.013169	-0.021077
EnglishExam	-0.199794	-0.196855	-0.139575	0.076182	-0.073195	0.020032	-0.020674	0.158243	0.098625	1.000000	...	-0.109084	0.036896	-0.126590	0.051351
EnglishProficiencyLevel	-0.125680	0.091712	0.204289	0.068812	0.075652	-0.075233	-0.000503	-0.115664	0.029963	0.092378	...	-0.041103	0.030224	-0.006600	-0.036391
CourseCode	-0.042992	0.209426	0.096737	-0.005154	0.009664	0.023149	-0.021948	-0.202400	0.046917	-0.044394	...	0.119770	0.152281	0.165134	-0.302739
GradeTotal	-0.016655	0.207483	0.195354	-0.207577	-0.075485	0.039947	-0.021146	-0.146131	-0.017723	-0.109084	...	1.000000	0.113706	0.269023	-0.101024
ClassRank	-0.088722	-0.269388	-0.185874	-0.048694	-0.064042	-0.022609	-0.016631	0.267155	0.073548	0.036896	...	0.113706	1.000000	0.035829	-0.065109
Presence	0.180908	0.308739	0.332176	-0.043954	0.029640	0.025589	0.010513	-0.279455	0.013169	-0.126590	...	0.269023	0.035829	1.000000	-0.198644
StudentCount	-0.014123	-0.280666	-0.268414	0.033656	-0.005679	-0.018970	0.018263	0.263209	-0.021077	0.051351	...	-0.101024	-0.065109	-0.198644	1.000000
IsMaleFaculty	-0.105636	0.115891	0.060314	0.005662	0.020916	0.010877	0.009619	-0.113121	-0.066978	-0.012404	...	0.018775	-0.032987	-0.013605	-0.034547
FacultyNationality	0.086930	-0.072770	-0.070337	0.023805	-0.013653	-0.005396	-0.005623	0.061561	0.136014	0.042821	...	-0.035771	-0.009234	-0.037733	0.000484
CourseCategory	-0.004490	-0.283939	-0.071846	-0.024542	-0.039983	-0.011810	0.005385	0.269028	-0.004604	0.079039	...	0.015031	0.020026	-0.025729	-0.045010
StudentAge	-0.012987	0.423242	0.289965	0.020116	0.203011	0.040593	-0.039851	-0.483152	0.139284	-0.033257	...	0.127804	0.014759	0.186971	-0.208462
FacultyAge	-0.001670	0.230059	0.172692	0.002669	0.048398	0.013573	0.016745	-0.222267	0.002026	-0.046081	...	0.041532	-0.023815	0.062753	-0.076781
BorderLineAndFailure	0.005250	-0.182799	-0.171947	0.182407	0.048864	-0.031486	0.011757	0.130169	-0.006069	0.093783	...	-0.798708	-0.099361	-0.243609	0.091088

	IsMaleFaculty	FacultyNationality	CourseCategory	StudentAge	FacultyAge	BorderLineAndFailure
StudentRefNo	-0.105636	0.086930	-0.004490	-0.012987	-0.001670	0.005250
DegreeName	0.115891	-0.072770	-0.283939	0.423242	0.230059	-0.182799
ProgramName	0.060314	-0.070337	-0.071846	0.289965	0.172692	-0.171947
IsMaleStudent	0.005662	0.023805	-0.024542	0.020116	0.002669	0.182407
Nationality	0.020916	-0.013653	-0.039983	0.203011	0.048398	0.048864
SchoolName	0.010877	-0.005396	-0.011810	0.040593	0.013573	-0.031486
SchoolCountry	0.009619	-0.005623	0.005385	-0.039851	0.016745	0.011757
SchoolGraduationRate	-0.113121	0.061561	0.269028	-0.483152	-0.222267	0.130169
AdmissionDescriptionEnglish	-0.066978	0.136014	-0.004604	0.139284	0.002026	-0.006369
EnglishExam	-0.012404	0.042821	0.079039	-0.033257	-0.046081	0.093783
EnglishProficiencyLevel	0.020226	0.027106	-0.058677	0.132555	0.015923	0.033126
CourseCode	0.020675	0.040281	-0.177758	0.205353	0.126410	-0.111843
GradeTotal	0.018775	-0.035771	0.015031	0.127804	0.041532	-0.798708
ClassRank	-0.032987	-0.009234	0.020026	0.014759	-0.023815	-0.099361
Presence	-0.013605	-0.037733	-0.025729	0.186971	0.062753	-0.243609
StudentCount	-0.034547	0.000484	-0.045010	-0.208462	-0.076781	0.091088
IsMaleFaculty	1.000000	-0.122720	-0.072797	0.068607	0.319674	-0.019876
FacultyNationality	-0.122720	1.000000	-0.007141	-0.043925	-0.025759	0.026920
CourseCategory	-0.072797	-0.007141	1.000000	-0.098123	-0.107670	-0.030341
StudentAge	0.068607	-0.043925	-0.098123	1.000000	0.128250	-0.121715
FacultyAge	0.319674	-0.025759	-0.107670	0.128250	1.000000	-0.042386
BorderLineAndFailure	-0.019876	0.026920	-0.030341	-0.121715	-0.042386	1.000000

Table 4.1-2 Correlation Matrix

The feature selection phase identified that the student reference number has the highest correlation impact on determining the student will be on "BorderLineOrFailure." The attribute of the Student Reference number is removed from the dataset.

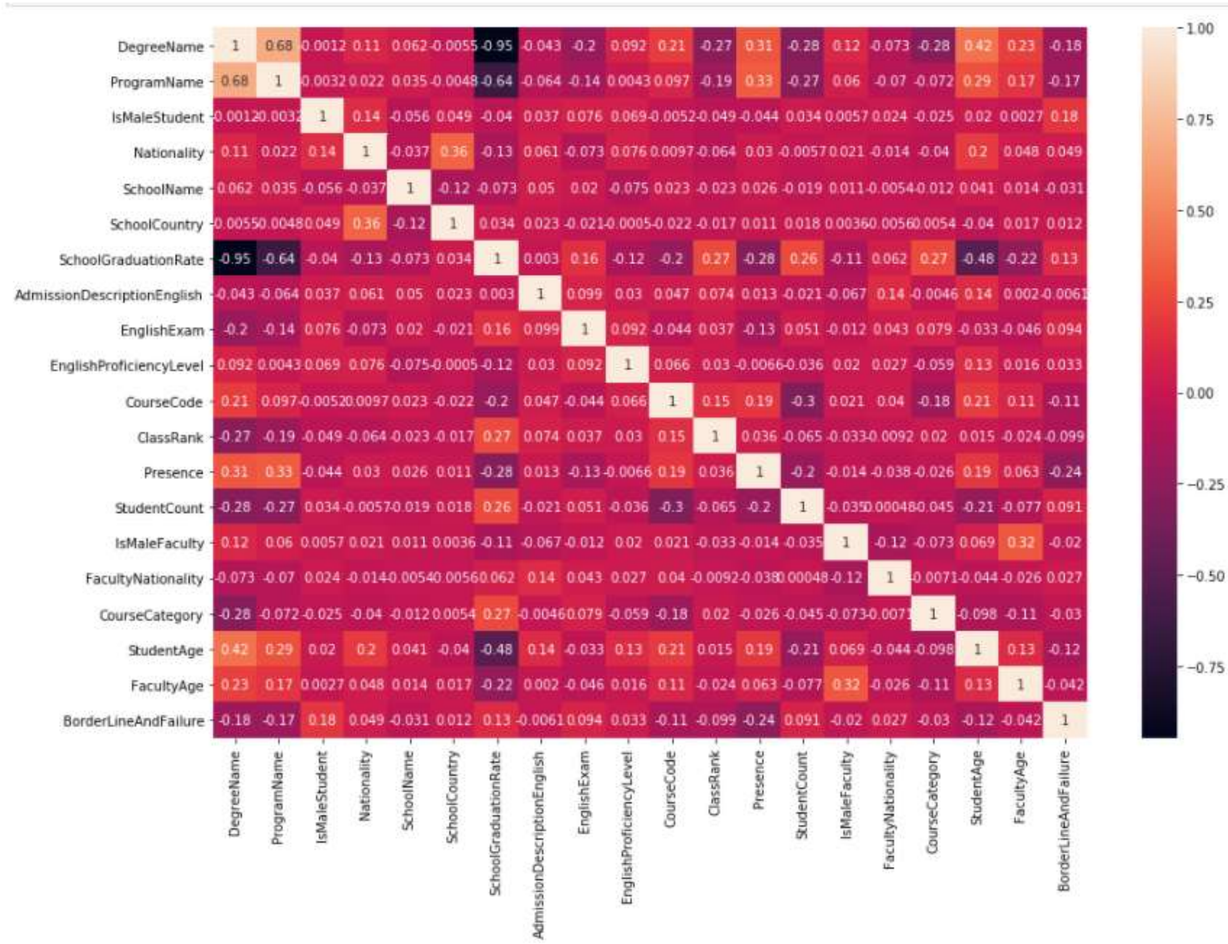


Figure 4.1-5 Visual Representation of Correlation Matrix

4.2 Machine learning implementation

This section of the chapter provides the results and analysis for the machine learning algorithms. The organization of the analysis shows the classification report, the evaluation metrics that are the nine evaluation measures, and lastly, examining the ROC AUC curve and Precision-Recall Curves.

4.2.1 K Nearest Neighbor (KNN)

The section provides the results and analysis for the KNN classifier.

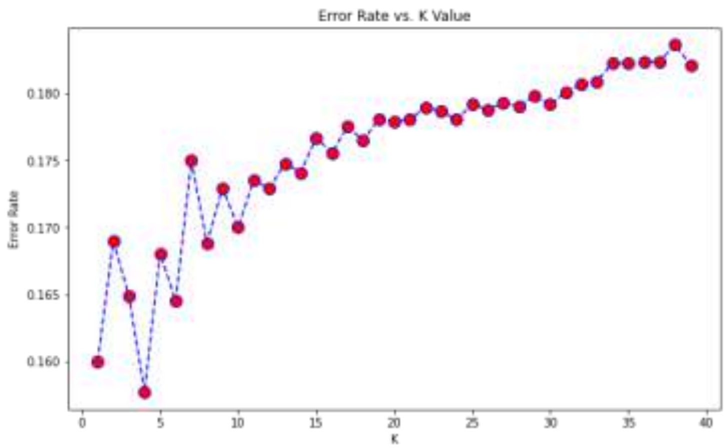


Figure 4.2-1 Best Value of K for KNN Classifier

The first step for the KNN classifier is to analyze the best value of K to be used by the modal. The above graph has plotted the value of K against the error rate. The minimum error was found to be 15.77% at the value of K=3.

Class Label	Precision	Recall	F1-Score
0	0.89	0.89	0.89
1	0.65	0.65	0.65

Table 4.2-1 KNN Classification Report

The above table shows the classification report generated by python code. The precision, recall the f1 score for predicting the students that are Not on the borderline or failure category shows better results as compared to the prediction of students on the borderline of failure category.

Accuracy	Precision	Recall	f1 score	ROC AUC SCORE	AUC	Mean Squared Error
83.51%	64.74%	65.23%	64.99%	77.17%	73.00%	16.49%

Table 4.2-2 KNN Evaluation Metrics

The KNN evaluation metrics show a better accuracy result of 83.51%, showing the model performed better in predicting the border line or failure Students (True Positive) and the regular students (True Negative). However, the false-positive predictions of students that should not be on borderline or failure is slightly higher that reduced the precision, recall, and f1 score to 64.74%,65.23%, and 64.99%, respectively. The mean squared error was calculated as 16.49%. The ROC and AUC score show reasonably better results.

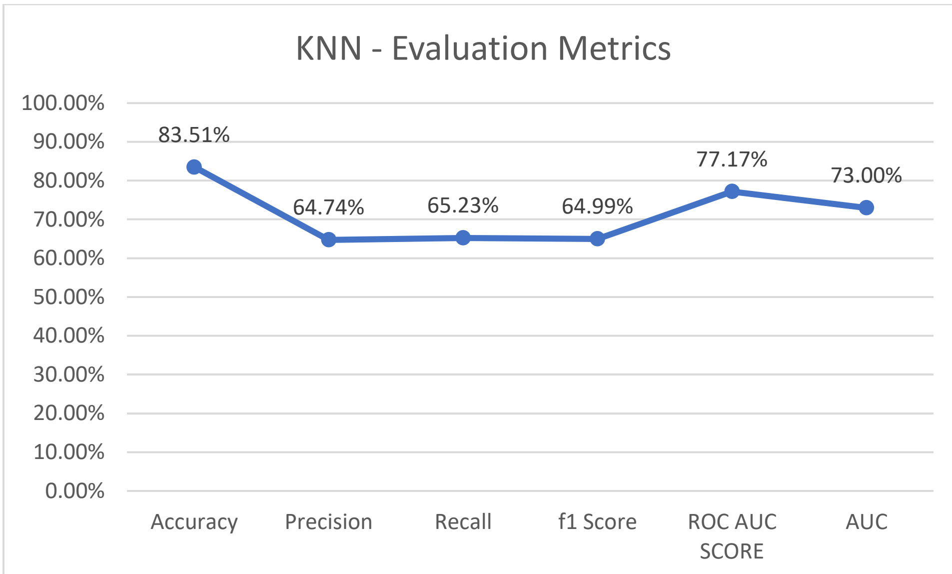
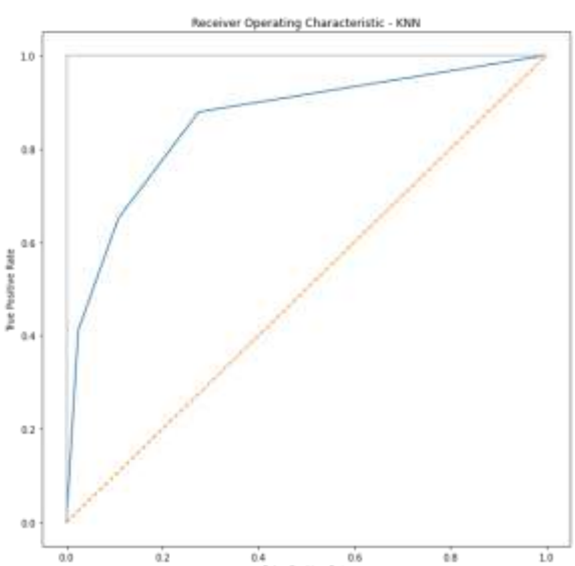


Figure 4.2-2 KNN Evaluation Metric Comparison

The above figure demonstrates the comparison of the metric evaluation result for the KNN Model. The model performed better is providing better accuracy. And the AUC and ROC score also suggests better confidence of the model in the prediction.



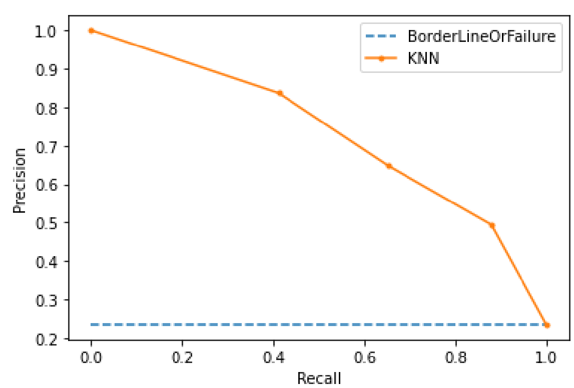


Figure 4.2-3 ROC AUC Curve – KNN

Figure 4.2-4 Precision-Recall Curve – KNN

The above figures are ROC AUC Curve and the precision-recall curve for the KNN model. The area under the ROC curve and the precision-recall curve should be higher for depicting the good performance of the machine learning model. The output suggested that the KNN model performed well in predicting the students that should be on borderline or failure.

4.2.2 Naïve Bayes:

The section provides the results and analysis for the Naïve Bayes classifier.

Class Label	Precision	Recall	F1-Score
0	0.78	0.98	0.87
1	0.63	0.09	0.16

Table 4.2-3 Naïve Bayes Classification Report

The classification report for Naïve Bayes Algorithms shows that in determining the Class label for the student that is not at Border Line or Failure are having higher precision, recall, and F1 Score measures as compared to the students that are predicted for the Border line or failure class label.

Accuracy	Precision	Recall	f1 score	ROC AUC SCORE	AUC	Mean Squared Error
77.43%	63.31%	8.96%	15.70%	53.69%	43.00%	22.57%

Table 4.2-4 Naïve Bayes Evaluation Metrics

The evaluation measure of recall value shows a low number indicating that the model did not perform better for predicting the students that should have been predicted on the border line or failure category. Since the f1 score is dependent on the recall and precision, the f1 score is also impacted. However, the model performed slightly better in predicting the true positives and true negatives, which resulted in 77.43% of accuracy. Further, the model was weak in predicting the false positive that reduced the precision score to 63.31%. The ROC AUC Score and AUC score of 53.69% and 43%, respectively, show that the model is confused in predicting the class label. The mean square error value showed 22.57% that will be compared with the other models.

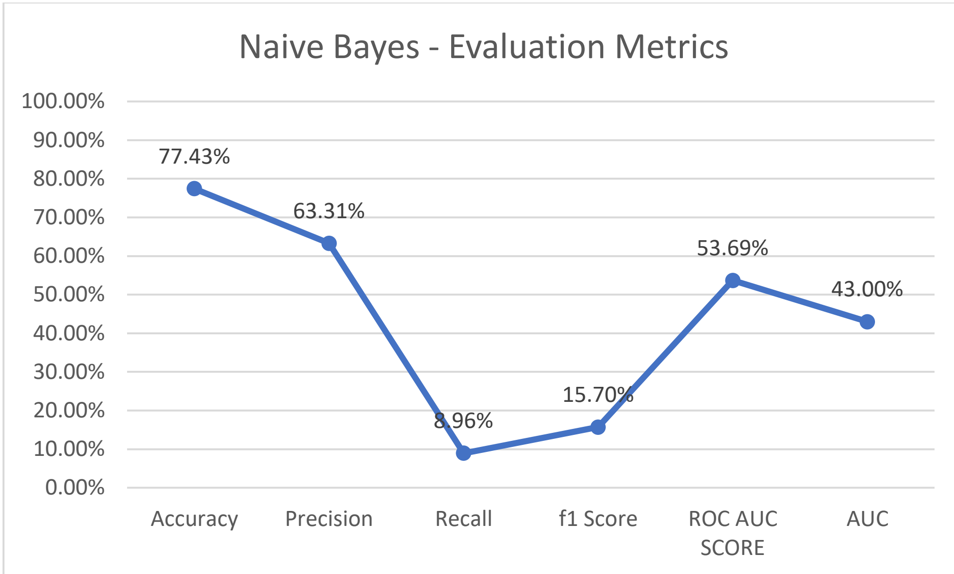


Figure 4.2-5 Naïve Bayes Evaluation Metric Comparison

The figure above demonstrates the comparisons of the evaluation metrics that show that the model provided a decent accuracy and precision simultaneously. The recall, f1 score, and AUC depict that the model did not perform better for the prediction.

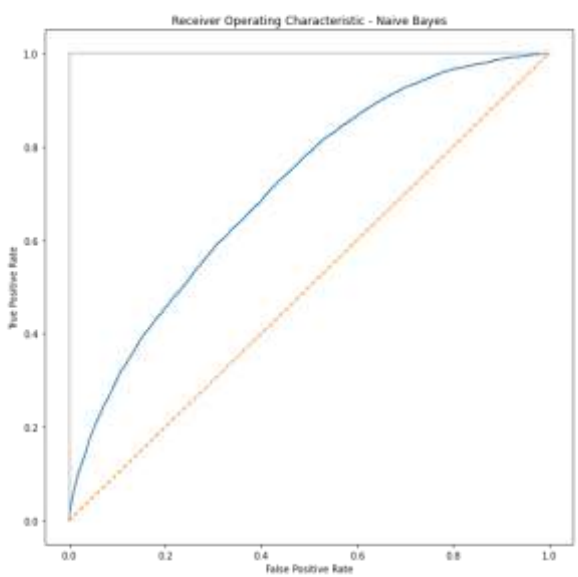


Figure 4.2-6 ROC AUC Curve – Naïve Bayes

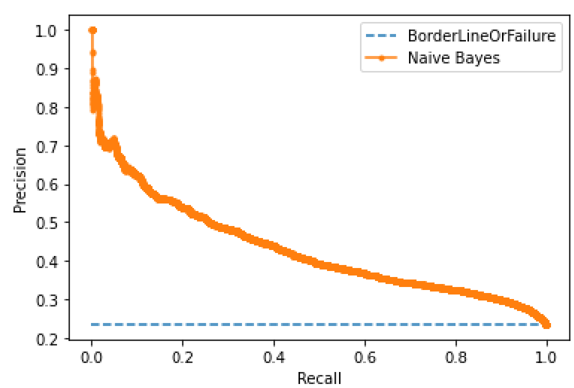


Figure 4.2-7 Precision Recall Curve – Naïve Bayes

The above figures are ROC AUC Curve and the precision-recall curve for the Naïve Bayes model. The area under the ROC curve and the precision-recall curve should be higher for depicting the good performance of the machine learning model. The output suggested that the Naïve Bayes model did not perform well in the prediction tasks.

4.2.3 Support Vector Machines (SVM)

The section provides the results and analysis for the Support Vector Machine (SVM) classifier.

Class Label	Precision	Recall	F1-Score
0	0.85	0.93	0.89
1	0.67	0.44	0.53

Table 4.2-5 SVM Classification Report

The classification report for SVM shows high precision, recall, and F1 scores of 85%, 93%, and 89%, respectively, for predicting the class label of regular students as compared to predicting the student for border line and failure.

Accuracy	Precision	Recall	f1 score	ROC AUC SCORE	AUC	Mean Squared Error
81.84%	67.12%	44.26%	53.34%	68.81%	63.3%	18.16%

Table 4.2-6 SVM Evaluation Metrics

The SVM model results showed a mean squared error of 18.16%. The accuracy rate of 81.84% shows that the True positive and True negative cases were predicted properly. However, the precision, recall, and f1 score indicate that the model is not performing decently in predicting the False Negative and False positive cases. The ROC and AUC scores also show average confidence in determining the student class label.

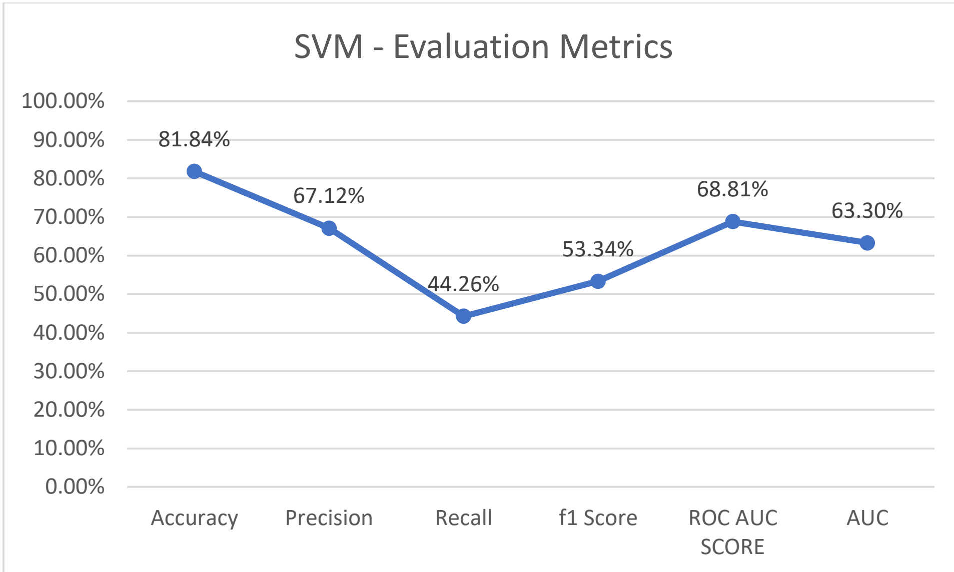
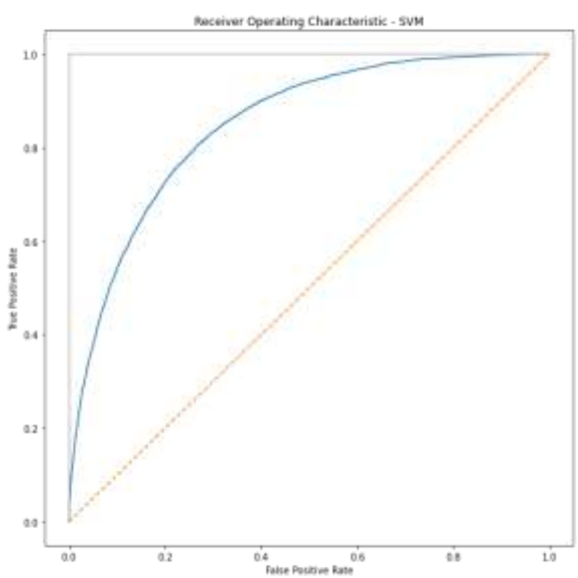


Figure 4.2-8 SVM Evaluation Metric Comparison

The figure above demonstrates the comparisons of the evaluation metrics from the SVM model that shows that the model provided better accuracy simultaneously the precision, recall, f1 score, and AUC depicts that the model did not perform better for the prediction.



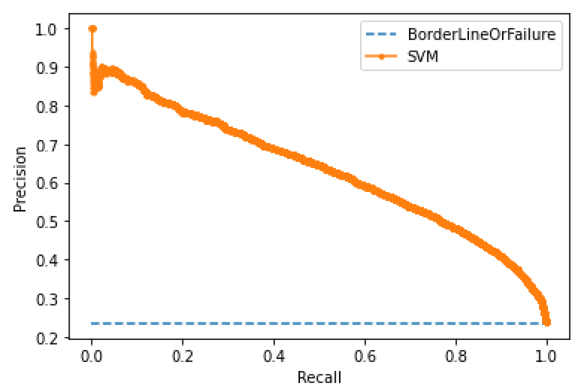


Figure 4.2-9 ROC AUC Curve – SVM

Figure 4.2-10 Precision-Recall Curve – SVM

The above figures are ROC AUC Curve and the precision-recall curve for the SVM model. The area under the ROC curve and the precision-recall curve should be higher for depicting the good performance of the machine learning model. The output suggested that the SVM model performed decently in the prediction tasks.

4.2.4 Logistic Regression

The section provides the results and analysis for the Logistic Regression classifier.

Class Label	Precision	Recall	F1-Score
0	0.85	0.93	0.89
1	0.67	0.44	0.53

Table 4.2-7 Logistic Regression Classification Report

The classification report for Logistic Regression shows high precision, recall, and F1 scores of 85%, 93%, and 89%, respectively, for predicting the class label of regular students as compared to predicting the student for border line and failure.

Accuracy	Precision	Recall	f1 score	ROC AUC SCORE	AUC	Mean Squared Error
81.57%	65.90%	44.45%	53.09%	68.70%	62.30%	18.42%

Table 4.2-8 Logistic Regression Evaluation Metrics

The SVM model results showed a mean squared error of 18.42%. The accuracy rate of 81.57% shows that the True positive and True negative cases were predicted properly. However, the precision, recall, and f1 score indicate that the model is not performing decently in predicting the False Negative and False positive cases. The ROC and AUC scores also show average confidence in determining the student class label.

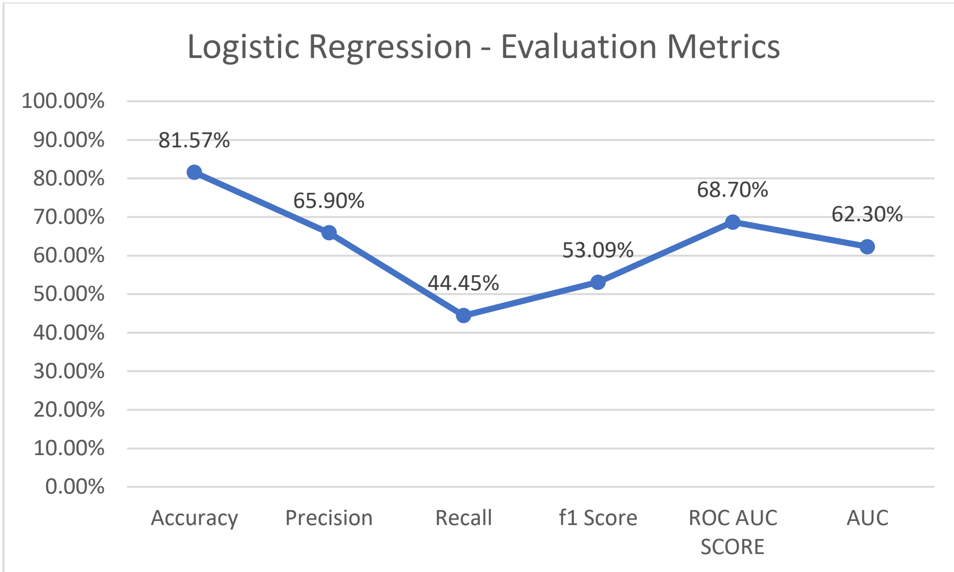


Figure 4.2-11 Logistic Regression Evaluation Metric Comparison

The figure above demonstrates the comparisons of the evaluation metrics from the logistic regression Model that shows that the model provided better accuracy simultaneously the precision, recall, f1 score, and AUC depicts that the model performed decently for the prediction.

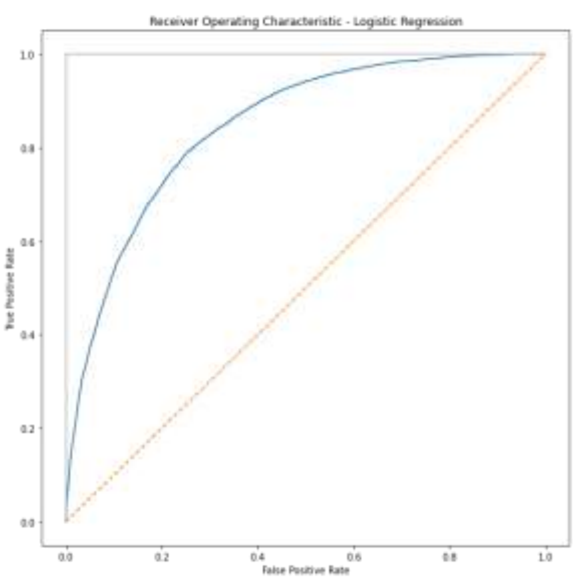


Figure 4.2-12 ROC AUC Curve – Logistic Regression

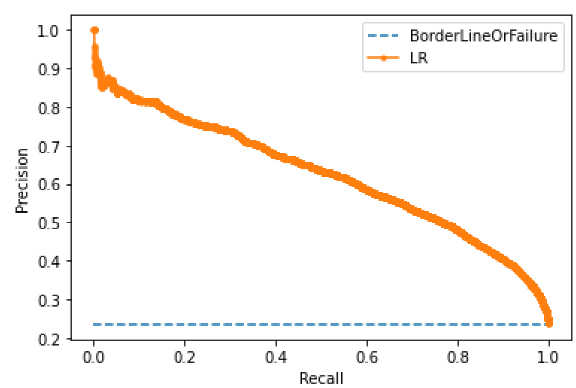


Figure 4.2-13 Precision-Recall Curve – Logistic Regression

The above figures are ROC AUC Curve and the precision-recall curve for the logistic regression model. The area under the ROC curve and the precision-recall curve should be higher for depicting the good performance of the machine learning model. The output suggested that the logistic model performed decently in the prediction tasks.

4.2.5 Decision Tree

The section provides the results and analysis for the Decision Tree classifier.

Class Label	Precision	Recall	F1-Score
0	0.86	0.94	0.9
1	0.72	0.52	0.6

Table 4.2-9 Decision Tree Classification Report

The classification report for the decision tree shows high precision, recall, and F1 scores of 86%, 94%, and 90% respectively for predicting the class label of regular students as compared to predicting the student for border line and failure where the precision, recall, and F1 Scores are 72%, 52%, and 60% respectively.

Accuracy	Precision	Recall	f1 score	ROC AUC SCORE	AUC	Mean Squared Error
83.92%	71.75%	51.90%	60.23%	72.82%	66.10%	16.07%

Table 4.2-10 Decision Tree Evaluation Metrics

The Decision Tree model results showed a mean squared error of 16.07%. The accuracy rate of 83.92% shows that the True positive and True negative cases were predicted properly. However, the precision Recall and f1 score indicate that score indicates that there were fewer false-positive detected compared to the false-negative cases. The model is performing better in terms of accuracy. The ROC and AUC scores also show average confidence in determining the student class label.

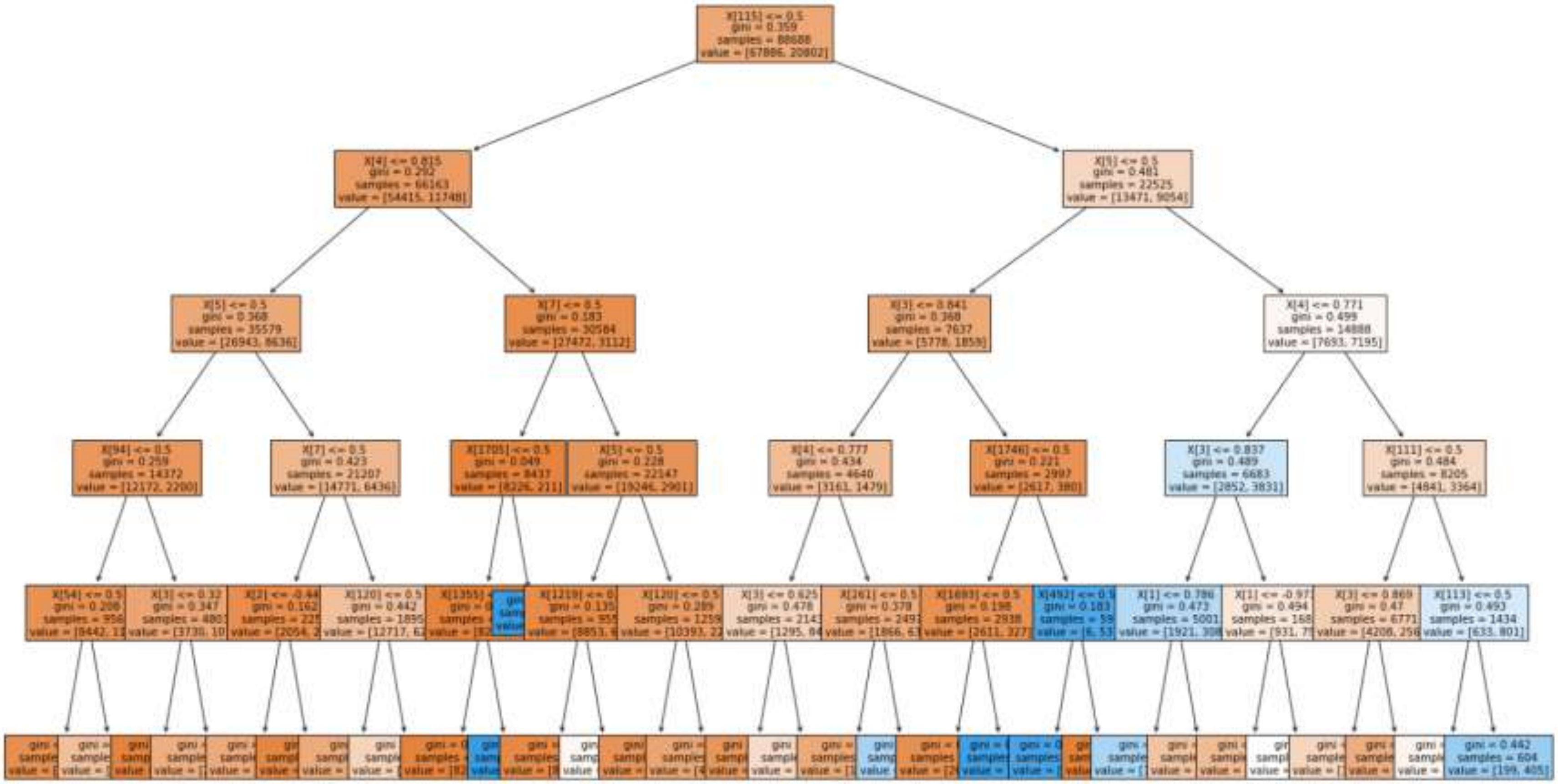


Figure 4.2-14 Visual Representation of Decision Tree

The above figure shows the visual representation of the decision tree model

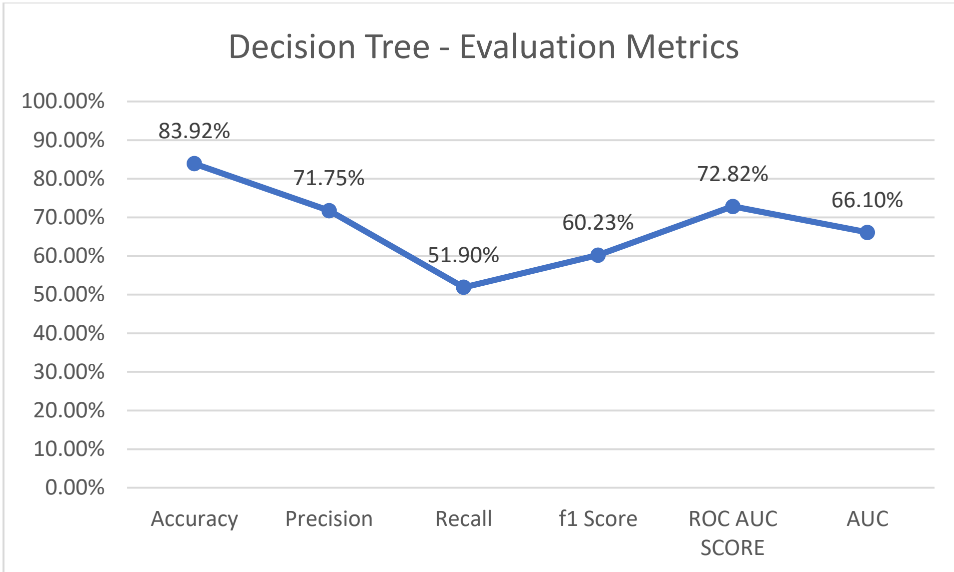
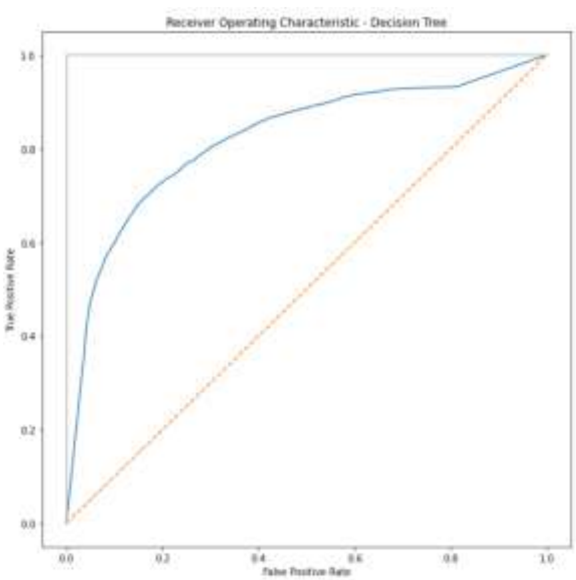


Figure 4.2-15 Decision Tree Evaluation Metric Comparison

The figure above demonstrates the comparisons of the evaluation metrics from the decision tree model that shows that the model provided better accuracy simultaneously the precision, recall, f1 score, and AUC depicts that the model performed decently for the prediction.



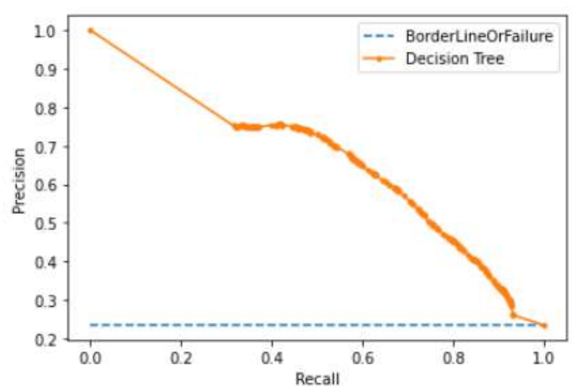


Figure 4.2-16 ROC AUC Curve – Decision Tree

Figure 4.2-17 Precision-Recall Curve – Decision Tree

The above figures are ROC AUC Curve and the precision-recall curve for the decision tree model. The area under the ROC curve and the precision-recall curve should be higher for depicting the good performance of the machine learning model. The output suggested that the decision tree model performed better in the prediction tasks.

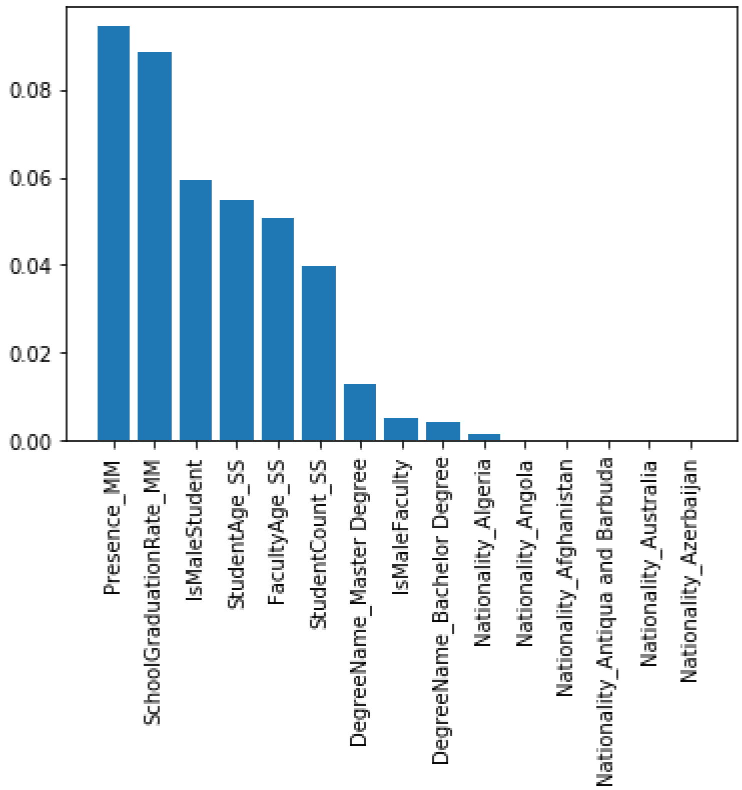


Figure 4.2-18 Feature Importance - Decision Tree

The above figure demonstrates the top fifteen (15) attributes that impacted the decision tree model to make decisions on the prediction of student performance. The highest influential attributes were Presence (Attendance Attribute) and School Graduation Rate (Grades Achieved in the last institution). The other important features include the student's gender, the age of the student, the age of the teaching faculty member, Student Count (Class Size), Degree attribute of master's degree and Bachelor's Degree-seeking students, gender of the teaching faculty member. The only significant nationality column identified was Algeria.

4.2.6 Random Forest

The section provides the results and analysis for the Random forest classifier.

Class Label	Precision	Recall	F1-Score
0	0.9	0.98	0.94
1	0.89	0.66	0.76

Table 4.2-11 Random Forest Classification Report

The classification report for random forest classification shows high precision, recall, and F1 scores of 90%, 98%, and 94%, respectively, for predicting the class label of regular students. The prediction of the student for border line and failure results also show good results where the precision, recall, and F1 Scores are 89%, 66%, and 76%, respectively.

Accuracy	Precision	Recall	f1 score	ROC AUC SCORE	AUC	Mean Squared Error
90.12%	89.14%	65.94%	75.81%	81.74%	87.90%	9.87%

Table 4.2-12 Random Forest Evaluation Metrics

The Random forest model results showed a mean squared error of 9.87%. The high accuracy rate and precision of 90.12% and 89.14%, respectively, shows that the True positive, True negative cases, and false-positive case were predicted properly. The recall and f1 score indicate good false-negative predictions. The model is performing better in terms of accuracy and precision. The ROC and AUC scores also show good confidence in determining the student class label.

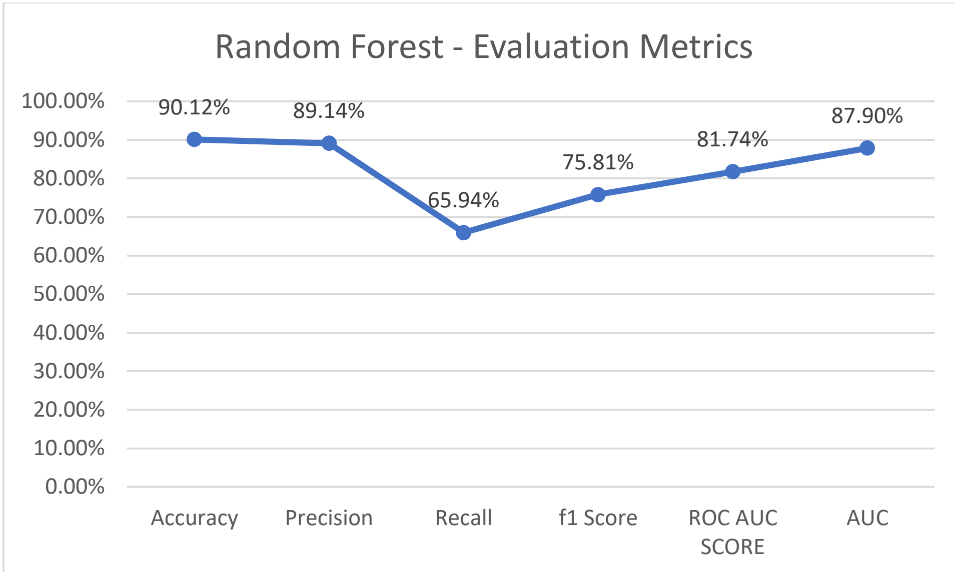
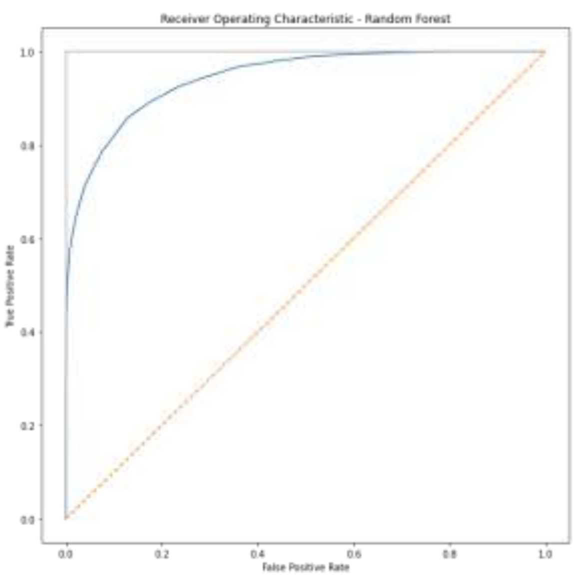


Figure 4.2-19 Random Forest Evaluation Metric Comparison

The figure above demonstrates the comparisons of the evaluation metrics from the decision tree model that shows that the model provided better accuracy, precision, and decent recall, f1 score. The AUC depicts that the model performed reasonably better for the prediction.



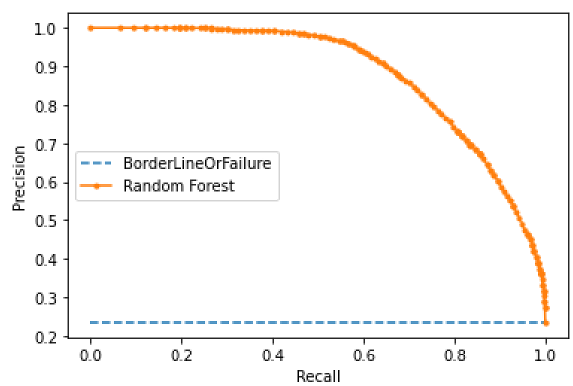


Figure 4.2-20 ROC AUC Curve – Random Forest

Figure 4.2-21 Precision-Recall Curve – Random Forest

The above figures are ROC AUC Curve and the precision-recall curve for the Random Forest model. The area under the ROC curve and the precision-recall curve should be higher for depicting the good performance of the machine learning model. The output suggested that the random forest model performed reasonably better in the prediction tasks.

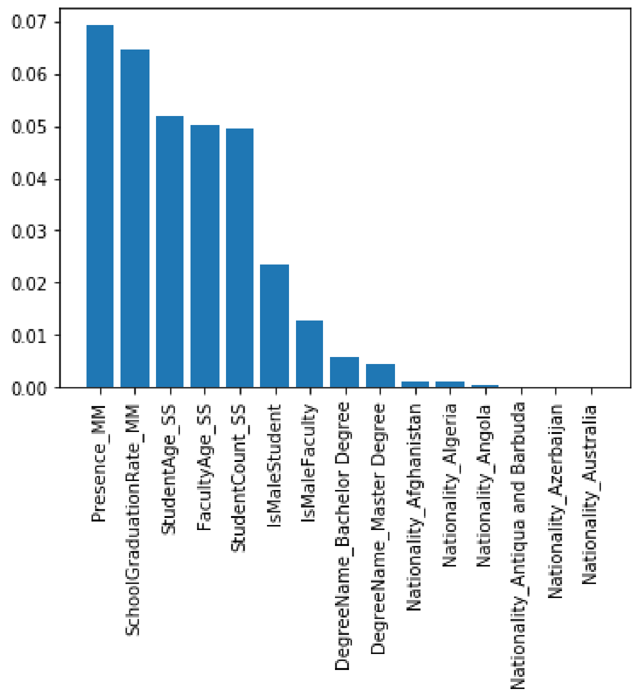


Figure 4.2-22 Feature Importance - Random Forest

The above figure demonstrates the top fifteen (15) attributes that impacted the Random Forest model to make decisions on the prediction of student performance for the class label. The highest influential attributes were Presence (Attendance Attribute) and School Graduation Rate (Grades Achieved in Last Institution). The other important features include the age of the student, the age of the teaching faculty member, Student Count (Class Size), gender of the student, gender of the teaching faculty member, Degree attribute of master's degree, and Bachelor's Degree-seeking students. The only significant nationality column identified were Afghanistan, Algeria, and Angola.

4.2.7 AdaBoost Classifier

The section provides the results and analysis for the AdaBoost classifier.

Class Label	Precision	Recall	F1-Score
0	0.82	0.94	0.87
1	0.61	0.32	0.42

Table 4.2-13 ADA Boost Classification Report

The classification report for ADA Boost Classifier shows high precision, recall, and F1 scores of 82%, 94%, and 87% respectively for predicting the class label of regular students as compared to predicting the student for border line and failure where the precision, recall and F1 Scores are 61%, 32%, and 42% respectively.

Accuracy	Precision	Recall	f1 score	ROC AUC SCORE	AUC	Mean Squared Error
79.29%	61.17%	32.14%	42.14%	62.94%	54.50%	20.70%

Table 4.2-14 ADA Boost Evaluation Metrics

The Random forest model results showed a mean squared error of 20.70%. The accuracy rate of 79.29% shows that the True positive and True negative cases were predicted decently. The precision, recall, and f1 score indicate that the model is not performing well for false-positive and false-negative cases. The ROC and AUC scores show low confidence in determining the student class label.

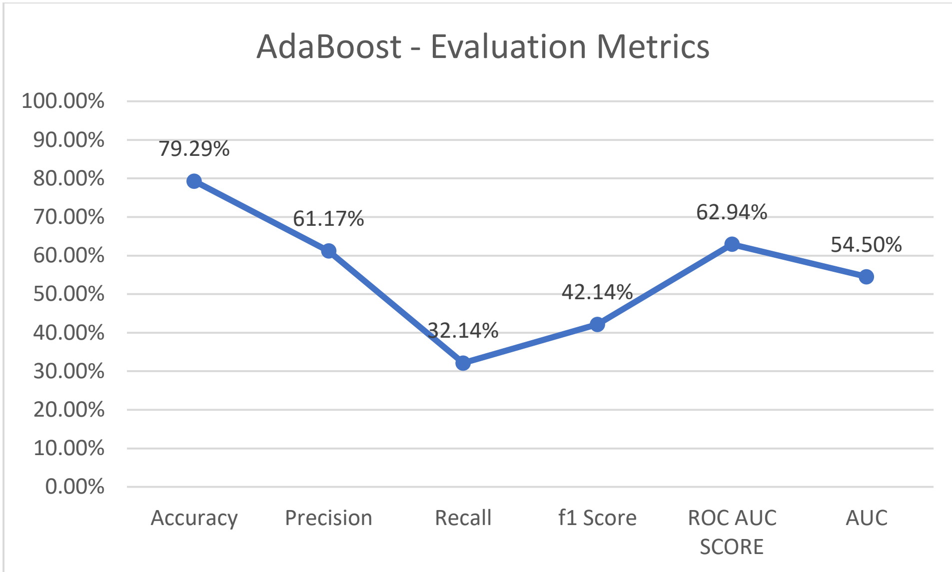
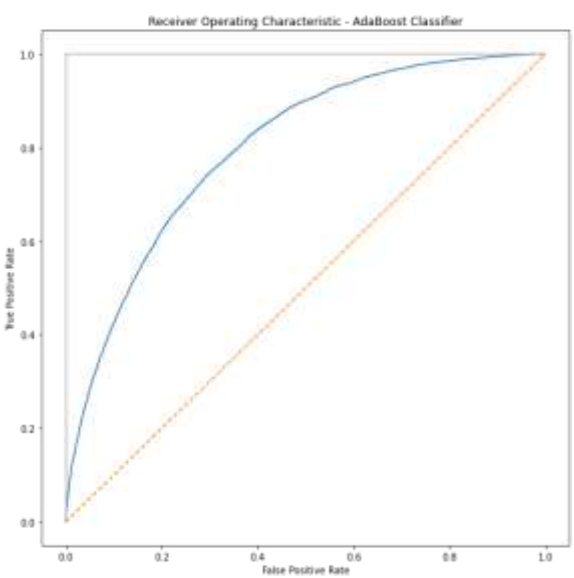


Figure 4.2-23 AdaBoost Evaluation Metric Comparison

The figure above demonstrates the comparisons of the evaluation metrics from the AdaBoost Model that shows that the model provided better accuracy. However, precision, recall, and f1 score were on the lower side. The AUC depicts that the model did not perform well for the prediction.



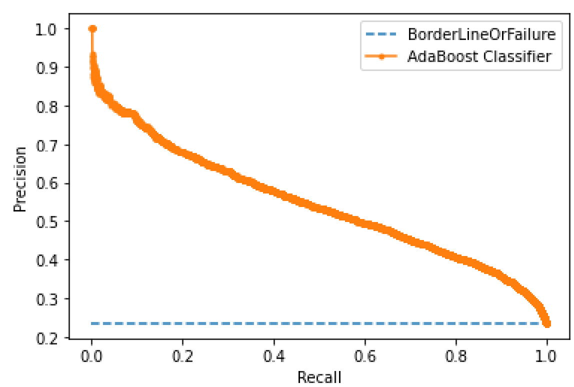


Figure 4.2-24 ROC AUC Curve – ADA Boost

Figure 4.2-25 Precision-Recall Curve – ADA Boost

The above figures are ROC AUC Curve and the precision-recall curve for the Ada Boost model. The area under the ROC curve and the precision-recall curve should be higher for depicting the good performance of the machine learning model. The output suggested that the Ada boost model performed averagely in the prediction tasks.

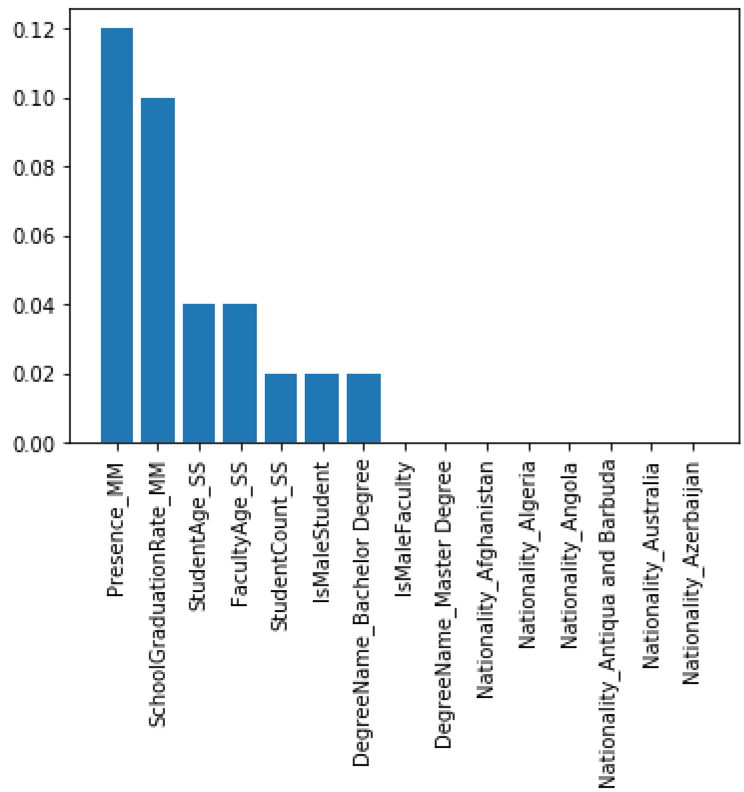


Figure 4.2-26 Feature Importance – AdaBoost

The above figure demonstrates the top fifteen (15) attributes that impacted the Ada Boost classifier model to make decisions on the prediction of student performance for the class label. The highest influential attributes were Presence (Attendance Attribute) and School Graduation Rate (Grades Achieved in Last Institution). The other important features include the age of the student, the age of the teaching faculty member, Student Count (Class Size), gender of the student, Degree attribute of Bachelor Degree-seeking students.

4.2.8 Bagging Classifier

The section provides the results and analysis for the Bagging classifier.

Class Label	Precision	Recall	F1-Score
0	0.9	0.96	0.93
1	0.82	0.66	0.73

Table 4.2-15 Bagging Classification Report

The classification report for bagging classifier shows high precision, recall, and F1 scores of 90%, 96%, and 93% respectively for predicting the class label of regular students as compared to predicting the student for border line and failure where the precision, recall, and F1 Scores are 82%, 66%, and 73% respectively.

Accuracy	Precision	Recall	f1 score	ROC AUC SCORE	AUC	Mean Squared Error
88.53%	81.88%	65.63%	72.86%	80.59%	82.60%	11.46%

Table 4.2-16 Bagging Evaluation Metrics

The Bagging classification model results showed a mean squared error of 11.46%. The high accuracy rate and precision of 88.53% and 81.88%, respectively, show that the True positive, True negative and false-positive cases were predicted well, whereas the Recall and f1 score indicates that the model is performing decently well for false-positive and false-negative cases. The ROC and AUC scores show good confidence in determining the student class label.

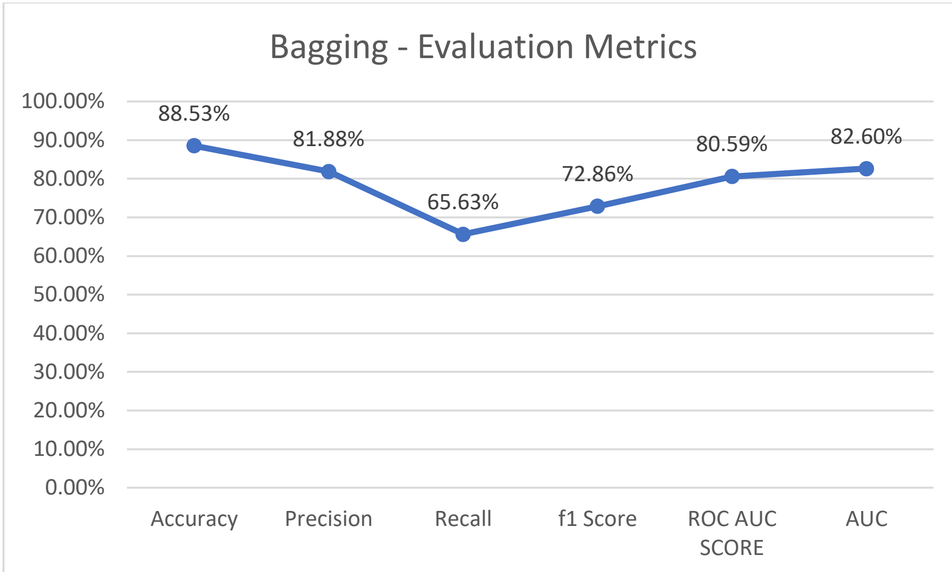


Figure 4.2-27 Bagging classifier Evaluation Metric Comparison

The figure above demonstrates the comparisons of the evaluation metrics from the bagging classifier Model that shows that the model provided better accuracy and precision. However, recall and f1 scores were decent scores. The AUC depicts that the model performed confidently for the prediction.

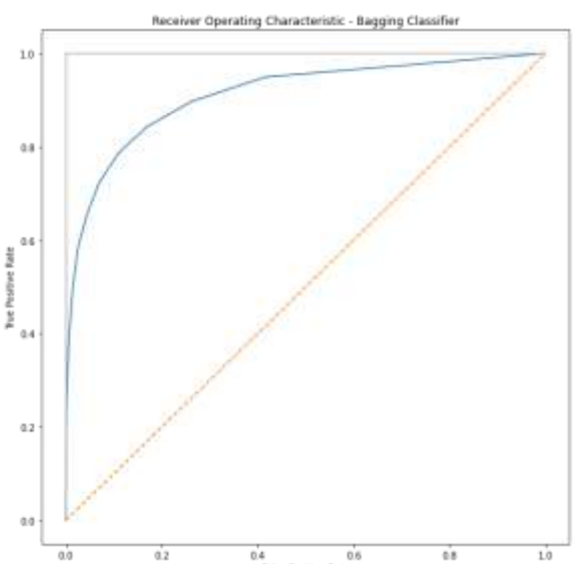


Figure 4.2-28 ROC AUC Curve – Bagging Classifier

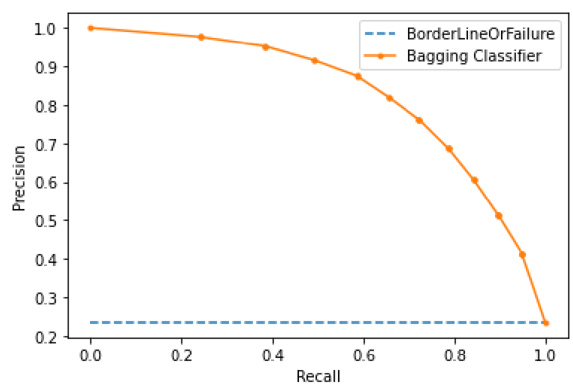


Figure 4.2-29 Precision-Recall Curve – Bagging Classifier

The above figures are ROC AUC Curve and the precision-recall curve for the bagging classifier model. The area under the ROC curve and the precision-recall curve should be higher for depicting the good performance of the machine learning model. The output suggested that the bagging model performed better in the prediction tasks.

4.2.9 Voting Classifier

The section provides the results and analysis for the Voting classifier.

Class Label	Precision	Recall	F1-Score
0	0.87	0.94	0.9
1	0.73	0.54	0.62

Table 4.2-17 Voting Classifier Classification Report

The classification report for voting Classifier shows high precision, recall, and F1 scores of 87%, 94%, and 90% respectively for predicting the class label of regular students as compared to predicting the student for border line and failure where the precision, recall, and F1 Scores are 73%, 54%, and 62% respectively.

Accuracy	Precision	Recall	f1 score	ROC AUC SCORE	AUC	Mean Squared Error
84.45%	72.74%	53.92%	61.93%	73.86%	71.00%	15.54%

Table 4.2-18 Voting Classifier Evaluation Metrics

The voting classification model results showed a mean squared error of 15.54%. The high accuracy rate and precision of 84.45% and 72.74%, respectively, show that the True positive, True negative and false-positive cases were predicted well, whereas the Recall and f1 score indicates that the model is performing decently well for false-positive and false-negative cases. The ROC and AUC scores show decent confidence in determining the student class label.

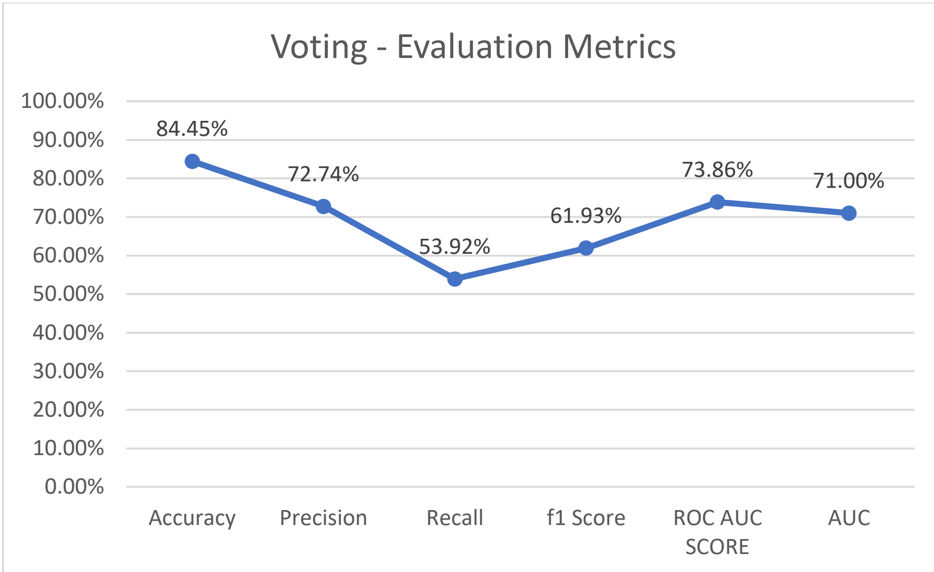


Figure 4.2-30 Voting classifier Evaluation Metric Comparison

The figure above demonstrates the comparisons of the evaluation metrics from the voting classifier Model that shows that the model provided better accuracy and precision. However, recall and f1 scores were decent scores. The AUC depicts that the model performed with average confidence for the prediction.

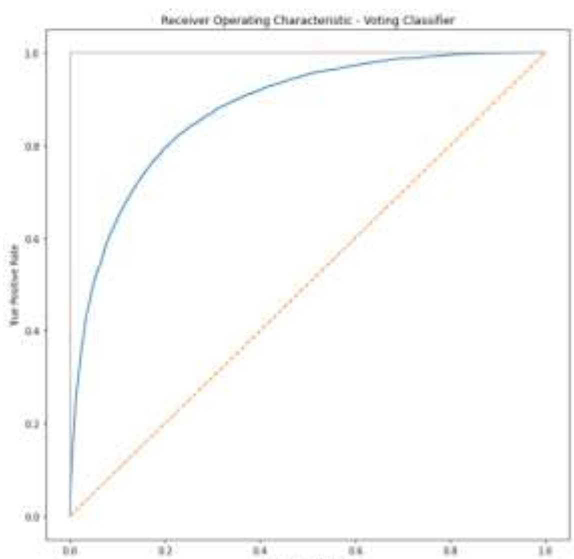


Figure 4.2-31 ROC AUC Curve – Voting Classifier

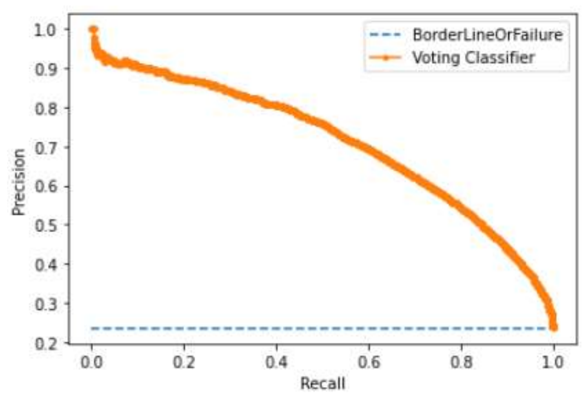


Figure 4.2-32 Precision-Recall Curve – Voting Classifier

The above figures are ROC AUC Curve and the precision-recall curve for the voting classifier model. The area under the ROC curve and the precision-recall curve should be higher for depicting the good performance of the machine learning model. The output suggested that the voting model performed decently better in the prediction tasks.

4.3 Summary

This section provides the summarized discussion for the metric evaluation for the nine machine learning algorithms, the identified important features, and the relation of the important feature to the grades achieved by the student.

4.3.1 Evaluation Metric Summary

Evaluation Metric	Machine Learning Algorithms								
	KNN	Naïve Bayes	SVM	Logistic Regression	Decision Tree	Random Forest	ADA boost	Bagging Classifier	Voting Classifier
Accuracy	83.51%	77.43%	81.84%	81.57%	83.92%	90.12%	79.29%	88.53%	84.45%
Precision	64.74%	63.31%	67.12%	65.90%	71.75%	89.14%	61.17%	81.88%	72.74%
Recall	65.23%	8.96%	44.26%	44.45%	51.90%	65.94%	32.14%	65.63%	53.92%
f1 Score	64.99%	15.70%	53.34%	53.09%	60.23%	75.81%	42.14%	72.86%	61.93%
ROC AUC SCORE	77.17%	53.69%	68.81%	68.70%	72.82%	81.74%	62.94%	80.59%	73.86%
AUC	73.00%	43.00%	63.30%	62.30%	66.10%	87.90%	54.50%	82.60%	71.00%
Mean Squared Error	16.49%	22.57%	18.16%	18.42%	16.07%	9.87%	20.70%	11.46%	15.54%

Table 4.3-1 Summary of Algorithm Evaluation Metrics

The above table is the summary of the performance of all the machine learning algorithms that were implemented for research purposes. The summary shows that the Random forest classifier performed better than other algorithms providing the least mean squared error.

4.3.2 Identified top fifteen Important Features

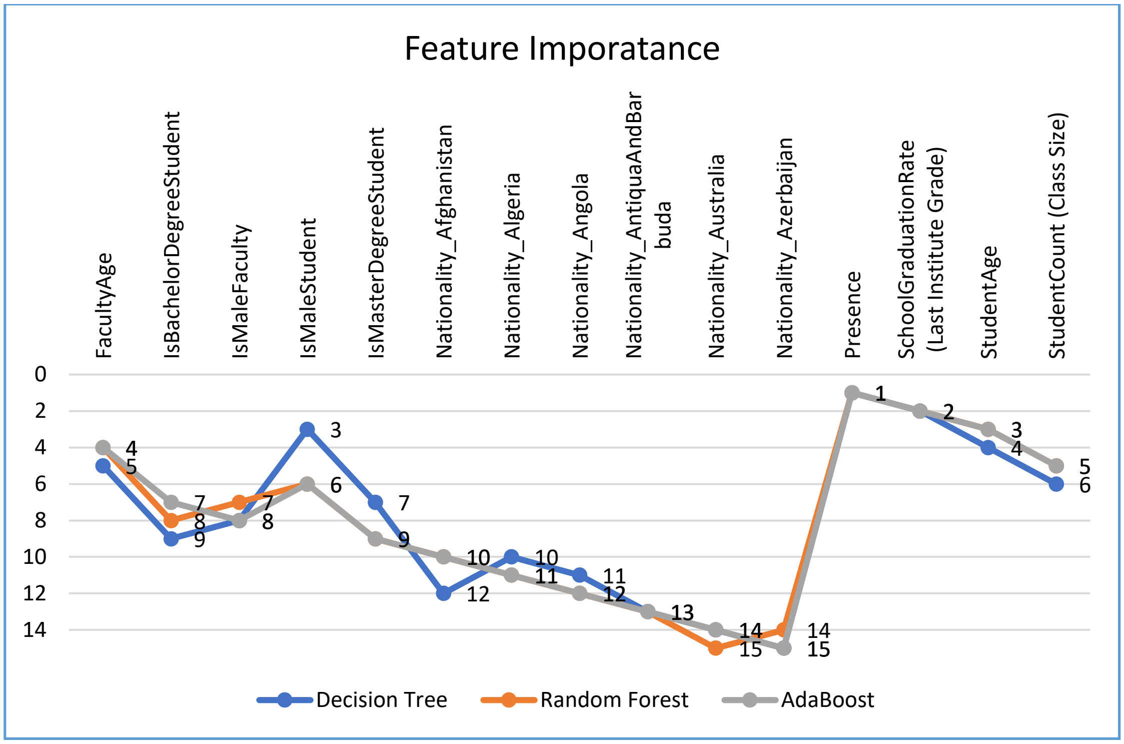


Figure 4.3-1 Summary of Feature Importance

The Above figure provides a summary of the top fifteen (15) important features ranked by a decision tree, random forest, and Adaboost classification algorithms. The top 15 features identified by the three classifiers are the same; however, they are slightly ranked differently. The above figure clearly demonstrates that presence in class is the top rank important feature followed by the grades achieved in the last institution. These two features are unanimously ranked top by all three classifiers. The ranking of the features based on the average of three classifiers are Presence (Attendance in the class), SchoolGraduationRate (Last Institute Grade), studentAge, FacultyAge, IsMaleStudent (Gender of Student), StudentCount (Class Size), IsMaleFaculty (Gender of Instructor), IsBachelorDegreeStudent, IsMasterDegreeStudent, Nationality_Afghanistan, Nationality_Algeria, Nationality_Angola, Nationality_AntiquaAndBarbuda, Nationality_Australia, and Nationality_Azerbaijan.

4.3.3 Relation of the Important Features with the grades of students

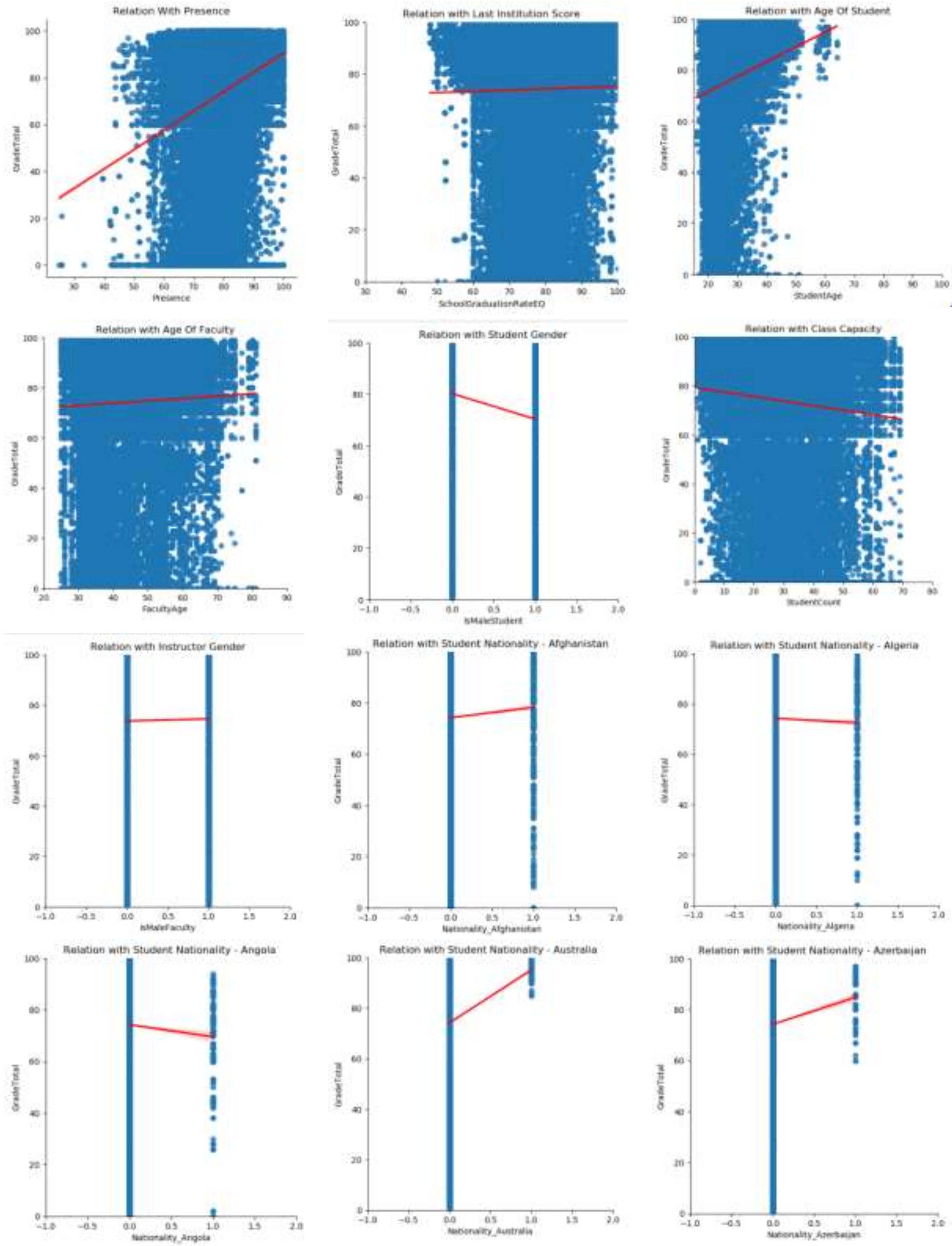


Figure 4.3-2 Relation of the Important Features with the grades of students

The above set of figures provides the correlation of the important features that contributes towards the accurate prediction of the student performance with the student grades. The attendance of the students in class have a strong positive relation with the grades. The last institution grades have a slight positive correlation that depicts that the past grade had a little impact on student performance. The student age showed strong positive correlation. The attribute of instructor age also depicts positive correlation. Another important attribute of class size showed negative correlation. As the number of the student increases the grades of the students are decreased. The figure also showed that female students are more likely to achieve good grades as compared to male students. Although the gender of the instructor is identified as important feature, the male faculty members tends to contribute slightly more towards student performance. Among the student nationality features Afghanistan, Australia and Azerbaijan showed positive correlation however, Algeria and Angola showed negative correlation.

5 Conclusion and recommendations for future work

This chapter forms the concluding section, including the conclusion, Contributions, limitations, and recommendations for future studies.

The data collected from the university were applied the data preprocessing steps in order to prepare for the application of machine learning algorithms. The record set was filtered to 126698 total records. The research utilized various tools for preparing and analysis of data. The tools used are SQL Queries, MS Excel, Power BI Desktop, and Jupyter Notebook to execute Python codes. According to the studied research papers, the most used machine learning algorithms were K-Nearest Neighbor (KNN), Naïve Bayes, Support Vector Machines, and Decision Trees. Nine machine algorithms were selected in the research that included K Nearest neighbor (KNN), Naïve Bayes, Decision Tree, SVM, Logistic Regression, Random Forest, AdaBoost Bagging Classifier, and Voting Classifier. The machine learning algorithms were evaluated based on seven (7) evaluation metrics that included Accuracy, Precision, Recall, F1 Score, ROC AUC score, AUC Score, and Mean Squared Error.

5.1 Conclusion

In this study, we primarily aim to predict the performance of the students studying in a private university in the United Arab Emirates and provide an intelligent framework using machine learning algorithms. This will enable the university to improve the quality of students, improve student grades, enhance student retention rates and enable the student to graduate on time.

The nature of the research topic shows that the primary evaluation metric that could be considered is accuracy, and other metrics can then be used to evaluate the algorithm in totality. By Analyzing the evaluation metrics of the algorithms, it shows that for this particular research data, Random forest performed better than other algorithms with an Accuracy of 90.12%, precision of 89.14%, recall of 65.94%, F1 score of 75.84%, ROC AUC area of 81.74%, AUC of 87.90%. The metric evaluation analysis also depicts that the Random forest model provided the least mean square error value of 9.87%. The most important feature contributing towards the better student performance is the attendance of student in class. The class size that is the number of students in class has a negative impact on the student performance. The university should implement optimum class size and encourage students to attend majority of the class to achieve higher success.

The future study can include more data attributes that provide more insights on the granular course assessments and student behavior. The research can further continue to enhance the study by adding a recommendation system for the predicted students at risk for additional training courses, or selecting elective courses suited to their strength to improve their overall performance.

5.2 Contributions

The research contributes to the practical essence of the academic institutions. The research provides a basis for retrieving useful insights to improve the overall quality of the learning ecosystem. The contribution includes

- To form a framework for enhancement in the advising process of students of the academic institutions.
- To take proactive actionable measures by academic institutions to prevent students from failing or scoring low marks.
- To improve the academic reputation as increasing the success for students impacts the quality of students.
- To improve the timely graduation rates for the academic institutions.

5.3 Limitations

This section explains the limitations of the study of the research.

- Firstly, the research data was collected from a specific educational institution that had built its own in-house learning management system.
- The data included in the research is the limited, restricted data information allowed for this research.
- The aspects of student behavior, financial implications, and scholarship information will also be useful for making predictions that may impact the performance of the student.
- Further, the academic institution follows the American style of education that may provide different results when applied to other curriculum styles.
- The research is primarily dependent on the summative grades, and there is no consideration of the formative assessments that also play an important part in the learning process.
- Limited computational power to run the algorithms for attempting to perform analysis for various parameters.

5.4 Recommendations for future work

The recommendation for future work is

- To study the student performance by applying more machine learning algorithms from other aspects including utilized resources for learning purposes, participation in extracurricular activities, Not limiting to summative grades but to enhance the scope to formative assessments, including student behavior in the class with classmates and instructors, participation in discussion groups, and in-class assignments.
- To study recommendation systems for academic institutions for providing specific training courses to improve the course understanding.

- To study recommendation systems for academic institutions for choosing the elective courses that would potentially provide them high scores and improve their grades.
- To predict student strength areas and recommend courses for career guidance.

References

[1] Yang, F. and Li, F.W., 2018. Study on student performance estimation, student progress analysis, and student potential prediction based on data mining. *Computers & education*, *123*, pp.97-108.

[2] Fok, W.W., He, Y.S., Yeung, H.A., Law, K.Y., Cheung, K.H., Ai, Y.Y. and Ho, P., 2018, May. Prediction model for students' future development by deep learning and tensorflow artificial intelligence engine. In *2018 4th international conference on information management (ICIM)* (pp. 103-106). IEEE.

[3] Xu, J., Moon, K.H. and Van Der Schaar, M., 2017. A machine learning approach for tracking and predicting student performance in degree programs. *IEEE Journal of Selected Topics in Signal Processing*, *11*(5), pp.742-753.

[4] Su, Y., Liu, Q., Liu, Q., Huang, Z., Yin, Y., Chen, E., Ding, C., Wei, S. and Hu, G., 2018, April. Exercise-enhanced sequential modeling for student performance prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 32, No. 1).

[5] Xu, J., Han, Y., Marcu, D. and Van Der Schaar, M., 2017, February. Progressive prediction of student performance in college programs. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 31, No. 1).

[6] Hamoud, A., Hashim, A.S. and Awadh, W.A., 2018. Predicting student performance in higher education institutions using decision tree analysis. *International Journal of Interactive Multimedia and Artificial Intelligence*, *5*, pp.26-31.

[7] Mohammadi, M., Dawodi, M., Tomohisa, W. and Ahmadi, N., 2019, February. Comparative study of supervised learning algorithms for student performance prediction. In *2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)* (pp. 124-127). IEEE.

[8] Oyerinde, O.D. and Chia, P.A., 2017. Predicting students' academic performances–A learning analytics approach using multiple linear regression. International Journal of Computer Applications (0975 – 8887) Volume 157 – No 4, January 2017

[9] Trakunphutthirak, R., Cheung, Y. and Lee, V.C., 2019, July. A study of educational data mining: Evidence from a thai university. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 33, No. 01, pp. 734-741).

[10] Kurniadi, D., Abdurachman, E., Warnars, H.L.H.S. and Suparta, W., 2019, December. A proposed framework in an intelligent recommender system for the college student. In 4th Annual Applied Science and Engineering Conference *Journal of Physics: Conference Series* (Vol. 1402, No. 6, p. 066100). IOP Publishing.

[11] Raut, A.B. and Nichat, M.A.A., 2017. Students performance prediction using decision tree. *International Journal of Computational Intelligence Research*, *13*(7), pp.1735-1741.

[12] Amra, I.A.A. and Maghari, A.Y., 2017, May. Students performance prediction using KNN and Naïve Bayesian. In *2017 8th International Conference on Information Technology (ICIT)* (pp. 909-913). IEEE.

[13] Kumar, M. and Salal, Y.K., 2019. Systematic review of predicting student's performance in academics. *Int. J. of Engineering and Advanced Technology*, *8*(3), pp.54-61.

[14] Bonde, S.N. and Kirange, D.K., 2018, April. Survey on evaluation of student's performance in educational data mining. In *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)* (pp. 209-213). IEEE.

[15] Rahman, M.H. and Islam, M.R., 2017, December. Predict Student's Academic Performance and Evaluate the Impact of Different Attributes on the Performance Using Data Mining Techniques. In *2017 2nd International Conference on Electrical & Electronic Engineering (ICEEE)* (pp. 1-4). IEEE.

[16] Cazarez, R.L.U. and Martin, C.L., 2018. Neural Networks for predicting student performance in online education. *IEEE Latin America Transactions*, *16*(7), pp.2053-2060.

[17] Vora, D.R. and Iyer, K., 2018. EDM–survey of performance factors and algorithms applied. *International Journal of Engineering & Technology*, *7*(2.6), pp.93-97.

[18] Liu, Q., Huang, Z., Yin, Y., Chen, E., Xiong, H., Su, Y. and Hu, G., 2019. Ekt: Exercise-aware knowledge tracing for student performance prediction. *IEEE Transactions on Knowledge and Data Engineering*, *33*(1), pp.100-115.

[19] Ajibade, S.S.M., Ahmad, N.B.B. and Shamsuddin, S.M., 2019, August. Educational data mining: enhancement of student performance model using ensemble methods. In *IOP Conference Series: Materials Science and*

Engineering (Vol. 551, No. 1, p. 012061) *Joint Conference on Green Engineering Technology & Applied Computing*. IOP Publishing.

[20] Hussain, S., Muhsion, Z.F., Salal, Y.K., Theodorou, P., Kurtoglu, F. and Hazarika, G.C., 2019. Prediction Model on Student Performance based on Internal Assessment using Deep Learning. *iJET*, *14*(8), pp.4-22.

[21] Zollanvari, A., Kizilirmak, R.C., Kho, Y.H. and Hernández-Torrano, D., 2017. Predicting students' GPA and developing intervention strategies based on self-regulatory learning behaviors. *IEEE Access*, *5*, pp.23792-23802.

[22] Shahiri, A.M. and Husain, W., 2015. A review on predicting student's performance using data mining techniques. *Procedia Computer Science*, *72*, pp.414-422.

[23] Wang, F., Liu, Q., Chen, E., Huang, Z., Chen, Y., Yin, Y., Huang, Z. and Wang, S., 2020, April. Neural cognitive diagnosis for intelligent education systems. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, No. 04, pp. 6153-6161).

[24] Geden, M., Emerson, A., Rowe, J., Azevedo, R. and Lester, J., 2020, April. Predictive student modeling in educational games with multi-task learning. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, No. 01, pp. 654-661).

[25] Harvey, J.L. and Kumar, S.A., 2019, December. A practical model for educators to predict student performance in K-12 education using machine learning. In *2019 IEEE Symposium Series on Computational Intelligence (SSCI)* (pp. 3004-3011). IEEE.

[26] Tarik, A., Aissa, H. and Yousef, F., 2021. Artificial intelligence and machine learning to predict student performance during the COVID-19. *Procedia Computer Science*, *184*, pp.835-840.

[27] Alamri, R. and Alharbi, B., 2021. Explainable student performance prediction models: a systematic review. *IEEE Access*.

[28] Kumar, A.D., Selvam, R.P. and Palanisamy, V., 2021, March. Hybrid classification algorithms for predicting student performance. In *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)* (pp. 1074-1079). IEEE.

[29] Guo, T., Xia, F., Zhen, S., Bai, X., Zhang, D., Liu, Z. and Tang, J., 2020, April. Graduate employment prediction with bias. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, No. 01, pp. 670-677).

[30] Luo, Y. and Pardos, Z., 2018, April. Diagnosing university student subject proficiency and predicting degree completion in vector space. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 32, No. 1).

[31] Smirnov, I., 2018, June. Predicting PISA scores from students' digital traces. In *Twelfth International AAAI Conference on Web and Social Media*.

[32] Naito, J., Baba, Y., Kashima, H., Takaki, T. and Funo, T., 2018, April. Predictive modeling of learning continuation in preschool education using temporal patterns of development tests. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 32, No. 1).

[33] Mengash, H.A., 2020. Using data mining techniques to predict student performance to support decision making in university admission systems. *IEEE Access*, *8*, pp.55462-55470.

[34] Zughoul, O., Momani, F., Almasri, O.H., Zaidan, A.A., Zaidan, B.B., Alsalem, M.A., Albahri, O.S., Albahri, A.S. and Hashim, M., 2018. Comprehensive insights into the criteria of student performance in various educational domains. *IEEE access*, *6*, pp.73245-73264.

[35] Hegde, V. and Prageeth, P.P., 2018, January. Higher education student drop-out prediction and analysis through educational data mining. In *2018 2nd International Conference on Inventive Systems and Control (ICISC)* (pp. 694-699). IEEE.

[36] Qu, S., Li, K., Zhang, S. and Wang, Y., 2018. Predicting achievement of students in smart campus. *IEEE Access*, *6*, pp.60264-60273.

[37] Petkovic, D., Sosnick-Pérez, M., Okada, K., Todtenhoefer, R., Huang, S., Miglani, N. and Vigil, A., 2016, October. Using the random forest classifier to assess and predict student learning of software engineering teamwork. In *2016 IEEE Frontiers in education conference (FIE)* (pp. 1-7). IEEE.

[38] Kondo, N., Okubo, M. and Hatanaka, T., 2017, July. Early detection of at-risk students using machine learning based on LMS log data. In *2017 6th IIAI international congress on advanced applied informatics (IIAI-AAI)* (pp. 198-201). IEEE.

[39] Yossy, E.H. and Heryadi, Y., 2019, December. Comparison of Data Mining Classification Algorithms for Student Performance. In *2019 IEEE International Conference on Engineering, Technology and Education (TALE)* (pp. 1-4). IEEE.

[40] Ahmed, N.S. and Sadiq, M.H., 2018, October. Clarify of the random forest algorithm in an educational field. In *2018 international conference on advanced science and engineering (ICOASE)* (pp. 179-184). IEEE.