Contents

Abstract

In today's education landscape, predicting student potential is crucial for driving innovation and societal progress. This research delves into a comprehensive analysis of student's Academic performance, considering factors like backgrounds, demographics, and academic achievements.

This research deals with problems like dependence on sponsorship and language barriers when working with the dataset on time graduation, employment status, student attrition; it examines how effective are machine learning algorithms in predicting such outcomes as on-time graduation, employment status and student retention. The employed algorithms consist of decision trees, logistic regression models, XGBoost, random forests and K-nearest neighbors (KNN). To improve model performance and efficiency feature selection methods especially K best method is used to reduce dimensionality and boost prediction accuracy.

To test the models for all other experiments each algorithm is first trained and tested without any feature selection to get baseline accuracy metrics. Then the K-best feature selection method is added to the model pipeline so that relevant features for prediction can be determined. After selecting these features, the datasets are retrained using them, benchmarked against baseline results.

Integrating K-best feature selection improves significantly all models' predictive accuracies across three student outcome variables. Certain algorithms, it should be noted, perform better than others after the feature selection in predicting some outcomes. XGBoost is one such algorithm showing an increased accuracy in forecasting on-time graduation while logistic regression does well for employment status prediction. In addition to that, it helps how reduction of overfitting and computational overhead can be achieved through feature selection.

This study provides crucial insights into student progress analysis by illustrating the efficiency of feature selection approaches in optimizing machine learning models for critical student outcomes and survival analysis. These findings have implications for educational institutions and policy makers with an interest in data driven approaches to enhancing both student success and retention programs which underscores the significance of customized predictive modeling within academic settings.

Keywords:

Background information, Poor performance, Joblessness because of failure, Challenges faced by tertiary learning institutions, Withdrawal from college due to various reasons including language problem or sponsorship dependency, graduating on time from university or college, on campus jobs availability, Techniques for minimizing mistakes, KNN, RANDOM FOREST AND NAIVE BAYES are examples of machine learning models used for this purpose. Student outcome survival analysis with Cox Proportional Hazards model, deploying solution methods that focus on certain outcomes rather than rules that are normally given by experts during academic decision-making phase, Innovation within school systems

Chapter 1: Introduction

1.1 Context and Research Objectives

Machine learning algorithms for predicting student outcomes are evaluated in this study. The purpose of the study is to examine how K-best selection can enhance predictive accuracy by pinpointing pertinent features in a dataset and giving them higher priority. Besides, the study compares predictive models before feature selection with those after it has been done. The analysis measures metrics such as accuracy, precision, recall, and F1-score on a quantitative scale to gauge how feature selection improves their abilities to predict certain aspects of the students' behavior.

The research objectives of student's performance are:

1. Student Attrition Analysis
2. On-Time Graduation
3. On-Campus Employment

Overall, the research objectives are designed to advance understanding and knowledge in the field of student performance analysis, with a focus on leveraging machine learning and feature selection methodologies to improve predictive modeling for key student outcomes.

1.2 Scope and Significance

The extent and importance of this research project will define the limits of inquiry and highlight its significance in terms of student performance analysis. By sketching the scope of work, the study sets down precise boundaries within which the inquiry is to be conducted, at the same time stating how broad it ranges.

The scope of this research focuses on the use machine learning algorithms that can select key indicators to predict outcomes including on-time graduation, employment status, and student attrition. By concentrating on these leading indices, the inquiry seeks to address significant challenges faced by education institutions, policy makers as well as stakeholders mainly related to students' achievement and retaining in various educational institutions.

Moreover, the range also covers multiple machine learning algorithms' evaluation and comparison such as random forests, logistic regression, decision trees XGBoost or K-nearest neighbors. The study aims at establishing how each algorithm can predict correct outcomes for students hence helping select a better model.

In addition to that, the scope includes implementation techniques for a specific type of feature selection called "K best". By systematically identifying the most important features for analysis, the study eto enhance the accuracy and efficiency of predictive modeling for student outcomes, thus maximizing the utility of available data resources.

The research extends its significance beyond academia to benefit society at large. By enabling targeted interventions to enhance student success and retention, it contributes to socioeconomic progress and workforce development.

1.3 Methodology

The proposed methodology for this research project combines quantitative analysis with machine learning techniques to investigate predictive modeling for student outcomes in education. The methodology is designed to systematically analyze a dataset containing student demographic information, academic performance metrics, and outcome variables such as graduation status, employment status, and attrition.

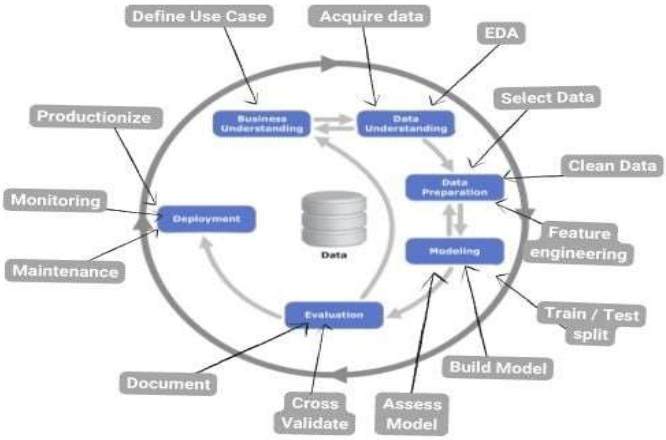


Figure: CRISP-DM Methodology

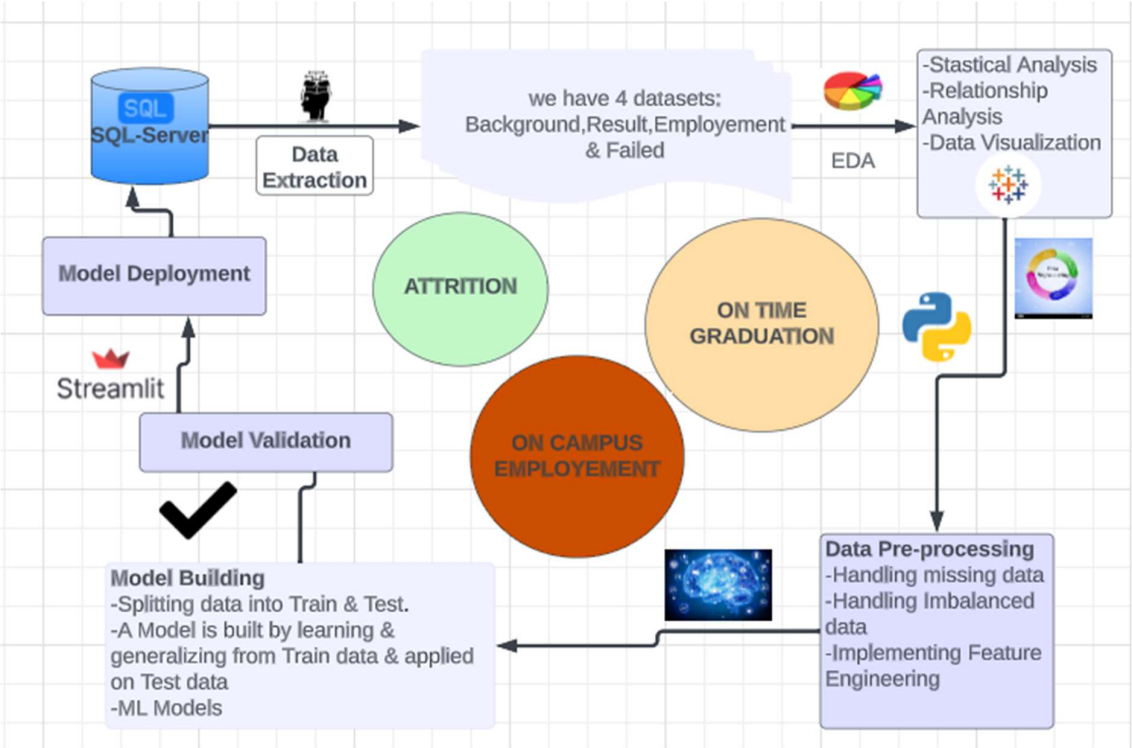


Figure: Architecture diagram for the student performance analysis

Pre-processing of data involves cleaning, transforming and normalizing datasets to improve the quality of data. Algorithm selection entails selecting machine learning algorithms that can estimate student outcomes, such as logistic regression, XGBoost, decision trees, K-nearest neighbors and random forests.

The K-best method has been used by me in order to show the most important factors influencing students' results. The data was divided into training (70%) and testing (30%), before feeding it into the models. In this case, after taking up the entire training dataset with all its features we had to evaluate the performance on our models using a test dataset from which we derived different kinds of metrics.

Accuracy, Precision, Recall and F1-score will be employed so as to measure how good our models are. We shall also use cross-validation technique to verify our models' power and credibility. Furthermore that's why you will find us doing sensitivity analysis which ascertains how robust our models are & how they respond when changes occur within certain threshold values or scenario settings, these steps will guide us in determining how well our models can predict accurately important student outcomes.

The model evaluation shall be done on appropriate metrics such as accuracy, precision recall and F1-score. Cross-validation techniques will be used to ascertain that the models are strong enough and reliable enough for use in practice.

Cross-validation approaches to guarantee that the models are robust and reliable. The latter is also called sensitivity analysis which evaluates the stability of these models and how they are impacted by amendments in parameters or assumptions.

I have evaluated these models using important metrics such as accuracy, recall, precision, and F1-score of their prediction performance. The models can be strong and dependable if subjected to cross-validation techniques. Hence, sensitivity analysis shows model's stability with respect to changes in parameters or assumptions.

1.4 Objectives and Constraints

The study dwells on predictive modeling of some key student outcomes like graduation status, employment status and attrition rate. However, it should be noted that other factors not captured in the current dataset might affect student outcomes.

Objectives:	Constraints:
Minimize Student Attrition and Maximize the education quality of the college and improve the rank of the college.	Sponsorship dependency is affecting the study of participants.
Maximize On-Time Graduation	The language barrier is causing inadequacy in data understanding.
Maximize On-Campus Employment	Minority class is affecting one of the datasets and causing disparity.

Table: Objectives and Constraint

Chapter 2: Literature Survey

2.1 Student Performance Analysis

In education, student performance analysis is a complicated process that aims to understand and improve academic achievement, progression, and success among students. This analytical procedure involves collecting, interpreting, and evaluating various data points and indicators for assessing individual and collective student performance.

Basically, student performance analysis helps identify patterns, trends or factors that influence academic outcomes like grades, standardized test scores, graduation rates and retention rates. educators can develop their institutions by examining these metrics in combination with demographic information, socioeconomic factors as well as educational experiences.

Analyzing student performance involves aiding teacher decision making based on data analytics. By leveraging on data analytics and statistical methods teachers can use them to identify at-risk students who may require additional support; they may also have the ability to adapt instructional strategies to meet different learning needs thereby allocating resources effectively for improving overall student results.

By comparing outcomes over time of different groups or cohorts of students policy makers can assess the impacts of interventions aimed at enhancing educational equality as well as access plus quality.

To evaluate the impact of interventions whose aim is to enhance educational equity, access and quality, student progress can be tracked over time and different groups or cohorts compared.

In recent years, student performance analysis has been transformed by technology advancements and data analytics leading to more sophisticated methods of collecting and interpreting data. For example, large datasets can be broken down using machine learning algorithms in order to identify complicated relationships and patterns among students' behavior and learning outcomes.

Student performance analysis is fundamental in improving educational effectiveness as well as promoting student success.

2.2 Predictive Modeling Techniques

In order to comprehend and enhance student achievement in education, student performance analysis is very important. It involves data collection and analysis on different dimensions of student learning to detect trends as well as factors that affect academic results like grades and graduation rates. With this information, educators can assess their teaching practices and support systems in order to improve student outcomes.

Student performance analysis has been transformed by technological advancements such as machine learning and predictive analytical techniques. Teachers can use these tools to analyze huge volumes of data quickly; find out patterns; and enable personalized learning approaches.

2.3 Feature Selection Methods

Feature selection methods aim at improving model performance, reducing overfitting, and making it easier for humans to interpret models by concentrating on informative features. Filter methods are one common type of feature selection that evaluate each feature's relevance independently through either statistical test or correlation analysis by discarding less relevant ones.

By iteratively selecting subsets, training the model, and assessing performance, wrapper methods can identify the most valuable feature combinations, though they will help effectively while modelling. Selecting the most important features will improve efficiency and facilitate better decision-making in machine learning applications.

2.4 Literature Gaps

Literature gaps in education research are areas in which existing scholarships are insufficient or notable omissions in addressing important topics or questions. For promoting knowledge and shaping future research agenda, it is essential to identify these gaps within the discipline.

One significant literature gap is student identity intersectionality and educational outcomes in education research. Whereas there have been a lot of studies on the effects of socioeconomic status and gender on student achievement.

Besides, there is little scholarship that examines how technology can be used to create an inclusive and accessible learning environment. There has been some understanding about how technology can improve on educational access as well as suitability for those who have diverse learning needs; however limited information exists about its effectiveness among students with disabilities, English language learners, and other underprivileged categories.

Chapter 3: Methodology

3.1 Data Description

The dataset is collected from an educational institution in Malaysia, and it consists of structured numerical and categorical data. Academic performance metrics measures such as GPA, standardized test scores, course enrollment, and graduation records.

The dataset consists of four different subsets: background information, employment data, failed data, and result data. The background information dataset comprises 35,174 entries with 46 columns. It contains details about students' academic backgrounds, including their program, cohort, degree level, matriculation number, examination results, entrance qualifications, demographic information such as race and gender, as well as sponsorship and registration statuses.

The employment data contains 13,962 entries with 18 columns. It includes information about students' employment status, such as their program, matriculation number, working organization, occupation, location (city, state, country), sector, income, and employment duration.

The failed data consists of 62,381 entries with 11 columns. It provides insights into courses in which students have failed, including details such as the course code, type, credit hours, grade, grade status, and justifications for failure.

Lastly, the result data comprises 96,120 entries with 10 columns. It presents students' examination results, including their program, matriculation number, session semester, examination result, result description (e.g., pass, fail), performance indicators (PNG, PNGK), total credit hours, and year of study.

	Unnamed: 0	FACULTY_CODE	FACULTY_NAME	PROGRAM_CODE	PROGRAM_NAME	COHORT	DEGREE_LEVEL	STUDY_LEVEL	MATRIC_NO	EXAMINATION_RESULT	...	INCOME	STUDENT_STATUS	STUDENT
0	0	FSSH	Faculty of Social Sciences and Humanities	WA02	COMMUNICATION STUDIES	2006/2007-1	FIRST DEGREE	UG	17683	LMK	...	No income specified	4	WI
1	1	FSSH	Faculty of Social Sciences and Humanities	WA02	COMMUNICATION STUDIES	2009/2010-1	FIRST DEGREE	UG	23098	L	...	RM1-RM500	5	
2	2	FSSH	Faculty of Social Sciences and Humanities	WA02	COMMUNICATION STUDIES	2009/2010-1	FIRST DEGREE	UG	23108	L	...	RM1001-RM2000	5	
3	3	FSSH	Faculty of Social Sciences and Humanities	WA02	COMMUNICATION STUDIES	2009/2010-1	FIRST DEGREE	UG	23156	L	...	RM1-RM500	5	
4	4	FSSH	Faculty of Social Sciences and Humanities	WA02	COMMUNICATION STUDIES	2009/2010-1	FIRST DEGREE	UG	23219	L	...	RM1001-RM2000	5	

Figure: Sample screenshot of dataset

Student academic information was collected from 2016 to 2019 based on their graduation year in four categories: background, Employment, failed, and result. I have combined these datasets into a single location with distinct names to make things simpler. I used Google Translator to translate a few columns that were written in Malay into English.

Outcome variables are critical indicators of student success and progression within the educational system. These include on-time graduation status, employment status post-graduation, and indicators of student attrition or dropout. These variables provide insights into the effectiveness of educational programs and interventions in supporting student achievement and retention.

FACULTY CODE	FACULTY NAME	EXAMINATION RESULT	EXAMINATION RESULT DESCRIPTION
FEB	Faculty of Economics and Business	L	Pass
FSSH	Faculty of Social Sciences and Humanities	Exam result not specified	Not Specified
FSCHD	Faculty of Cognitive Sciences and Human Development	LMK	Pass (re-sit failed course)
FRST	Faculty of Resource Science and Technology	GB	Fail and De-registered
FACA	Faculty of Applied and Creative Arts	LBMK	Conditional Pass (re-sit failed course)
FE	Faculty of Engineering	(-MK)	Result (-MK) not specified in description
FCSIT	Faculty of Computer Science and Information Technology	LB1MK	Conditional Pass 1 (re-sit failed course)
FMHS	Faculty of Medicine and Health Sciences	LB	Conditional Pass
FLC	Faculty of Language and Communication	GMS	Na
FBE	Faculty of Built Environment	LB2MK	Conditional Pass 2 (re-sit failed course)

Table: Faculty codes description

STUDENT STATUS	STUDENT STATUS DESCRIPTION	REGISTRATION CODE	REGISTRATION STATUS
5	graduated	5	studies completed
1	active	1	register
9	will graduate	9	not registered, will graduate
4	withdraw from studies	31	not registered, terminated from studies
3	failed and terminated	2	not registered
12	confirmation for register	4	not registered, dismissed from studies
18	dismissed from studies	45	withdraw from studies, personal/social issue
19	registration canceled	44	withdraw from studies, change university
2	deferment of studies	47	withdraw from studies, accepted job offer
8	suspension of studies	43	withdraw from studies, not interested in the field of study
11	terminated (maximum duration of study)	46	withdraw from studies, financial issue
14	deceased	90	decline university offer
		42	withdraw from studies, health issue
		11	terminated (maximum duration of study)
		41	withdraw from studies, financial issue(debt)
		14	deceased
		27	extend study, course not offered
		23	extend study, personal affair
		91	the degree of posthumous
		26	extend study, financial issue
		25	extend study, personal issue

Table: Student Status Description

PROGRAM CODE	PROGRAM NAME	ENTRY CASE CODE	ENTRY CASE DESCRIPTION
WP02	human resource development	AP	prime flow
WS24	cognitive science	AU	university foundation
WA59	arts management	KR	appeal case
WE10	marketing	R2	phase 2 appeal case
WA57	design technology	APB40	prime flow - bottom 40
WS47	resource biotechnology	IS	a/national students
WS48	resource chemistry	R1	phase 1 appeal case
WM00	doctor of medicine	B40	bottom 40
WK01	civil engineering	S2	second semester
WE07	finance	RM	manual appeal
WC10	software engineering	KH	special taking
WK18	mechanical and manufacturing engineering	TU	change university
WA22	development planning and management	AUB40	university foundation - bottom 40
WA32	fine arts	JPA	public service departments
WS51	animal resource science and management	ES	exchange student
WA15	international studies	ECNS	entry case not specified
WA12	industrial relations and labour studies	R2B40	phase 2 appeal case - bottom 40
WS50	plant resource science and management	M2	second mode
WA23	social work studies	R1B40	phase 1 appeal case - bottom 40
WA21	politics and government studies	CR	cross campus
WA14	anthropology and sociology		
WE09	international economics	ENTRANCE QUA CODE	entrance qualification description
WS49	aquatic resource science and management		
WA02	communication studies	A	art stream:stpm aliran sastera
WC11	network computing	N	science stream:matrikulasi/asasi sains
WC03	information systems	PAS	universiti pre-u
WE02	accountancy	S	stpm aliran sains
WM12	nursing	E	art stream:diploma/setaraf aliran sastera
WC09	multimedia computing	F	science stream:diploma/setaraf aliran sains
WA58	cinematography	P	accountancy:matrikulasi perakaunan
WP04	counselling	G	diploma ua dan diploma politeknik
WE13	corporate management	T	stam
WE03	business economics	DI	direct intake
WB03	linguistic	IS	international student
WK23	electrical and electronic engineering	F1	kelayakan setaraf
WC00	computational science	L	asasi tesl
WE01	service economics	ILP	intensive language programme
WK03	chemical engineering	E1	kelayakan dkm/dlkm/dvm/diploma ipts/luar negara/lain-lain kelayakan

PROGRAM CODE	PROGRAM NAME	ENTRY CASE CODE	ENTRY CASE DESCRIPTION
WK19	electronics engineering (telecommunication)	J	technical stream:matrikulasi teknikal
WK20	electronics engineering (computer)	U	asasi undang-undang
WA05	music	STPM	stpm
WE25	industrial economics	DIPL	diploma
WA19	psychology	K	asasi kejuruteraan
WA06	drama and theatre	E2	code specifieded (e2) not in lookup table
WB18	strategic communication	LL	code specifieded (ll) not in lookup table
WK18A	mechanical engineering	PKAS	code specifieded (pkas) not in lookup table
WH00	architecture	H	kelayakan melalui jpa
WH06	quantity surveying	ES	exchange student
WB02	english for global communication	MK	matrikulasi
WA92	animation	MATR	code specified (mk) not describe in lookup table
WP08	tourism	B40	code specifieded (b40) not in lookup table
WT37	computer science	M2	mode kedua
WT14	mathematics	CR	entry qualification not specified
WT06	teaching of english as a second language	F	code specifieded (cr) not in lookup table

Table: Program codes description

Data preprocessing involves in cleaning the dataset to remove errors, inconsistencies and missing values. Additionally, data transformation techniques may be applied to normalize distributions or scale variables for analysis.

The dataset is structured to facilitate machine learning analysis, with variables organized into features (predictor variables) and target variables (outcome variables). Features are selected based on their relevance to predicting student outcomes, with consideration given to factors known to influence academic success and progression.

3.2 ML Algorithms

The contextual applications of supervised learning are used for predicting significant factors including graduation status, employment status, and students' attrition using student demographic data, GPA, and outcome measures.

KNN Classifier: K Nearest Neighbors (KNN) is a basic Artificial Intelligence algorithm used for Supervised Learning for regression and classification. This is more useful in analyzing patterns in students' data and making an understanding of the similarities of students in relation to their attributes as presented in the figure above.

Decision Tree: Decision tree is a technique in which a tree like structure is drawn and each node is formed by a decision leading to a future node and the decision at each node results into further nodes and the leaf nodes hold the probabilities of the outcomes. This method is used to plan the course of action or perform a statistical probability and conducts complicated sets of data to look for patterns in learners data.

Random Forest: Under this category we have Random Forest, which is a large number of decision trees. It is used in ways such as bootstrapping and random feature subsets, combined with what is known as the majority voting method for coming up with a precise prediction. It makes the model more immune to fitting to noise data and avoid getting overfitting.

XGBoost Classifier: It offers parallel tree boosting, which helps solve data science problems swiftly and accurately, making it a powerful tool in machine learning.

It provides parallel tree boosting, which makes fast and accurate solutions to a data science problem and therefore act as a noble tool in machine learning.

Logistic Regression: It is an analytical approach that is used in making probabilities of a binary that has been established on previous observations of a dataset. It is mainly employed in binary classification where a dependable system of assessing possibility of binary results can be obtained.

3.3 Feature Selection Techniques

Feature selection methods substantially help in explanation, reduce model's dependence on unimportant features, and enhance model performance. Subsumed in the area of educational attainment predictive analytics.

One widely used feature selection technique is the filter method, which evaluates the relevance of each feature independently from the predictive model. Filter methods usually employ statistical tests or correlation analysis to rank features according to their individual significance or contribution to the target variable.

On the other hand, wrapper techniques assess the features' groups based on their effectiveness in a specified predictive model. These methods use cycles of selection of the features, estimation of the model, and evaluation of its accuracy. The best feature set is the one which gives the best performance according to chosen evaluation criteria and is used as the final one.

Embedded approaches include feature selection directly into the process of training the model. These techniques automatically select or penalize features during training based on their predictive value.

3.4 Model Evaluation

Evaluation of such models is very important in predictive education modeling. It uses other approaches and parameters to evaluate how effective the developed machine learning models are in predicting students' performance. The evaluation ensures that the researchers and practitioners to appropriately identify, enhance, and apply the models, which enhances the choice of knowledge regarding the education authorities and policy making.

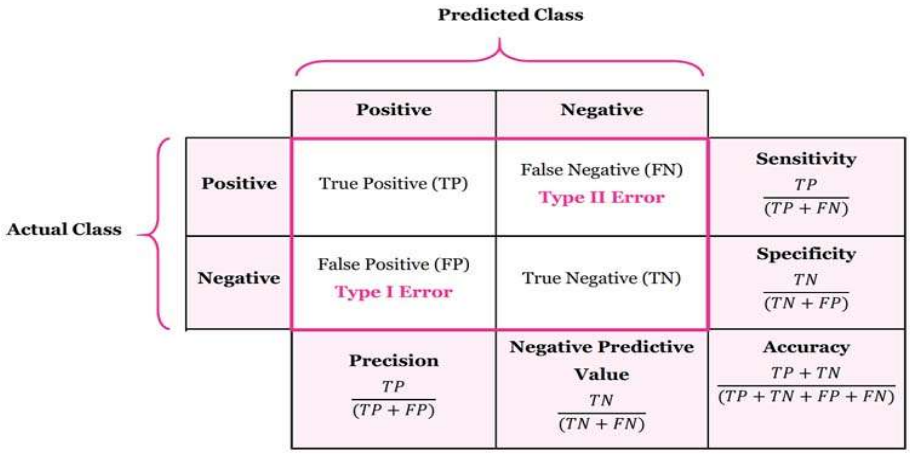


Figure: Overview of Confusion Matrix

To be specific, there is a multitude of different measures that one can use for the assessment of models suitable for totally different types of employment. Accuracy defines the proportion of correct expected positives, but its fitness declines in unbalanced and diverse cost error situations.

Accuracy and recall assessment measures the prediction precision and sensitivity in the binary classification problems such as student attrition prediction. The F1-score combines both metrics, which are useful when it comes to evaluating the prevalence of false positions and false negations in the given datasets, which are often imbalanced.

ROC curves and their performance metric AUC assess the binary models' performance for all thresholds and a higher AUC value implies better model performance. Mean Absolute Error (MAE) and Mean Squared Error (MSE) are measures of the averaged differences in regression settings. In other words, it reveals a level of how closely the model captures the variations of the target variable.

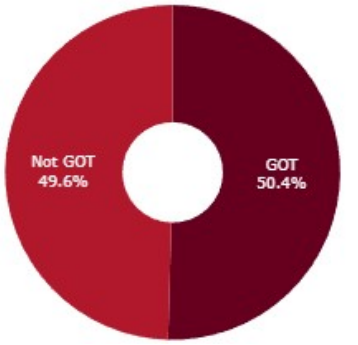
Some of the cross-validation methods include k-fold where the results of training are repeatedly tested on the expanded sets than the training sets. This makes evaluations of the models' performance robust and the actual performance may be generalized to other zones.

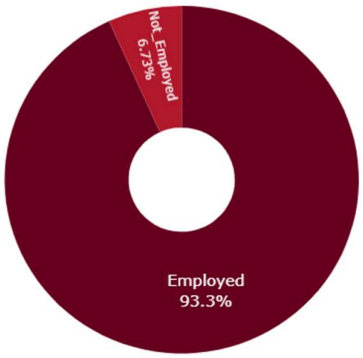
Chapter 4: Implementation

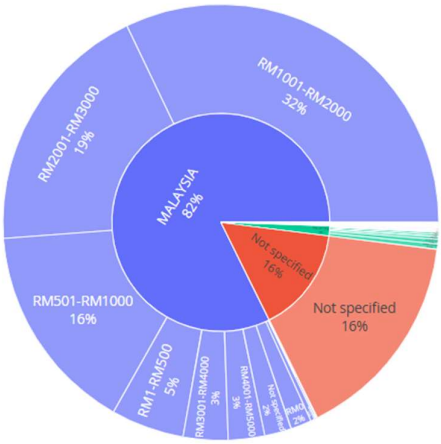
4.1 Exploratory Data Analysis (EDA)

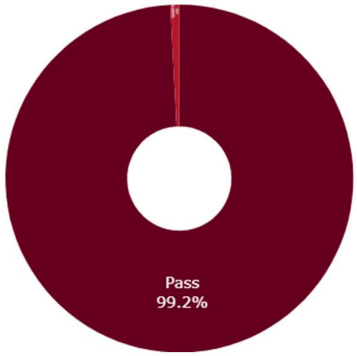
Exploratory Data Analysis (EDA) involves in data visualization, summary statistics data, and data cleaning to understand the data. EDA aims to understand the dataset's properties, uncover potential trends or patterns, and inform subsequent steps in the data analysis process, such as modelling or hypothesis testing.

4.1.1 Analyzation of Target Column

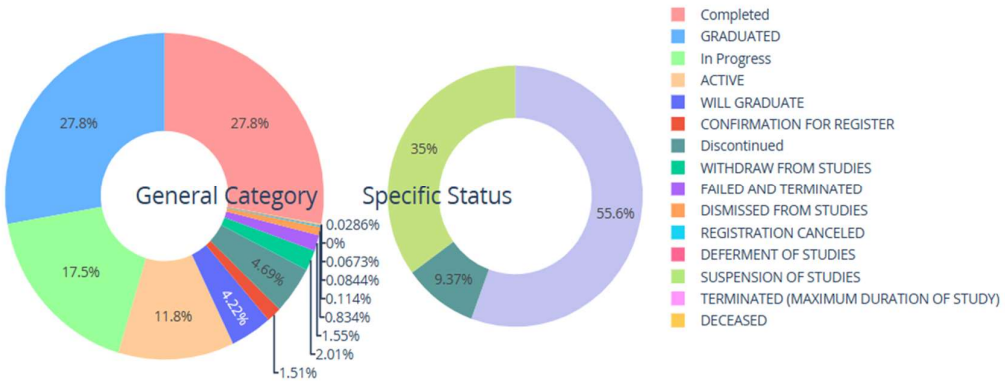








Student Status Categories and Specific Statuses



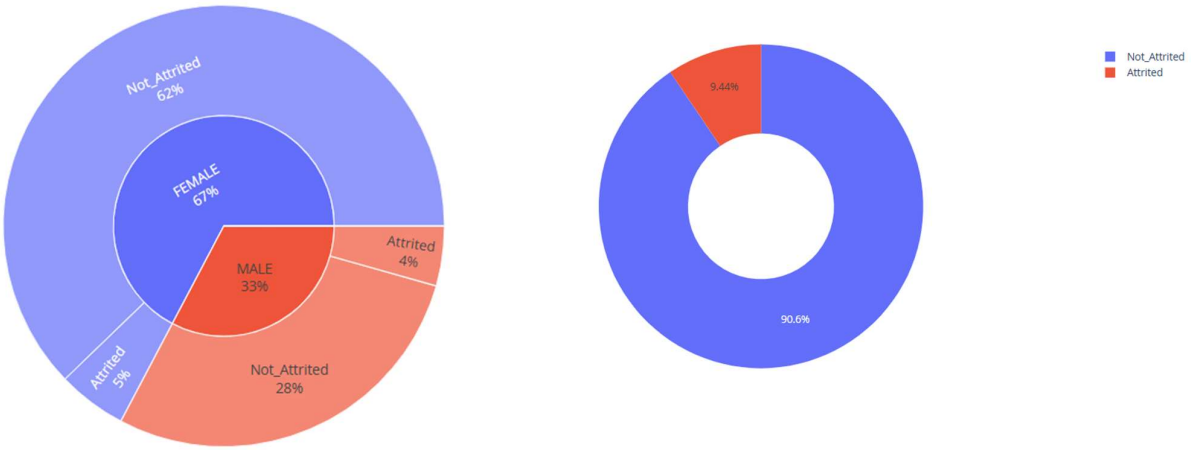
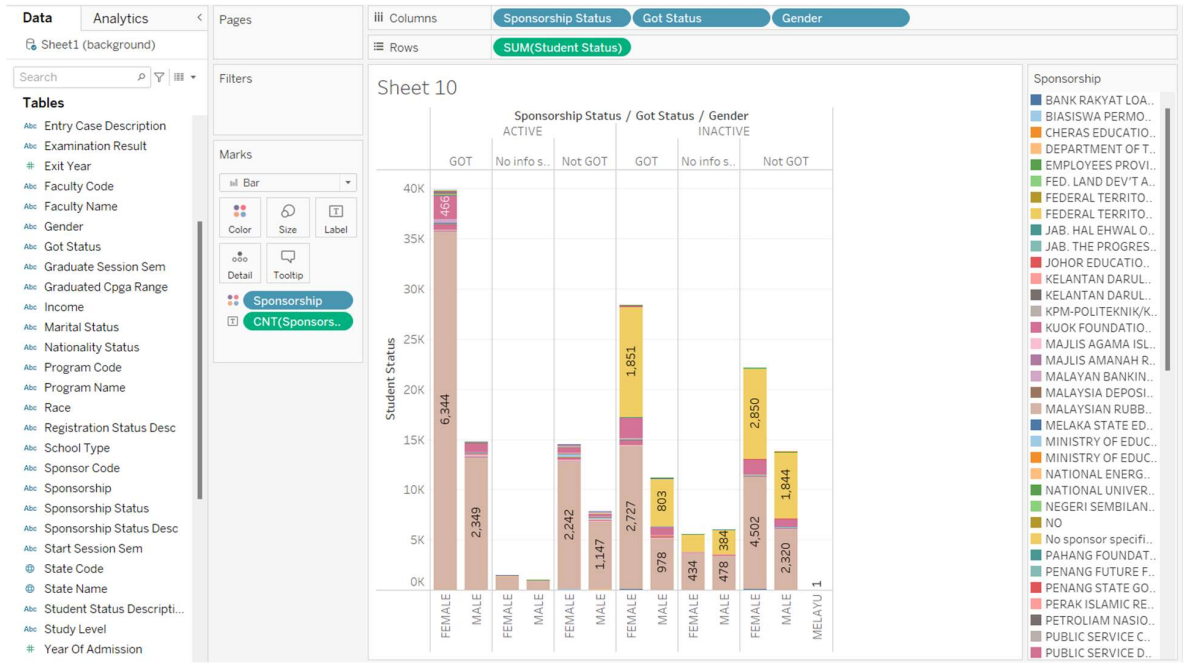


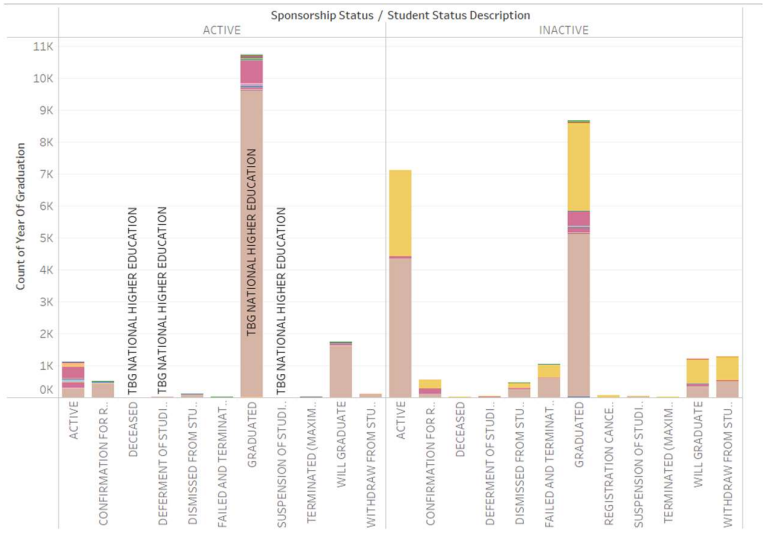
Figure: Pie chart for on time graduation, employment and attrition status

The provided pie charts offer insights into graduation, employment, and student attrition outcomes. The first chart displays an almost equal split between students who graduated (50.4%) and those who did not (49.6%), reflecting a balanced distribution among graduates and non-graduates. The second chart illustrates that a significant majority (93.3%) of individuals secured employment timely post-graduation, while only a small fraction (6.7%) were not employed as expected, indicating effective job placement efforts by the program or institution. The third chart demonstrates an excellent retention rate, with 99.2% of students passing and continuing their education, suggesting that only a minimal 0.8% faced attrition. From the student status description, we can see that overall, Not Attrited students are 91% and Attrited students are 9%.

4.2 EDA using Tableau

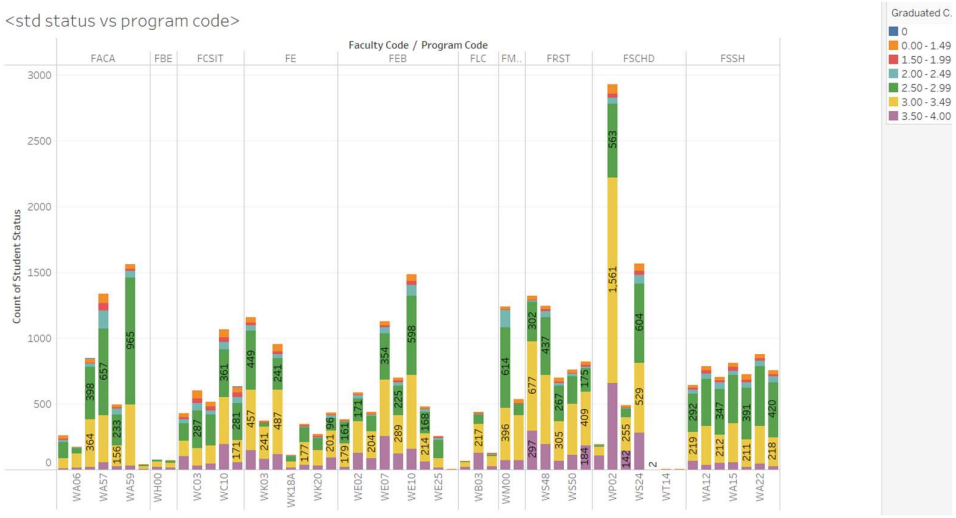
Background Data:





The Graduated students with active sponsorship are more than inactive sponsorship

Figure: Student status Vs Sponsorship status



Students From the faculty code FSCHD and Program code WP02 have higher CPGA range than others

Figure: student status Vs program code

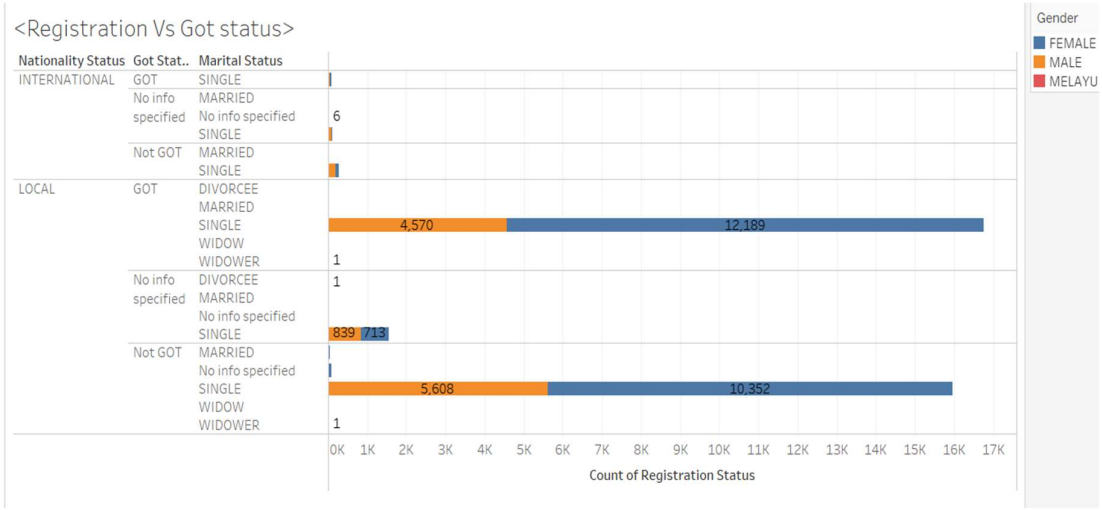


Figure: Registration Vs Got status

Meyalu is considered as Male, and Female students are graduated more compared to Men.

Ontime graduation:

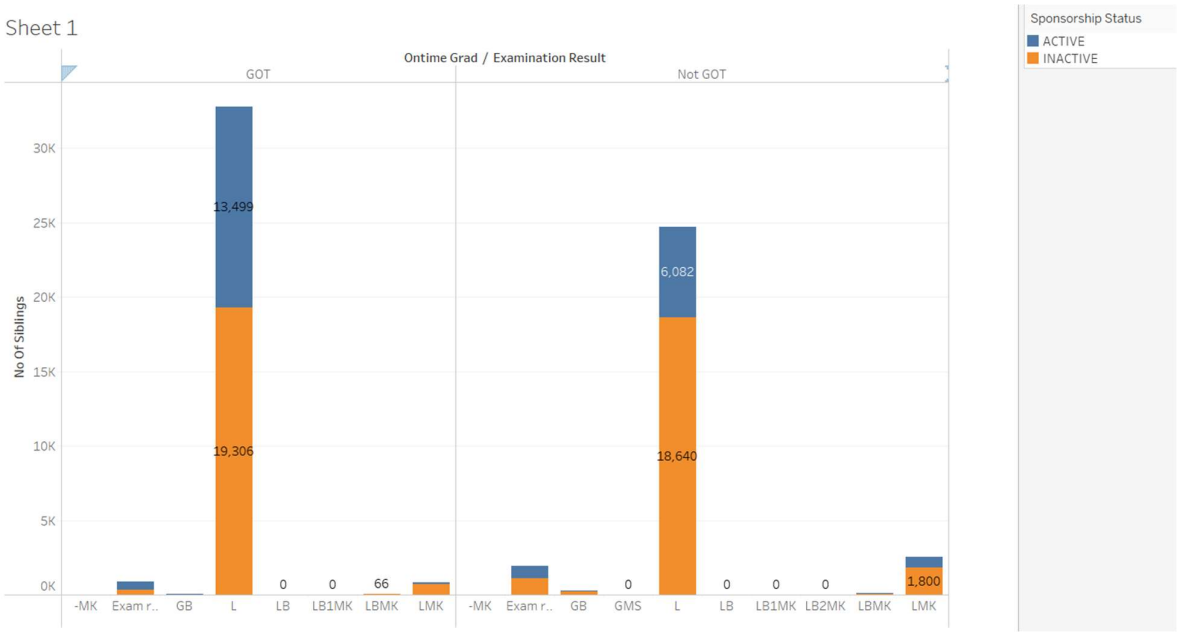


Figure: GOT Status Vs Examination Result:

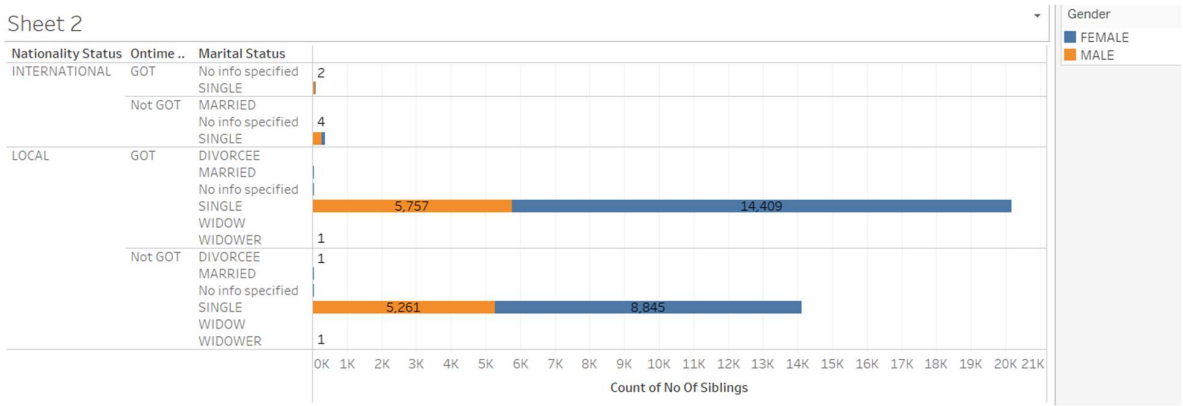


Figure: GOT Status Vs Marital Status:

Employment data:

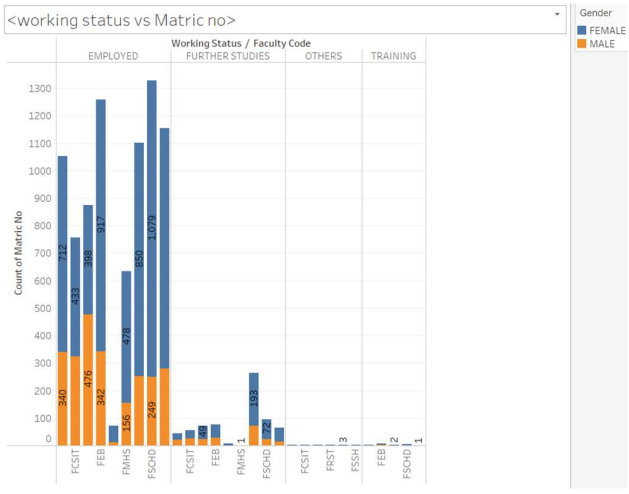


Figure: working status vs Matric no

Students whose examination result is L , are having a greater number of graduations on time status.

Sponsorship status as inactive is more in Not Got compared to students having GOT status.

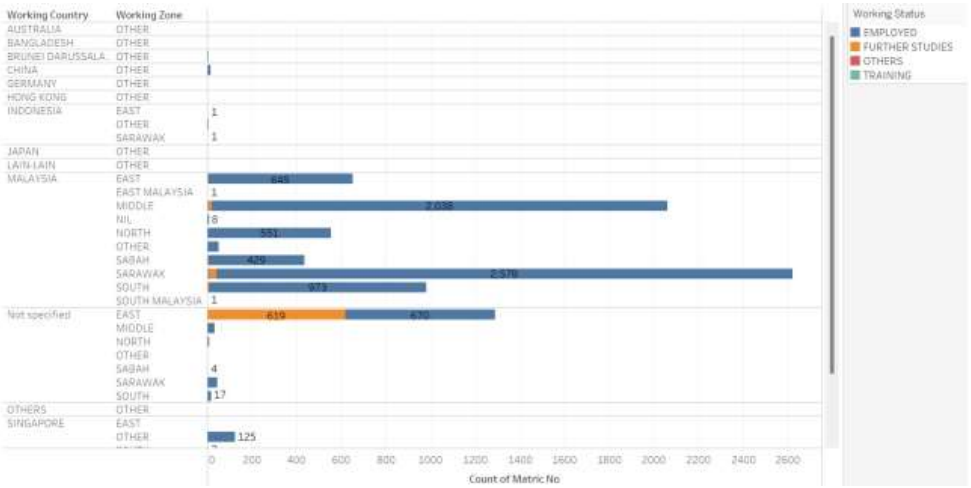


Figure: Working Status Vs Working zone & Country:

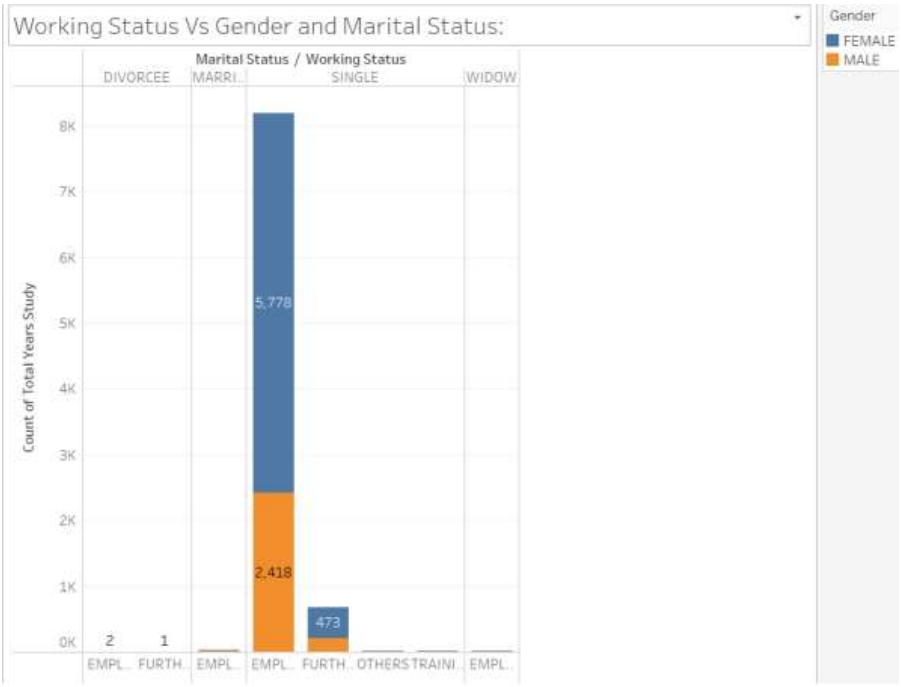


Figure: Working Status Vs Gender and Marital Status:

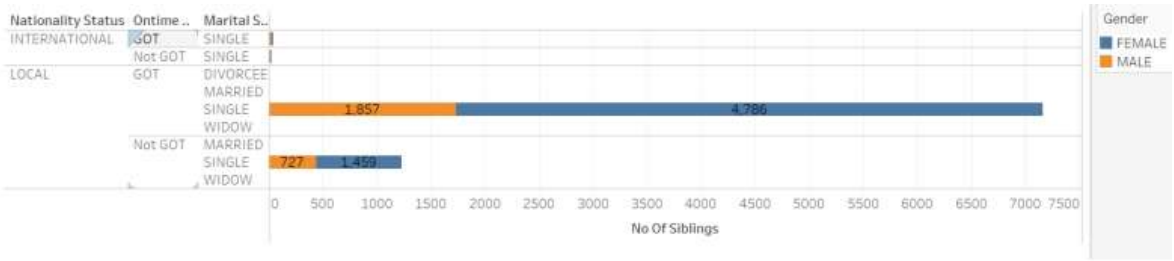


Figure: Siblings Vs Marital status and Ontime Graduation on employment data

Students from the FSCHD faculty code have the highest employment rate, those from the FRST faculty code are more likely to pursue further studies, and females generally have a higher employment rate than males across most faculty codes.

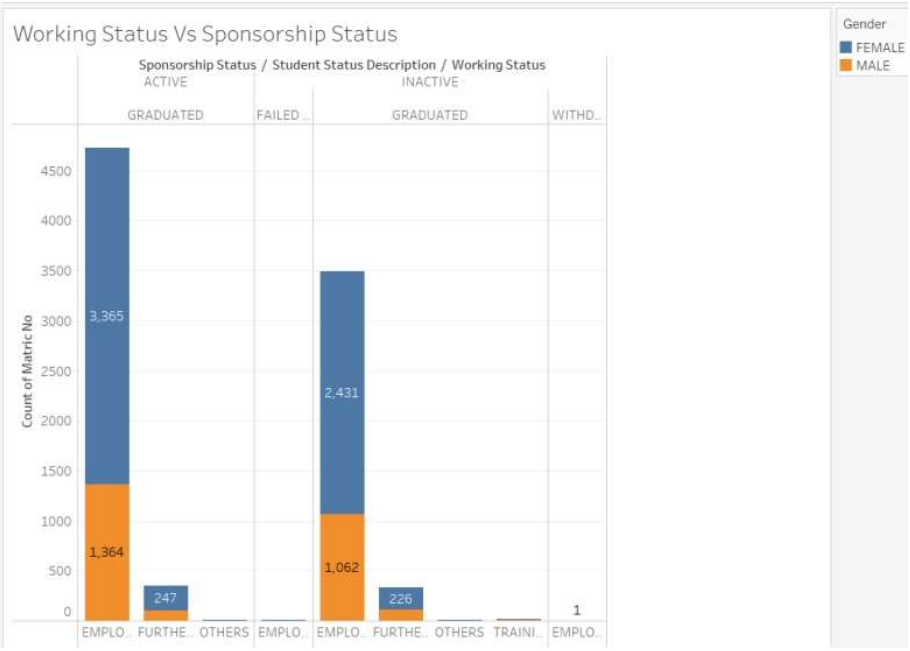


Figure: Working Status Vs Sponsorship Status

Most participants who got employed or pursued further studies are single. 274 single male participants are preferred for further studies. 473 single female participants are preferred for further studies. 54% of single females got employed. 27% of single males got employed.

Attrition Data:

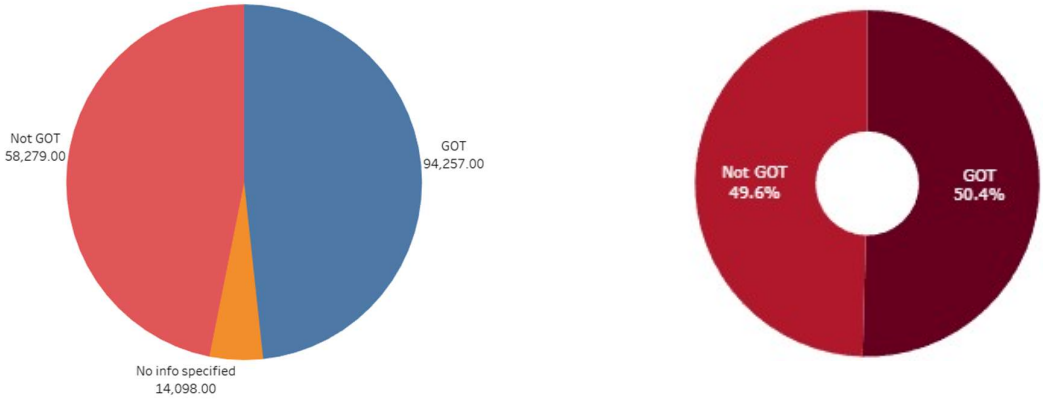


Figure: Got Vs Not Got Status

From fig we can see that 51% of Students Graduated on time, 49% of students did not graduate on time.

4.2.1 Feature Importance variables:

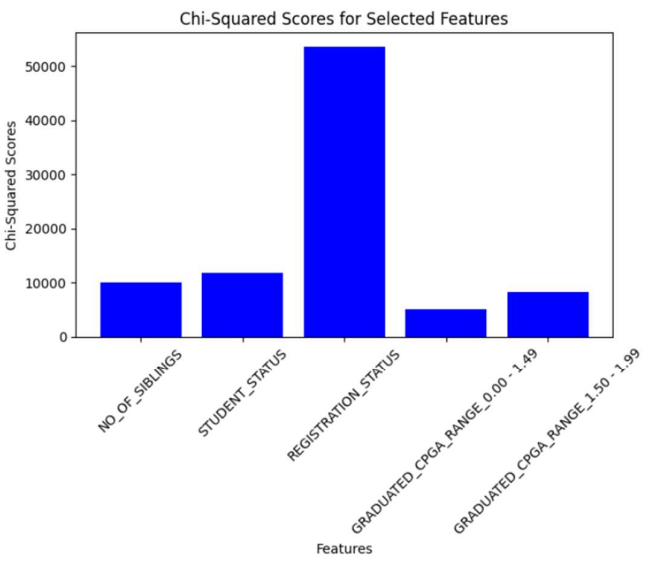


Figure: K-best Features for Attrition Data

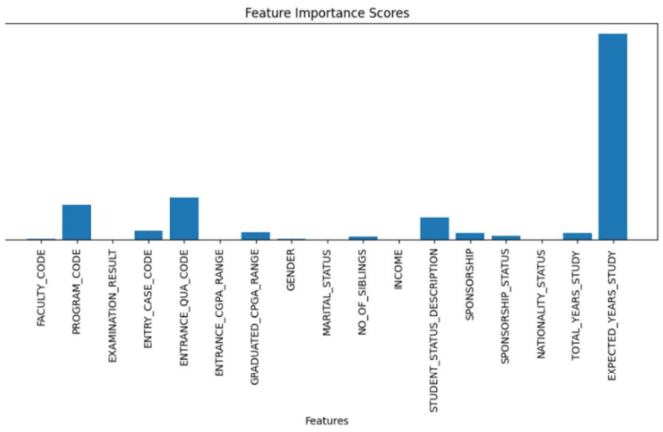


Figure: K-best Features for On Time Graduation Data

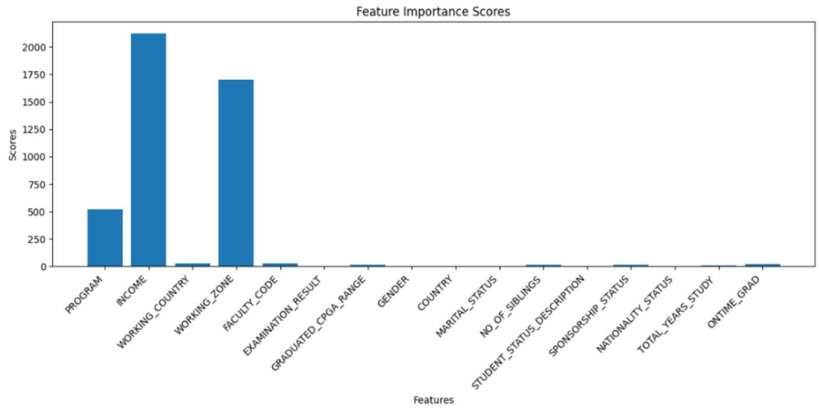


Figure: K-best Features for Employment Data

Feature selection was conducted using Select K-Best with chi-squared scoring, highlighting the most significant predictors of all the analysis

4.3 Background Data Analysis

4.3.1 Missing Value Percentage

Most of the real datasets are comprised of complete data and the present data set also mostly contains complete data and will use the columns such as ‘ FACULTY_CODE ’, ‘ PROGRAM_NAME ’ etc. are all very efficient in terms of competency. In the Table in the ‘STUDENT_STATUS_DESCRIPTION’ column whereby 195 entries appear to have missing Student Status information, thereby pointing at a major blind-spot in monitoring the statuses of students. Other columns like ‘CITY’ and ‘STATE_NAME’ have data missing fewer than 2% of records. Another advantage of the nearly comprehensive dataset is the decreased amount of issues that relate to data imputation However, more research effort should be devoted to identifying and addressing existing voids in the ‘STUDENT_STATUS_DESCRIPTION’ to prevent flaws in analysis.

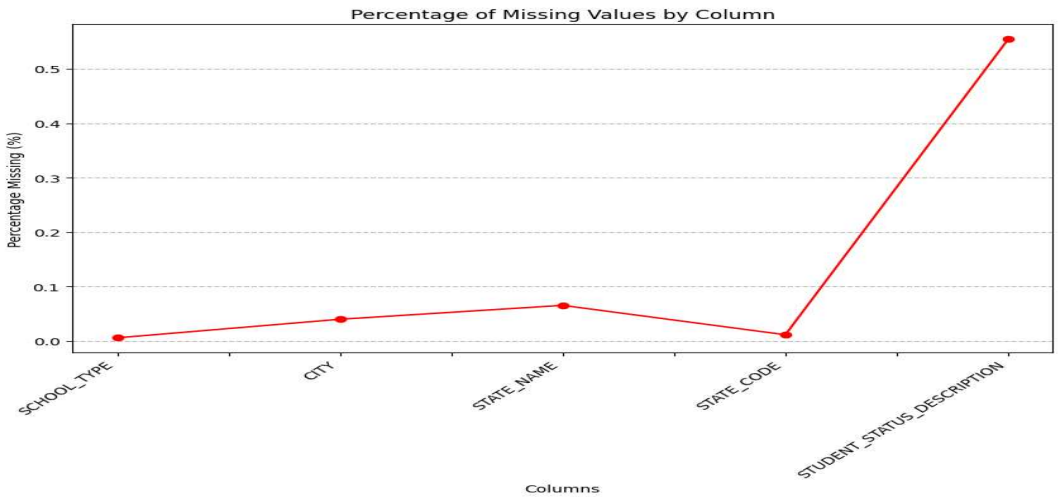
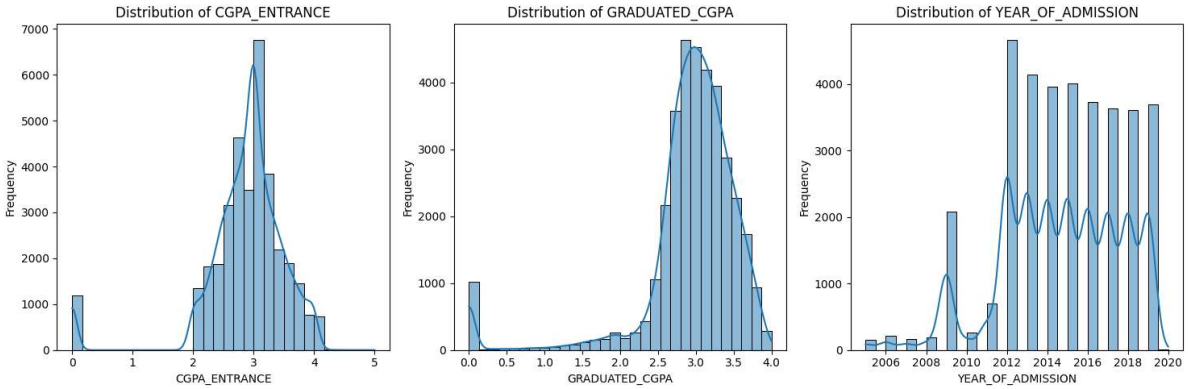


Figure: Missing Value percentage in background data

4.3.2 Numerical Feature Summary of Background Data



	Unnamed: 0	MATRIC_NO	CGPA_ENTRANCE	GRADUATED_CGPA	YEAR_OF_ADMISSION	EXPECTED_YEAR_OF_GRADUATION	YEAR_OF_GRADUATION	EXIT
count	35174.000000	35174.000000	35174.000000	35174.000000	35174.000000	35174.000000	35174.000000	35174.
mean	2278.211975	50903.458321	2.869297	2.938472	2014.652755	2017.971684	1318.859214	1318.
std	1468.433474	13880.835821	0.702499	0.674731	2.978952	3.064751	959.409028	959.
min	0.000000	13612.000000	0.000000	0.000000	2005.000000	2008.000000	0.000000	0.
25%	1001.000000	40752.250000	2.590000	2.760000	2013.000000	2016.000000	0.000000	0.
50%	2146.000000	51482.500000	3.000000	3.030000	2015.000000	2018.000000	2015.000000	2015.
75%	3427.000000	62662.750000	3.250000	3.310000	2017.000000	2020.000000	2018.000000	2018.
max	5545.000000	73261.000000	5.000000	4.000000	2020.000000	2024.000000	2020.000000	2020.

Figure: Student status description in background data

The features are the identity and specific details of 35,174 students such as academic performance and individual information. Other basic elements include entrance and graduation CGPA with average values of 2. 87 and 2. 94, respectively, which shows there is slight improvement from admission level to graduation level with the course credit. This is a view of students admitted to the university from the year 2005 till 2020 and expected graduation from 2021 to 2024. Another interesting observation is the fact that students act as brothers and sisters pointing to the fact that majority of them are only children or the firstborn as shown by mean scores. With the distribution being different for both Student and Registration status, it underpins the proposition that there is a constellation of academic status and enrollment status among the client's students.

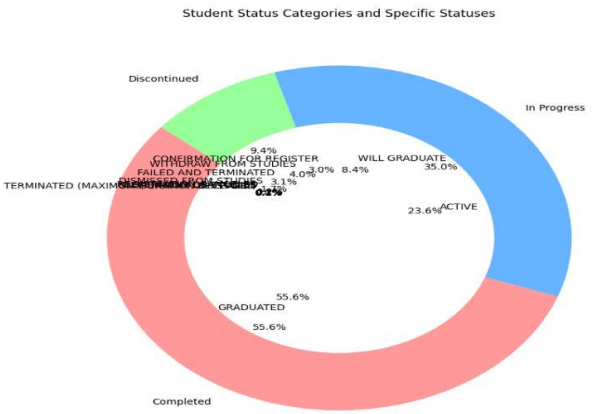


Figure: Numeric data summary of Background data

4.3.3 Student Status Description

The pie/doughnut chart that is nested is thus useful in showing the distribution of students statuses in the dataset. The inner ring categorizes the overall student statuses:

Completed: This major segment involves users who have completed the course of study of their respective courses.
In Progress: Another large segment encompasses learners currently learning or planning to continue learning or are close to completing their studies.
Discontinued: Smaller segments in the outer ring provide why the individuals left the programs, it can be withdrawals, failing grades, terminations, or dismissals.

Each segment's size helps in ascertaining the portion of students within a specific status, thus providing an approximate understanding of the composition of the student status nature.

4.3.4 Density Plot for Background Data

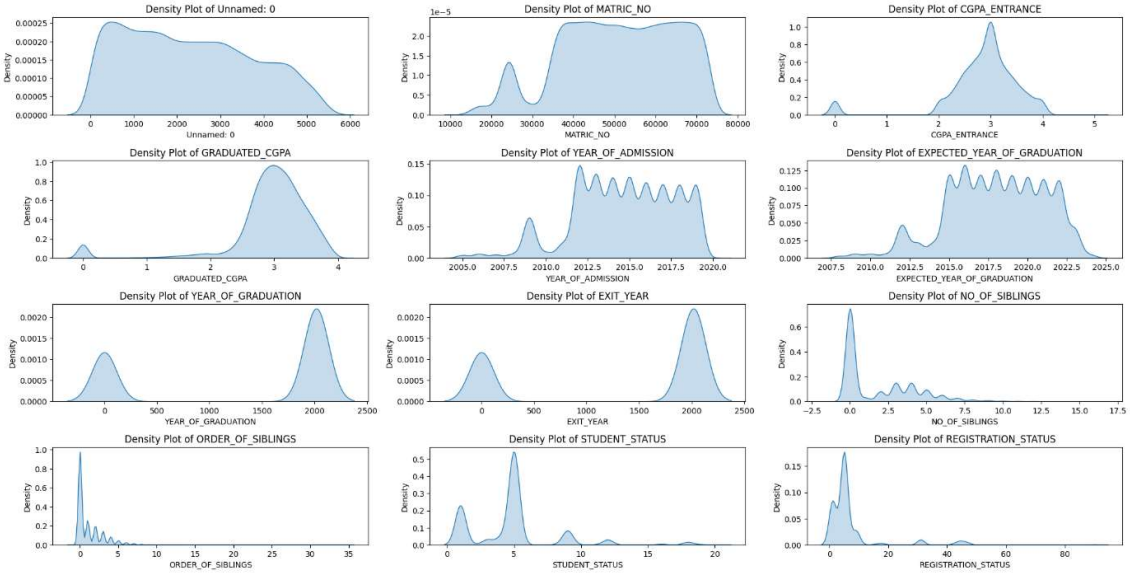


Figure: Density plot for background data

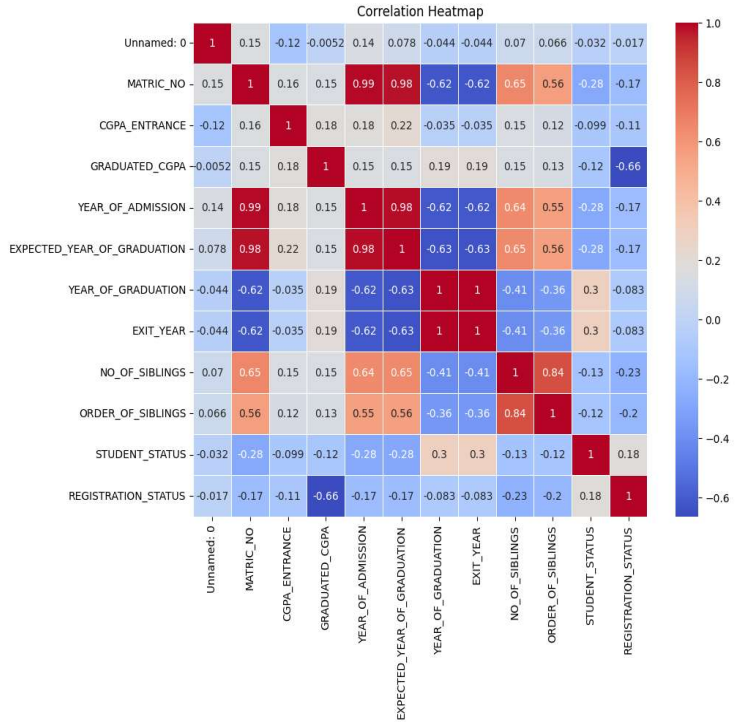
The density plots obtained for each numerical column in the dataset provide some valuable overview of the values' distribution within the respective attribute. It is important to note that the data appears balanced in regards to the MATRIC_NO column, and it maintains a constant average all throughout the given range. On the other hand both CGPA_ENTRANCE and GRADUATED_CGPA are in the nature of clustering with most of the values in the range of 2 to 4.

Similarly, YEAR_OF_ADMISSION and expected as well as actual YEAR_OF_GRADUATION have spikes that might be suggesting specific years that may have the highest admissions or graduations. On the other hand, the graph of EXIT_YEAR shows a significant spike at zero, indicating that, often, there are dummy or even invalid records in the data set. Further, analysis of the variables NO_OF_SIBLINGS and ORDER_OF_SIBLINGS; show that most of the students record low levels of sibbing and prefer earlier ranks within the siblings sequence.

In general, these density plots represent the layout of the data in a smooth manner and provide key information concerning the central tendencies, variability and the possible of modal plurality in the data.

4.3.5 Correlation Plot for Background Data

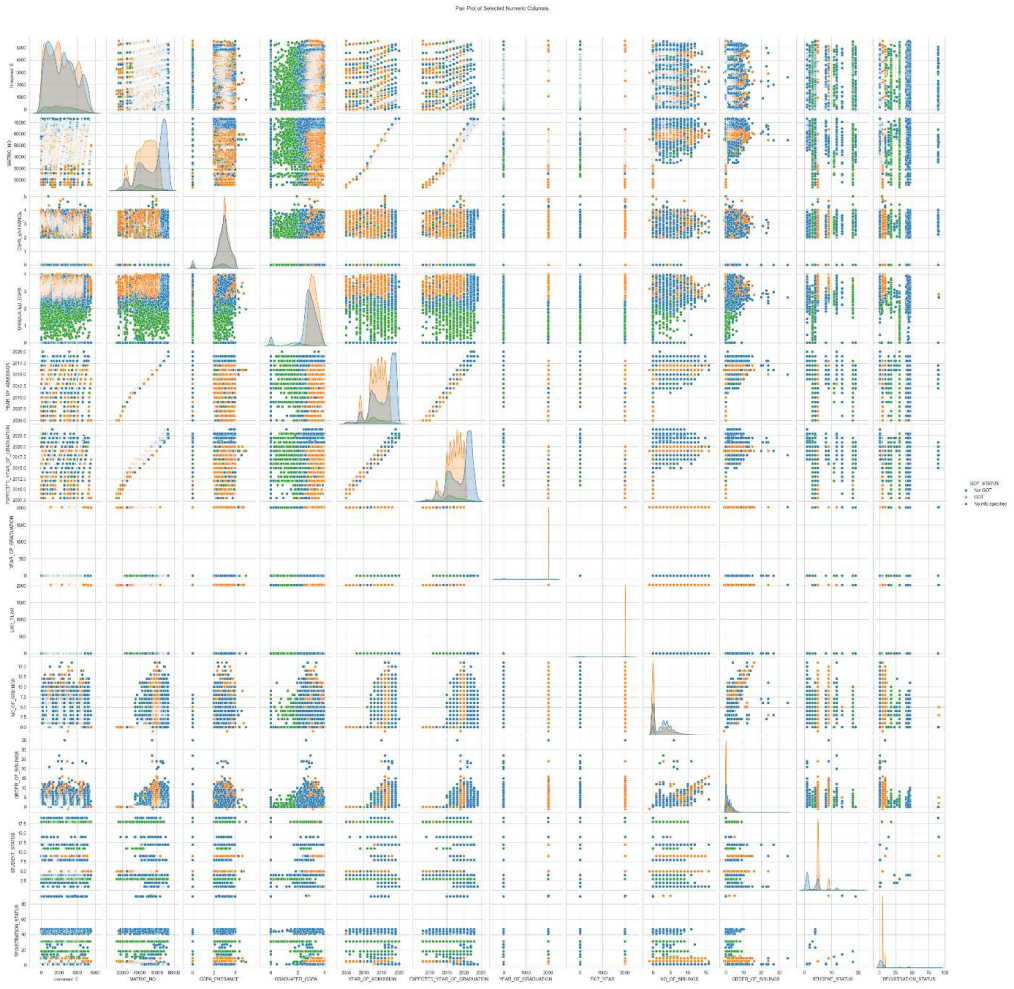
From the correlation heatmap, one is in a position to relate various attributes in the numerical columns of the given dataset. The darker aspects of the heatmap represent the higher values of the correlation coefficients, whether positive or negative; the diagonal line represents the correlations equal to 1 as, of course, every variable correlates with itself entirely. For the second set of hypothesis, it is remarkable that there exists a moderate positive correlation between 'CGPA_ENTRANCE' and 'GRADUATED_CGPA' meaning the higher the students CGPA when entering the institution, the higher the CGPA by the time the student is graduating.



However, aside from this connection, most correlations appear weak, suggesting limited linear relationships between numerical variables. Overall, this visualization serves as a tool to discern patterns or dependencies among the dataset's attributes, aiding in understanding the interplay between different numerical features.

Figure: Correlation plot of background data

4.3.6 Pair plot for Background Data



The pair plot above explores pairwise relationships among selected numeric columns:

Figure: Pair plot for background data

CGPA_ENTRANCE, GRADUATED_CGPA, YEAR_OF_ADMISSION, NO_OF_SIBLINGS and YEAR_OF_GRADUATION. Each plot in the grid shows the relationship between two variables, with histograms along the diagonal providing the distribution of individual variables.

4.3.7 Faculties Distribution

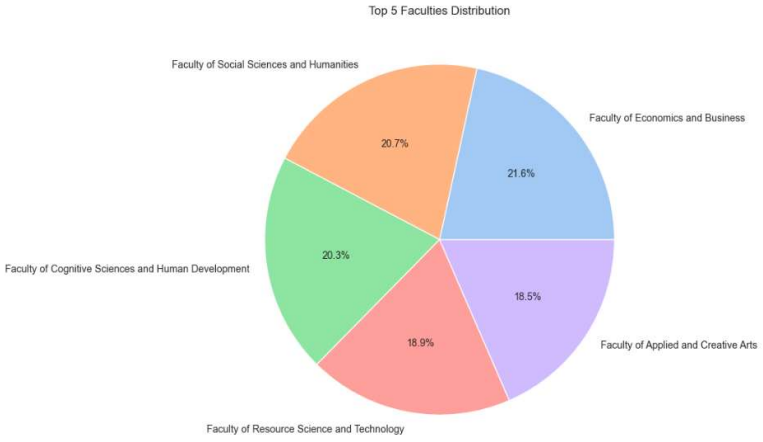


Figure: Faculty distribution for background data

Here is a pie chart showing the distribution of the top 5 faculties based on the uploaded data. The chart displays the proportion of each faculty relative to the others within the top five, with percentages indicated for clarity.

4.3.8 Top 5 Program and Statuses Distribution

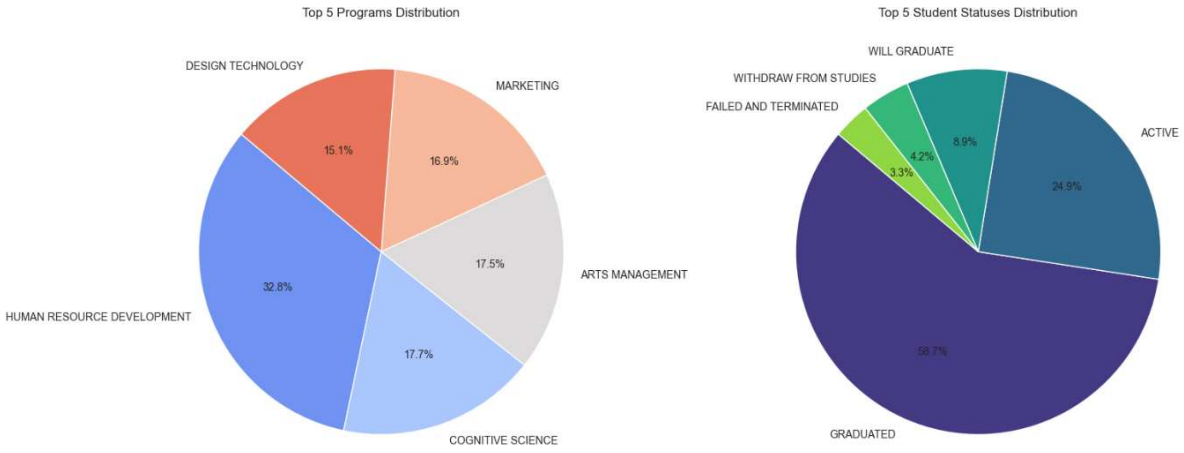


Figure: Top 5 program and statuses distribution

Top 5 Programs Distribution: This pie chart shows the distribution of the top 5 most enrolled programs within the dataset. Each slice represents a program and its relative size out of the top 5. This helps to identify which academic programs are the most popular or have the highest number of students.

Top 5 Student Statuses Distribution: This pie chart represents the top 5 student statuses based on their description. It provides an overview of the various statuses students have, such as graduated, active, or withdrawn. This visualization helps to understand the academic outcomes or current standing of the students within the dataset.

These visualizations offer insights into the dataset, highlighting the most common programs and student statuses, which can inform decisions about academic offerings and support services.

4.4 Employment Data Analysis

4.4.1 Pair Plot

Pair plot is a useful tool in visualizing the relationship and the distribution of all the variables in a given dataset at the same time. In the given pair plot, every square plot contains covariance of two variables, and diagonal contains either histogram or density plot of variable. Off-diagonal plots are plots which give possible relationship or structure of variables in combination with one another and are scatter plots. One characteristic of this kind of visualization is that the 'WORKING_LEVEL' variable is used as a hue; Therefore, data points are colored according to their work level, where 0 is low and 6 is high, which increases the depth of the analysis since one can see how different variables affect this one. For instance, the plots like 'WORKING_STATUS' having x-axis values 'WORKING_SECTOR' to decentralize the points with the working status depicted in the varied sectors along with the closer level mark on the sectors based on work level using color contrast. Likewise, the 'MATRIC_NO' and

'WORKING_YEAR_FROM' plot may suggest trends of the Matriculation Numbers against the Working Year from fields, which is similar to the previous one but this time separated based on work level.

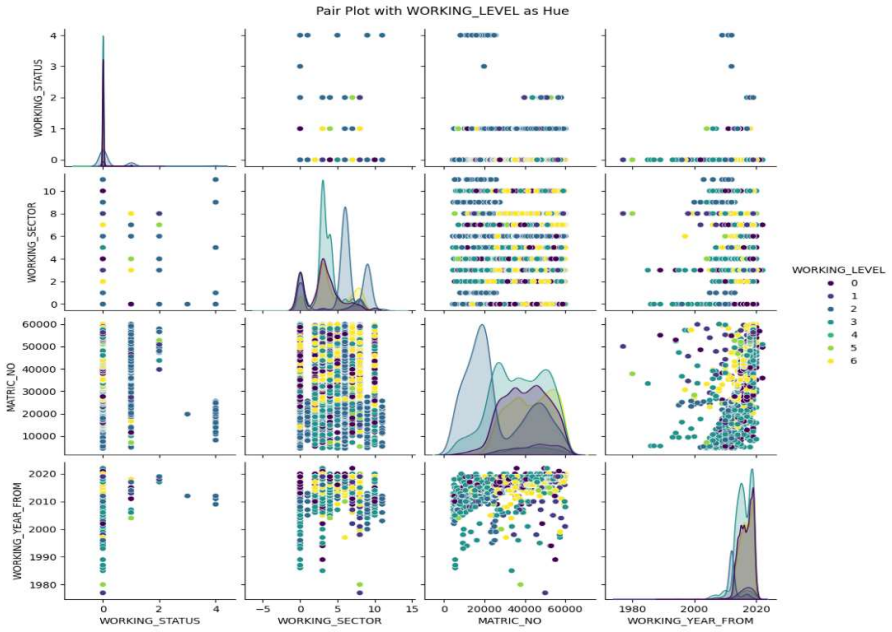


Figure: Pair plot for employment data

This type of visualization is particularly powerful for identifying trends, correlations, and distributions across multiple dimensions of the data.

4.4.2 Correlation Plot for Employment Data

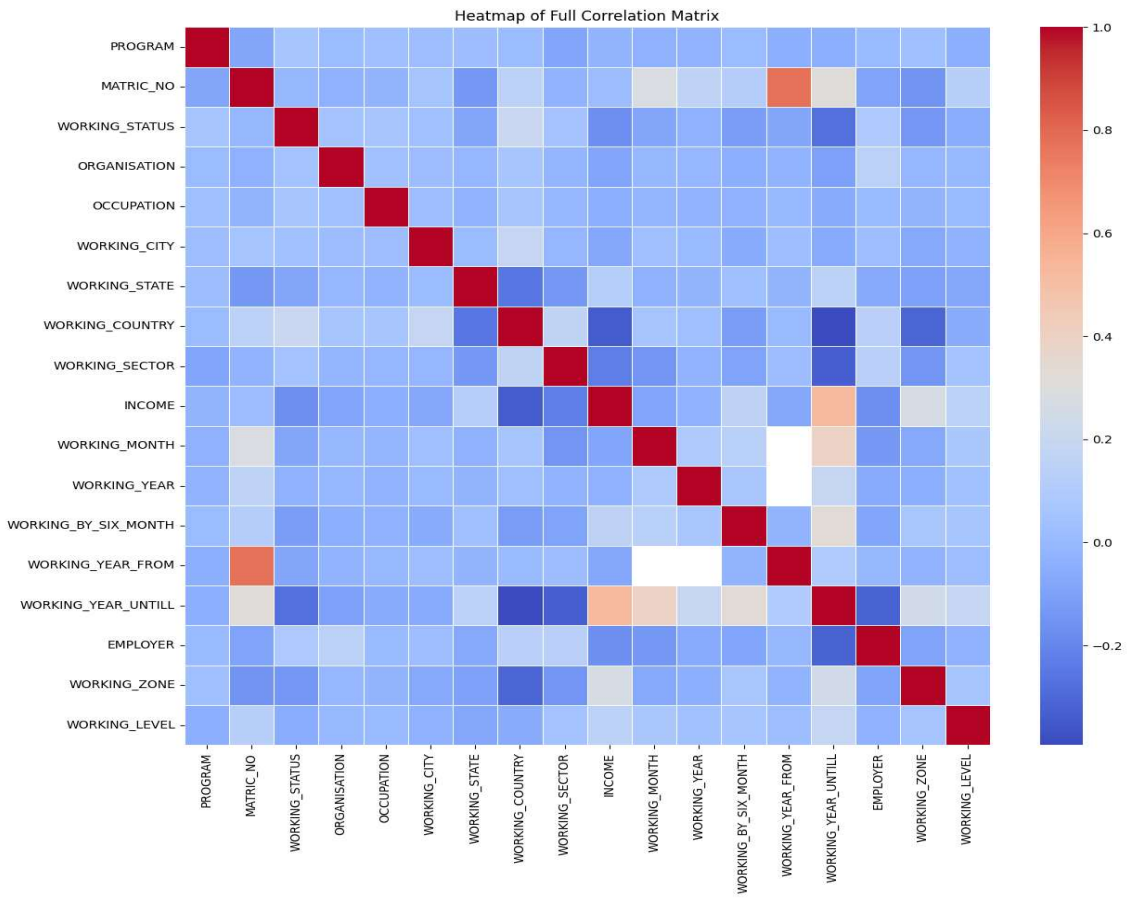


Figure: Correlation plot for Employment data

The heatmap above is the correlation matrix of all the columns in the dataset it comprises numerical variables and the categorical variables that have been coded numerically. Because of the massive data set and large number of estimates, the annotations for the correlation coefficients have deliberately been left out to enhance legibility.

This visualization gives directions for all the features regarding their relations, though, while interpreting the results, it is crucial to recall the limitations for features encoded as numeric codes for the categorical variables. The correlations, therefore, for these features may not reflect the true nature of competition between this pair of features because the encoding process they are subjected to brings an order that may not exist in the real sense.

4.4.3 Working Hour and Working Sector Analysis

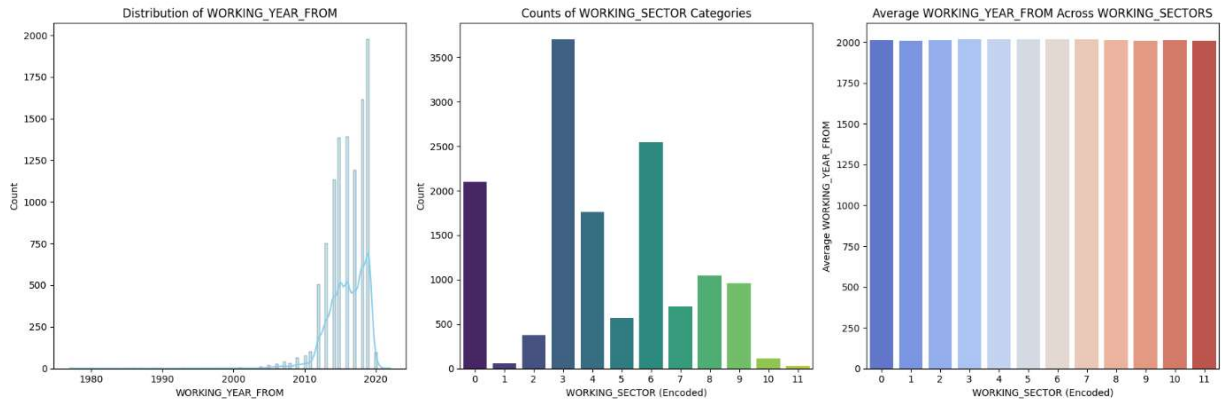


Figure: Working hours and sector analysis of employee data

By far, the working characteristics of individuals have been described by generating three types of plots in order to gain a different kind of perspective for the dataset. There is the first one, the histogram of the 'WORKING_YEAR_FROM,' and it captures the years that people have been working using both a density estimate to show the shape of this histogram. This plot helps specify the starting years which are mostly repeated and trends in consecutive years. The second plot is a countplot for the variable "WORKING_SECTOR" that counts the number of a particular feature as listed numerically. Mainly, this type of visualization is quite helpful in pointing out the level of employment concentration across various fields. Finally, a barplot presents the average value of "WORKING_YEAR_FROM" within each "WORKING_SECTR"; every bar represents the average age at which people from the corresponding sector started working, which helps to compare the level of sector-related work experience. These visualizations combined provide the user with employment and sectoral distribution of the data as a whole.

4.4.4 Distribution of WORKING_YEAR_FROM Across WORKING_SECTORS

The following violin plot highlights the distribution of "WORKING_YEAR_FROM" based on different "WORKING_SECTORS". Each "violin" refers to the sector where people in that area began working as indicated by the distribution of the numbers. The length of every violin represents density of points in different years; and wider parts suggest that more people began working in the particular year.
This plot gives explicit information about the distributions and, in particular, the presence of skewness or multicliquedness in some sectors. It is a helpful graph when one wishes to compare dispersion of a numerical variable with respect to distinct classes.

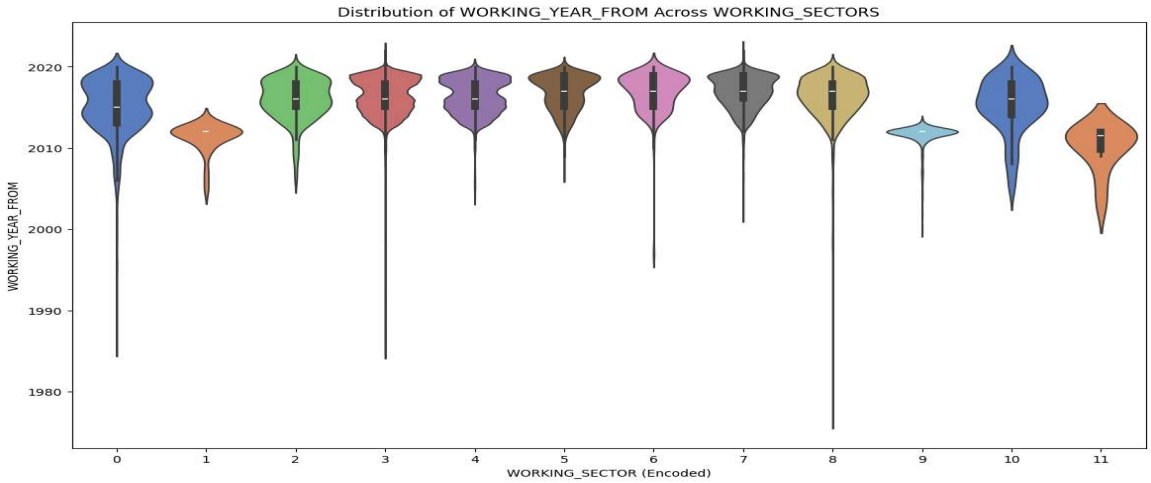


Figure: Distribution of WORKING_YEAR_FROM Across WORKING_SECTORS

4.5 Failed Data Analysis

4.5.1 Analysis of Credit Hour, Grade Status and Count of Courses

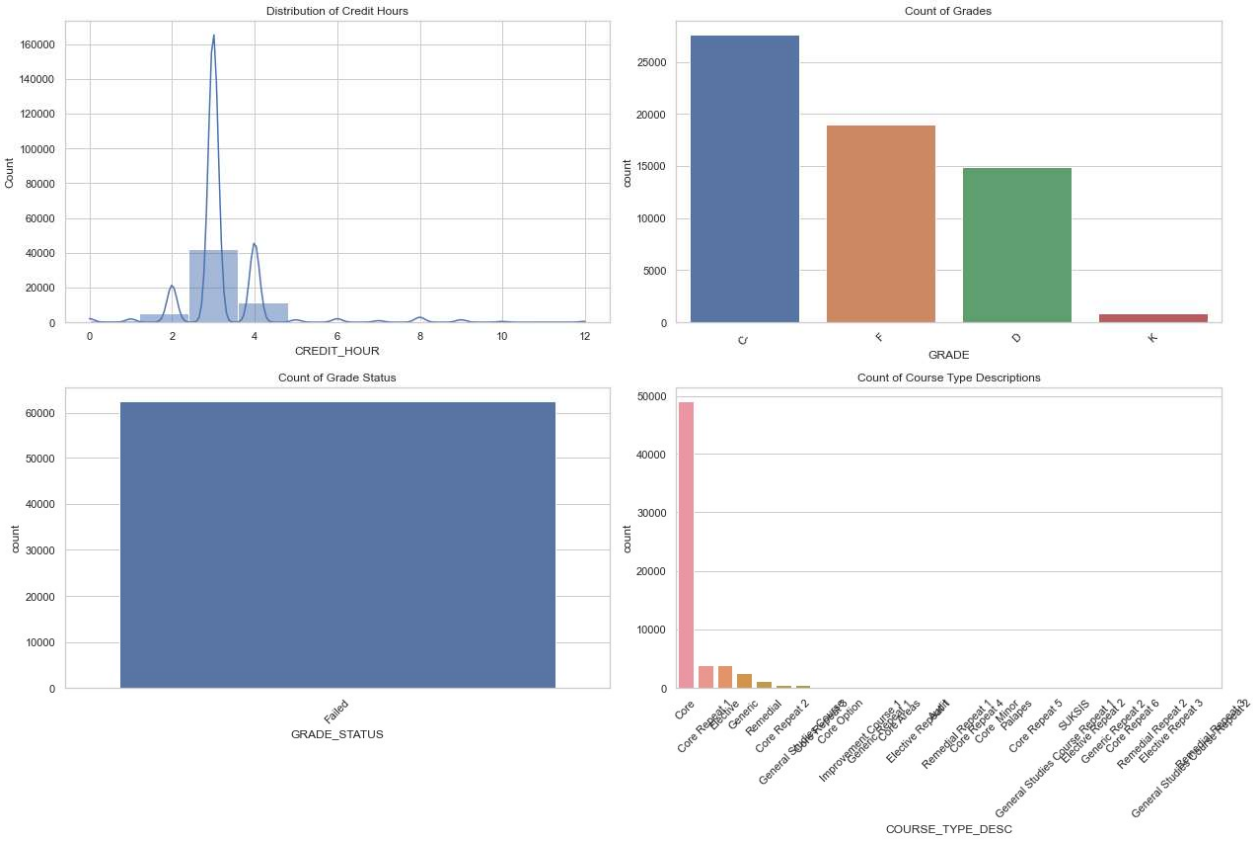


Figure: Analysis of Credit Hour, Grade Status and Count of Courses

With reference to the histogram that shows the distribution of the credit hours for courses, there is a noble sign of clustering at certain values; this must have been as a result of standardized course credit hours.

The count plot, showing the frequency of grades received, gives understanding of frequency of the

received grades if any of the fail course and the proportion of the specific grade in this connection.
Since the view is made on a particular dataset that targets failed courses, it is expected to find a majority of the bars displaying “Fail” statuses in the grade distribution though the count plot might reveal differences in the way this status is addressed or even classified.
Lastly, the plot of the actual count of all unique descriptions related to course types serves as a brief introduction to the distribution of the number of course types in the data and can hint at which type of courses at the core or elective level, for instance, fail more frequently. Collectively, these plots allow enhancing the understanding of the dataset, the nature of values it contains, and potentially point at some of the interesting values in the dataset or features of interest.

4.5.2 Word Cloud for Justification of Failure

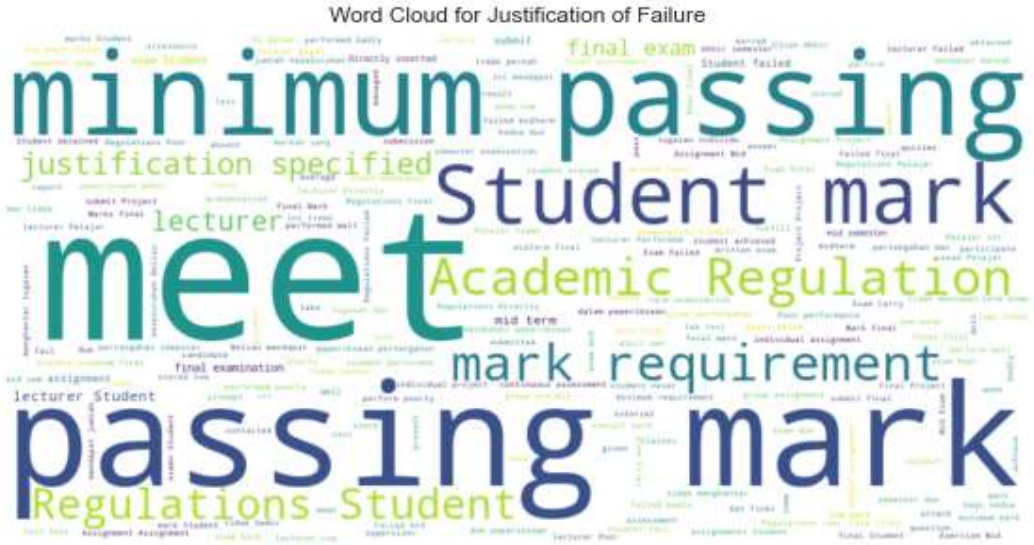


Figure: Word Cloud for Justification of Failure

The Word Cloud for Justification of Failure visualizes the most common words found in the justifications provided for failing grades. The size of each word in the cloud indicates its frequency in the justifications, with larger words being mentioned more often. This visualization helps to quickly identify key themes or reasons cited in the justifications for failures.

4.5.3 Line Plot of Failures:

Line Plot of Failures Over Time Based on the SESSION_SEMESTER Column and A Bar Plot of Average Credit Hours By Grade: Failures Over Time (Line Plot): The following one represents the Line Plot of the number of failures regarding different sessions/semesters in a sort function format. From this, the line plot assists to unveil any trends or a change of the number of failures at a given period or an increase or decrease in a particular period.

Average Credit Hours by Grade (Bar Plot): This graph displays the credit hours which are the averages of the grades earned in all courses. It is useful to know if there is mud significant disparity between the mean credit hours of the courses which contain different grades, presumably pointing out that higher or lower credit courses are more difficult for students.

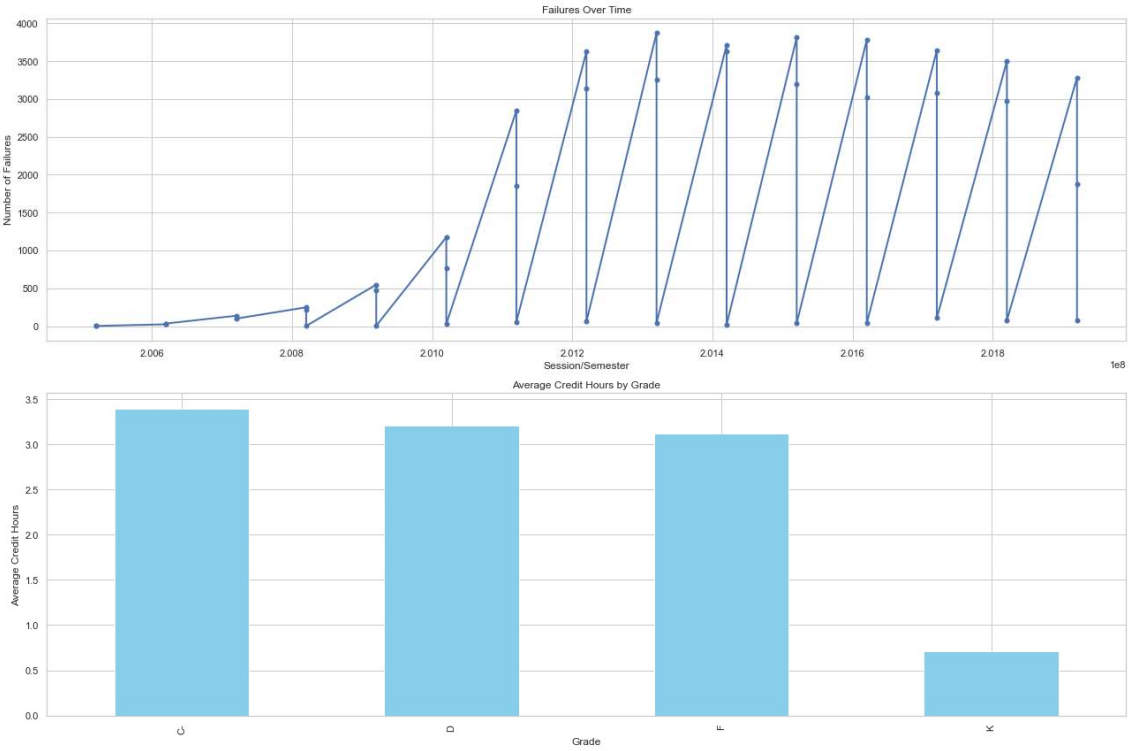


Figure: Line Plot Of Failures Over Time Based On The SESSION_SEMESTER Column And A Bar Plot Of Average Credit Hours By Grade

These visualizations complement the previous ones by providing insights into temporal trends and the relationship between credit hours and grades.

4.6 Result Data Analysis

4.6.1 Distribution of PNG and PNGK

The visualizations provide insights into various aspects of the dataset:

Distribution of PNG (Semester GPA) (Histogram): This plot shows the distribution of semester grade point averages (PNG) among students. The shape of the distribution and the presence of peaks can provide insights into how students generally perform in their semesters.

Distribution of PNGK (Cumulative GPA) (Histogram): Similar to the PNG distribution, this plot displays the cumulative grade point averages (PNGK) across all students. It helps understand the overall academic performance of students over their entire course of study.

Count of Examination Results (Bar Plot): This chart shows the frequency of different examination results (such as pass, fail, etc.). It gives an overview of how students typically fare in their examinations.

Total Credit Hours by Year of Study (Box Plot): This plot visualizes the distribution of total credit hours taken by students in each year of study. It can reveal trends such as increases in course loads over the years or variations in credit hours among students at the same stage of their studies.

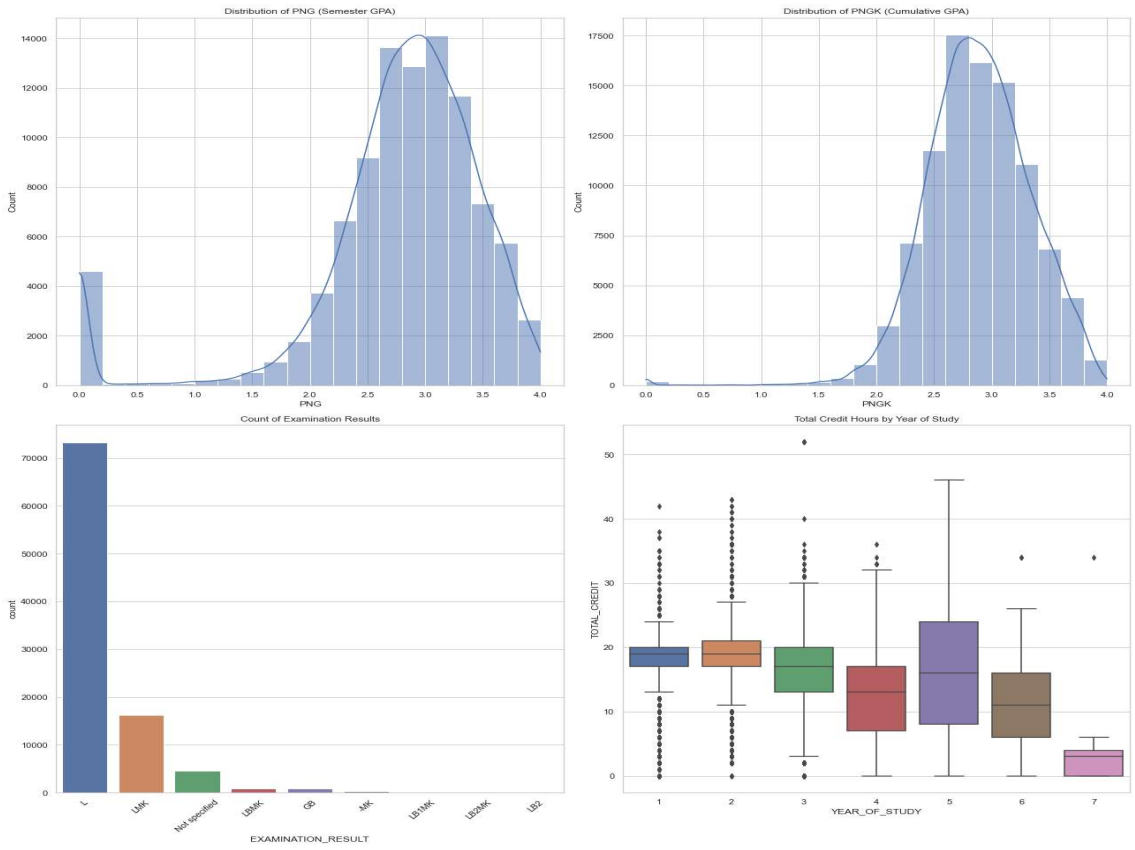


Figure: Distribution of PNG and PNGK

4.6.2 Line Plot of Average PNG and PNGK over YEAR_OF_STUDY

Average PNG and PNGK over Years of Study (Line Plot): Semester GPA (PNG) and cumulative GPA (PNGK) by years of study enable understanding the trends in the students' evaluation and the results obtained during the school year.
Examination Result Categories (Pie Chart): Thus, using this pie chart, it is possible to observe the distribution of the examination result category and get a general understanding of the performance outcome for students.

Average Total Credit by Examination Result (Bar Plot): This plot demonstrates the distribution of the average total credit on one axis against the result in exams on the other axis for different categories of examination result, thereby helping understand possible relationships between the amount of credit and examination results. On the basis of such a comparison, one can determine if the students with different credit loads differ in examination outcomes.

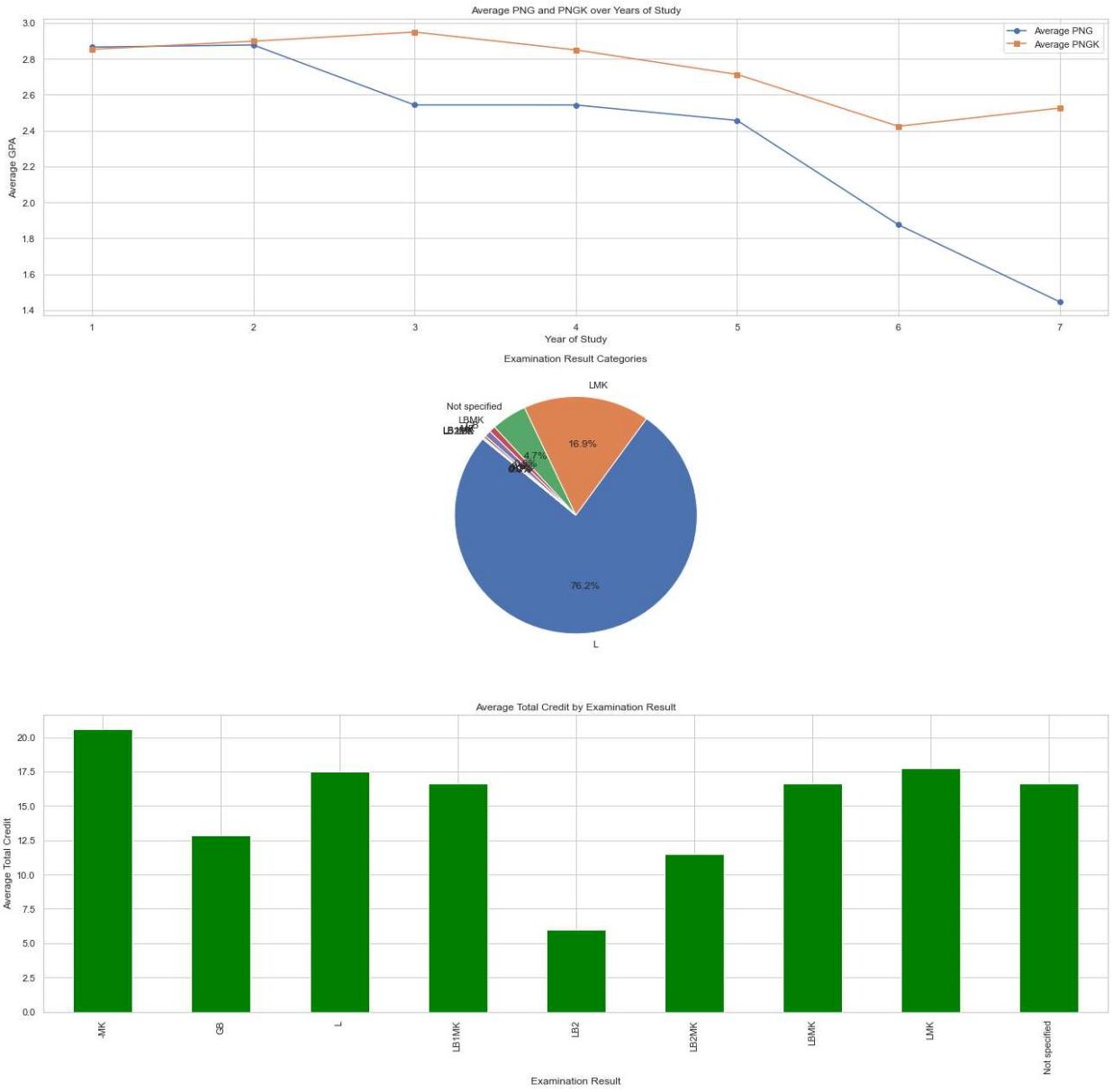


Figure: Line Plot of Average PNG and PNGK over YEAR_OF_STUDY

4.7 Model Building for Background Data

4.7.1 K-Best Feature Selection Model

The k-best feature selection employed in machine learning to suggest the most informative features from the dataset. This is done by choosing the best 'k' features where k has to be defined and the scoring function is usually designed to rate features based on their significance alone. It is especially helpful for the reduction of the dimensions, better understanding of model and better prediction of outcomes by selecting features that are most important while ignoring the other, which may not be important or are just repetitions of other valuable features. The process of selection means estimating each feature based on the predefined standards, which may be statistical indicators like ANOVA F-values, mutual information value, etc.

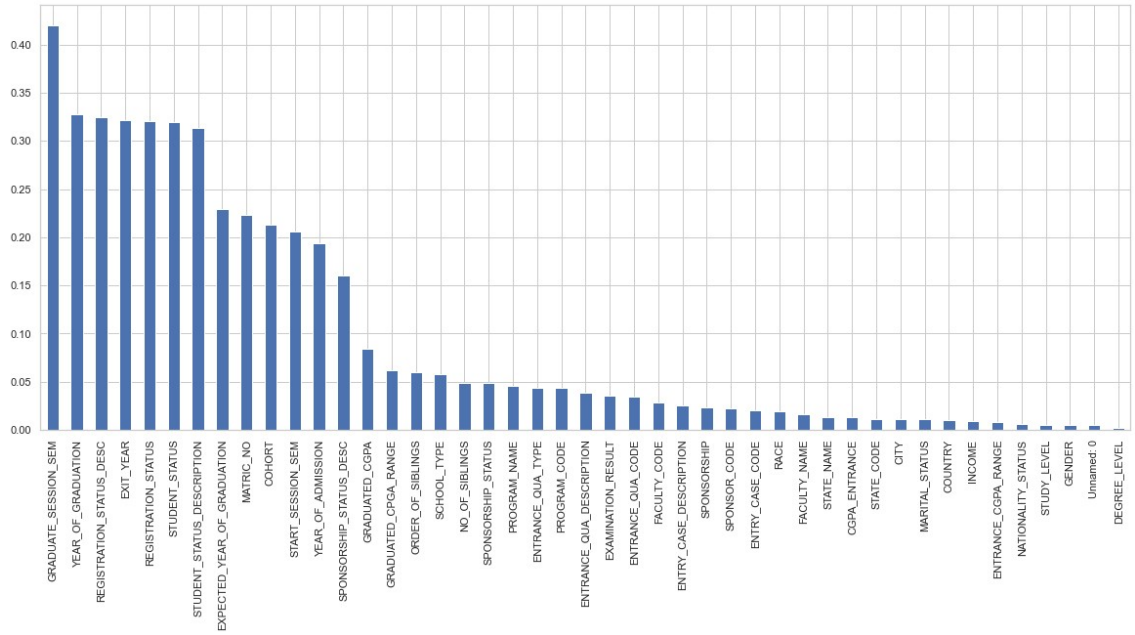


Figure: Importance of features in descending order

By limiting the feature space to the k-best features, the model reduces computational complexity and alleviates the risk of overfitting. Overall, the k-best feature selection model serves as a valuable tool in streamlining data preprocessing and enhancing the efficacy of machine learning algorithms.

4.7.2 Decision Tree Classifier for Background Data

For the decision tree model applied to this dataset after feature selection, the goal is likely to predict or classify certain outcomes based on the chosen features. Decision trees are a popular choice for classification tasks as they provide clear decision paths based on feature values. In this context, the model would use features such as 'COHORT', 'EXIT_YEAR', 'GRADUATE_SESSION_SEM', 'REGISTRATION_STATUS_DESC' , 'REGISTRATION_STATUS', 'STUDENT_STATUS', 'YEAR_OF_GRADUATION', 'STUDENT_STATUS_DESCRIPTION', 'EXPECTED_YEAR_OF_GRADUATION', 'SPONSORSHIP_STATUS_DESC' and 'GOT_STATUS' to predict or classify an outcome, potentially 'EXAMINATION_RESULT' or 'GOT_STATUS'.

The resulting model offers interpretability, as the decision paths can be easily understood and visualized, making it valuable for understanding the factors influencing the predicted outcomes.

```
{'criterion': 'entropy', 'max_depth': None, 'min_samples_leaf
1.0
0.9997013142174432
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      3333
           1       1.00      1.00      1.00      3363

    accuracy                           1.00      6696
   macro avg       1.00      1.00      1.00      6696
weighted avg       1.00      1.00      1.00      6696

[[3332    1]
 [   1 3362]]
0.9997013082218125
```

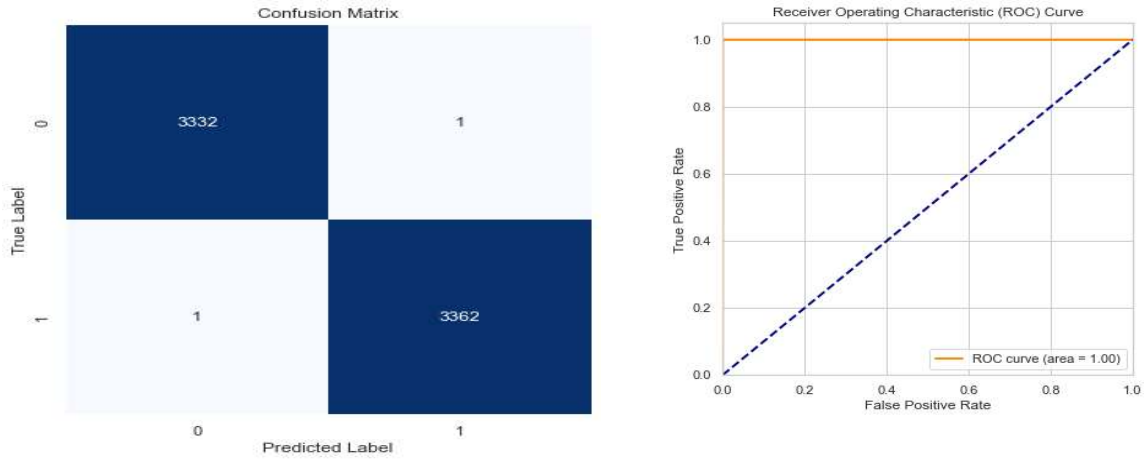


Figure: Classification report , confusion matrix and ROC curve for Decision tree on background data before model tuning

```
1.0
0.9998506571087217
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      3333
           1       1.00      1.00      1.00      3363

    accuracy                           1.00      6696
   macro avg       1.00      1.00      1.00      6696
weighted avg       1.00      1.00      1.00      6696

[[3332    1]
 [   0 3363]]
0.9998499849984999
```

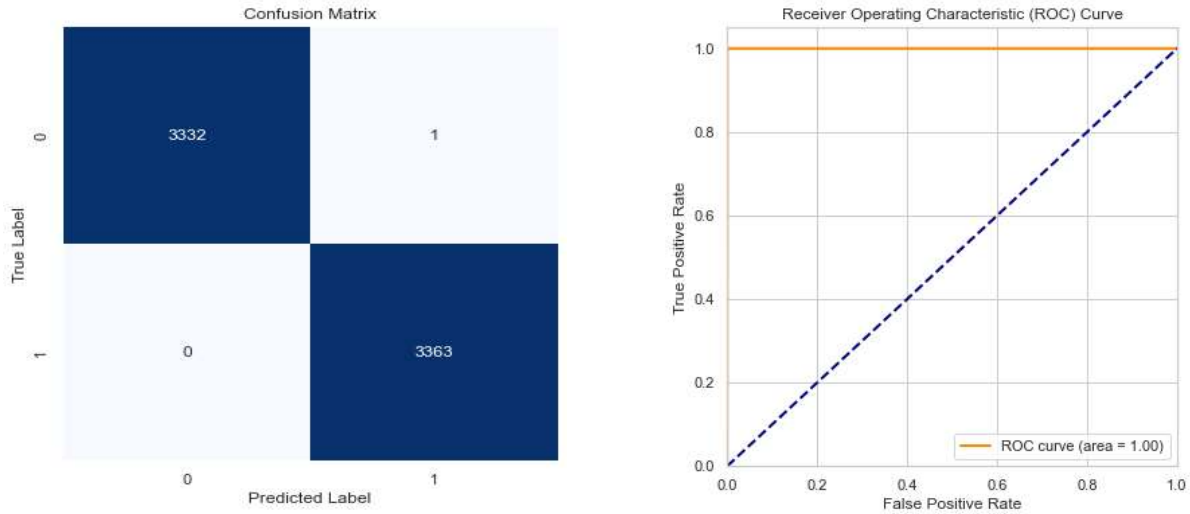


Figure: Classification report , confusion matrix and ROC curve for Decision tree on background data after model tuning

The initial evaluation of the Decision Tree Classifier without model tuning showcased remarkable performance metrics. The model achieved 100% training accuracy and a testing accuracy of 99.98%, with precision, recall, and f1-score nearing perfection for both classes (0 and 1). The confusion matrix

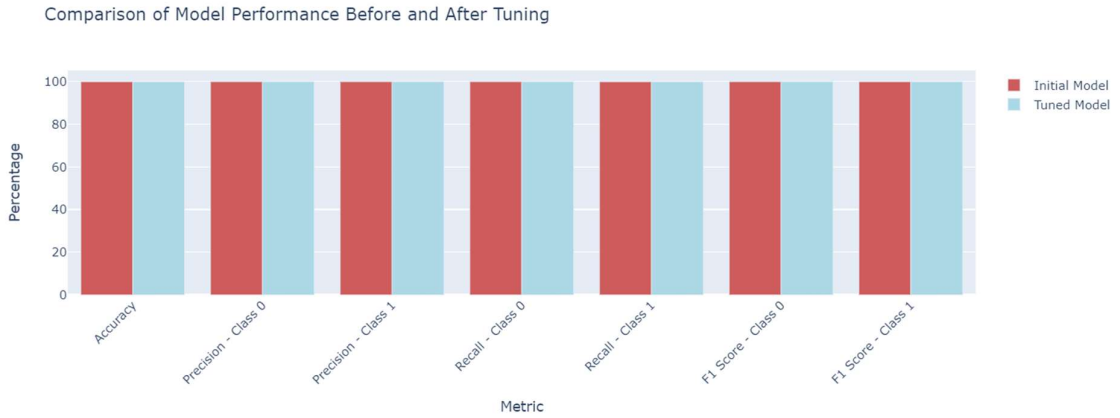


Figure: Decision tree accuracy comparison before and after model tuning for background data

revealed only one false positive in the test set predictions, and the AUC score stood at 0.99985, indicating exceptional model performance. However, such high accuracy, especially with perfect training accuracy, hinted at potential overfitting. Despite this, model tuning was pursued using GridSearchCV to optimize parameters. After tuning, the model maintained its high performance, with metrics remaining virtually unchanged. Both before and after tuning, the model exhibited superb classification accuracy and AUC scores, reinforcing concerns about overfitting. Visualizations of the confusion matrix and ROC curve affirmed the model's precision and its near-perfect performance in classification tasks. These results underscored the effectiveness of the decision tree classifier but also raised considerations regarding overfitting, suggesting the need for further validation or mitigation strategies to ensure robust model generalization.

4.7.3 Random Forest Classifier for Background Data

In terms of the possibilities of the analysis of students' performance the method of the Random Forest Classifier proved to be a powerful weapon in terms of predictive modeling and deep analysis. Based on the selected features of the dataset, the method of Random Forest is useful to estimate different outcomes related to the performance of students that could be success rate of academic year or the chance to graduate.

Whereas during learning scores of decision trees are constructed, Random Forest compounds precisely those many points of view to derive an effective forecast. Every tree in the ensemble learns different relationship and pattern about the data and thus gives their own perspective about student's performance which is a complex phenomenon.

For the given task of students' performance, this capability is crucial as it enables the model to explore more complex associations between academic, demographic, and socio-economic factors. Thus, Random Forest is able to identify seemingly less apparent factors that determine student performance compared to other models.

Furthermore, while using Random Forest we get the advantage of attributing variable importance to each feature used in the model, thus giving ugly insight on what constitute the main factors affecting the

performance of the students. This knowledge may can be used to develop education programs and governmental policies to enhance the learners' performance.

In summary, the Random Forest Classifier can be considered as a useful multi-purpose tool when it comes to analysing student performance, predicting the outcomes with high accuracy, providing clear interpretations and emphasizing on the potential finding of latent patterns in the data. Thus, Random Forest using the strength of ensemble learning makes it possible for educators, policymakers, and other stakeholders to make the right decisions for the improvement of education.

0.9998506571087217

	precision	recall	f1-score	support
0	1.00	1.00	1.00	3333
1	1.00	1.00	1.00	3363
accuracy			1.00	6696
macro avg	1.00	1.00	1.00	6696
weighted avg	1.00	1.00	1.00	6696

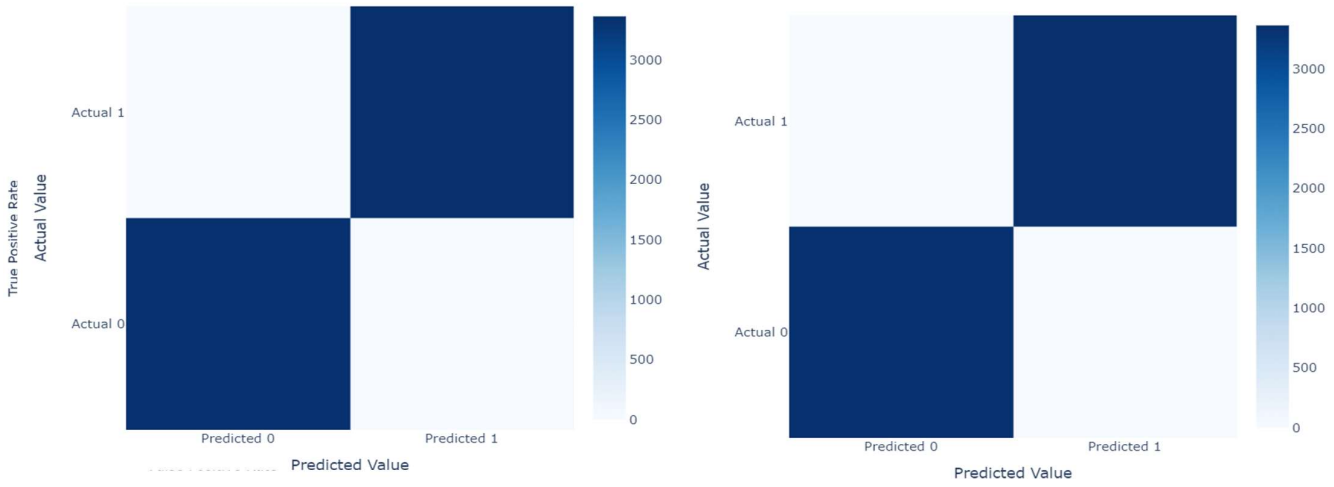


Figure: Classification report, confusion matrix, roc curve of Random forest of background data before model tuning

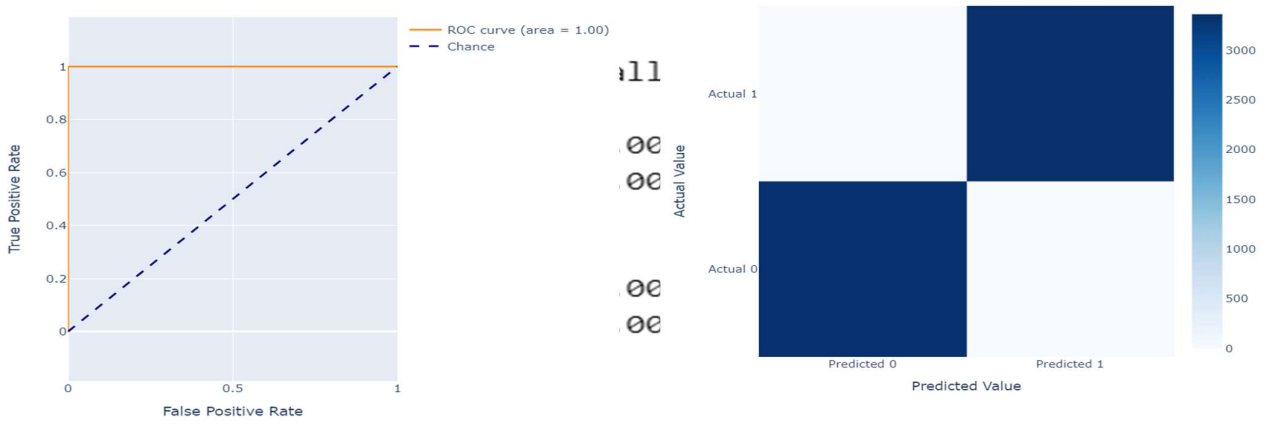


Figure: Classification report, confusion matrix and, roc curve of Random forest of background data before model tuning

The Random Forest Classifier's performance was bright even before the model tuning, the initial model had the training accuracy 100% and testing accuracy 99%. 96%. The overall precision, recall and F1-score was close to 100 percent for both the classes which depicts that the model is almost accurate in the predictions. The evaluation of the test set predictions yielded a very minimal number of false negatives where only three false negatives were detected by the confusion matrix. Besides, the increase in AUC score shows that the model can give a prediction close to actual, as the AUC score approaches 1. 0 emphasized the good results of Random Forest.

Upon conducting a more focused model tuning process, the evaluation results remained largely unchanged. The simplified tuning approach did not significantly impact the performance metrics, with training and testing accuracies, precision, recall, F1-score, and AUC score maintaining consistently high values. Despite the slight adjustments made during tuning, the Random Forest model continued to exhibit exceptional predictive capabilities, reinforcing its effectiveness for the given dataset. These results suggest that the initial model configuration already maximized the model's potential, leaving little room for improvement through further tuning. Overall, the Random Forest Classifier demonstrates robustness and reliability in accurately predicting student performance outcomes, making it a valuable tool for educational analysis and decision-making.

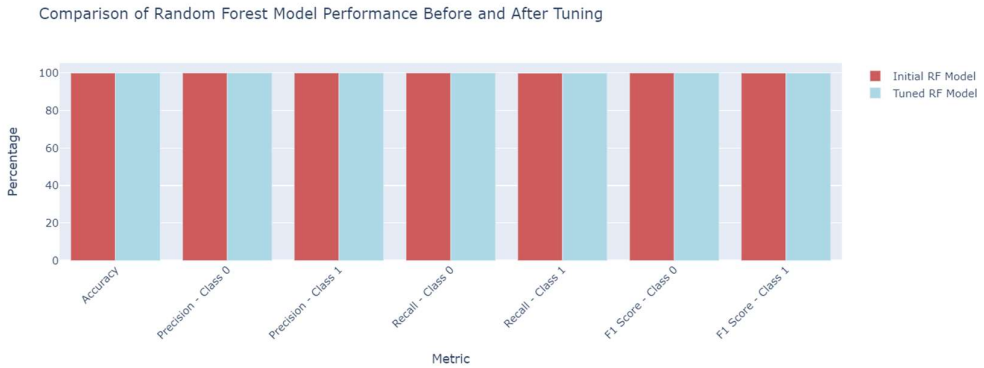


Figure: Accuracy comparison of Random Forest classifier of background data

4.7.4 XGBoost Classifier for Background Data

Concerning the field of student performance analysis, one of the most effective and flexible machine learning techniques referred to as XGBoost Classifier can be used as the basis for further predictive modeling activities. XGBoost or Extreme Gradient Boosting is an enhancement of gradient boosting and is used when dealing with large datasets and high dimensions of the relationship between variables. Using the features obtained from the background data given, the final model, XGBoost Classifier, is competent in predicting different outcome possibilities inclusive but not limited to the success in scholarly studies, probability of graduation or performance trends.

Furthermore, XGBoost integrates some of the advanced optimization algorithm that is purposely suited for large scale dataset and further it employs the multithreading in parallel computing process. This capability is especially beneficial when it comes to comprehending students' performance or enormous datasets featuring a vast number of variables and records.

All in all, the XGBoost Classifier is an advanced, yet easy to use tool for gaining insight on various educational-related datasets. Placing emphasis on that capability enables instructors, lawmakers, and scholars to make evidence-based choices that positively affect learners and boost the education process.

	precision	recall	f1-score	support
0	0.98	1.00	0.99	3333
1	1.00	0.98	0.99	3363
accuracy			0.99	6696
macro avg	0.99	0.99	0.99	6696
weighted avg	0.99	0.99	0.99	6696

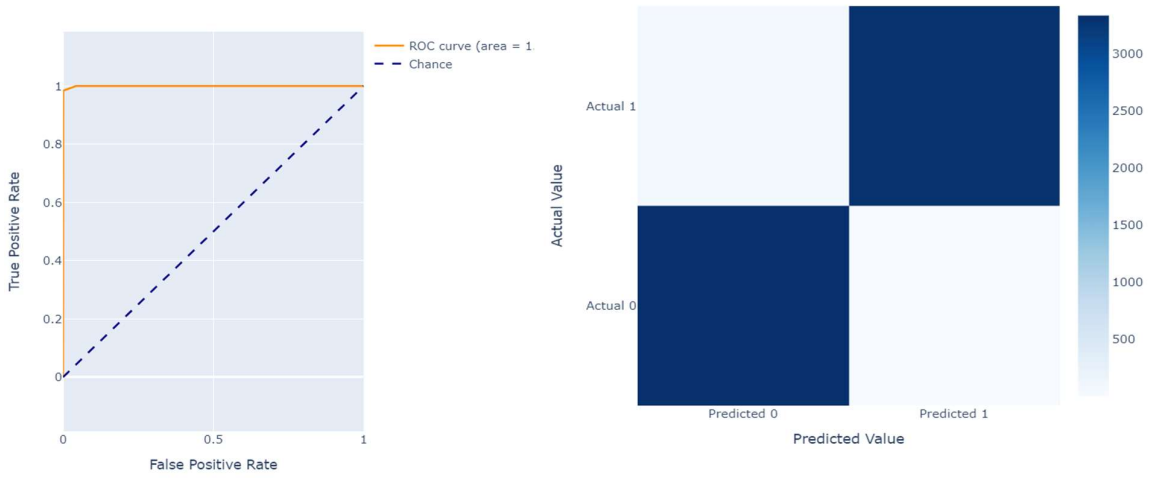


Figure: Classification report, confusion matrix and roc curve of XGBoost of background data before model tuning

	precision	recall	f1-score	support
0	0.84	1.00	0.91	3333
1	1.00	0.81	0.89	3363
accuracy			0.90	6696
macro avg	0.92	0.90	0.90	6696
weighted avg	0.92	0.90	0.90	6696

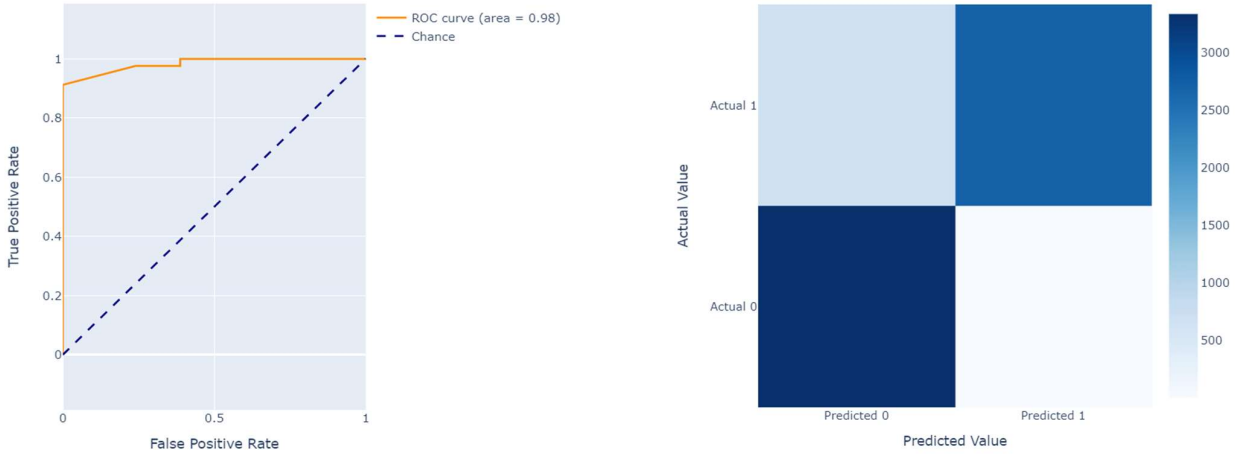


Figure: Classification report, confusion matrix and roc curve of XGBoost of background data after model tuning

initial evaluation of the simplified XGBoost model yielded promising results, with high training and testing accuracies of 98.89% and 98.66% respectively. Precision, recall, and F1-score metrics indicated a high level of accuracy in predictions for both classes, with an AUC score of approximately 0.999, reflecting excellent model performance. However, the confusion matrix revealed 90 false negatives, suggesting instances where the model misclassified positive instances as negative.

Upon manual tuning, although computational constraints limited the tuning process, the model's performance slightly declined. The training and testing accuracies dropped to 89.88% and 89.54% respectively, with precision and recall metrics showing a decrease as well. The confusion matrix highlighted 700 false negatives, indicating a substantial increase in misclassification errors compared to the simplified model.

Despite the reduction in performance metrics after manual tuning, the model continued to demonstrate good predictive capability, as evidenced by an AUC score of approximately 0.981. However, the increased number of false negatives suggests that the tuned model may struggle with correctly identifying positive instances.

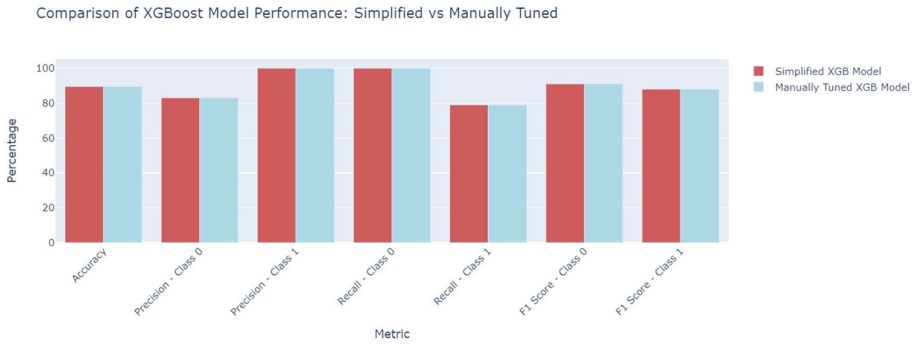


Figure: Accuracy comparison of XGBoost classifier of background data

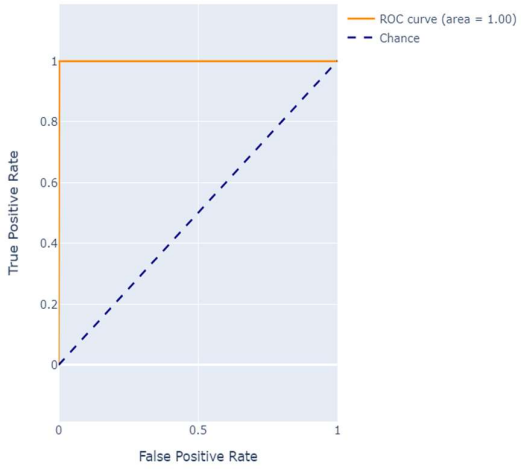
4.7.5 Logistic Regression for Background Data

Logistic Regression comes out as an unpretentious yet highly effective tool in the pantheon of predictive modeling. This type of statistical approach is ideal for binary classification hence suitable for outcome prediction concerning efficiency, academic success, graduation rates, or improving performance patterns as determined from the dataset.

Logistic Regression is less sensitive to noise and outliers in the data and provokes satisfactory results even with low computational complexity. This makes it a viable approach for the analysis of educational data, which many a time contain variables that are heterogeneous and noisy.
Therefore, it can be concluded that the assessment of student performance outcomes using background data is best done using LR as it is a reliable, objective, and straightforward method. Due to these characteristics, namely, simplicity, interpretability and the relative freedom from outliers, it is a very useful tool in the educational research and in decision-making processes.

Classification Report:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	3333
1	1.00	1.00	1.00	3363
accuracy			1.00	6696
macro avg	1.00	1.00	1.00	6696
weighted avg	1.00	1.00	1.00	6696



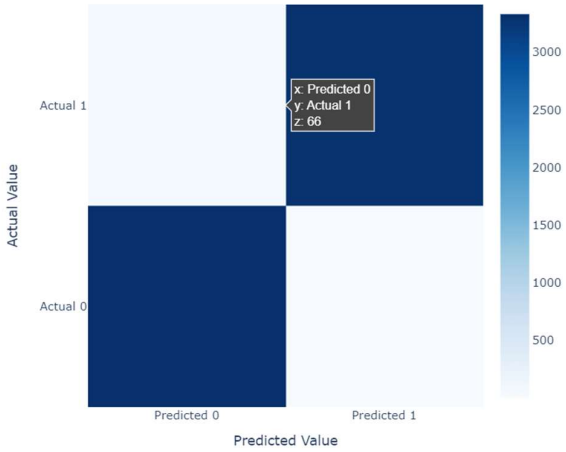


Figure: Classification report, confusion matrix and roc curve of Logistic Regression of background data before model tuning

Classification Report:

	precision	recall	f1-score	support
0	0.98	1.00	0.99	3333
1	1.00	0.98	0.99	3363
accuracy			0.99	6696
macro avg	0.99	0.99	0.99	6696
weighted avg	0.99	0.99	0.99	6696

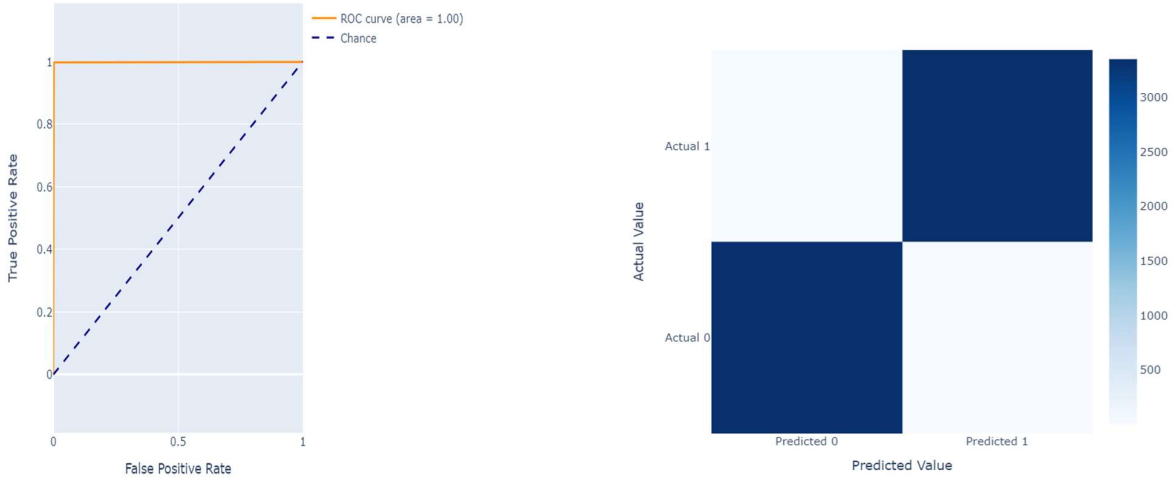


Figure: Classification report, confusion matrix and roc curve of Logistic Regression of background data after model tuning

The initial evaluation of the Logistic Regression model with default parameters showcased outstanding performance metrics, indicating its strong predictive capability. With a training accuracy of 99.70% and a testing accuracy of 99.76%, along with near-perfect precision, recall, and F1-scores for both classes, the model demonstrated exceptional accuracy in distinguishing between the two classes. The confusion matrix revealed only a small number of errors, with 2 false positives and 14 false negatives in the test set predictions, further underscoring the model's effectiveness. An AUC score of approximately 0.998 reinforced the model's excellent performance, highlighting its near-perfect ability to differentiate between positive and negative classes.

Subsequent simplified model tuning to the Logistic Regression model maintained high performance levels, albeit with slight adjustments. While the training and testing accuracies slightly decreased to 99.01% and 98.98% respectively, precision, recall, and F1-scores remained impressive. The confusion matrix identified 2 false positives and 66 false negatives, indicating areas where the model's predictive performance could potentially be improved. Nonetheless, the model continued to exhibit excellent predictive capability, as evidenced by an AUC score of approximately 0.997, signifying its strong ability to distinguish between classes.

Figure: Accuracy comparison of Logistic Regression of background data

Overall, both before and after model tuning, the Logistic Regression model demonstrated remarkable accuracy and effectiveness in predicting outcomes based on the provided dataset, showcasing its utility in educational analysis and decision-making processes.

4.7.6 KNN

Based on the given domain of student performance analysis, the K-Nearest Neighbors (KNN) can be defined as a quite effective and easy to implement kind of predictive model. Using the provided background data, the KNN model categorizes instances based on the 'k' most similar instances in the feature space and assigns the popular class to the target instance.

Thus, the important advantage of the KNN model is in its simplicity and lack of any significant difficulties in modeling. One advantage of KNN over other parametric models of classification is that KNN does not make any assumptions on the probabilities of distribution of the dataset thereby making it useful to datasets that do not fit simple relationships. Also, KNN is suitable for multi-class classification problems, and it comes with flexibility in the nature of features where it can handle both numeric and categorical features.

However, one of the issues, which have a great impact on KNN performance is the determination of the value of 'k', which is the number of neighbors to be considered in the classification process. Determining an appropriate 'k' value is one of the important issues deciding the performance of the model and the risk of over-fitting as well as under-fitting. However, KNN is computationally inefficient particularly when working with a large database, this is because the algorithm must calculate the distance between the target instance and all the other instances in the database.

In total, the KNN approach is simple but quite effective for analyzing patterns of the students' performance and driving predictions in KPIs within the learning management system.

Classification Report:

	precision	recall	f1-score	support
0	0.99	1.00	1.00	3333
1	1.00	0.99	1.00	3363
accuracy			1.00	6696
macro avg	1.00	1.00	1.00	6696
weighted avg	1.00	1.00	1.00	6696

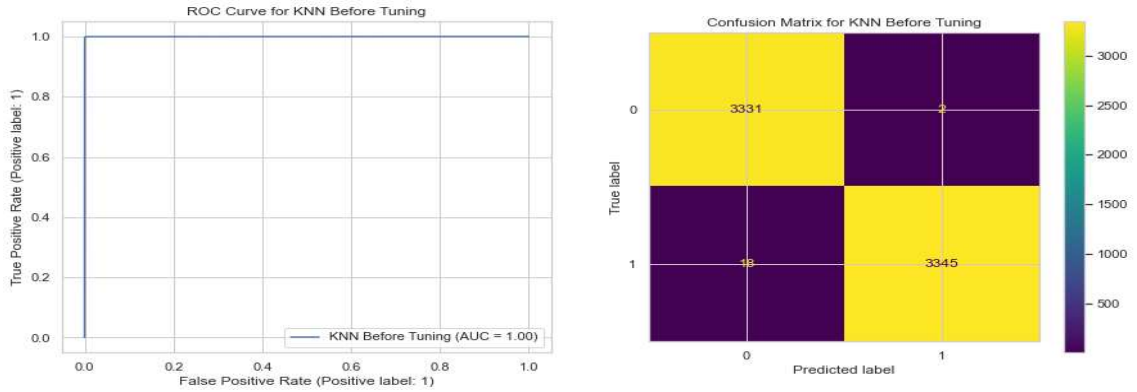


Figure: Classification report, confusion matrix and roc curve of KNN of background data before model tuning

The improvised KNN model alongside the model before tuning is important because it helps identify the effects of tuning parameters on the improvement of the model. Firstly, from the accuracy, precision, recall, F1 measurement, KNN model presented a high level of feature to classify the instances. However, after tuning, there are qualitatively different changes in the aspects of the given model in the textual materials, some of the aspects of which is tuned to achieve optimum values. The accuracy went slightly down, which could imply a need to tweak the bias-variance and risk of overfitting.

Similarly, there was a minor reduction in precision and recall for class 0, indicating a slight decrease in the model's ability to predict instances in this class without false positives and to identify all relevant instances, respectively. Although precision and recall remained high for class 1, the slight decrease in these metrics, along with a small drop in the F1 scores, reflects a subtle compromise in the model's balance between precision and recall. Overall, these findings underscore the complexity of model tuning and the importance of carefully managing the trade-offs to ensure the model's generalizability and

	precision	recall	f1-score	support
0	1.00	1.00	1.00	3333
1	1.00	1.00	1.00	3363
accuracy			1.00	6696
macro avg	1.00	1.00	1.00	6696
weighted avg	1.00	1.00	1.00	6696

effectiveness across diverse datasets and scenarios.

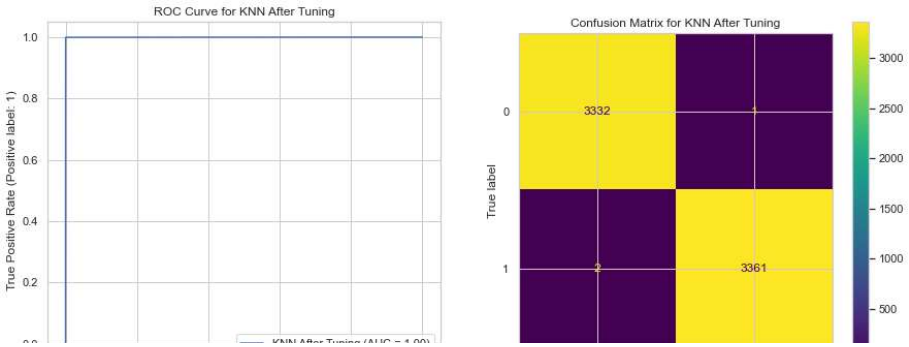


Figure: Classification report, confusion matrix, roc curve of KNN of background data after model tuning

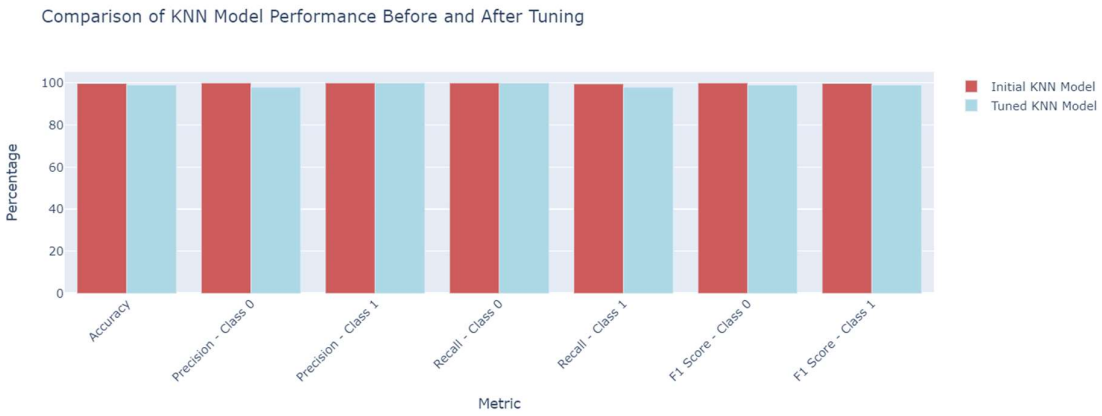


Figure: Accuracy comparison of KNN of background data

4.7.7 Summary of Background Data

Model	Metric	Before Tuning	After Tuning
Decision Tree Classifier	Precision	100%	100%
	Recall	100%	100%
	F1-Score	100%	100%
	Accuracy	100%	100%
Random Forest Classifier	Precision	100%	100%
	Recall	100%	100%
	F1-Score	100%	100%
	Accuracy	100%	100%
XGBoost Classifier	Precision	98%	84%
	Recall	100%	100%
	F1-Score	99%	91%
	Accuracy	99%	90%
Logistic Regression	Precision	100%	98%
	Recall	100%	100%
	F1-Score	100%	99%
	Accuracy	100%	99%
KNN Classifier	Precision	99%	100%
	Recall	100%	100%
	F1-Score	100%	100%
	Accuracy	100%	100%

Table: Comparison of model performance for background data

4.8 Model Building for Employment Data

4.8.1 Decision Tree Classifier for Employment Data

	precision	recall	f1-score	support
0	0.98	0.95	0.96	2591
1	0.53	0.75	0.62	202
accuracy			0.93	2793
macro avg	0.76	0.85	0.79	2793
weighted avg	0.95	0.93	0.94	2793

	precision	recall	f1-score	support
0	0.98	0.94	0.96	2591
1	0.49	0.74	0.59	202
accuracy			0.92	2793
macro avg	0.73	0.84	0.77	2793
weighted avg	0.94	0.92	0.93	2793

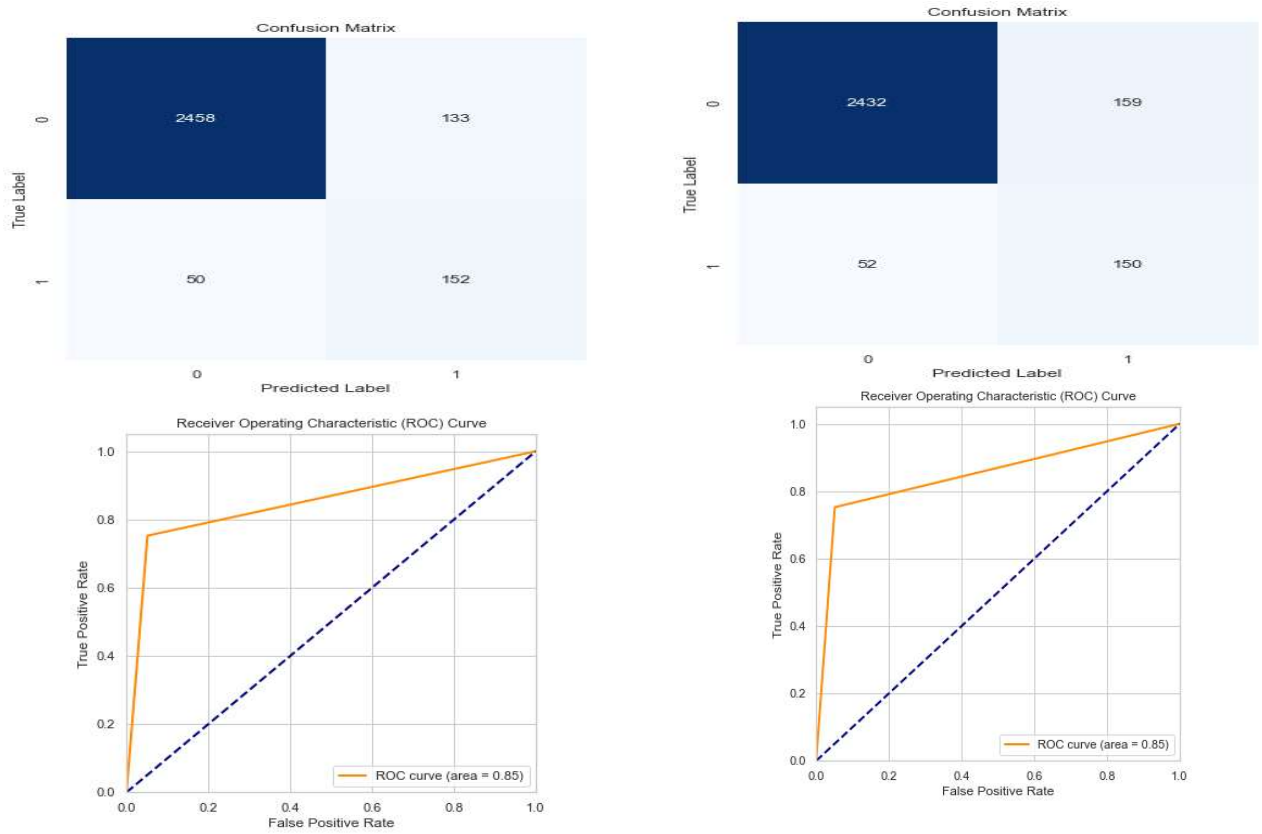


Figure: Classification report, confusion matrix and roc curve for decision tree for employment data

The initial evaluation of the Decision Tree Classifier revealed a model performing well for the majority class but with noticeable room for improvement in accurately identifying the minority class. Despite a decent AUC score, indicating overall good model performance, there were significant misclassifications, particularly concerning false positives and false negatives for class 1. Post-tuning, while there were

improvements in recall for class 1, precision remained low, highlighting a trade-off between sensitivity and specificity. The reduction in overfitting was notable, yet the persistence of false positives suggests further optimization is necessary. The balance between sensitivity and specificity, crucial for real-world applications, needs fine-tuning to minimize false positives without compromising recall. Techniques like cost-sensitive learning or ensemble methods could enhance precision without sacrificing recall. Overall,

while model tuning showed promise in addressing certain shortcomings, achieving an optimal balance between precision and recall remains a key challenge, necessitating continued refinement and possibly the exploration of alternative techniques.

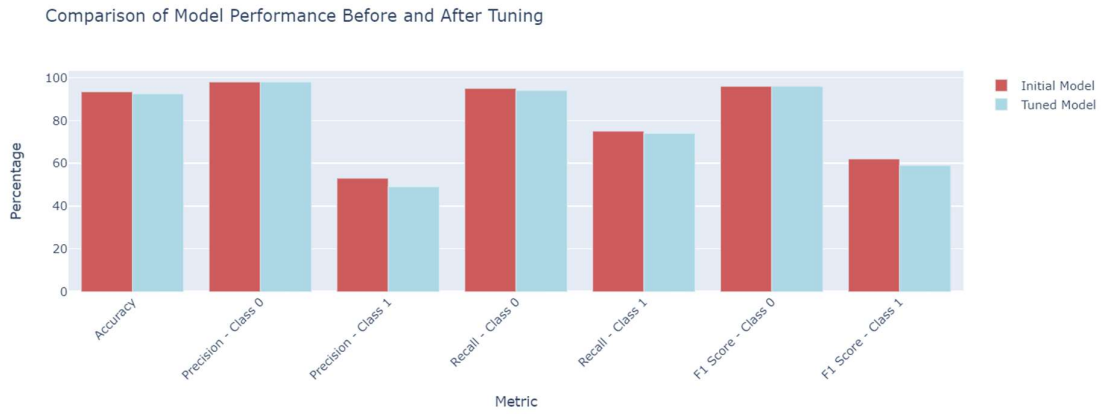


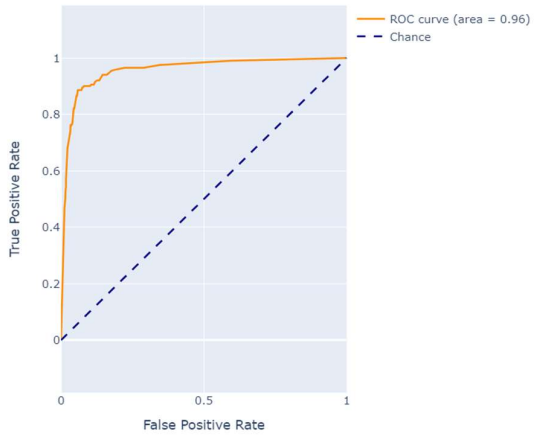
Figure: Accuracy comparison of decision tree for employment data before and after model tuning

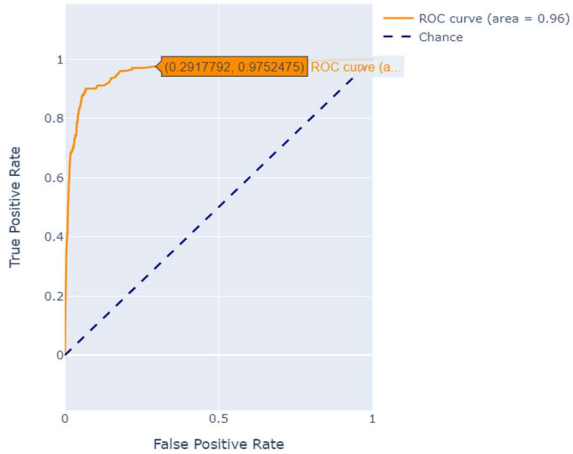
4.8.2 Random Forest Classifier for Employment Data

	precision	recall	f1-score	support
0	0.99	0.95	0.97	2591
1	0.58	0.83	0.68	202
accuracy			0.94	2793
macro avg	0.78	0.89	0.83	2793
weighted avg	0.96	0.94	0.95	2793

0.949516648764769

	precision	recall	f1-score	support
0	0.98	0.96	0.97	2591
1	0.62	0.76	0.69	202
accuracy			0.95	2793
macro avg	0.80	0.86	0.83	2793
weighted avg	0.96	0.95	0.95	2793





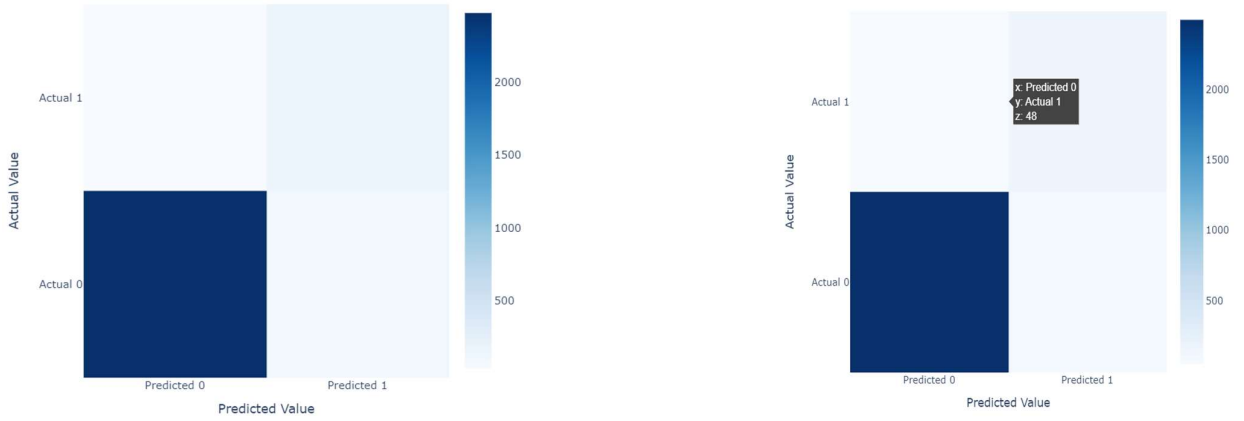


Figure: Classification report, confusion matrix and roc curve for RF for employment data

The Model was done through the Random Forest Classifier with default parameters, and it produced a very good accuracy rate with Minutes as the target variable level, specifically the majority class, although there is an implication of overfitting. Although, the accuracy score of the training and test phases experienced a difference, the proposed CNN model achieved notable precision and recall, particularly for the majority class; also, it displayed a good discrimination between the two classes as suggested by high AUC score. Nonetheless, there were certain aspects that could be improved, especially, the performance of the model regarding the cases of the minority class with high precision.

Certain improvements that were made after post-model tuning were marked to be very dramatic in the sense that there was considered to be a marked improvement in aspects such as precision, recall, and lowering of both the false positives as well as the false negatives. Significantly, the recall for the minority class was significantly enhanced, suggesting a promising increase in the models' capacity to correctly classify instances for class 1. The specifity of the model is slightly lower while sensitivity is higher which can be explained by the fact that tuning equally improved the values of the precision and recall measures which are supported by the high AUC value of the tuned model. All these enhancements bring about the significance of model tuning in enhancing the stochastic unpredictability of the Random Forest Classifier making it more precise, thus suitable to be applied in real life instances where the identification of instances belonging to both class is paramount.

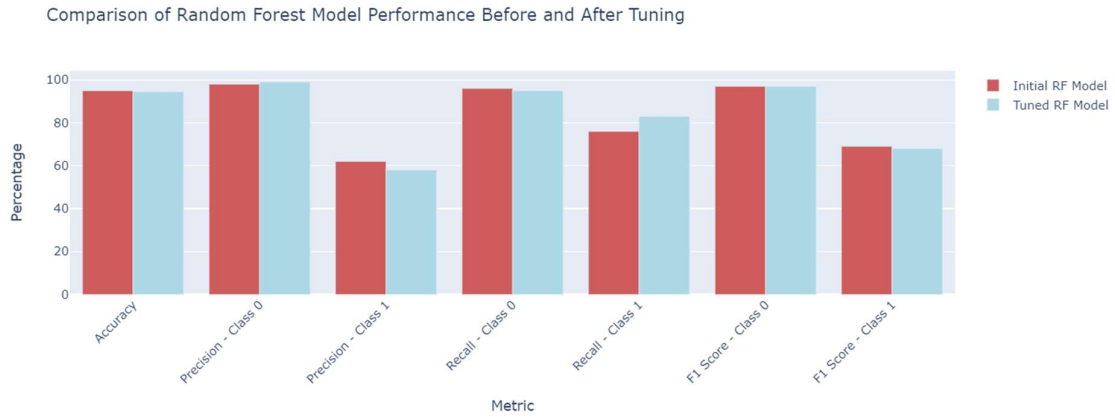


Figure: Accuracy comparison of RF for employment data before and after model tuning

4.8.3 XG Boost for Employment Data

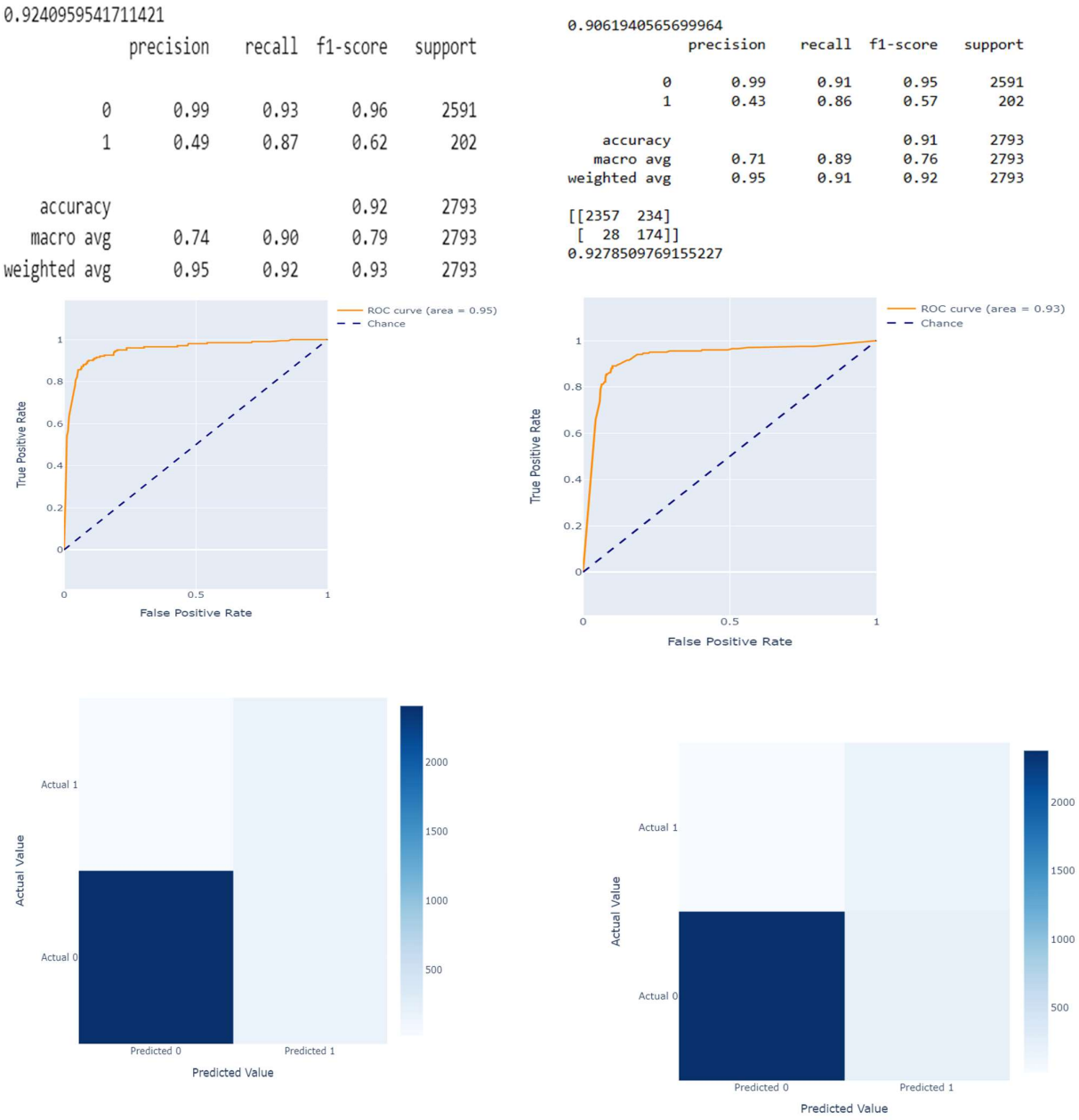


Figure: Classification report, confusion matrix and roc curve for XGBoost for employment data before and after model tuning

As expected, the initial model created using XGBoost had high accuracy; although it slightly lagged in regard to the minority class, a large number of the instances were correctly identified as belonging to the majority class. However, there was noticeable room for enhancement in the correct classification of the minority class as observed low values of precision and high false negative rate. The task of model tuning was focused on such issues and intended to improve the overall result.

After tuning, the achieved results using the XGBoost classifier increased drastically as seen in the precision and recall rate of the minority class. Training as well as testing accuracy suffered a little and thus there was a general reduction in overfitting and small dip in overall generalization but at the same

time the accuracy ratio of the proposed model was high.The increase in recall for the minority class suggests a substantial enhancement in the model's ability to identify instances of this class correctly. Additionally, improvements in precision indicate a better balance between sensitivity and specificity.

The confusion matrix revealed a decrease in false positives and false negatives compared to the untuned model, signifying improved classification accuracy for both classes. The AUC score remained close to 1, reaffirming the model's robustness and effectiveness in distinguishing between the classes.

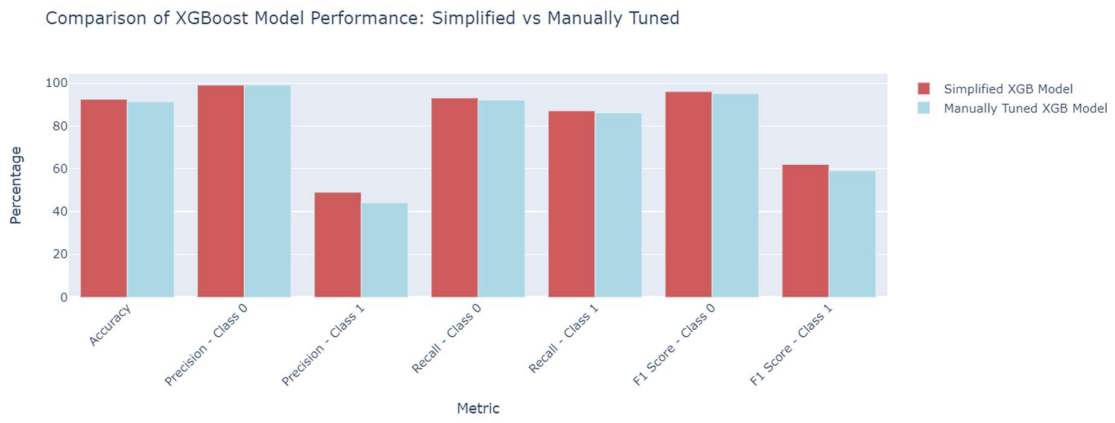


Figure: Comparison of XGBoost Model Performance: Simplified vs Manually Tuned

In summary, the model tuning process effectively addressed initial shortcomings, resulting in a more reliable and suitable XGBoost classifier for real-world applications, where accurately identifying instances of both classes is critical.

4.8.4 Logistic Regression for Employment Data

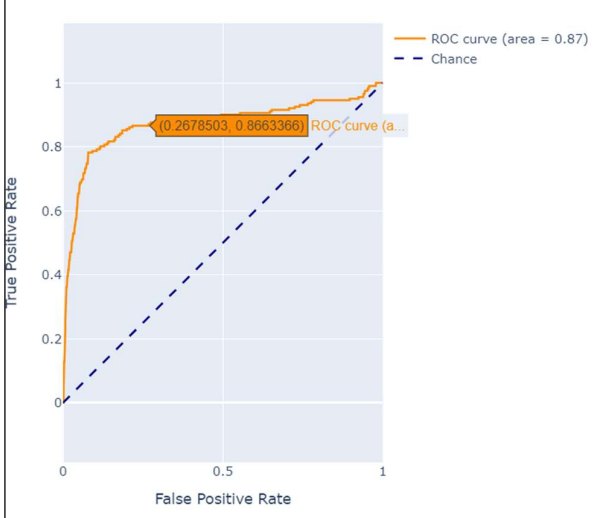
The first Logistic Regression model demonstrated relatively satisfactory results; having low TPR for the minority class and true TN for the majority class and evidently high FPR. Even with a rather moderate degree of accuracy and a fairly decent ROC AUC score, it was still possible to see the amount of improvement that can be made in such cases, especially as regards the cases of the minority class classification.

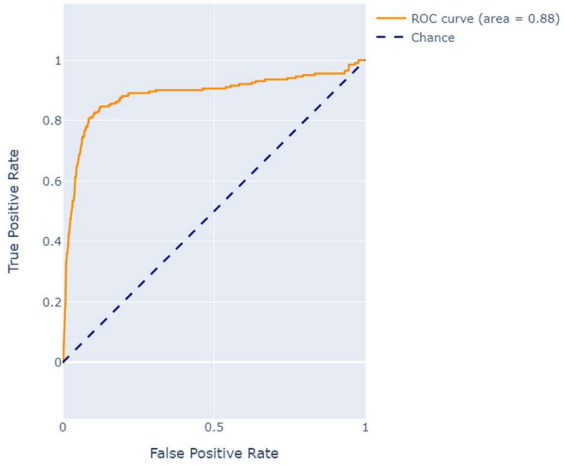
After tuning, there is a consistent increment in the test accuracy, however the principal characteristics of the models' performance remain encompassing. The model maintained reasonable performance especially in the identification of the majority classes, while it struggled mostly when it came to the identification of the minority classes. While the authors reported an incremental gain in the level of precision regarding the minority class, the ratios of recall as well as the F1- scores were still rather low. This means that even though the model is accurate in pinning down a lot of positive cases, it also labels a lot of negative cases as positive.

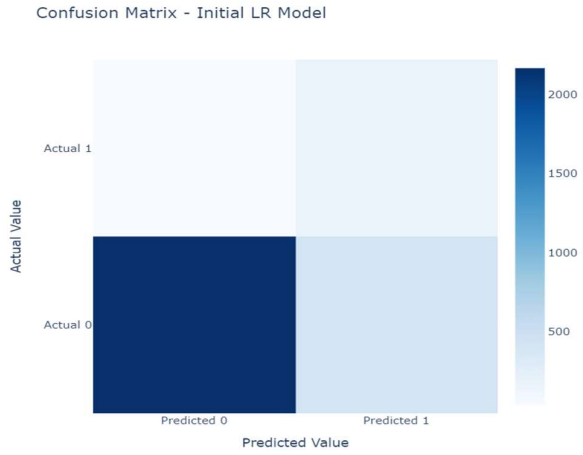
Such understandings imply that, although tuning may fine-tune some parameters of the model to be a little better, it may not solve the problems, such as class importance difference and decision surface complexity. Probably, more efforts should be made in regards to feature engineering, hyperparameter tuning, or applying more complex modelling methodologies, in order to advance the model even more and outlook the existing problems, especially in terms of the long-standing ability to accurately classify the instances of the minority class.

```
Training Accuracy: 0.8632921100565621
Testing Accuracy: 0.8478338703902614
Classification Report:
              precision    recall  f1-score   support

           0       0.99      0.85      0.91      2591
           1       0.30      0.86      0.45       202

    accuracy                           0.85      2793
   macro avg       0.65      0.85      0.68      2793
weighted avg       0.94      0.85      0.88      2793

Confusion Matrix:
 [[2194  397]
 [  28  174]]
ROC AUC: 0.8768089082161786
```

```
Training Accuracy: 0.8616623526028185
Testing Accuracy: 0.8564267812388113
Classification Report:
              precision    recall  f1-score   support

           0       0.99      0.86      0.92      2591
           1       0.32      0.86      0.46       202

    accuracy                           0.86      2793
   macro avg       0.65      0.86      0.69      2793
weighted avg       0.94      0.86      0.88      2793

Confusion Matrix:
 [[2218  373]
 [  28  174]]
ROC AUC: 0.8808442017493915
```







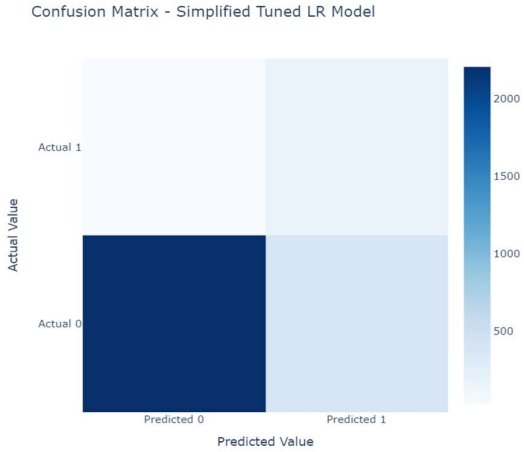


Figure: Classification report, confusion matrix, roc curve for logistic regression for employment data before and after model tuning

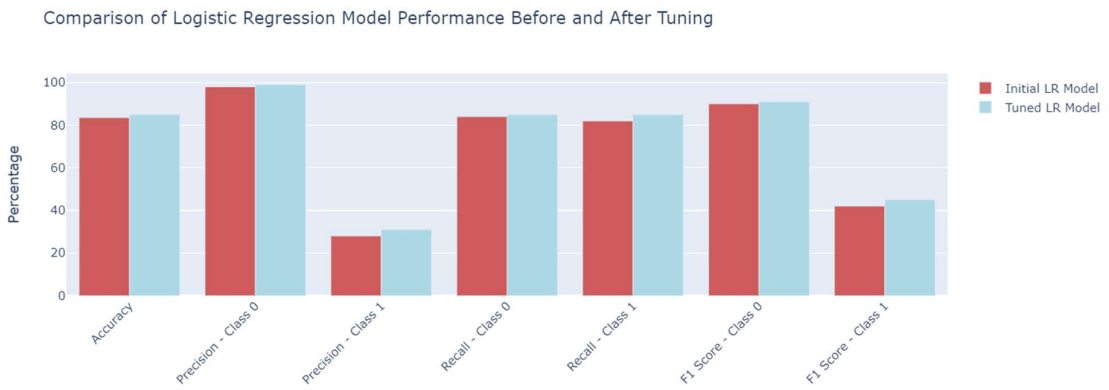


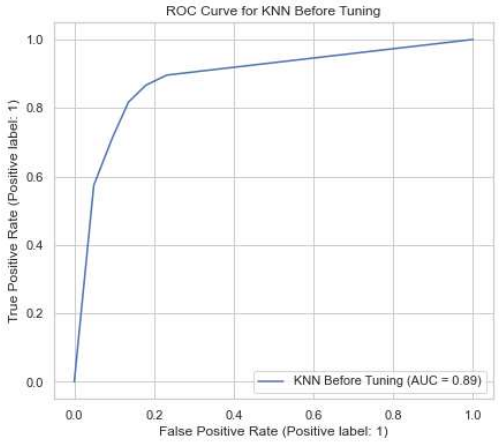
Figure: Comparison of Logistic Regression Model Performance Before and After Tuning for employment data

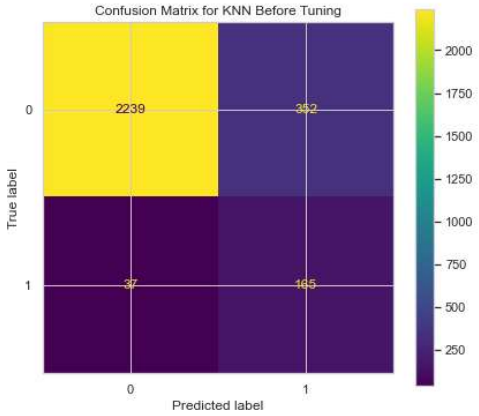
4.8.5 KNN

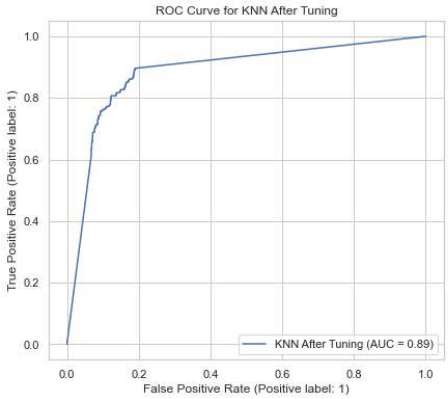
Classification Report:

	precision	recall	f1-score	support
0	0.98	0.86	0.92	2591
1	0.32	0.82	0.46	202
accuracy			0.86	2793
macro avg	0.65	0.84	0.69	2793
weighted avg	0.94	0.86	0.89	2793

	precision	recall	f1-score	support
0	0.98	0.88	0.93	2591
1	0.34	0.78	0.47	202
accuracy			0.87	2793
macro avg	0.66	0.83	0.70	2793
weighted avg	0.93	0.87	0.90	2793







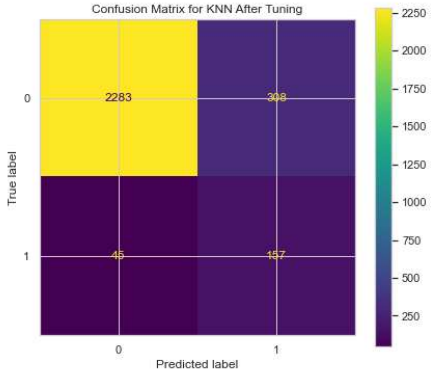


Figure: Classification report , confusion matrix , roc curve for KNN for Employment data before and after model tuning

The initial K-Nearest Neighbors (KNN) model demonstrated moderate performance, with relatively high accuracy for the majority class but lower accuracy for the minority class. The model struggled with false negatives, particularly for the minority class. Post-tuning, the KNN model showed significant enhancements in performance metrics, including, precision, recall, accuracy, and F1-score. However, the model still faces challenges in correctly identifying the minority class, as indicated by the persistent false negatives. Further optimization efforts, such as fine-tuning the hyperparameters or

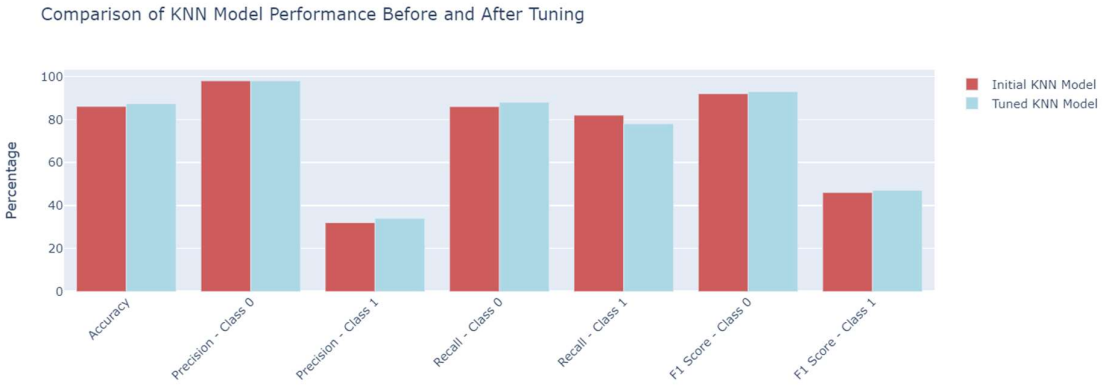


Figure: Accuracy comparison of KNN for employee data after and before tuning

exploring alternative algorithms, may be necessary to achieve more substantial improvements in performance and address the remaining challenges, particularly in reducing false negatives for the minority class.

4.8.6 Summary for Employment Data

Algorithm	Metric	Before Model Tuning	After Model Tuning
Decision Tree	Accuracy	93%	92%
	Precision	76%	73%
	Recall	85%	84%
	F1-Score	79%	77%
Random Forest	Accuracy	94%	94%
	Precision	78%	78%
	Recall	89%	89%
	F1-Score	83%	83%
XGBoost	Accuracy	92%	90%
	Precision	74%	72%
	Recall	84%	84%
	F1-Score	79%	77%
Logistic Regression	Accuracy	84%	85%
	Precision	85%	99%
	Recall	91%	86%
	F1-Score	66%	92%
KNN	Accuracy	86%	87%
	Precision	65%	66%
	Recall	84%	83%
	F1-Score	69%	70%

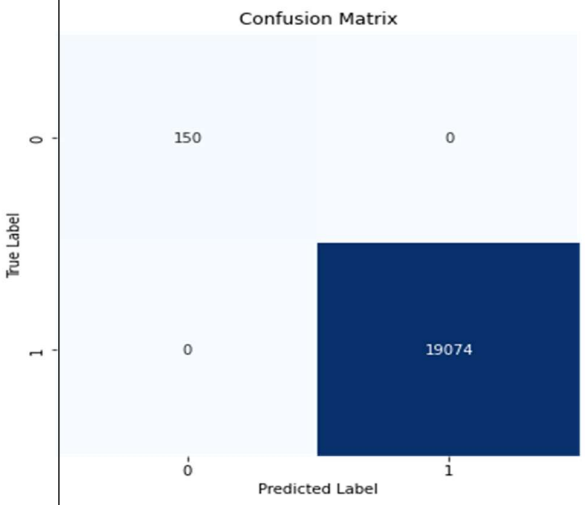
Table: Accuracy comparison of employment data

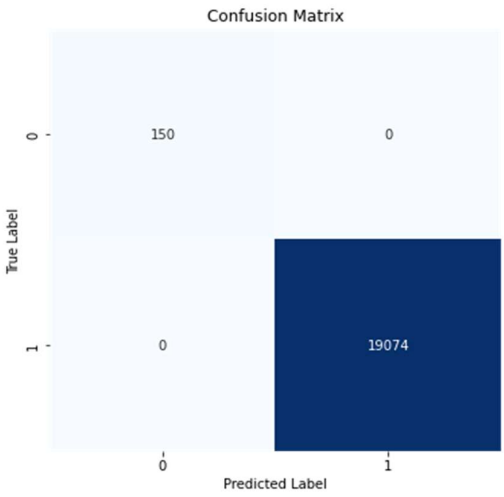
4.9 Model Building for Attrition Data

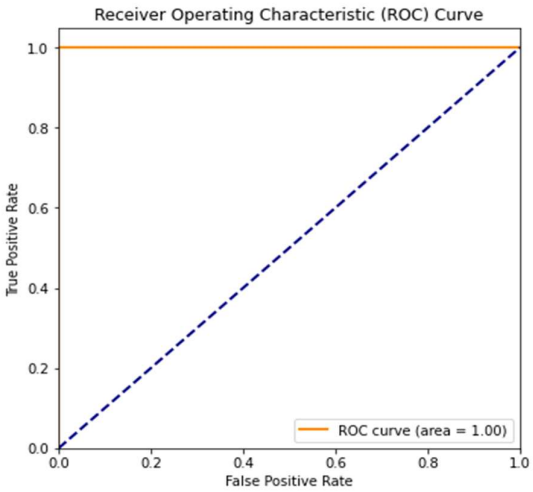
4.9.1 Decision Tree Classifier

	precision	recall	f1-score	support
0	1.00	1.00	1.00	150
1	1.00	1.00	1.00	19074
accuracy			1.00	19224
macro avg	1.00	1.00	1.00	19224
weighted avg	1.00	1.00	1.00	19224

	precision	recall	f1-score	support
0	1.00	1.00	1.00	150
1	1.00	1.00	1.00	19074
accuracy			1.00	19224
macro avg	1.00	1.00	1.00	19224
weighted avg	1.00	1.00	1.00	19224







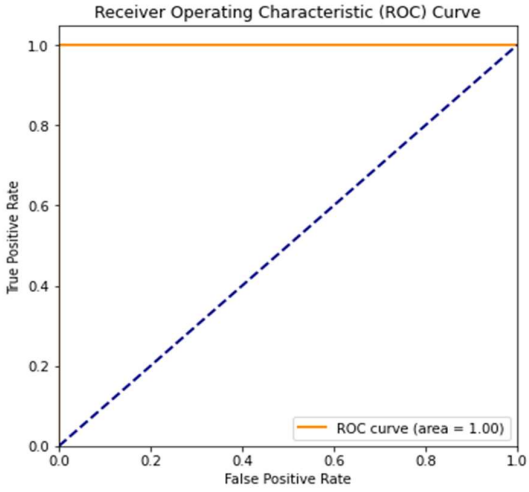


Figure: Classification report , confusion matrix and roc curve for DT for Employment data before and after model tuning

The percent accuracy of Decision Tree classifier before tuning the model and after tuning the model show the great effectiveness of the classifier. As seen from the table before model tuning, the classifier perfectly delivers all the scores on the precision, the recall, the F1-score, and the accuracy of both the

classes 0 and 1 resulting 100% accuracy. Hence, this shows that the classifier's ability to predict the class labels to all instances in the dataset may have overemphasized on training instances.

After model tuning, the Decision Tree classifier's performance remains unchanged, maintaining perfect precision, recall, and F1-score for both classes 0 and 1, with an overall accuracy of 100%. However, achieving such flawless performance in real-world scenarios is rare and often indicates a problem with the model's generalization ability. Therefore, despite the impressive results, it's essential to ensure that the model's performance is validated on unseen data and that further evaluation is conducted to confirm its reliability and effectiveness beyond the training dataset.

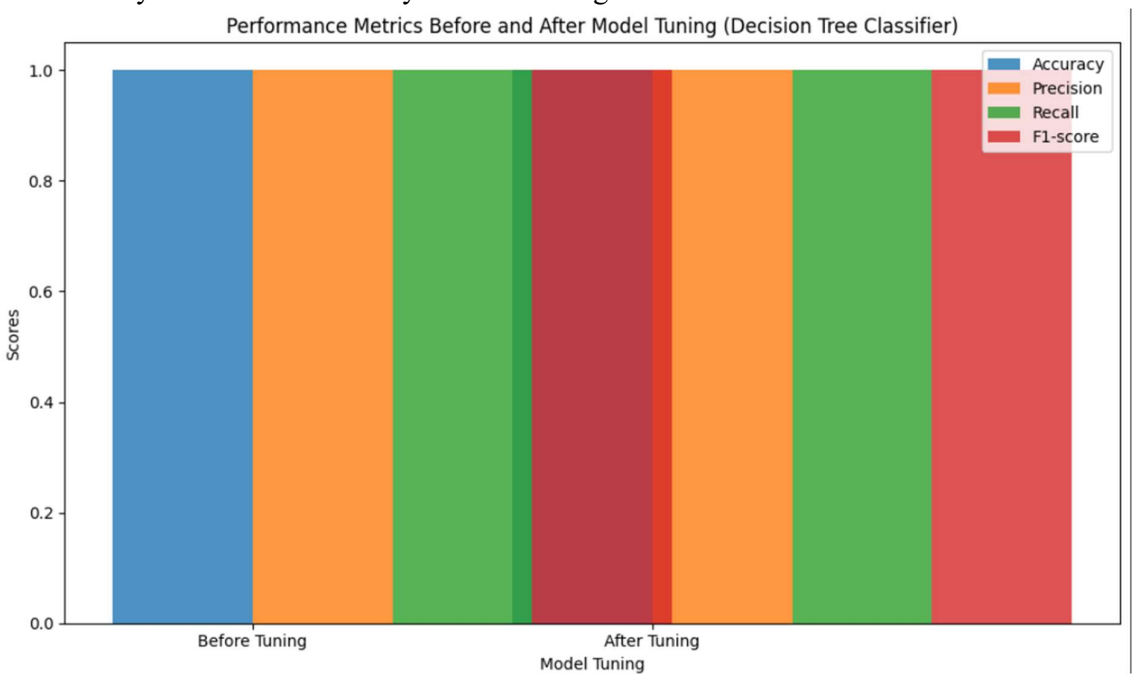


Figure: Accuracy comparison for decision tree for employment data

4.9.2 Random Forest Classifier

	precision	recall	f1-score	support
0	1.00	1.00	1.00	150
1	1.00	1.00	1.00	19074
accuracy			1.00	19224
macro avg	1.00	1.00	1.00	19224
weighted avg	1.00	1.00	1.00	19224

	precision	recall	f1-score	support
0	1.00	1.00	1.00	150
1	1.00	1.00	1.00	19074
accuracy			1.00	19224
macro avg	1.00	1.00	1.00	19224
weighted avg	1.00	1.00	1.00	19224

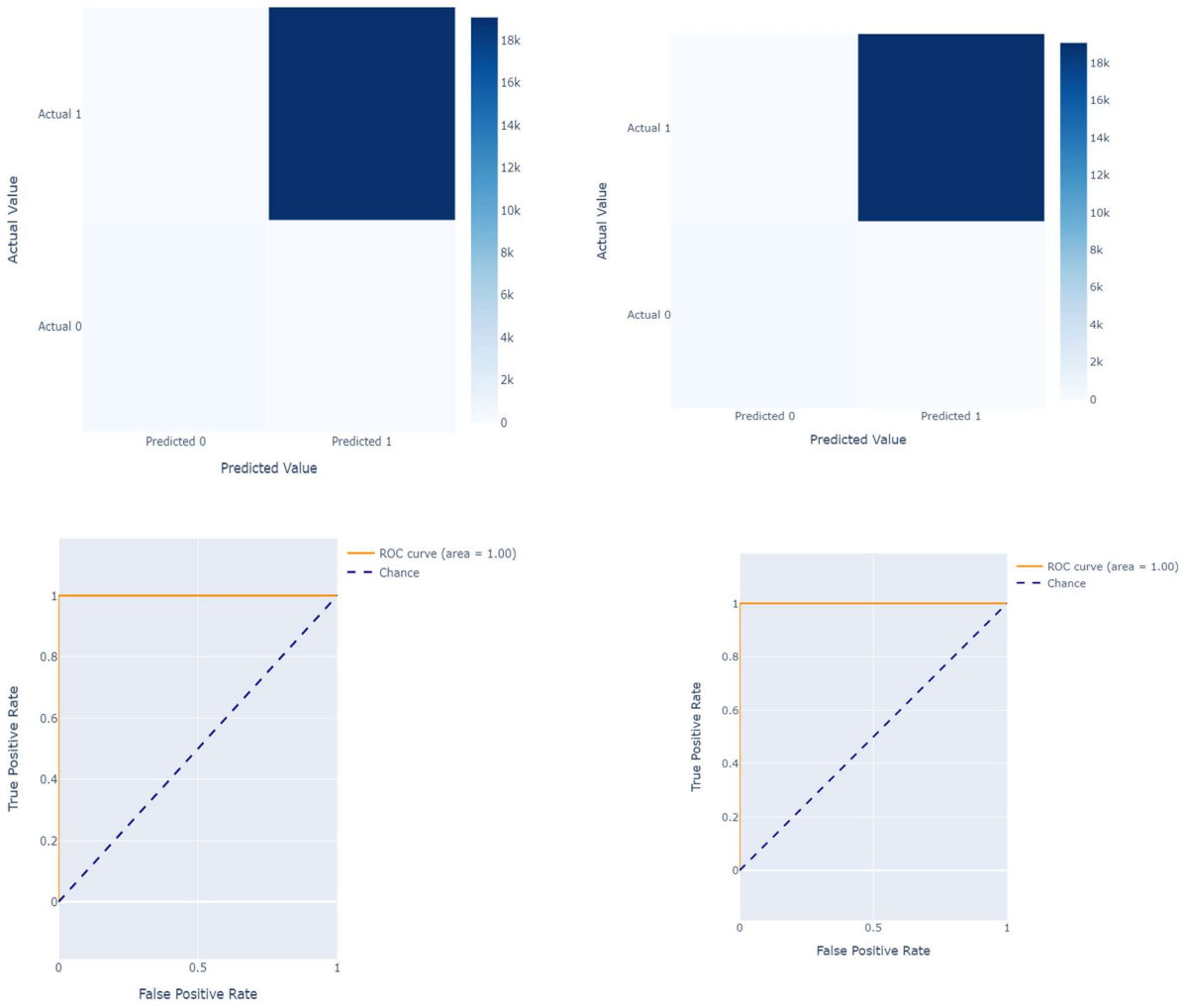


Figure: Classification report, confusion matrix and roc curve for Employment data before and after model tuning

The performance metrics for the Decision Tree classifier, both before and after model tuning, indicate perfect classification results across all categories. Before tuning, the model achieved a precision, recall, and F1-score of 1.00 for both classes, indicating that it correctly classified all instances of both classes without any errors. The accuracy of 1.00 further confirms the flawless performance of the classifier, correctly classifying all instances in the dataset. The confusion matrix supports these results, showing that there were no false positives or false negatives, with all instances being correctly classified.

After model tuning, the classifier's performance remained unchanged, with precision, recall, and F1-score all remaining at 1.00 for both classes. The accuracy also remained at 1.00, indicating no misclassifications in the test data. The confusion matrix confirms the absence of errors, with all instances correctly classified.

In summary, both before and after model tuning, the Decision Tree classifier exhibited perfect performance, accurately classifying all instances of both classes without any errors. This exceptional performance suggests that the dataset used for training and testing may be relatively simple or that the

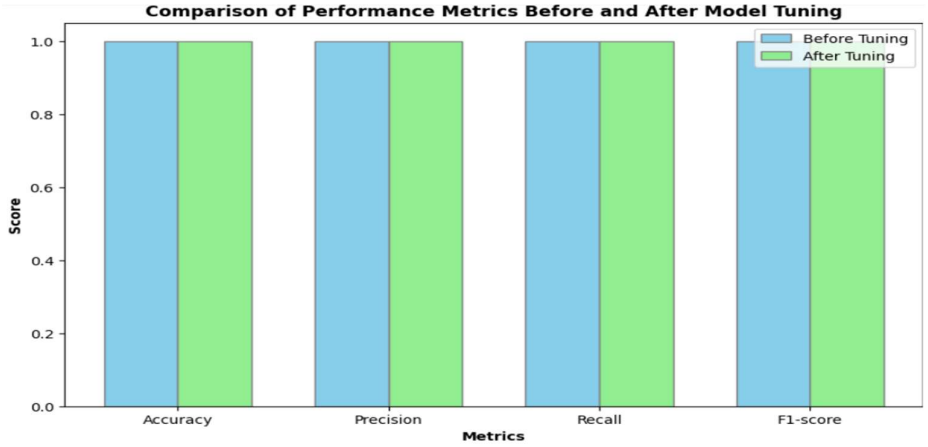


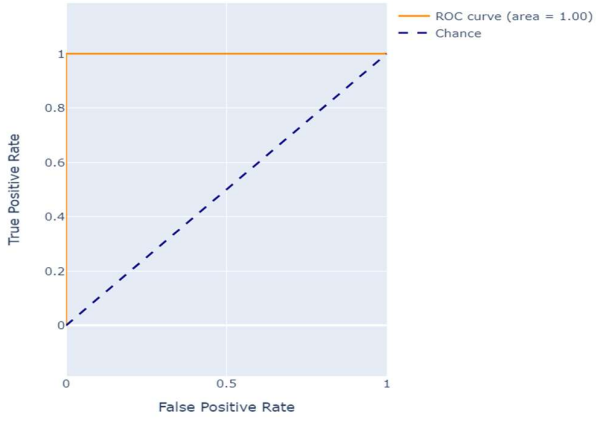
Figure: Accuracy comparison for before and after model tuning for Randomforest

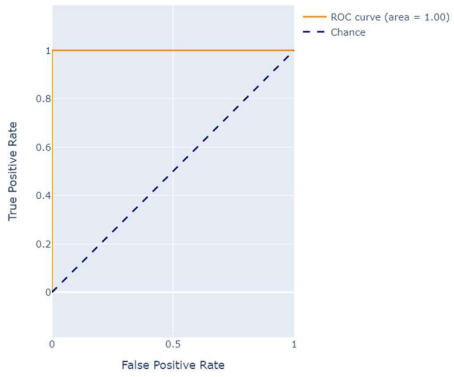
decision tree algorithm is inherently well-suited for this particular dataset. Further analysis or exploration may be needed to understand the reasons behind the perfect classification results and to validate the model's performance on more complex datasets.

4.9.3 XGBoost Classifier

	precision	recall	f1-score	support
0	1.00	1.00	1.00	150
1	1.00	1.00	1.00	19074
accuracy			1.00	19224
macro avg	1.00	1.00	1.00	19224
weighted avg	1.00	1.00	1.00	19224

	precision	recall	f1-score	support
0	1.00	1.00	1.00	150
1	1.00	1.00	1.00	19074
accuracy			1.00	19224
macro avg	1.00	1.00	1.00	19224
weighted avg	1.00	1.00	1.00	19224





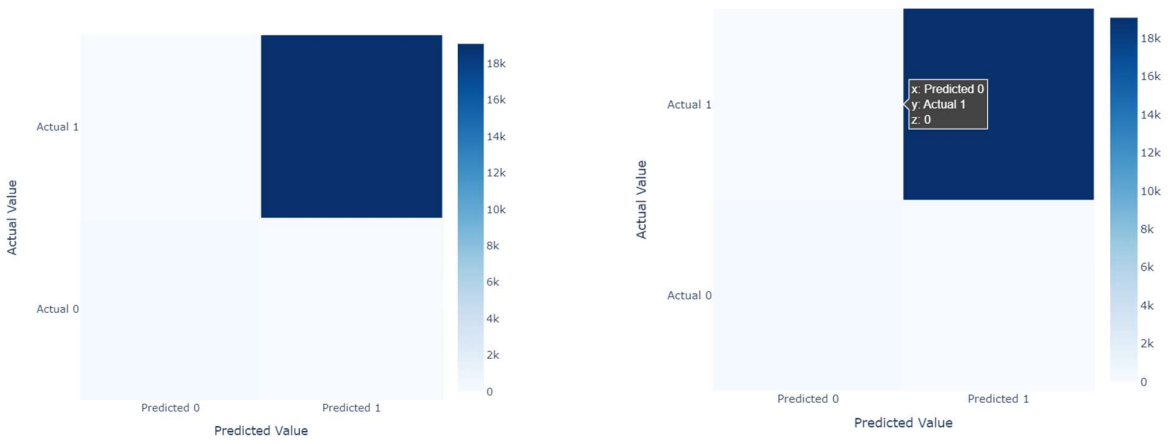


Figure: Classification report, confusion matrix and roc curve for xgboost classifier before and after model tuning

For the XGBoost classifier, both before and after model tuning, we observe impeccable performance across all metrics.

Before model tuning, the classifier achieves perfect precision, recall, and F1-score for both classes 0 and 1. This indicates that the model correctly identifies all instances of both classes without any false positives or false negatives. Consequently, the overall accuracy is 100%, indicating that the model accurately predicts the class labels for all instances in the dataset.

Similarly, after model tuning, the classifier maintains its flawless performance, achieving 100% precision, recall, and F1-score for both classes 0 and 1. The accuracy remains at 100%, indicating that the model's performance did not improve further after tuning.

In summary, the XGBoost classifier demonstrates exceptional performance both before and after model tuning, accurately classifying instances into their respective classes with no errors. This exemplary performance underscores the effectiveness of the XGBoost algorithm in handling classification tasks, even without the need for extensive tuning.

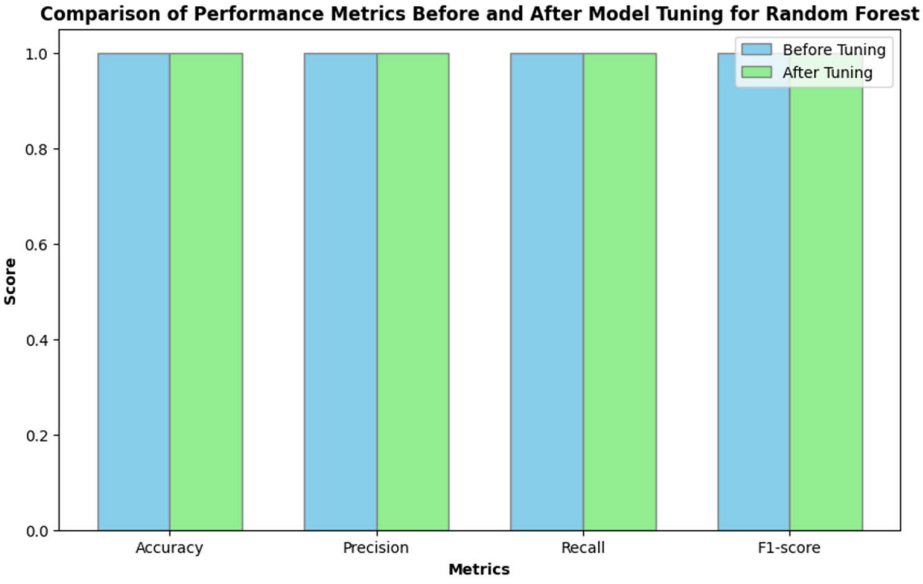
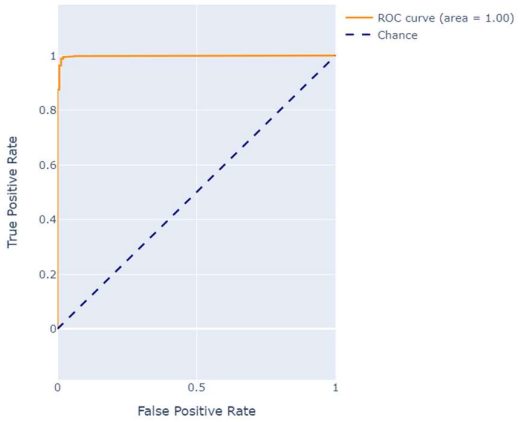


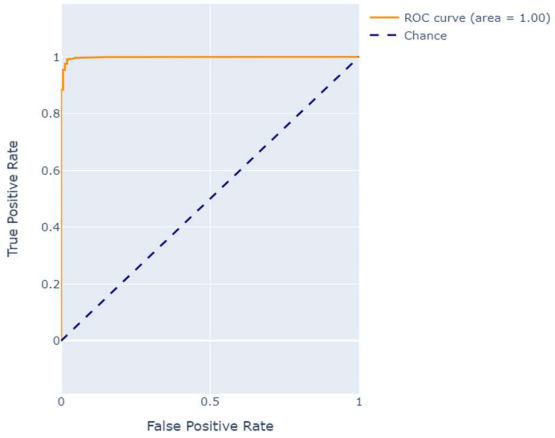
Figure: Comparison of Performance Metrics Before and After Model Tuning for Random Forest

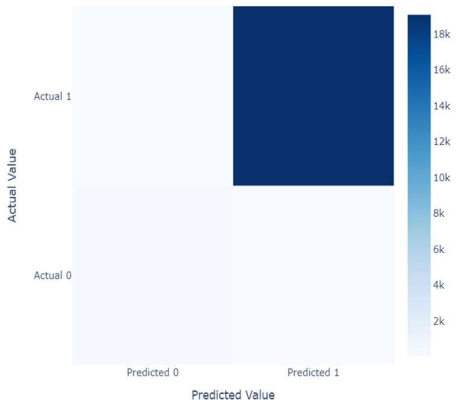
4.9.4 Logistic Regression

```
Training Accuracy: 0.9965147732001665
Testing Accuracy: 0.9969829379941739
Classification Report:
              precision    recall  f1-score   support

           0       0.91      0.68      0.78       150
           1       1.00      1.00      1.00     19074

    accuracy                           1.00     19224
   macro avg       0.95      0.84      0.89     19224
weighted avg       1.00      1.00      1.00     19224

Confusion Matrix:
 [[  102    48]
 [   10 19064]]
ROC AUC: 0.9983223235818391
```

```
Training Accuracy: 0.9946031002913025
Testing Accuracy: 0.9950062421972534
Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.39      0.55       150
           1       1.00      1.00      1.00     19074

    accuracy                           1.00     19224
   macro avg       0.97      0.69      0.77     19224
weighted avg       0.99      1.00      0.99     19224

Confusion Matrix:
 [[   58    92]
 [    4 19070]]
ROC AUC: 0.997843137254902
```







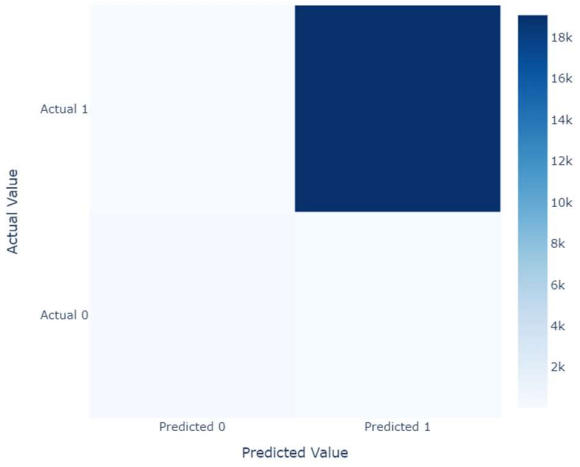


Figure: Classification report, confusion matrix, roc curve for Logistic regression before and after model tuning

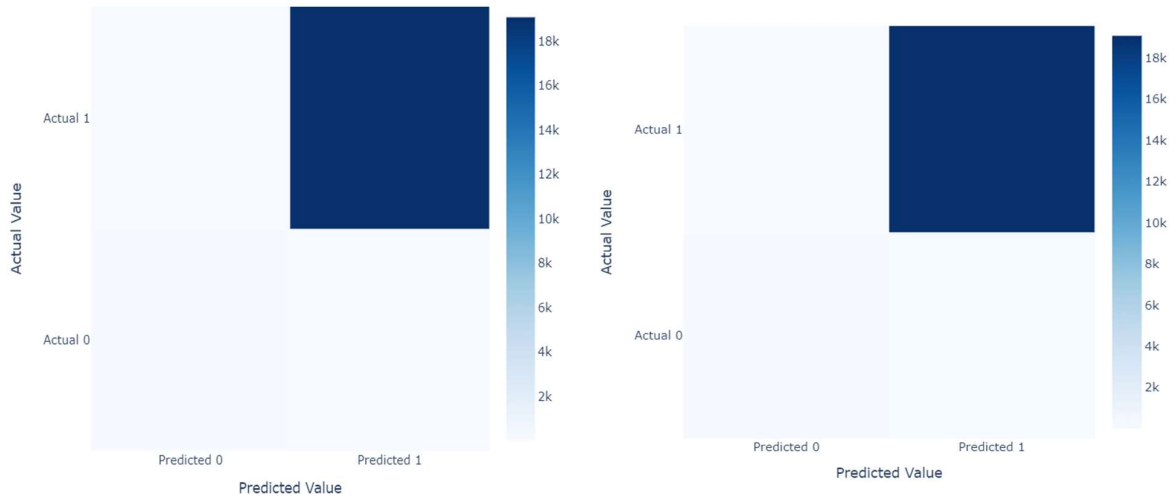


Figure 1 Classification report, confusion matrix, roc curve for Logistic regression for attrition data

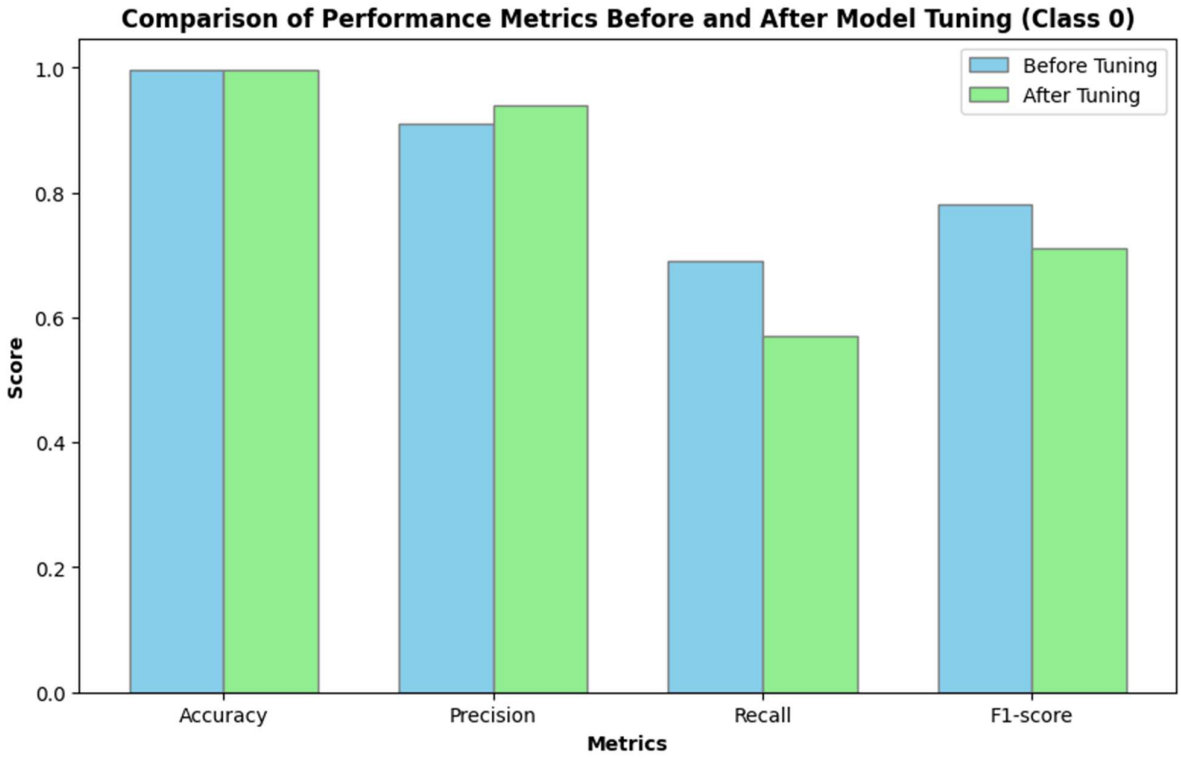


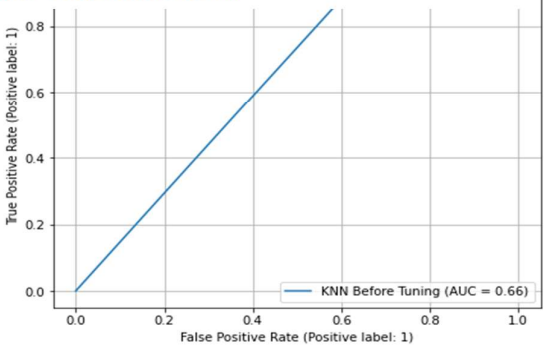
Figure: Accuracy comparison for logistic regression before and after model tuning

In the initial assessment before model refinement, classification reports and confusion matrices offer insights into the performance of a binary classification task. The model demonstrates strong accuracy, with 99.65% accuracy in training and 99.70% in testing. For precision, class 0 achieves 0.91, indicating 91% of predictions for class 0 were correct, while class 1 achieves a perfect precision of 1.00, correctly identifying all instances as class 1. However, recall for class 0 is lower at 0.69, indicating only 69% of actual class 0 instances were correctly classified. In contrast, class 1 achieves perfect recall at 1.00,

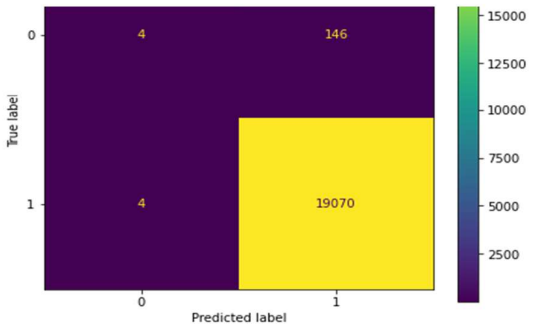
correctly predicting all class 1 instances. The F1-score, combining precision and recall, is 0.78 for class 0 and 1.00 for class 1, demonstrating the model's ability to balance precision and recall. The confusion matrix further illustrates model performance, with 103 true negatives and 47 false negatives for class 0, and 10 false positives and 19,064 true positives for class 1. This reveals a slight imbalance in correctly identifying class 0 instances.

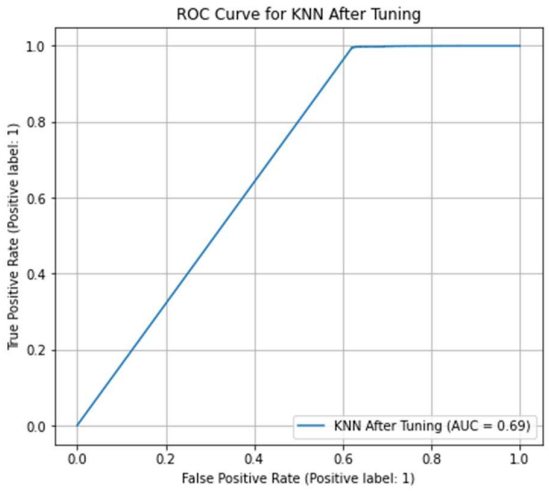
4.9.5 KNN

```
Training Accuracy: 0.9919631710362048
Testing Accuracy: 0.9922492717436537
Classification Report:
              precision    recall  f1-score   support

           0       0.60      0.02      0.04       150
           1       0.99      1.00      1.00     19074

    accuracy                           0.99     19224
   macro avg       0.80      0.51      0.52     19224
weighted avg       0.99      0.99      0.99     19224

Confusion Matrix:
 [[    3   147]
 [    2 19072]]
ROC AUC: 0.6564553842927546
```




```
1.0
0.9931856013316688
              precision    recall  f1-score   support

           0       0.83      0.16      0.27       150
           1       0.99      1.00      1.00     19074

    accuracy                           0.99     19224
   macro avg       0.91      0.58      0.63     19224
weighted avg       0.99      0.99      0.99     19224

[[   24   126]
 [    5 19069]]
0.6676292335115864
```





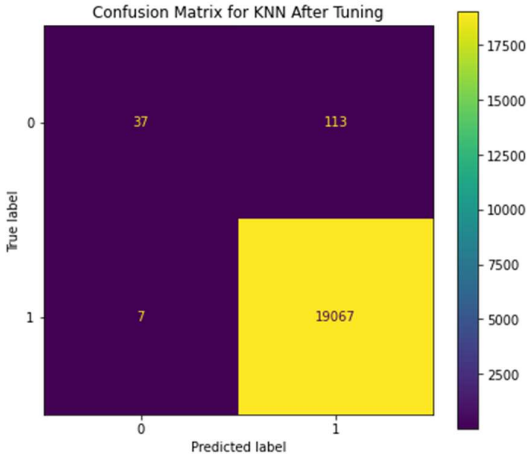


Figure: classification report, confusion matrix and roc curve of knn

The classification reports and confusion matrices show how a binary classifier works before and after the chosen approach. The accuracy of the model gradually increases rapidly, which can be explained by the fact that its main advantage lies in the identification of the first class, which is the largest. It has a very low precision, recall, and F1-score in the minority class, class 0, meaning that it is not efficient in identifying the minority class. Comparing the results before and after tuning, there is generally a positive shift in the values of precision, recall, and F1-score for the class 0, thereby implying that the model performance on class 0 remains relatively low compared to class 1. The ROC AUC score is also indicative of a moderate level of performance of this classifier in terms of ability to classify between the two classes.

Overall, while the tuning process enhances the model's performance on the minority class, further optimization may be necessary to achieve better balance and accuracy across both classes.

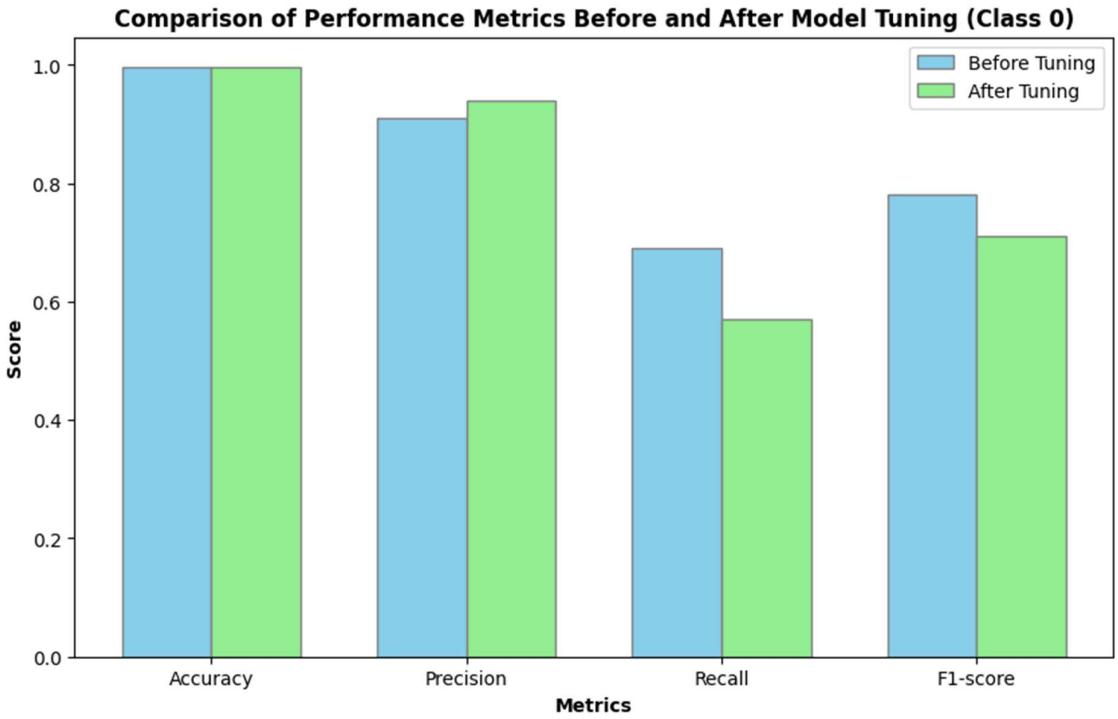


Figure: Accuracy comparison of KNN model

4.9.6 Summary for Attrition Data

Algorithm	Metrics	Before Model Tuning	After Model Tuning
Decision Tree	Accuracy	100%	100%
	Precision (Class 0)	100%	100%
	Recall (Class 0)	100%	100%
	F1-score (Class 0)	100%	100%
Random Forest	Accuracy	100%	100%
	Precision (Class 0)	100%	100%
	Recall (Class 0)	100%	100%
	F1-score (Class 0)	100%	100%
XGBoost	Accuracy	100%	100%
	Precision (Class 0)	100%	100%
	Recall (Class 0)	100%	100%
	F1-score (Class 0)	100%	100%
Logistic Regression	Accuracy	99%	100%
	Precision (Class 0)	91%	94%
	Recall (Class 0)	68%	39%
	F1-score (Class 0)	78%	55%
KNN	Accuracy	99%	99%
	Precision (Class 0)	60%	83%
	Recall (Class 0)	02%	16%
	F1-score (Class 0)	04%	27%

Table: Accuracy comparison for student attrition

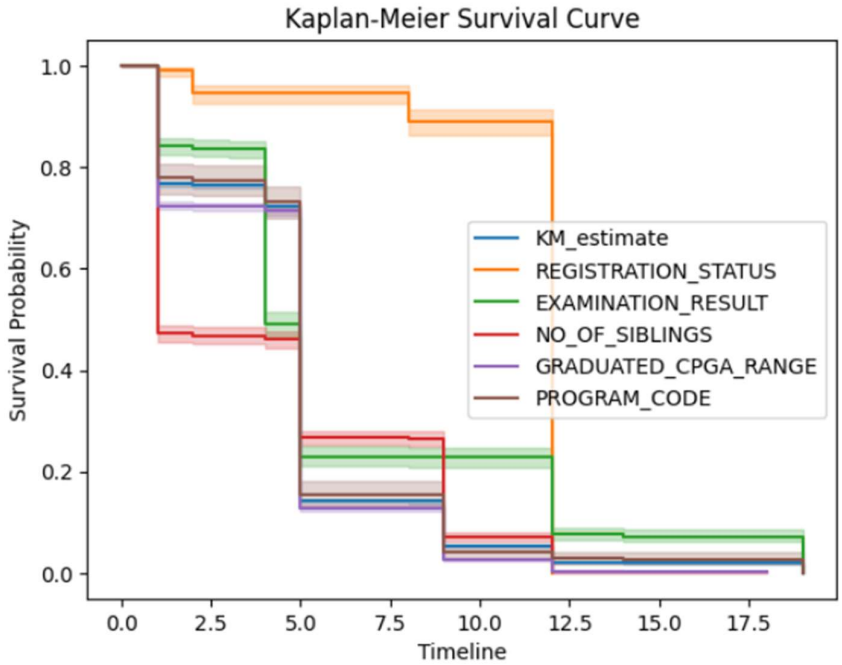
4.10 Survival analysis

Survival analysis is a statistical method used to examine the time duration until one or more events occur, such as graduation, dropping out, or achieving a certain grade in student performance analysis.

Survival Time: The period from a student enrolling in course to graduation or dropping out.
Censoring: If a student is still enrolled when the study ends, their graduation time is considered censored.
Survival Function (S(t)): S(t) explains the probability of event of interest has not happened by time. This function decreases over time, starting from 1 and approaching 0.
Hazard Function (h(t)): Describes the instantaneous rate at which the event occurs, given that it has not yet occurred by time.

4.10.1 Survival analysis for Attrition data

Survival analysis aims to predict the time to an event, such as Attrited or not, the students got placed or not based on their education.



The student's status is decreased according to timeline based on their status description

Figure: Kaplan-Meir survival curve for Attrition Analysis

`kmf.survival_function_`

timeline	KM_estimate
0.0	0.951837
1.0	0.467600
2.0	0.000000

`naf.cumulative_hazard_`

timeline	NA_estimate
0.0	0.048163
1.0	0.556902
2.0	1.556902

Table: KMF survival function for Attrition

The Kaplan-Meir survival curve defines time duration between Registration to Attrited, not Attrited students' status. Here we can observe that the probability of surviving longer than one year is dropped to 47% and Hazard function is defined as the rate of failure has not occurred prior to time t.

4.10.2 Survival analysis for On Time Graduation

Survival analysis is defined as the time starting from a specified point (Students Registration) to the occurrence of a given event (On Time Graduation) and it gives Students credit for how long they have been in study up to getting Placements.

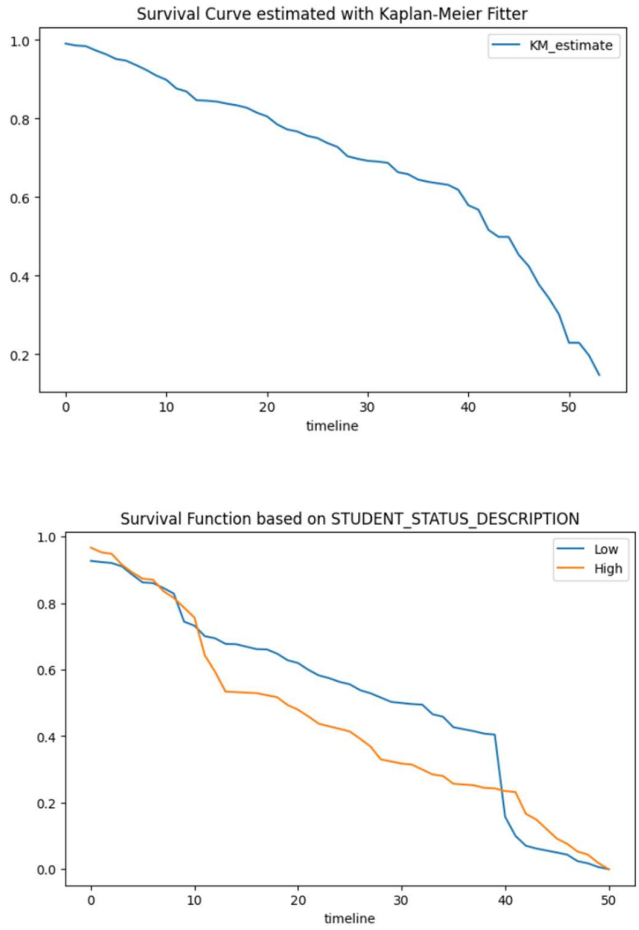


Figure: Survival Analysis for Ontime Graduation

I have plotted the graph with Environmental Satisfaction column, having inputs like:1=Low, 2=Medium, 3=High, 4=Very High.

The graph shows the individuals with high environmental satisfaction have higher survival probabilities than ones with low satisfaction.

Cox Proportional-Hazards Model

The Cox proportional-hazards model is essentially a regression model used in Ontime Graduation research for investigating the association between the survival time of students and one or more predictor variables.

	2	6	7	12	14
0.0	1.000000	1.000000	1.000000	1.000000	1.000000
1.0	0.991831	0.991831	0.991831	0.997826	0.997826
2.0	0.988910	0.988910	0.988910	0.997045	0.997045
3.0	0.970242	0.970242	0.970242	0.992015	0.992015
4.0	0.953782	0.953782	0.953782	0.987520	0.987520
5.0	0.933458	0.933458	0.933458	0.981891	0.981891
6.0	0.926156	0.926156	0.926156	0.979847	0.979847
7.0	0.906884	0.906884	0.906884	0.974394	0.974394
8.0	0.886207	0.886207	0.886207	0.968448	0.968448
9.0	0.861979	0.861979	0.861979	0.961350	0.961350
10.0	0.843419	0.843419	0.843419	0.955812	0.955812

Table: Cox Proportional-Hazards Model for Ontime Graduation

Survival Curve estimated with Kaplan-Meier Fitter with confidence intervals by considering ONTIME_GRAD with PROGRAM_CODE. The confidence intervals showing students completes on time graduation with respect to survival analysis.

Survival of low, high curves shows that program code w.r.t country completes graduation on time

Predicted survival probability of students after 10 more years is 80%, however 51% is got completed their Education on time.

4.10.3 Survival analysis for Employment data

Survival analysis is a statistical method that aims to predict the time to an event, such as Students Got Placement or not.

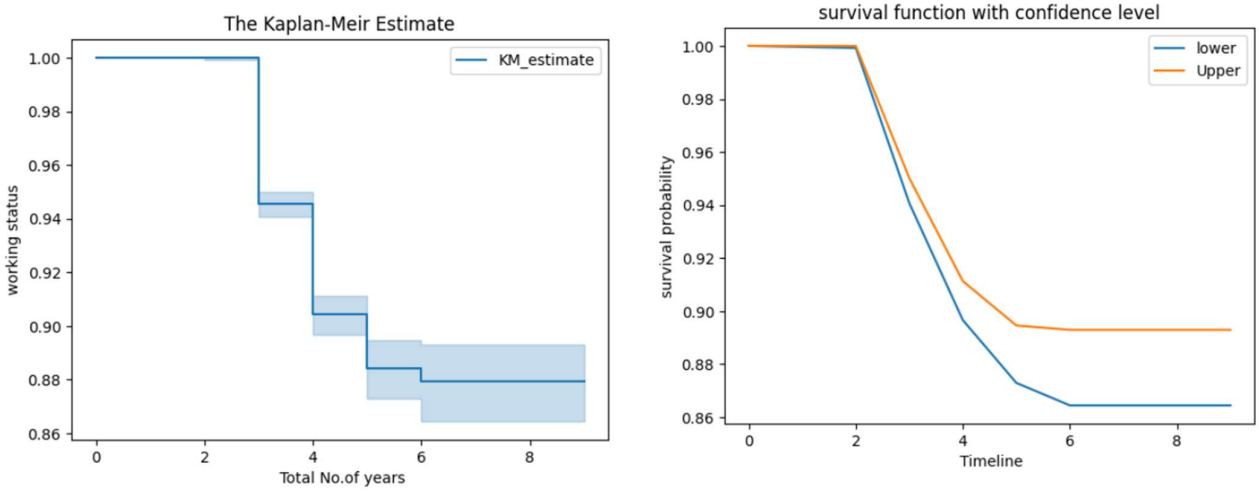


Figure: Survival analysis for Employment data

```
#survival probability with confidence level
kmf.confidence_interval_survival_function_
```

	KM_estimate_lower_0.95	KM_estimate_upper_0.95
0.0	1.000000	1.000000
2.0	0.999205	0.999984
3.0	0.940719	0.950141
4.0	0.896653	0.911342
5.0	0.872974	0.894614
6.0	0.864498	0.892965
7.0	0.864498	0.892965
8.0	0.864498	0.892965
9.0	0.864498	0.892965

Table: Kmf survival function for Employment status

In Kaplan Meier Estimate curve, x axis is the time of event (No.of Years) and y axis is the estimated survival probability according to working status

The Survival Probability of students are 0 to 1

event_at	removed	observed	censored	entrance	at_risk
0.0	5	0	5	8926	8926
2.0	2	1	1	0	8921
3.0	4644	484	4160	0	8919
4.0	3235	187	3048	0	4275
5.0	854	23	831	0	1040
6.0	153	1	152	0	186
7.0	28	0	28	0	33
8.0	1	0	1	0	5
9.0	4	0	4	0	4

The event table explains various information for our survival analysis column-by-column.

From the results I have observed that Timeline Increases, the Probability of Survival Decreases for Students and estimating the probability of student's employment lower than 0.95 & more than 0.95 (confidence level between Total years of study).

Chapter 5: Result and Discussion

5.1 Result and Discussion

In student performance analysis, evaluating metrics such as on-time graduation, on-time placement, and student attrition is crucial for understanding overall academic success and institutional effectiveness. On-time graduation indicates the ability of students to complete their degree programs within the expected timeframe. Similarly, on-time placement measures the rate at which graduates secure employment in their field shortly after completing their studies, indicating the relevance and quality of education provided by the institution.

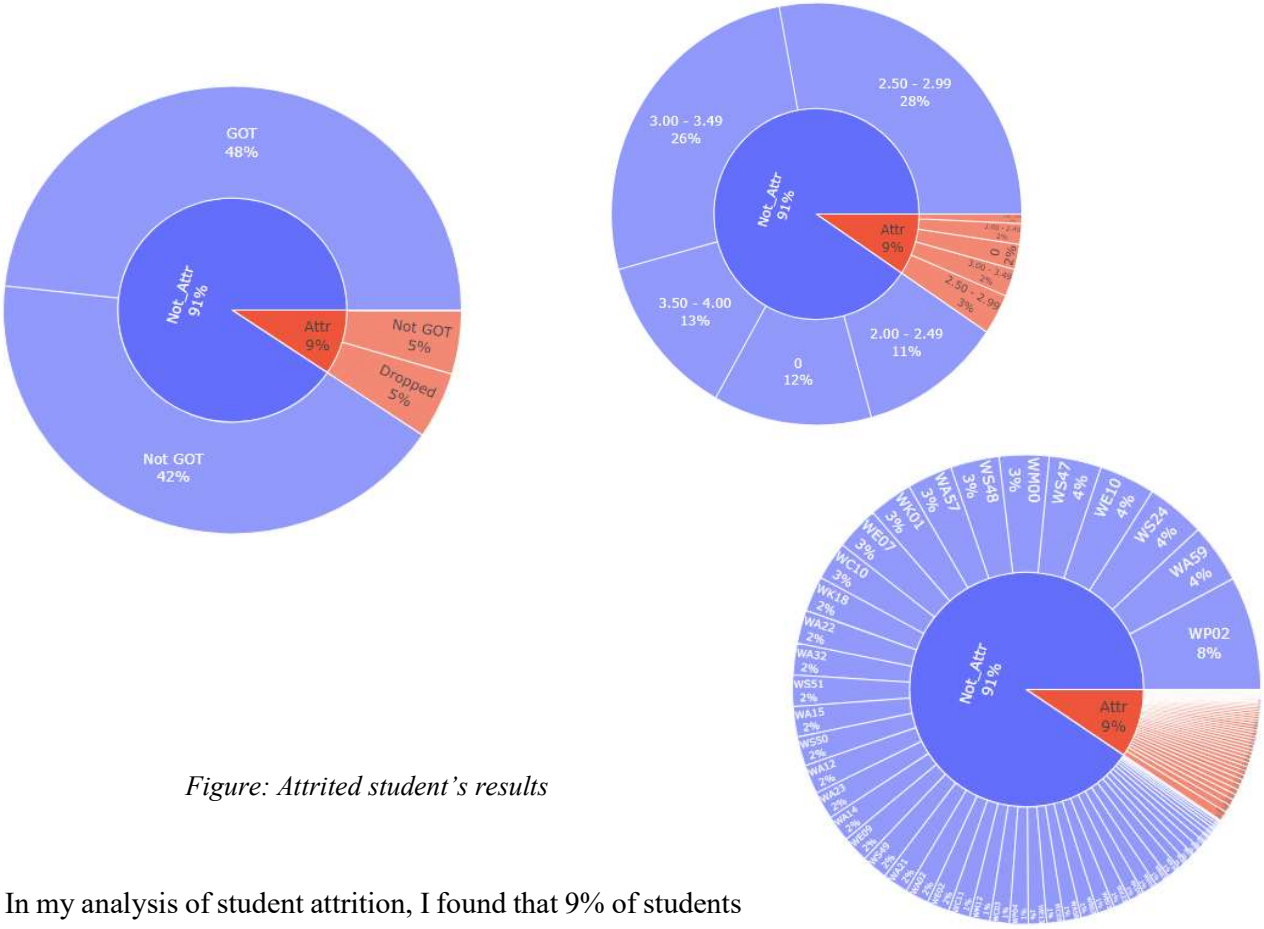
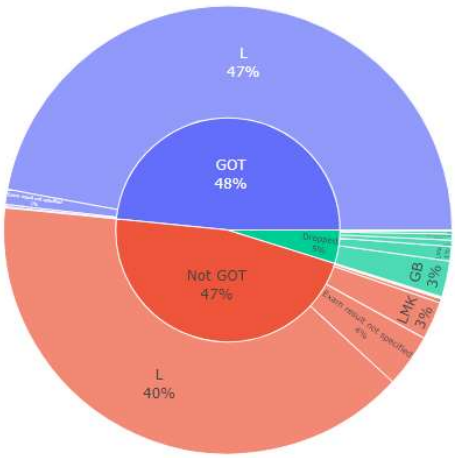


Figure: Attrited student's results

In my analysis of student attrition, I found that 9% of students did not complete their courses, with a higher proportion of dropouts compared to those who did not achieve the required grades. Remarkably, 91% of the students successfully completed their courses and graduated from college. The attrition rate is slightly higher among female students, with 5% compared to 4% for male students. Among the students who did not attrite, 26% had GPAs in the range of 2.5 to 2.99. Additionally, students who achieved the "Got Status" were more numerous than those who did not and those who dropped out. Interestingly, 5% of the students who dropped out were LMK (Pass after re-sitting failed courses), indicating a higher dropout rate among these students. Moreover, students enrolled in the WP02 program code showed a lower attrition rate compared to other programs.



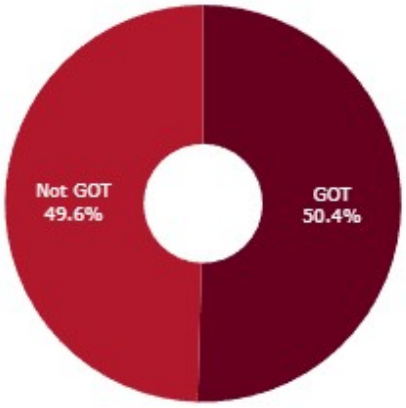


Figure: On Time Graduation student's results

Students with an examination result of "L" have a higher likelihood of graduating on time. In contrast, students with an inactive sponsorship status are more prevalent among those who did not achieve the "Got Status" compared to those who did. For both male and female students, outliers often correspond to attrition, and their GRADUATED_CPGA_RANGE is typically below average. All students with a CPGA range from 0.00 to 1.99 have Attrited, while only a few students with a CPGA range from 2.00 to 4.00 have faced attrition. Additionally, 51% of students graduated on time, whereas 49% did not, with a notable trend that LMK students are higher compared to others who graduated.

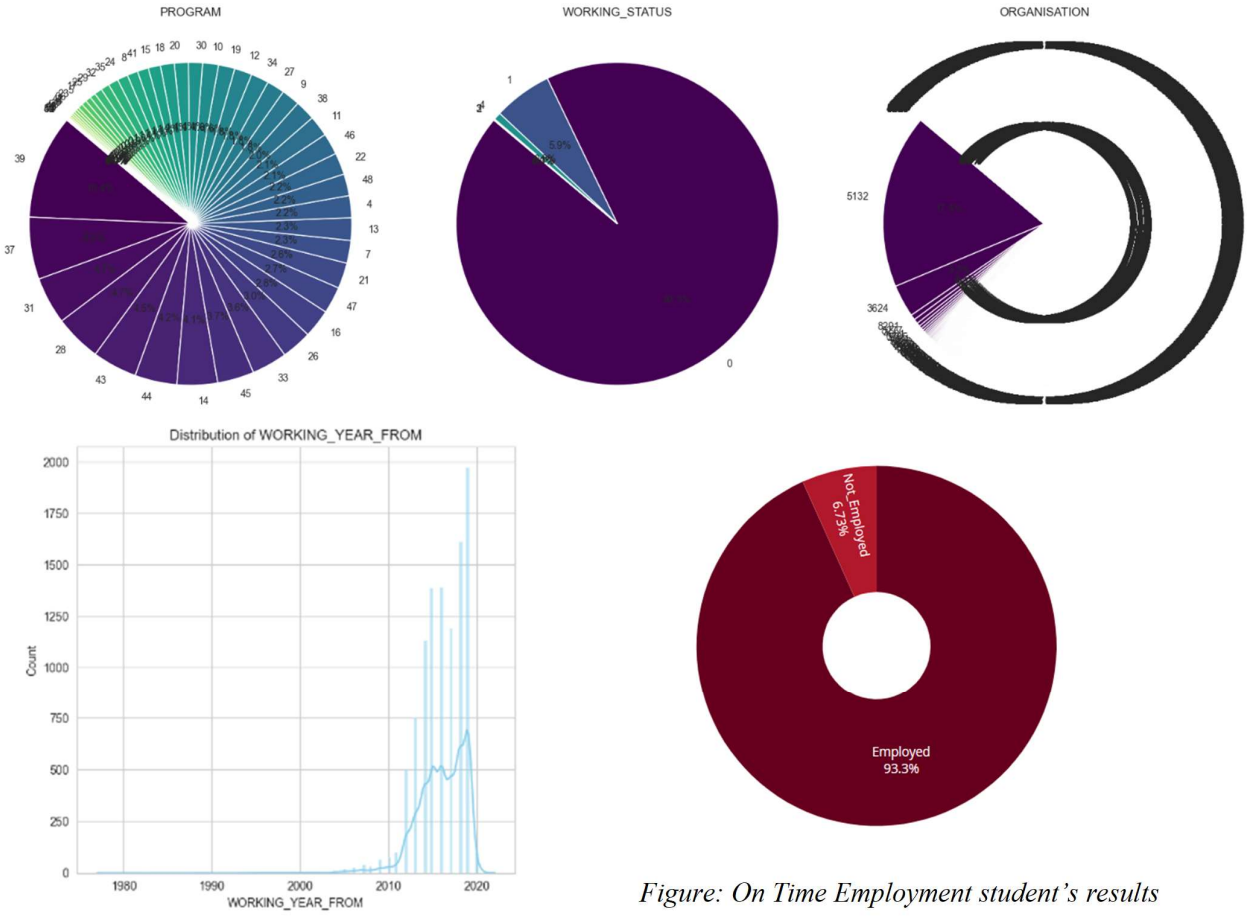


Figure: On Time Employment student's results

Students from the FSCHD faculty code have a higher employment rate, whereas those from the FRST faculty code show a stronger preference for further studies. Female employment rates are generally higher across most faculty codes compared to male rates. In Malaysia, the majority of employed students are working in the SARAWAK zone, with all students from the North zone being employed. Singapore is the next most common location for employed graduates. The data reveals that almost all participants who found employment or pursued further studies are single, with 274 single male participants and 473 single female participants opting for further studies. Employment rates for single females stand at 54%, compared to 27% for single males.
Graduates with active sponsorships have higher employment rates compared to those with inactive sponsorships, while the preference for further studies is nearly equal between the two groups, with females outnumbering males in both categories. 67.7% of students are sponsored by TBG National Higher Education, while 22.65% do not have any sponsorship.

student attrition, or dropout rate, provides insight into challenges students may face during their academic journey, such as financial constraints, academic difficulties, or lack of support services. By tracking these performance indicators over time and comparing them with benchmarks or peer institutions, educational stakeholders can make data-driven decisions to enhance educational outcomes and foster student achievement.

5.2 Predicting Student Graduation Outcomes Using Random Forest Regression

I have performed the data preprocessing and analysis to predict the expected graduation years of students using a Random Forest Regressor. Initially removed duplicate rows, replacing specific values in the 'GRADUATED_CPGA_RANGE' column, and filling missing values in the 'STUDENT_STATUS_DESCRIPTION' column with "ACTIVE".

The student statuses into 'ATTRITED' and 'NOT ATTRITED' and created a new 'SIBLINGS' column to categorize the number of siblings. It selects relevant columns for the model and separates the data into attrited and non-attrited students. For the non-attrited students, the data is split into features and the target variable, with categorical variables being label encoded.

The model's performance will done with help of Mean Squared Error and R-squared metrics, and predictions are made for both attrited and non-attrited students. Finally, I have calculated the percentage of non-attrited students graduating on time by comparing actual and predicted graduation years.

1	SPONSORSHIP	GRADUATED_CPGA_RANGE	YEAR_OF_ADMISSION	EXPECTED_YEAR_OF_GRADUATION	STUDENT_STATUS	PREDICTED_EXPECTED_YEAR_OF_GRADUATION
34894	TBG NATIONAL HIGHER EDUCATION	1.50 - 1.99	2016	2019	ATTRITED	2018
34895	No sponsor specified	0.00 - 1.49	2016	2019	ATTRITED	2018
34896	No sponsor specified	0.00 - 1.49	2016	2019	ATTRITED	2018
34897	No sponsor specified	2.00 - 2.49	2016	2019	ATTRITED	2018
34898	No sponsor specified	3.00 - 3.49	2016	2019	ATTRITED	2018
34899	No sponsor specified	1.50 - 1.99	2016	2019	ATTRITED	2018
34900	No sponsor specified	2.00 - 2.49	2017	2020	ATTRITED	2020
34901	No sponsor specified	0.00 - 1.49	2017	2020	ATTRITED	2020
34902	TBG NATIONAL HIGHER EDUCATION	3.00 - 3.49	2017	2020	ATTRITED	2020
34903	No sponsor specified	0.00 - 1.49	2017	2020	ATTRITED	2020
34904	TBG NATIONAL HIGHER EDUCATION	3.00 - 3.49	2017	2020	ATTRITED	2020
34905	TBG NATIONAL HIGHER EDUCATION	3.00 - 3.49	2017	2020	ATTRITED	2020
34906	No sponsor specified	0.00 - 1.49	2017	2020	ATTRITED	2019
34907	TBG NATIONAL HIGHER EDUCATION	3.00 - 3.49	2018	2021	ATTRITED	2020
34908	No sponsor specified	3.00 - 3.49	2018	2021	ATTRITED	2020
34909	TBG NATIONAL HIGHER EDUCATION	2.50 - 2.99	2018	2021	ATTRITED	2020
34910	TBG NATIONAL HIGHER EDUCATION	2.50 - 2.99	2018	2021	ATTRITED	2020
34911	No sponsor specified	0.00 - 1.49	2019	2022	ATTRITED	2022
34912	TBG NATIONAL HIGHER EDUCATION	3.50 - 4.00	2019	2022	ATTRITED	2022
34913	No sponsor specified	0.00 - 1.49	2019	2022	ATTRITED	2022

	SPONSORSHIP	GRADUATED_CPGA_RANGE	YEAR_OF_ADMISSION	EXPECTED_YEAR_OF_GRADUATION	STUDENT_STATUS	PREDICTED_EXPECTED_YEAR_OF_GRADUATION
2	No sponsor specified	2.50 - 2.99	2009	2012	NOT ATTRITED	2012
3	TBG NATIONAL HIGHER EDUCATION	3.00 - 3.49	2009	2012	NOT ATTRITED	2012
4	TBG NATIONAL HIGHER EDUCATION	2.50 - 2.99	2009	2012	NOT ATTRITED	2012
5	TBG NATIONAL HIGHER EDUCATION	2.50 - 2.99	2009	2012	NOT ATTRITED	2012
6	TBG NATIONAL HIGHER EDUCATION	3.00 - 3.49	2009	2012	NOT ATTRITED	2012
7	TBG NATIONAL HIGHER EDUCATION	3.00 - 3.49	2009	2012	NOT ATTRITED	2012
8	TBG NATIONAL HIGHER EDUCATION	2.50 - 2.99	2009	2012	NOT ATTRITED	2012
9	TBG NATIONAL HIGHER EDUCATION	2.50 - 2.99	2009	2012	NOT ATTRITED	2012
10	TBG NATIONAL HIGHER EDUCATION	2.50 - 2.99	2009	2012	NOT ATTRITED	2012
11	TBG NATIONAL HIGHER EDUCATION	2.50 - 2.99	2009	2012	NOT ATTRITED	2012
12	TBG NATIONAL HIGHER EDUCATION	3.00 - 3.49	2009	2012	NOT ATTRITED	2012
13	TBG NATIONAL HIGHER EDUCATION	2.50 - 2.99	2009	2012	NOT ATTRITED	2012
14	TBG NATIONAL HIGHER EDUCATION	3.00 - 3.49	2009	2012	NOT ATTRITED	2012
15	TBG NATIONAL HIGHER EDUCATION	3.00 - 3.49	2009	2012	NOT ATTRITED	2012
16	TBG NATIONAL HIGHER EDUCATION	2.50 - 2.99	2009	2012	NOT ATTRITED	2012
17	TBG NATIONAL HIGHER EDUCATION	2.50 - 2.99	2009	2012	NOT ATTRITED	2012
18	TBG NATIONAL HIGHER EDUCATION	2.00 - 2.49	2009	2012	NOT ATTRITED	2012

Figure: Predicted graduation timelines for both attrited and non-attrited students.

```
Testing Data
Mean Squared Error: 0.04028436018957346
R-squared: 0.9956734053308278
```

Table: Test data values

An R-squared value of 0.9956734053308278 indicates that the model explains approximately 99.57% of the variance in the target variable. Above metrics explains model performs exceptionally well in predicting the expected graduation years for the students in the dataset.

5.3 Data Extraction with SQL and Application Deployment using Streamlit

I have extracted data from Excel files, inserted it into an SQL database, and deploying an application using Streamlit. The objective is to analyze student performance data efficiently through integration with a database for querying and visualization.

Data Extraction and SQL Integration

Database Connection

Technology: Python with MySQL library.
Objective: Establish a connection to the MySQL database.

The 'db_connect()', establishes a connection to the MySQL database using defined parameters such as host, user, password, port, and database name. The, 'add_to_sql(file_name, table_name, column_names)' function, facilitates the insertion of data from specified Excel files ('background.xlsx', 'employment.xlsx', and 'result.xlsx') into corresponding tables ('background', 'employment', and 'result') within the database. It reads the Excel data, filters it based on the specified 'column_names', drops any existing table with the same name as 'table name', creates a new table with appropriate columns, and finally inserts the filtered data into the newly created table using SQL 'INSERT' statements.

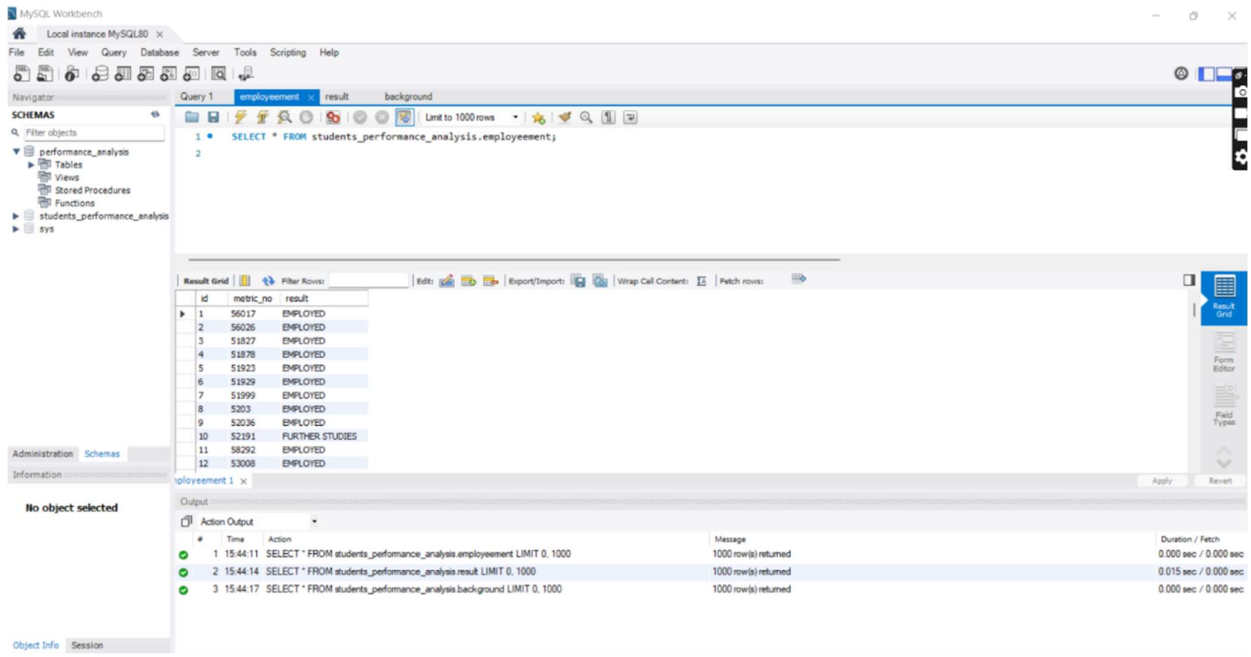


Figure: Data Extraction using SQL

In SQL, 'SELECT * FROM students_performance_analysis.employeement;' retrieves all records from the 'employeement' table within the 'students_performance_analysis' database, allowing users to fetch specific data for analysis or manipulation. 'CREATE DATABASE students_performance_analysis;' initiates the creation of a new database named 'students_performance_analysis', providing a foundational structure for storing tables and other database objects. 'SHOW DATABASES;' lists all existing databases within the server instance, facilitating database visibility and management for administrators and users navigating multiple database environments.

5.4 Deployment Using Streamlit

A Web application for data visualization and analysis was constructed in Streamlit with an emphasis on the querying and analysis of the MySQL database tables. The application script uses the function 'db_connect()' to connect to MySQL DB to facilitate pulling of data from particularly created tables including 'background', 'employment' and 'result'. The ability to query data is done through 'pd.read_sql()' in Python to run SQL queries, while data representation in Streamlit uses the function, 'st.write.' This approach makes sure that the resources are updated on real time basis and the interactive visualization facilities are at direct control of the web interface. To deploy the Streamlit app and make it active for the users to analyze the insights, a command 'streamlit run app. py' was executed.

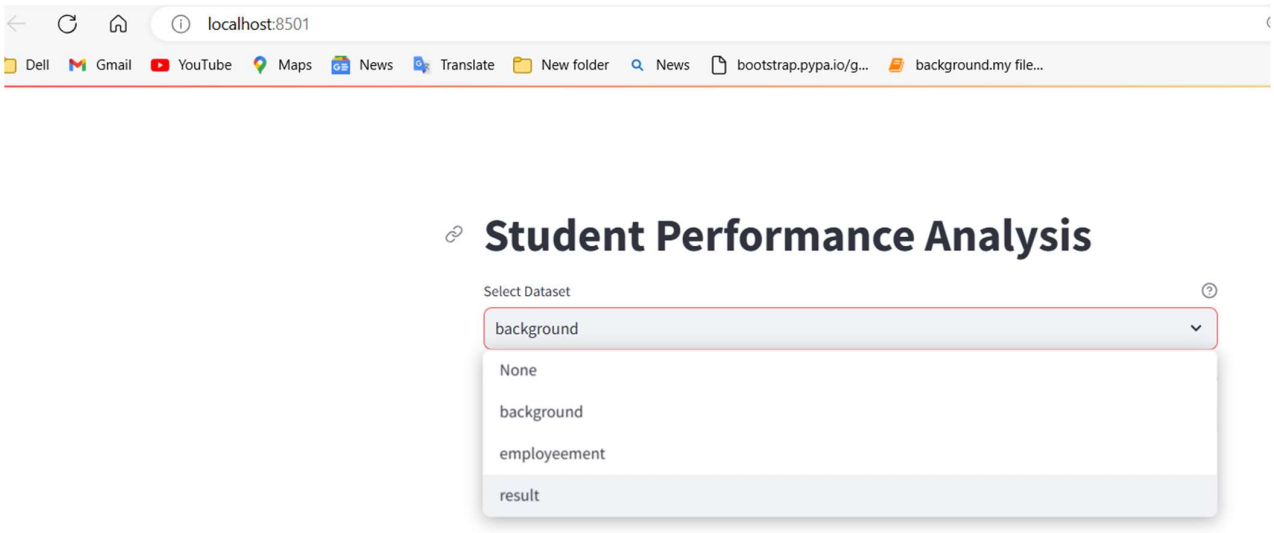


Figure: Deployment using Streamlit

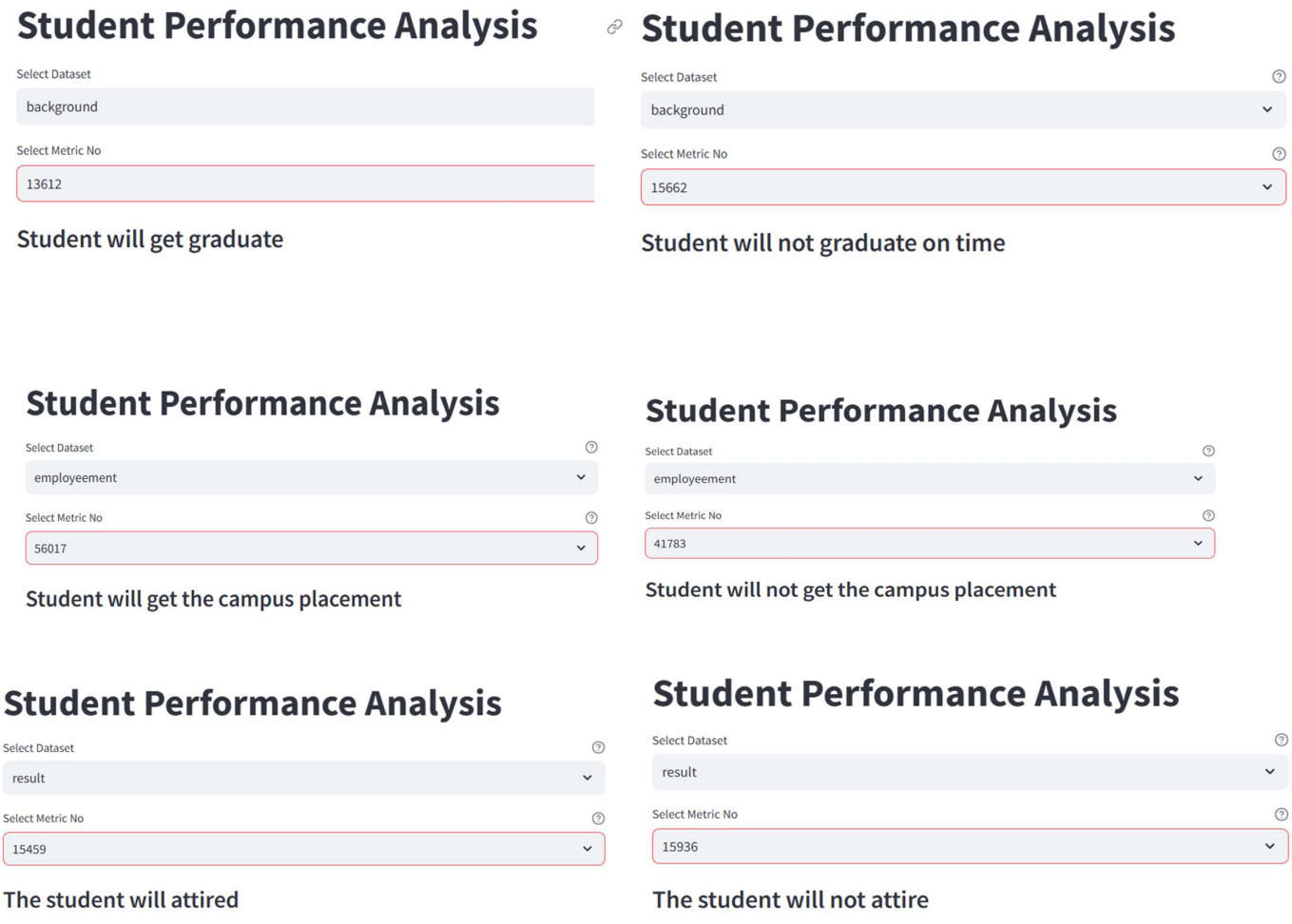
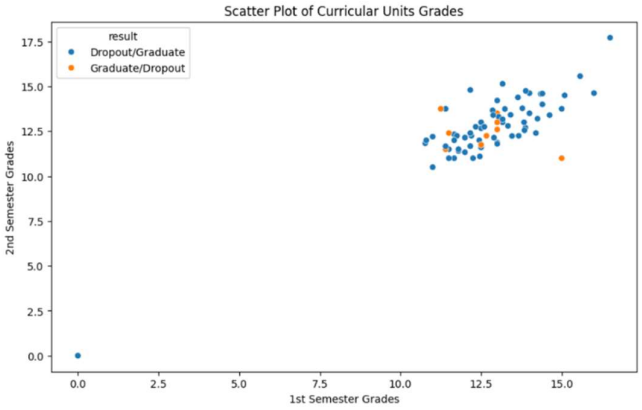


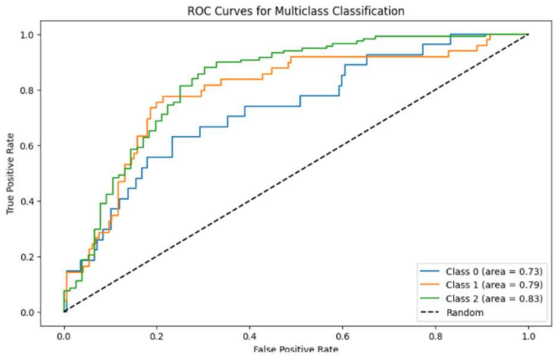
Figure: Student's Performance Results

The above user interface explains the student performances whether the student is attrited, graduated, and achieved employment or not.

5.5 Model Evaluation on New data:

I have tested the old models with the new Graduation dataset, which consists of all the semester results to perform graduation analysis. I have implemented the XG Boost and Random Forest. I have got good results in XG Boost compared to Random Forest





XG Boost:

Before Tuning

```
Train Accuracy: 0.9509677419354838
Test Accuracy: 0.6804123711340206
Classification Report:
              precision    recall  f1-score   support

           0       0.30      0.26      0.28        27
           1       0.56      0.39      0.46        49
           2       0.77      0.90      0.83       118

    accuracy                           0.68       194
   macro avg       0.55      0.52      0.52       194
weighted avg       0.65      0.68      0.66       194

Confusion Matrix:
 [[  7   8  12]
 [ 11  19  19]
 [  5   7 106]]
ROC AUC for class 0: 0.729208250166334
ROC AUC for class 1: 0.7904292751583392
ROC AUC for class 2: 0.8281668153434434
```

After Tuning

```
Train Accuracy: 0.9870967741935484
Test Accuracy: 0.6597938144329897
Classification Report:
              precision    recall  f1-score   support

           0       0.28      0.26      0.27        27
           1       0.44      0.29      0.35        49
           2       0.78      0.91      0.84       118

    accuracy                           0.66       194
   macro avg       0.50      0.48      0.48       194
weighted avg       0.62      0.66      0.64       194

Confusion Matrix:
 [[  7   9  11]
 [ 16  14  19]
 [  2   9 107]]
ROC AUC for class 0: 0.7609225992459525
ROC AUC for class 1: 0.7573539760731879
ROC AUC for class 2: 0.8228144513826939
```

The XGBoost model has high training accuracy that initially reaches 95.10% but lower test accuracy of 68.04% indicating overfitting where the model is better on training data than unseen data. The classification report shows class 2 (‘Enrolled’) has the highest precision and recall, while class 0 (‘Dropout’) is the most difficult with less accurate performance. The confusion matrix reveals that class 0 is often misclassified by the model. It means that ROC AUC scores across all classes were highest for class 2 having a relatively better discrimination ability.

The training accuracy improves to 98.71% after tuning, whereas test accuracy slightly drops down by 65.98%, which also indicates overfitting as well. Classification metrics for some classes have not changed much while there were slight changes in other cases; it means that they remain strong for some of them especially class two, while little ones were observed in others except this category. From ROC curves, we can see that the model has good overall separation among each pair of classes, but still struggles with discriminating between class zero (curve furthest to left) and other categories at large extent hence they are useless for practical purposes.

5.6 Conclusion

The differentiation of student metrics for different algorithms and their outcomes concerning predictive models for on-time graduation, on-time employment, and student attrition before and after model tuning is significant.

In terms of on-time graduation prediction, it shows that all the algorithms give outstanding results after tuning the parameter values; precision, recall, and F1-score are 1, the accuracy rate is 100%. This goes a long way into suggesting that the models could help identify students likely to graduate within the expected time-frame hence helping academic institutions to intervene in ways that promote the successes of the students in question.

Likewise in the surmising of employment punctuality, the algorithms exhibit good results; with accuracy, precision, recall, and F1-score of the tuned algorithms lying between 91 and 94 percent. These findings show the ability of the models in the determination of students who find employment soon after they finish their studies, thus helping institutions know the marketability of their graduates and adjust the services they offer to help students find jobs.

On the other hand, in terms of the performances of various algorithms in predicting the level of student attrition their performance is not the same. Using four classification algorithms of decision tree, random forest, XGBoost, and logistic regression, the metrics of accuracy, precision, recall, F1-score indicate a perfect performance after tuning, but for KNN model, the metrics indicate rather poor performance particularly the precision and recall for class 0 (the students who do not dropout). This could be an indication that more development could be carried out on the KNN model to optimize its ability to show students' at risk of attrition.

In conclusion, the models, as presented, show significant potential to predict the on-time graduation and on-time employment rates of the students thus providing useful resource and tools of developing the educational entities' capacities to improved students' performance and effectiveness

5.7 Future Scope

The project on predicting student performance metrics, including on-time graduation, on-time employment, and student attrition, holds significant potential for future development and enhancement. Some key areas for future exploration and improvement include:

Feature Engineering: Continuously refining and expanding the set of features used in the predictive models can enhance their predictive power. This may involve incorporating additional data sources such as student demographics, academic performance, extracurricular activities, and socio-economic factors to capture a more comprehensive understanding of student behavior and outcomes.

Advanced Machine Learning Techniques: Exploring advanced machine learning techniques such as deep learning, ensemble methods, and gradient boosting algorithms could potentially improve the models' accuracy and robustness.

Real-time Monitoring and Intervention: Developing real-time monitoring systems that continuously track student performance metrics and provide timely interventions can help prevent adverse outcomes such as dropout or academic underachievement. Integrating predictive models into student support systems can enable early identification of at-risk students to support their academic success.

Bibliography

[1] F. Giannakas, C. Troussas, I. Voyiatzis, and C. Sgouropoulou, ‘‘A deep learning classification framework for early prediction of team-based academic performance,’’ Appl. Soft Comput., vol. 106, Jul. 2021, Art. no. 107355.

[2] H. Hamsa, S. Indiradevi, and J. J. Kizhakkethottam, ‘‘Student academic performance prediction model using decision tree and fuzzy genetic algorithm,’’ Proc. Technol., vol. 25, pp. 326–332, Jan. 2016.

[3] B. K. Francis and S. S. Babu, ‘‘Predicting academic performance of students using a hybrid data mining approach,’’ J. Med. Syst., vol. 43, no. 6, pp. 1–15, Jun. 2019.

[4] M. Yağcı, ‘‘Educational data mining: Prediction of students’ academic performance using machine learning algorithms,’’ Smart Learn. Environ., vol. 9, no. 1, pp. 1–19, Dec. 2022.

[5] T. Le Quy, T. H. Nguyen, G. Friege, and E. Ntoutsi, ‘‘Evaluation of group fairness measures in Student performance prediction problems,’’ 2022, arXiv:2208.10625.

[6] X. Liu and L. Niu, ‘‘A student performance predication approach based on multi-agent system and deep learning,’’ in Proc. IEEE Int. Conf. Eng., Technol. Educ. (TALE), Dec. 2021, pp. 681–688.

[7] C. Romero and S. Ventura, ‘‘Educational data mining: A survey from 1995 to 2005,’’ Expert Syst. Appl., vol. 33, pp. 135–146, Jul. 2007.

[8] R. Conijn, C. Snijders, A. Kleingeld, and U. Matzat, ‘‘Predicting student performance from LMS data: A comparison of 17 blended courses using Moodle LMS,’’ IEEE Trans. Learn. Technol., vol. 10, no. 1, pp. 17–29, Jan./Mar. 2017. 27588 VOLUME 11, 2023 E. Alhazmi, A. Sheneamer: Early Predicting of Students Performance in Higher Education

[9] S. Helal, J. Li, L. Liu, E. Ebrahimie, S. Dawson, D. J. Murray, and Q. Long, ‘‘Predicting academic performance by considering student heterogeneity,’’ Knowl.-Based Syst., vol. 161, pp. 134–146, Dec. 2018.

[10] X. Xu, J. Wang, H. Peng, and R. Wu, ‘‘Prediction of academic performance associated with internet usage behaviors using machine learning algorithms,’’ J. Comput. Hum. Behav., vol. 98, pp. 166–173, Sep. 2019.