Name: Drishti Harshit Doshi

Student ID: 2203898

Module Code: CS5811

Module Title: Distributed Data Analysis

Academic Year: 2022-2023

Section1: Data Description and Research Question

Dataset:

When students drop out of school or fail to complete their education at higher standards, it becomes a barrier to their economic growth, employment, competitiveness, and productivity. This directly impacts students, families, higher education institutions, and society. The dataset describes demographic, socioeconomic, macroeconomic, and academic variables that will help analyze predictors of student dropout and academic success. The data was collected at the end of the first and second semesters. This project will provide valuable insights to the tutoring team, researchers, and institutions for formulating interventions that will help in student retention.

The dataset: Predict students' dropout and academic success has been collected from Kaggle**.** https://www.kaggle.com/datasets/thedevastator/higher-education-predictors-of-student-retention

Metadata:

This dataset contains data from a higher education institution on various features related to undergraduate students, including demographics, social-economic factors, and academic performance, to investigate the impact of these factors on student dropout and academic success. The dataset has a total of 35 features and 4425 rows.

Demographic Data:

Marital Status	The marital status of the student.	Categorical
Nationality	The nationality of the student.	Categorical
Displaced	Whether the student is a displaced person.	Categorical
Gender	The gender of the student.	Categorical
Age at enrollment	The age of the student at the time of enrollment.	Numerical
International	Whether the student is an international student.	Categorical

Socio-Economic Data:

Mother's qualification	The qualification of the student's mother	Categorical
Father's qualification	The qualification of the student's father	Categorical
Mother's occupation	The occupation of the student's mother.	Categorical
Father's occupation	The occupation of the student's father.	Categorical
Educational special needs	Whether the student has any special educational needs.	Categorical
Debtor	Whether the student is a debtor.	Categorical
Tuition fees up to date	Whether the student's tuition fees are up to date.	Categorical
Scholarship holder	Whether the student is a scholarship holder.	Categorical

Macro-Economic Data:

Unemployment rate	The number of unemployed people in the population.	Numerical
Inflation rate	The rate of increase in prices over a given period.	Numerical
GDP	Measures the total value of all the goods made, and services provided, during a specific period of time.	Numerical

Academic Data at Enrollment:

Application mode	The method of application used by the student.	Categorical
Application order	The order in which the student applied.	Numerical
Course	The course taken by the student.	Categorical
Daytime/evening attendance	Whether the student attends classes during the day or in the evening.	Categorical
Previous qualification	The qualification obtained by the student before enrolling in higher education.	Categorical

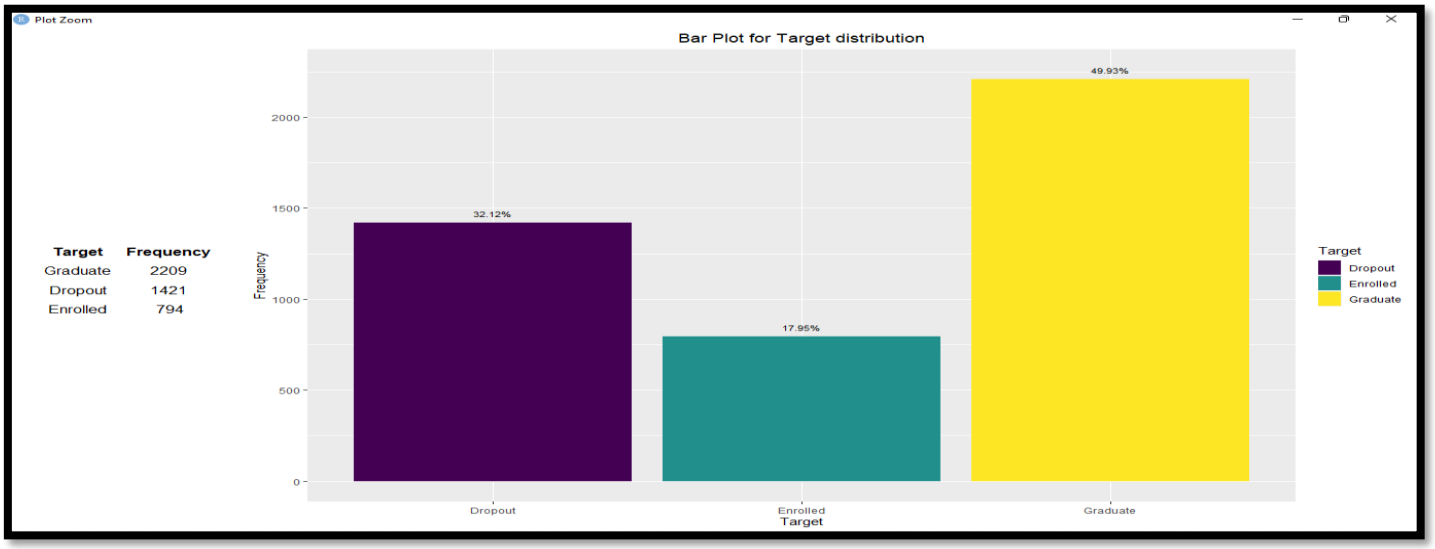
Academic data at the end of 1st semester:

Curricular units 1st sem (credited)	The number of curricular units credited by the student in the first semester.	Numerical
Curricular units 1st sem (enrolled)	The number of curricular units enrolled by the student in the first semester.	Numerical
Curricular units 1st sem (evaluations):	The number of curricular units evaluated by the student in the first semester.	Numerical
Curricular units 1st sem (approved)	The number of curricular units approved by the student in the first semester.	Numerical
Curricular units 1st sem (grade)	The number of curricular units graded in the first semester.	Numerical
Curricular units 1st sem (without evaluations)	The number of curricular units without evaluations in the second semester.	Numerical

Academic data at the end of 2nd semester:

Curricular units 2ndt sem (credited)	The number of curricular units credited by the student in the second semester.	Numerical
Curricular units 2nd sem (enrolled)	The number of curricular units enrolled by the student in the second semester.	Numerical
Curricular units 2nd sem (evaluations):	The number of curricular units evaluated by the student in the second semester.	Numerical
Curricular units 2nd sem (approved)	The number of curricular units approved by the student in the second semester.	Numerical
Curricular units 2nd sem (grade)	The number of curricular units graded in the second semester.	Numerical
Curricular units 2nd sem (without evaluations)	The number of curricular units without evaluations in the second semester.	Numerical

Target Variable: ***a) Dropout:*** Students leaving before completing their course ***b) Enrolled:*** Students currently registered for the course. ***c) Graduate:*** *Students who completed their studies.*



Research Question:

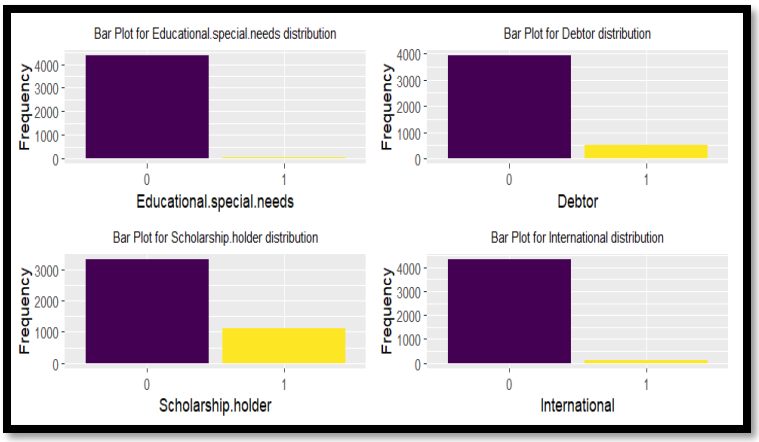
How does the student's demographic, socioeconomic, and academic data available at different points during the course affect the likelihood of student dropout or graduation at the end of the course?

Section2: Data Preparation and Cleaning

Data Preparation:

The data frame *students.df* has 4424 rows and 35 columns. The target variable has 3 unique values i.e., Dropout, Enrolled, and Graduate. The data frame has no missing values and no duplicated rows. Removing the outliers reduces the model's efficiency hence they are kept as it is. The presence of outliers depicts the real-time situation of students' life scenarios and contributes to a better understanding of the research problem. The features in the dataset have been defined as categorical and numerical based on the data quality checks and the metadata. **The categorical features are converted to factors.** Univariate and Bivariate Visual analysis was performed using the *custom function()* in the R studio.

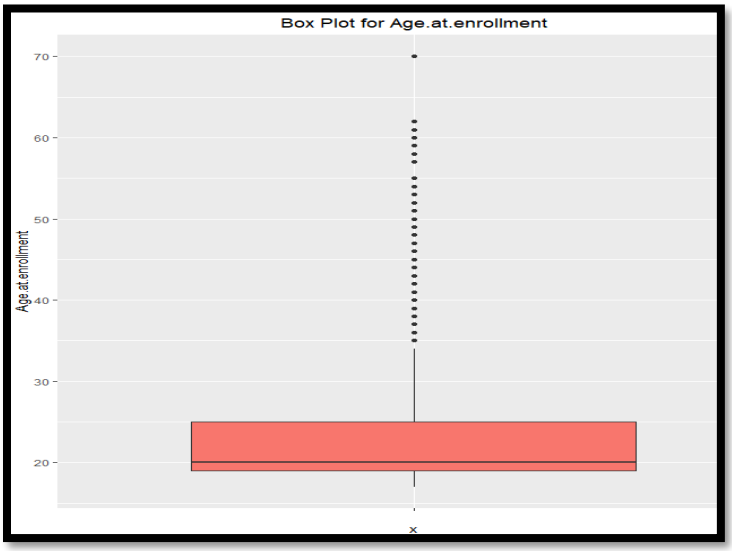
Uni-variate Analysis: Categorical Variables



Inference: It can be inferred from the univariate analysis for **categorical variables** that the maximum number of students were single, applying for the first time without requiring any special education needs, and preferred attending morning classes. 97% of students were Portuguese nationals out of which nearly half of them were displaced. The frequent qualifications of parents ranged from 4th grade of basic education to general commerce. Mothers were involved in unskilled work professions whereas fathers worked as admin staff, skilled workers(army) and unskilled work professions as well. There is a problem of **multi-collinearity** seen between Nationals and Internationals, Debtor and Tuition fees up to date.

Uni-variate Analysis: Numerical Variables

Analyzing the Macro-Economic Data: The *p0*(0th percentile) for GDP and the Inflation rate have negative values which indicate an **economic recession** in the country. There is a nearly equal distribution of the Unemployment rate for much of the population. GDP is negatively skewed. The mode value for the inflation rate is 1.4



Enrollment Age of Students: The given box plot represents the age of students at the time of enrollment. It can be seen that there are **multiple outliers** present, which indicates that people of different ages are pursuing studies. The most common age of student enrollment is **20 years.** Since the median is less than the mean, the plot is **positively skewed**.

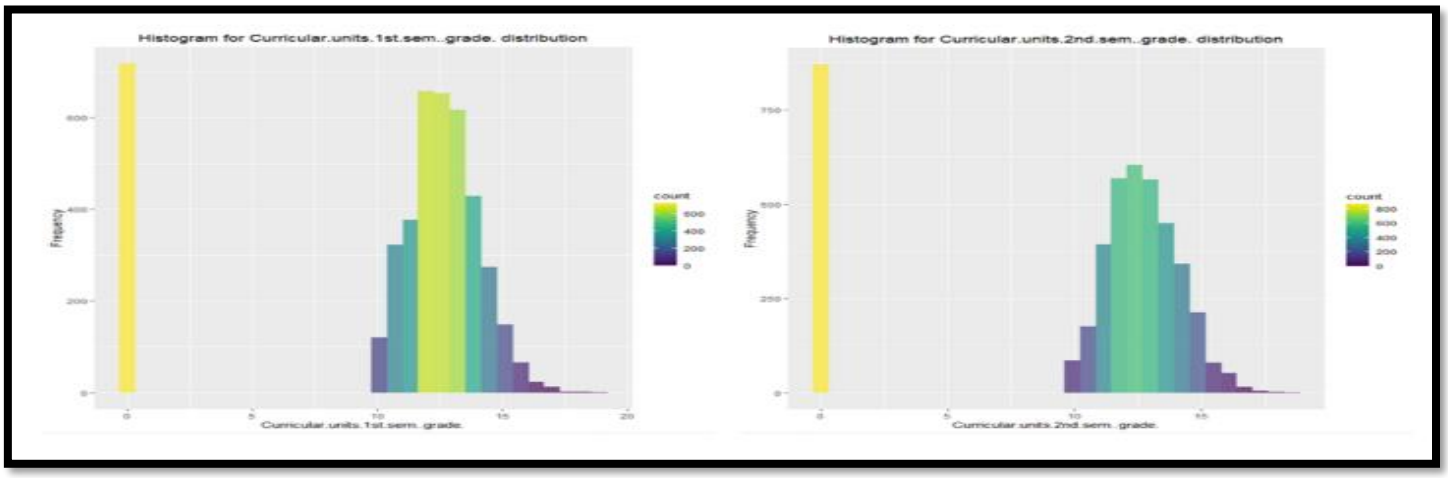
The below table compares different variables of marks between 2 semesters i.e. 1st Semester and 2nd Semester:

Inference: The overall academic performance of Students in both semesters is **nearly the same** with very minor improvement in the 2nd Semester.

Variables for 1st Sem:	Common Variables:	Variables for 2nd Sem:
Highly Positive skewed with a mean of 0.7 units	Curricular Units - **Credited** *No major difference noticed*	Positive skewed with a mean of 0.5 units
Slight positive skewed with a median of 6. Maximum value:26	Curricular Units - **Enrolled** *The range of units decreased slightly in 2nd Sem, and data became more distributed*	Near normal distribution with mean:6.23, median:6 and outliers present after 10 units till 23 units
Maximum score:45 units with a mean of 8.3. Approximately 400 students with a score of 0	Curricular Units - **Evaluation** *Maximum score value slightly decreased in 2nd Sem, and data became more distributed. There was no change in the frequency of 0 scores.*	Maximum score:33 with a mean value of 8. Near normal distribution. Approximately 400 students with a score of 0

Ranging from 0-10 units. Nearly 700 students with score of 0	Curricular Units - **Approved** *Range score and the number of students with 0 score increased in the 2nd Sem*	Ranging from 0-13 units. Nearly 800 students with a score of 0.
Negatively skewed. Value of mode: 600	Curricular Units - **Grade** *Distribution became more uniform in the 2nd Sem*	Slightly less negative skewness Value of mode:400

Histogram for Curricular units of 1st Sem and 2nd Sem Grades *:*



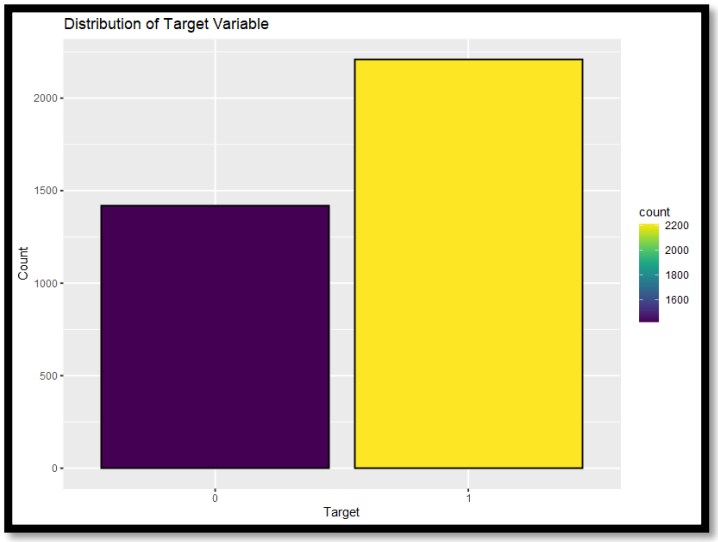
Data Cleaning:

The data was cleaned using RMarkdown. The variable Application Order had a **0 value,** which was **replaced with mode**. All the categories with frequencies less than 150 were grouped together to reduce the complexity of the data, improving the statistical performance and for better visualization. The spelling of the variable named Nationality was corrected. Removing "Enrolled" from Target Variable because the research question aims to predict student academic success and dropouts. Converting the Target Variable used for prediction, assigning 0 for Graduate and 1 for Dropout. Converting all categorical variables into factors for Exploratory Data Analysis.

[1] "For Marital.status variable, -1 (Other) constitutes 2.84810126582278 % & combines categories 3, 4, 5, 6"

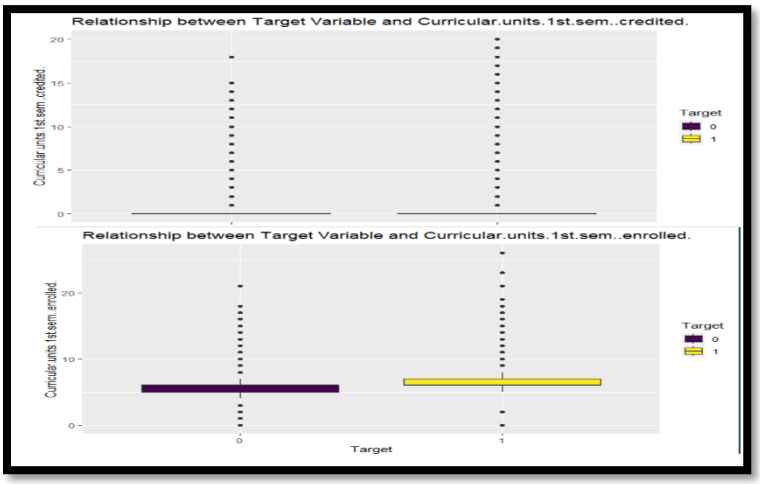
```
##                         V1              V2              V3
## Variable  Marital.status Marital.status Marital.status
## Category                -1               1               2
## Frequency              126            3919             379
## Percent              2.85%          88.58%          8.57%
```

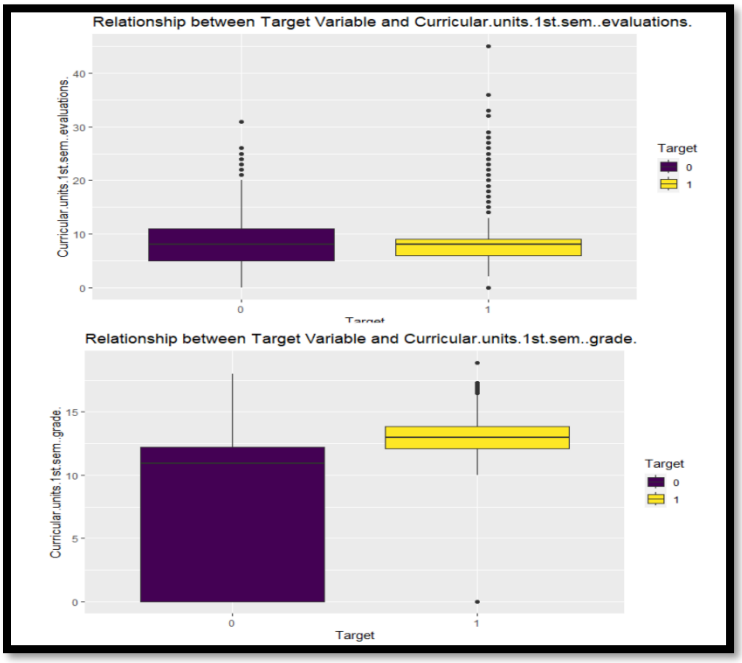
Section3: Exploratory Data Analysis



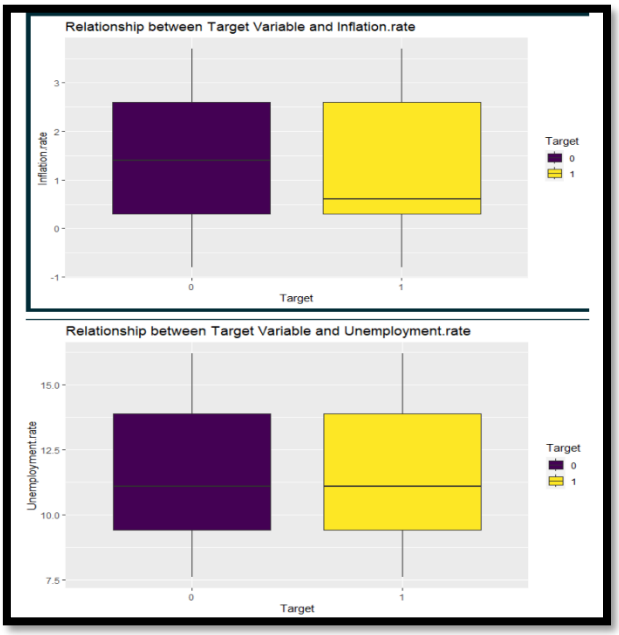
The dataset*, students.df.cleaned* has "**Target"** as a Response variable and the rest of the other are Exploratory variables. The distribution of the Target Variables is as follows:**0 i.e. Dropout = 1421 and 1 i.e. Graduate = 2209.** The distribution of the target variable was analyzed using Boxplot against all other exploratory variables. ***Insights:*** Many of the students who were drop-outs belonged to ages 20-30 and the graduated students were mostly enrolled at the age of 20. The elderly population, i.e., outliers, had equal distribution. For the academic variables ie curricular units, there was an ascending trend of association noticed from credited<enrolled<evaluations<approved<grade

The target variable had no uniform pattern of distribution for credited curricular units, whereas it had a maximum impact on the grade of curricular units in both semesters.

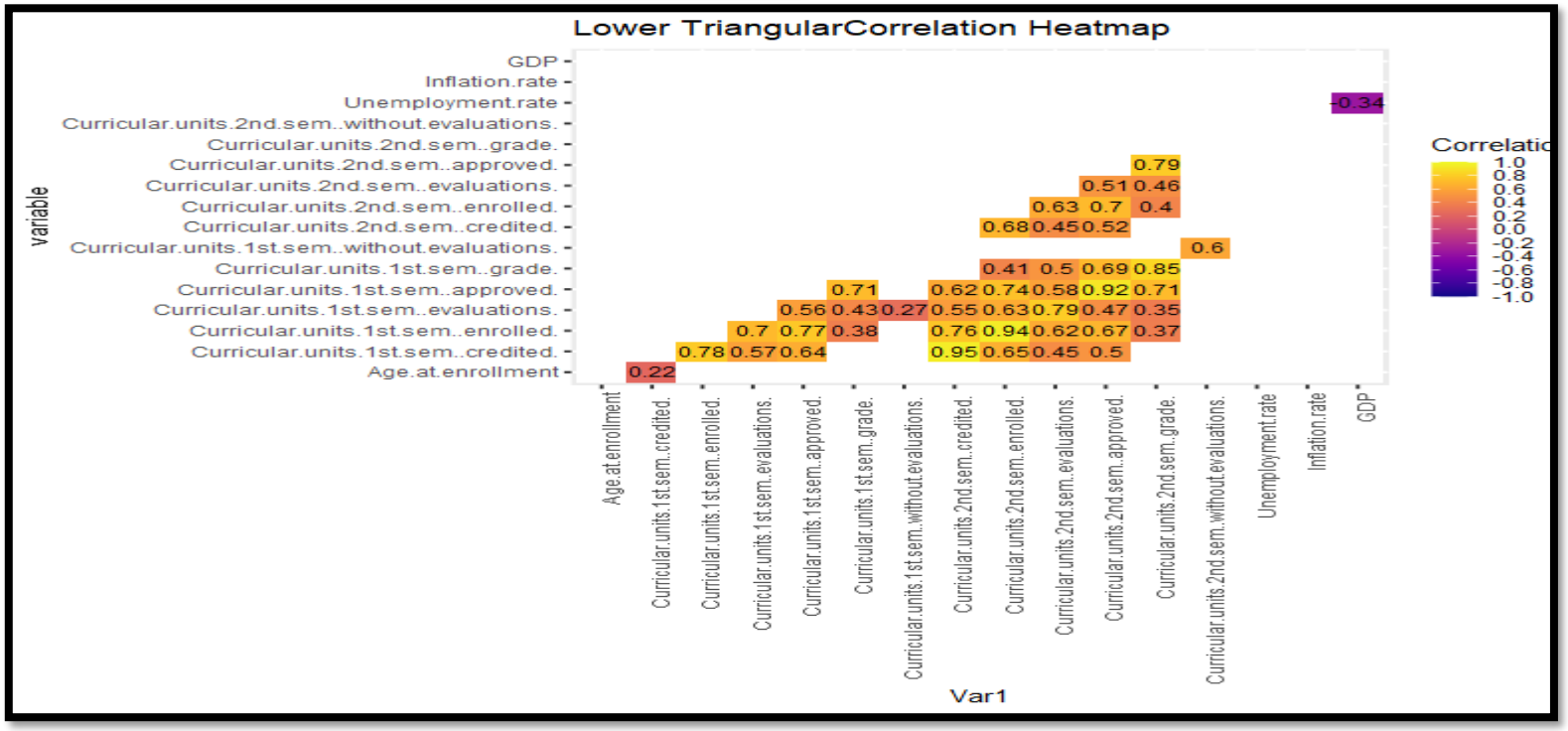




The **economic indicators** had nearly the same influence on the student's dropout and success. Most of the students dropping out of academia had an inflation rate of 1.5 and 0.5 for graduates. However, the unemployment rate for both groups was found to be the same.

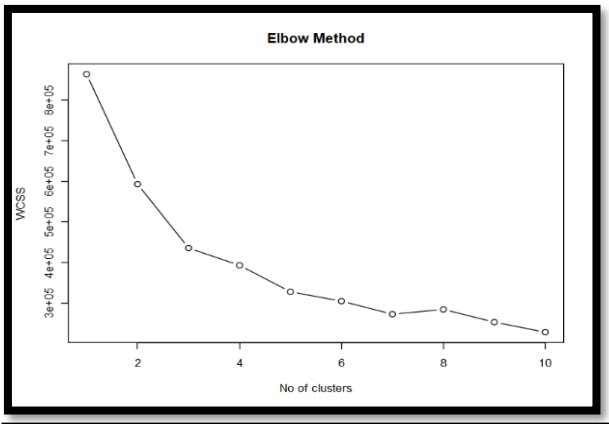


For Numerical Variables: **Multicollinearity:** correlated variables can affect the integrity of the model. This will affect the accuracy and precision of the target predictions. Hence, the **Heat-Map** was used to understand the **correlation matrix** between the Numerical variables.



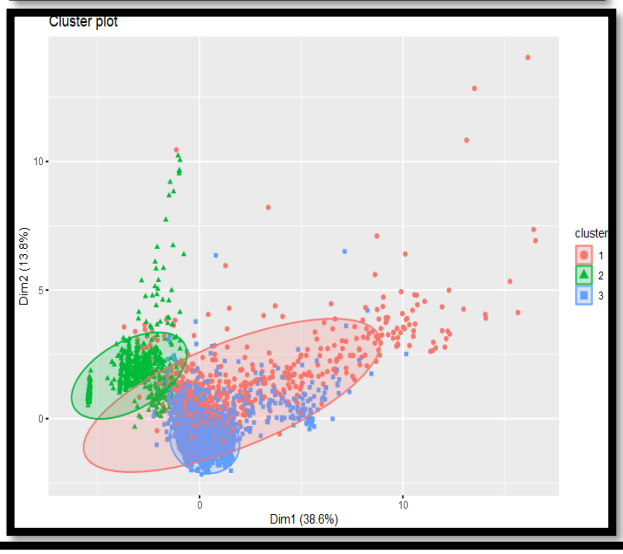
Inference: The curricular units of students in the first and second semesters were found to be correlated. This can be because they had similar curricular units in both semesters. Also, GDP has a negative correlation with the Unemployment rate and Inflation rate.

Unsupervised Machine Learning: K-Means Clustering

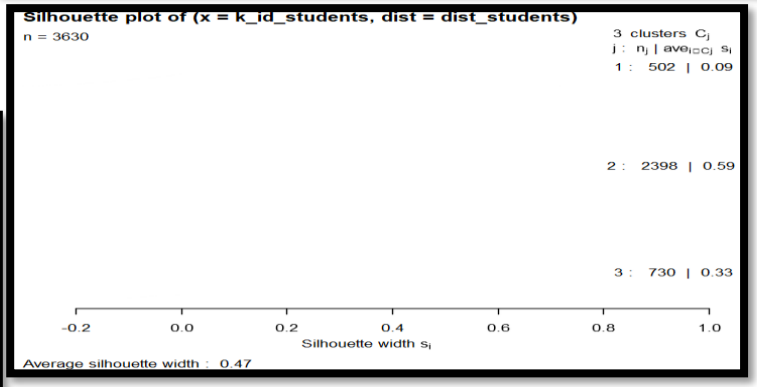


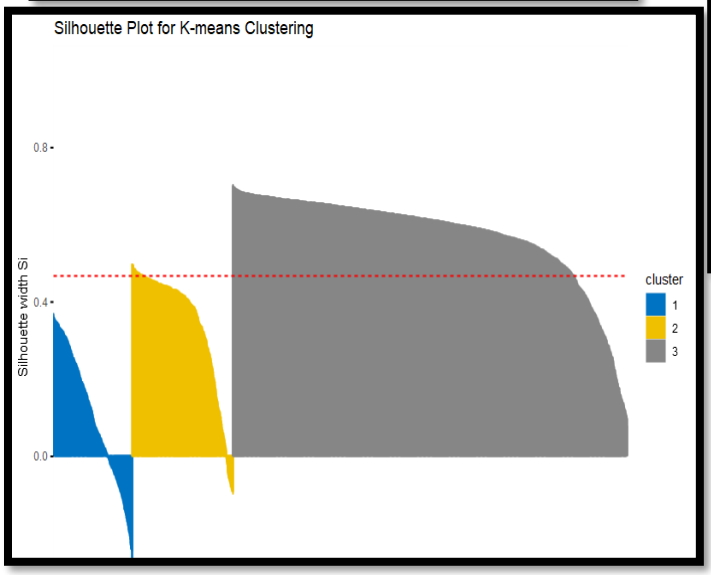
The Unsupervised learning method is useful in finding the homogenous sub-groups in complex datasets. This is based on the distance measure between two data instances. **Euclidean distance** is commonly used in cluster analysis. The unsupervised machine learning was performed on the *students.df.cleaned* dataset using the K-means clustering.

Insights: The **elbow method** (within the sum of squares) revealed **three clusters** as the optimal number of clusters in the chosen dataset. The k-means cluster analysis was performed using k=3 and nstart=25 ie using 3 clusters and iterating the datapoints for 25 times till the **total sum of the squares** score is minimum(totss =757015) The K-means cluster plot was cluttered with overlapping values, probably due to large datasets. Reducing the dimensions of the data-frame would certainly help in getting better distinct clusters. **Evaluating** cluster analysis using **Silhouette plot**. Since kmeans require k as input, it doesnot learn from the training data, and there is no accurate prediction of the best results. However, domain knowledge can help better in finding the optimal algorithm. The **Silhouette score** was **maximum for cluster 2** and the **highest number of data points** belong to **cluster 2**. The score for cluster 1 and cluster 3 was similar, ie **close to zero** indicating that the clusters are very close to the neighboring clusters **(overlapping).**



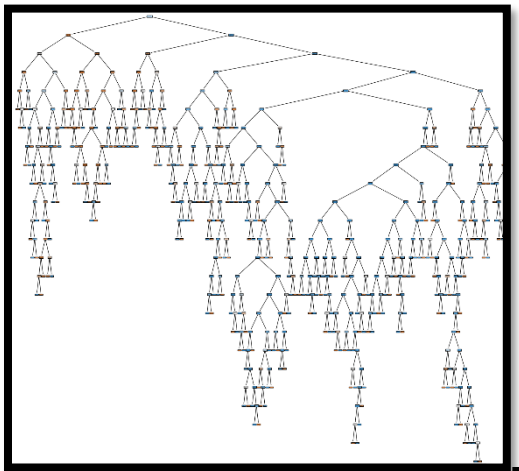
	cluster <fctr>	size <int>	ave.sil.width <dbl>
1	1	608	0.10
2	2	776	0.34
3	3	3040	0.57



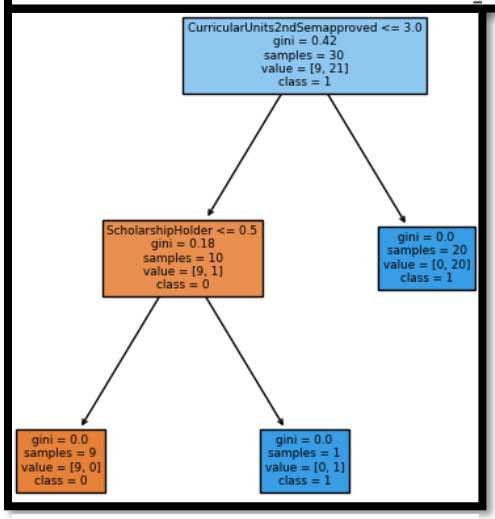


Drawbacks of K- means clustering: K- means clustering cannot be applied to complex geometrical shapes. Using multi-variate data can reduce the effectiveness of k-means due to increasing intra-variations between the data points. K-means gives more weightage to bigger clusters. Unlike hierarchical clustering, k-means does not have the option of linkages ie single, maximum, complete, ward, and distance methods. However, Keans have a better computational advantage over hierarchical clustering and better visuals over dendrograms for complex datasets.

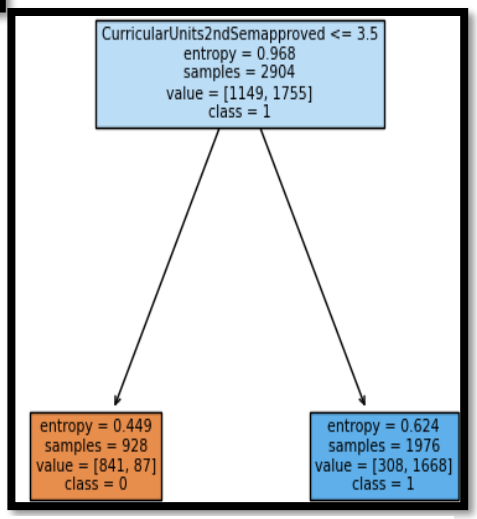
Section4: Machine Learning Predictions Decision Tree (Classification)



Supervised Machine Learning predictions were performed using a Decision tree for categorical Target variables using the classification method in **Jupyter Notebook using Python Language.** We have used the **"Target"** variable as the response variable which gives an answer to our research question: **Target variable**: 0 = Dropout implies the student will fail to complete the academic studies in the future and 1 = Graduate implies the student will successfully complete his academia. Exploratory features were of categorical and numerical datatypes. The response variable was categorical, hence for better results, **Label-Encoding** was performed on the **"Target"** variable. **Python librar**ies like Sklearn, Pandas, and Matplotlib were used for forming decision trees, hyper-parameterizing, performance metrics, and visualizations. **Tree pruning** was performed to reduce the complexity of the algorithm. Different permutations and combinations of the **Hyper-parameters** were used to find the optimal model. **GridSearchCV** could have been an alternative option. It was not used due to computational difficulties. An **Unpruned Decision tree** had an **accuracy of 87%** with the measure of **impurity Gini =0.42**, and **CurricularUnits2ndSemapproved** as **Root Node**. Further, the Decision tree model was **pruned** using different Hyper-parameters from SKLearn Library using the following parameters: **dt_entropy = DecisionTreeClassifier(criterion ="entropy", max_depth=3, ccp_alpha=0.2)** The model accuracy was coming to 88.5%.

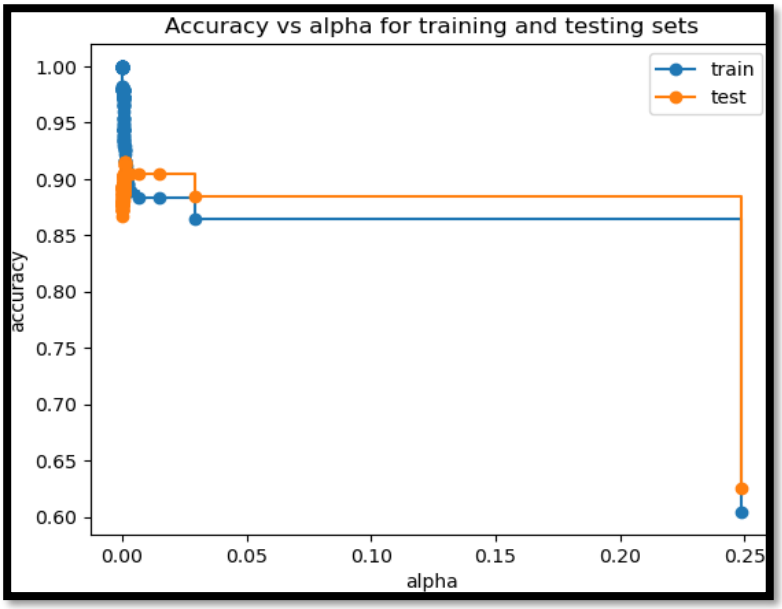


Unpruned Decision Tree (Sample)



Pruned Decision Tree

Post-pruning Method: Finding different values of CCP-Alpha by iterating through decision trees helps to subset only those trees that have been indexed by Alpha ie having less Cost complexity Pruning. This lessens the work and helps to find an optimal model. This was followed by finding training and test scores of Decision trees by using different values of CCP-Alpha. The Accuracy score with Alpha value was plotting. It was noted that, when the alpha value was 0.01, the accuracy score of train and test model was nearly 90%. Hence CCP-Alpha =0.01 was taken as one of the model hyper-parameter for model tuning.




```
path                                        =
dt_model.cost_complexity_pruning_path(x , y)
ccp_alpha = path.ccp_alphas
dt_model2 = []
for ccp in ccp_alpha :
    dt_m =DecisionTreeClassifier(ccp_alpha =
ccp)
    dt_m.fit(x_train , y_train)
    dt_model2.append(dt_m)
train_score = [i.score(x_train , y_train) for i in dt_model2 ]
test_score = [i.score(x_test , y_test) for i in dt_model2]
```

Decision Trees: **Pruning** the decision tree can reduce the complexity and over-fitting issues of the basic algorithm. Visualizing a pruned tree is user-friendly and easily comprehendible. **Hyper-Parameters: Reducing Entropy,** i.e., the measure of impurity, and **Gini Index** i.e., the probability of randomly selected features classified as incorrect can **contribute** to **minimizing the cross-validation error**. Reducing the **maximum depth** helps to reduce the dimensionality and complexity of the tree. It also facilitates better interpretability. **Information Gain** helps to understand important features impacting the target variable. The **Optimal Decision tree model** had the following parameters:

```
dt_model_optimal = DecisionTreeClassifier(ccp_alpha=0.01 , criterion='entropy' ,splitter =
'best' ,max_depth=5
```

Section4: High Performance Computational Implementation:

My research study involved the use of **RStudio and R-programming language** for data cleaning and exploratory data analysis**. Python-Jupyter Notebook** for Machine learning models and optimization.

Apache Spark was used for the high-performance computational method. The dataset was imported into **Google Collab** followed by EDA and **ML - Logistic Regression model** since the target variable was binary. The performance metrics of the Logistic Regression model will be compared with other models in the next section of the report.

The dataset used in the study has a size of (356,352 bytes). Initially, an attempt was made to perform Decision Tree machine learning in the R programming language. But, because of the size of the dataset, there were computational errors in running the algorithms. A **sample** of the original dataset took 85 seconds to run. Hence, Jupyter Notebook and Spark were used for the same reason. The performance time for running different ML algorithms was less than 10 seconds in Spark.

One Hot Encoding in Spark: was used to map all the categorical variables to integers with help of dummy variables.

First, the String Indexer was used to convert the categorical into numerical form followed by OneHotEncoderEstimator for encoding multiple columns of the dataset. The Vector Assembler combines all the feature columns into a single vector column. Since the dataset **does not** have any **ordinal** feature, **ordinal encoding** was not performed. This helped to reduce the issue of multi-collinearity in the dataset. Since the decision tree model does not require input values as numbers, **label-encoding** was performed only for the target variable (**jupyter notebook.)**

Machine Learning using Spark: maintained the integrity of the dataset and prevented sampling needs. The computation efficiency was fluent. It is a python friendly programming language hence preferable over JAVA MapReduce. However, the syntax of Spark is different from the native Python language. Importing and processing big datafile was relatively easy in Spark.

The distributed, dynamic nature of Spark makes it useful for combining data files from different sources for analysis i.e. Parallel batch processing. Apache Spark can be used for real-time streaming and analysis of data, unlike HADOOP which takes time for analysis. The Machine Learning pipelines run several algorithms and help in quicker data analysis.

Section5: Performance Evaluation and Comparison of Methods

The **Decision Tree Mode**l had the following parameters: F1 Score= 0.92661555 accuracy= 0.9077134, precision =0.92156863, and recall score= 0.93171806, Kappa score = 80%

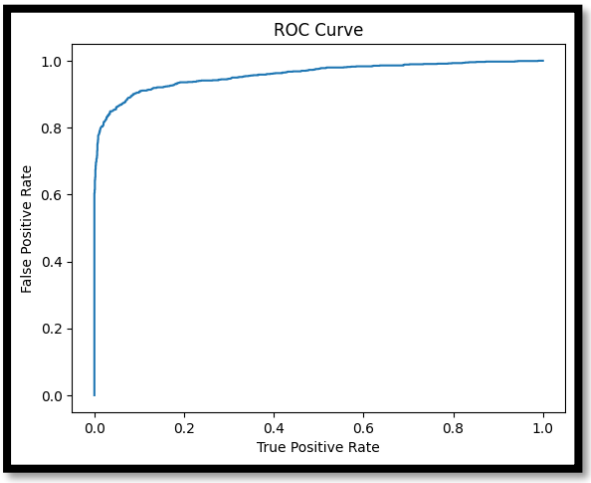
For the **Random Forest Classifier**: the data was cleaned and PCA was performed. The best-performing model was RF OverSampled with the following performance metrics = 0.90934066 Accuracy = 0.906593410.92 Precision = 0.88873626 Recall=0.92572215 F1=0.96423659]

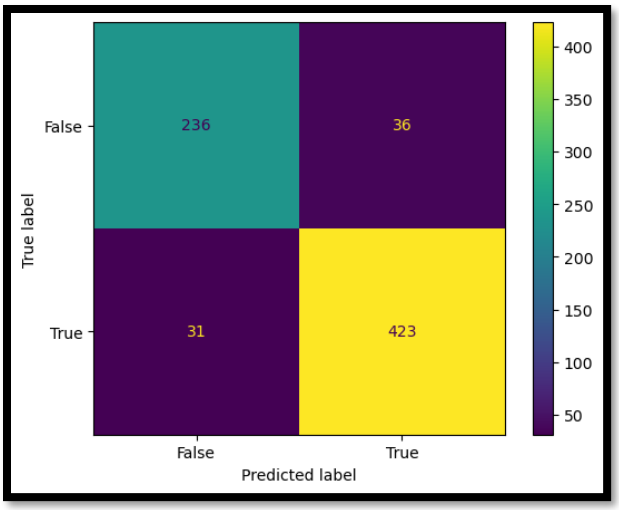
The Support Vector Machines : The accuracy was 87% with kappa value = 71.69%

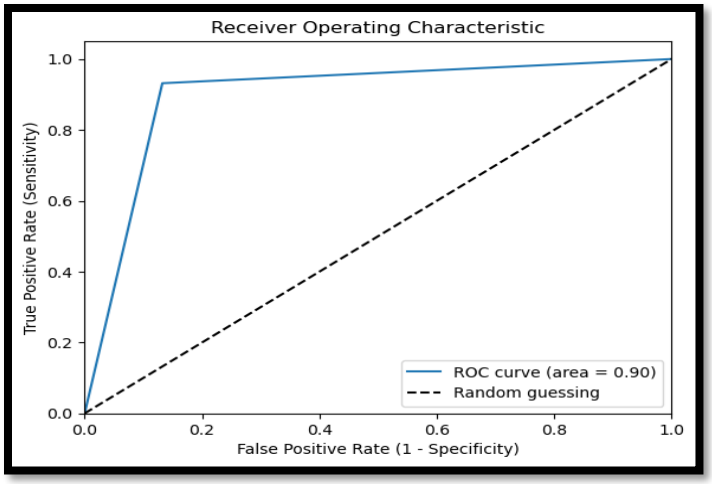
The logistic regression model .had precision of 93.4, recall of 83.4, accuracy = 0.91

Hence, it can be inferred from the above score that the logistic regression model performed in Spark had the highest accuracy. The target variable is a binary categorical variable which was converted into the numerical variable which accounts for its accuracy.

The area under the RoC curve was 96% for the Logistic regression model.

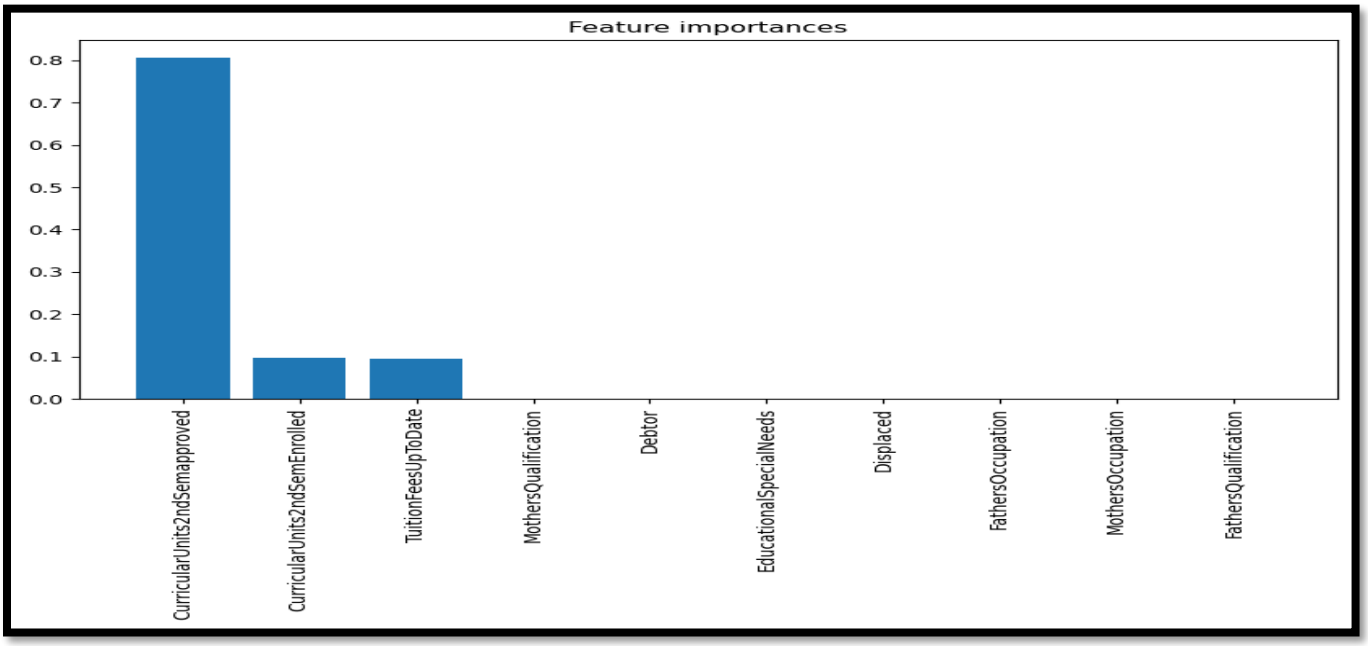






Confusion Matrix for Decision tree

Decision tree ROC curve



The Top 10 important features for the Decision Tree Model

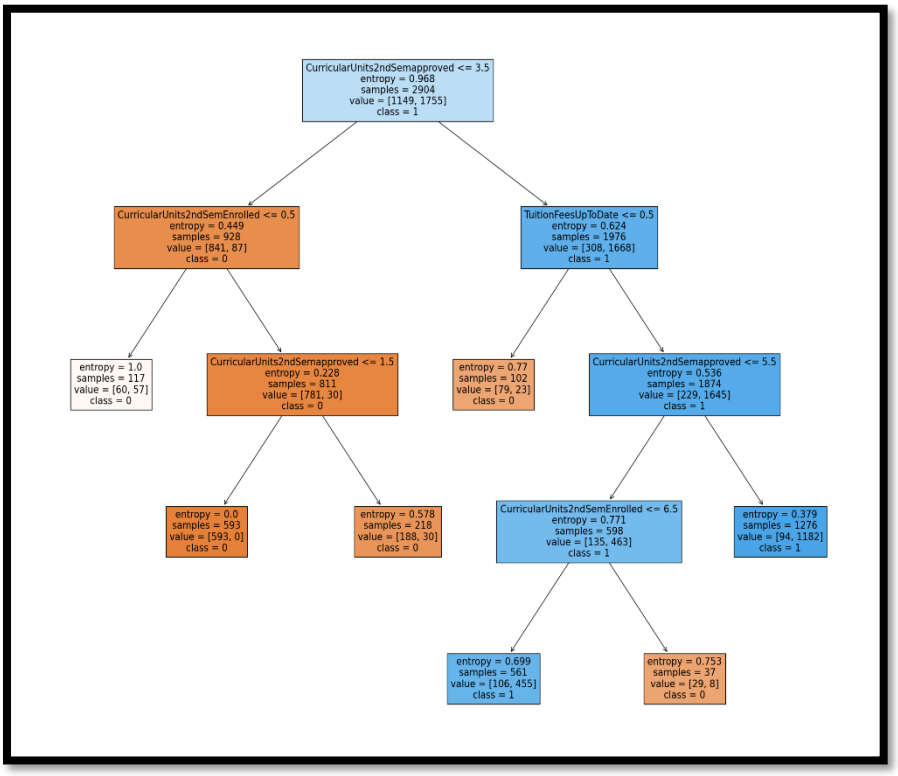
Section6: Discussion of the findings:

The CurricularUnits for the second sem was a leaf node. A Decision tree (CART) can be used for both the Classification and Regression problems of Supervised Machine learning. Additionally, the Decision tree takes less processing time as compared to other methods like Random Forest, and SVM. Feature Engineering steps like data normalization, transformation, and scaling of the data can be skipped for classification problems. Whereas, the Logistic Regression model is better suited for generating probability scores, regularization can help to deal with multi-collinearity.It worked well with spark. The processing time was extremely less. However, it does not perform well in case of high dimesntionality data.It cannot perform well in case of multiple categorical variables in the data. One hot encoding, converted categories hence it worked well. As the name suggests, linear regression can not be used for the non linear data. In that scenario, Support Vector machines work better by increasing the dimensions with the help of kernel function. SVM work well with binary classification. They are applicable for both regression and classification. SVM has kernel complexity which makes it difficult to train for larger datasets.

The Random Forest: doesnot require pruning of trees. It works well with the high dimensionality data.It is not affected by the outliers and missing values due to the aggregation strategy. Since Random forests and decision trees split from the leaf node, they cannot be predicted beyond the range of the response variable.

This research study, predicting Students academic success and dropouts was modelled with 90% of maximum accuracy. But reducing the catgeories in the data, removing noise can help to improve the prediction further. Overall, the accuracy of all the models was nearly similar.

Decision Tree Optimal model



Appendix:

Authorship Contribution Statement:

Sujeet Sharma designed the Data collection and Research question formulation, Drishti Doshi performed the Data quality check, Data cleaning was collectively performed as a group, Naman Pandey performed Exploratory data analysis, Sujeet Sharma implemented and applied Support vector machines, Drishti Doshi implemented and applied Decision Tree predictor and Logistic Regression, Naman Pandey implemented and applied Random Forest. , Naman Pandey implemented and applied Deep Learning models

Data Management Plan

1. Overview

Researcher: Drishti Doshi, Naman Pandey, Sujeet Sharma
Project title: Forcasting Student Dropout and Academic Success
Project duration: 3 months
Project context: Distributed data analysis using machine learning and Hpci techniques.

2. Defining your data/research sources

2.1 Where will your data/research sources come from? The data is collected from an online open source called "Kaggle". This dataset is in .CSV file which holds 4424 rows and 35 columns which are of numerical and categorical type.
2.2 How often will you get new data? The data was only retrieved once and was not updated throughout the analysis.
2.3 How much data/information will you generate? *The original size of the dataset is 460KB.* *All the group members are working on the same dataset.* *The size of the dataset remains constant throughout the complete process.*
2.4 What file formats will you use? This dataset can be extracted from online open source "Kaggle". https://www.kaggle.com/datasets/thedevastator/higher-education-predictors-of-student-retention This dataset is in .CSV format.

3. Organising your data

3.1How will you structure and name your folders and files? **The parent dataset file is named "student_dropout.csv".** **After cleaning we will name the data "students_df_cleaned.csv".**
3.2 What additional information is required to understand each data file? **This dataset is a direct extraction from the online open source "Kaggle" and we don't need any tool/information to reproduce this dataset.**

3.3 What different versions of each data file or source will your create? **All the group members will be working on the same data file, but if we have to denote the dataset with different versions we would like to go ahead with .xlsx format with nomenclature like "student_dropout_v1.xlsx" etc.**

4. Looking after your data

4.1 Where will you store your data? **We have created a shared repository on google drive where we have rested our dataset also we can save it on the local system, onedrive, and portable storage devices.**
4.2 How will your data be backed up? **We have saved two duplicate copies on our local systems, onedrive, and portable storage device which is not being updated throughout the process as it's a static dataset.**
4.3 How will you test whether you can restore from your backups? **For easy & quick backup we have hosted our dataset on google drive which can be accessed from any location.**

5. Sharing your data

5.1 Who owns the data you generate? **Kaggle owns this dataset.**
5.2 Who else has a right to see or use this data? **Supervisor(Alaa Marshan, Stasha Lauria, Alessandro panini), coursework marker, group members, Keggle users.**
5.3 Who else should reasonably have access to this data when you share it? **Readers of my coursework such as module leader, and assessor.**
5.4 What should/shouldn't be shared and why? **This dataset is extracted from open source so there shouldn't be any legal, ethical, and commercial limitations on the dataset.**

6. Archiving your data

6.1 **What should be archived beyond the end of your project?** **There's no requirement to archive this dataset beyond the end of this project as it's not part of my dissertation, but if any group member or assessor needs this data for further usage they can archive the .csv file and rest it at any data storage platform.**
6.2 For how long should it be stored? **There's no requirement to store this dataset as we are not using it anywhere else apart from CS5811 Coursework.**
6.3 When will files be moved into the data archive/repository? **Once we submit the coursework on wise-flow we will move the complete project to a shared repository (google drive) for future reference.**
6.4 Where will the data be stored? **We will move the complete project with the dataset to a shared repository (google drive) for future reference.**
6.5 Who is responsible for moving data to the data archive and maintaining it? **All the group members are responsible for moving data to the archive.**
6.6 Who should have access and under what conditions?

There are no restrictions to accessing the dataset.

7. Executing your plan

7.1 **Who is responsible for making sure this plan is followed?** **All the group members are responsible for making sure this plan is followed.**
7.2 How often will this plan be reviewed and updated? **This plan was well discussed and reviewed with all the team members via various group meetings.**
7.3 What actions have you identified from the rest of this plan? • **Data exploration and finalization: February 20th, 2023.** • **Brainstorming on data & data cleaning: March 10th, 2023.** • **EDA result sharing: March 20th, 2023.** • **Machine learning & HPCI result sharing: April 5th, 2023.** • **Discussion on DMP: April 10th, 2023.**
7.4 What further information do you need to carry out these actions? **To carry out mentioned actions we need dataset and project logs. we can find these details on the shared repository and Kaggle.**

Appendix A

Table A1. Marital status values.		
Attribute	Values	
Marital status	1	Single
	2	Married
	3	Widower
	4	Divorced
	5	Facto union
	6	Legally separate

Table A2. Nationality values.		
Attribute	Values	
Nationality	1	Portuguese
	2	German
	3	Spanish
	4	Italian
	5	Dutch
	6	English
	7	Lithuanian
	8	Angolan
	9	Cape Verdean
	10	Guinean
	11	Mozambican
	12	Santomean
	13	Turkish

	14	Brazilian
	15	Romanian
	16	Moldova (Republic of)
	17	Mexican
	18	Ukrainian
	19	Russian
	20	Cuban
	21	Colombian

Table A3. Application mode values.		
Attribute	Values	
Application mode	1	1st phase-general contingent
	2	Ordinance No. 612/93
	3	1st phase-special contingent (Azores Island)
	4	Holders of other higher courses
	5	Ordinance No. 854-B/99
	6	International student (bachelor)
	7	1st phase-special contingent (Madeira Island)
	8	2nd phase-general contingent
	9	3rd phase-general contingent
	10	Ordinance No. 533-A/99, item b2) (Different Plan)
	11	Ordinance No. 533-A/99, item b3 (Other Institution)
	12	Over 23 years old
	13	Transfer
	14	Change in course
	15	Technological specialization diploma holders
	16	Change in institution/course
	17	Short cycle diploma holders
	18	Change in institution/course (International)

Table A4. Course values.		
Attribute	Values	
Course	1	Biofuel Production Technologies
	2	Animation and Multimedia Design
	3	Social Service (evening attendance)
	4	Agronomy
	5	Communication Design
	6	Veterinary Nursing
	7	Informatics Engineering
	8	Equiniculture
	9	Management
	10	Social Service
	11	Tourism
	12	Nursing
	13	Oral Hygiene
	14	Advertising and Marketing Management
	15	Journalism and Communication
	16	Basic Education

	17	Management (evening attendance)

Table A5. Previous qualification values.		
Attribute	Values	
Previous qualification	1	Secondary education
	2	Higher education bachelor's degree
	3	Higher education degree
	4	Higher education master's degree
	5	Higher education doctorate
	6	Frequency of higher education
	7	12th year of schooling not completed
	8	11th year of schooling not completed
	9	Other 11th year of schooling
	10	10th year of schooling
	11	10th year of schooling not completed
	12	Basic education 3rd cycle (9th/10th/11th year) or equivalent
	13	Basic education 2nd cycle (6th/7th/8th year) or equivalent
	14	Technological specialization course
	15	Higher education degree (1st cycle)
	16	Professional higher technical course
	17	Higher education master's degree (2nd cycle)

Table A6. Mother's and Father's values.		
Attribute	Values	
Mother's & Father's qualification	1	Secondary Education 12th Year of Schooling or Equivalent
	2	Higher Education bachelor's degree
	3	Higher Education degree
	4	Higher Education master's degree
	5	Higher Education doctorate
	6	Frequency of Higher Education
	7	12th Year of Schooling not completed
	8	11th Year of Schooling not completed
	9	7th Year (Old)

	10	Other 11th Year of Schooling
	11	2nd year complementary high school course
	12	10th Year of Schooling
	13	General commerce course
	14	Basic Education 3rd Cycle (9th/10th/11th Year) or Equivalent
	15	Complementary High School Course
	16	Technical-professional course
	17	Complementary High School Course not concluded
	18	7th year of schooling
	19	2nd cycle of the general high school course
	20	9th Year of Schooling not completed
	21	8th year of schooling
	22	General Course of Administration and Commerce
	23	Supplementary Accounting and Administration
	24	Unknown
	25	Cannot read or write
	26	Can read without having a 4th year of schooling
	27	Basic education 1st cycle (4th/5th year) or equivalent
	28	Basic Education 2nd Cycle (6th/7th/8th Year) or equivalent
	29	Technological specialization course
	30	Higher education degree (1st cycle)
	31	Specialized higher studies course
	32	Professional higher technical course
	33	Higher Education master's degree (2nd cycle)
	34	Higher Education doctorate (3rd cycle)

Table A7. Mother's and Father's occupation.		
Attribute	Values	
Mother's & Father's occupation	1	Student
	2	Representatives of the Legislative Power and Executive Bodies, Directors, Directors and Executive Managers
	3	Specialists in Intellectual and Scientific Activities
	4	Intermediate Level Technicians and Professions
	5	Administrative staff
	6	Personal Services, Security and Safety Workers, and Sellers
	7	Farmers and Skilled Workers in Agriculture, Fisheries, and Forestry
	8	Skilled Workers in Industry, Construction, and Craftsmen
	9	Installation and Machine Operators and Assembly Workers
	10	Unskilled Workers
	11	Armed Forces Professions
	12	Other Situation; 13
	14	Armed Forces Officers
	15	Armed Forces Sergeants
	16	Other Armed Forces personnel
	17	Directors of administrative and commercial services

18	Hotel, catering, trade, and other services directors
19	Specialists in the physical sciences, mathematics, engineering, and related techniques
20	Health professionals
21	Teachers
22	Specialists in finance, accounting, administrative organization, and public and commercial relations
23	Intermediate level science and engineering technicians and professions
24	Technicians and professionals of intermediate level of health
25	Intermediate level technicians from legal, social, sports, cultural, and similar services
26	Information and communication technology technicians
27	Office workers, secretaries in general, and data processing operators
28	Data, accounting, statistical, financial services, and registry-related operators
29	Other administrative support staff
30	Personal service workers
31	Sellers
32	Personal care workers and the like
33	Protection and security services personnel
34	Market-oriented farmers and skilled agricultural and animal production workers
35	Farmers, livestock keepers, fishermen, hunters and gatherers, and subsistence
36	Skilled construction workers and the like, except electricians
37	Skilled workers in metallurgy, metalworking, and similar
38	Skilled workers in electricity and electronics
39	Workers in food processing, woodworking, and clothing and other industries and crafts
40	Fixed plant and machine operators
41	Assembly workers
42	Vehicle drivers and mobile equipment operators
43	Unskilled workers in agriculture, animal production, and fisheries and forestry
44	Unskilled workers in extractive industry, construction, manufacturing, and transport4
45	Meal preparation assistants
46	Street vendors (except food) and street service providers

Table A8. Gender values.		
Attribute	Values	
Gender	1	male
	0	female

Table A9. Attendance regime values.		
Attribute	Values	
Daytime/evening attendance	1	daytime
	0	evening

Table A10. Yes/No attributes.		
Attribute	Values	
Displaced	1	yes
Educational special needs		
Debtor		
Tuition fees up to date	0	no
Scholarship holder		
International		

Categorical Variables	**Univariate analysis: Insights**
Marital Status	89% of students are **single** and 8% of students are married
Application mode	Unevenly distributed, 40% belong to 1st phase contingent
Application order	70% of students applied for the **first time**
Course distribution	Uniform distribution, mode -18% of **Nursing course**
Attendance session	90% of students preferred **daytime** classes over an evening
Previous qualification	Maximum students completed **secondary education prior**
Nationality	97% of students were **Portuguese** nationals
Mother's qualification	A nearly equal proportion of **general commerce** and **secondary education**
Father's qualification	The most common qualifications seen were **4th-9th grade** of basic education and **secondary education**
Mother's occupation	The maximum frequency was of **unskilled occupation**
Father's occupation	It varied from **skilled workers to admin staff** and 22% were **unskilled workers**
Displaced	55% of students were **displaced** and 45% were not displaced
Education Special needs	99% of students had **no education special needs**

Debtor	88% of students were **not debtors**, rest 22% were debtors
Tuition fees up to date	88% of students **paid fees on time**
Gender	2/3rd of students were **females**, 1/3rd of students were **males**
Scholarship holder	Only 25% of students were **scholarship holders**
International	Only 3% of students were i**nternational**
Target	**Enrolled:32% Dropout:18% Graduate: 50%**

Data Cleaning:

Correcting the variable: Application order

Finding the mode of the Application.order column and replace 0 values with the mode

```
mode_value <- as.numeric(names(sort(-table(students.df.cleaned$Application.order)))[1])
students.df.cleaned$Application.order[students.df.cleaned$Application.order == 0] <- mode_value
```

Define the minimum frequency for combining categories

```
min_freq <- 150
```

Combine categories with low frequency

```
combined_cats_dict <- list()
for (col in categorical_cols) {

#Check if the variable is not binary
  if (length(unique(students.df.cleaned[, col])) > 2) {

# Count the frequency of each category
    counts <- table(students.df.cleaned[, col])

# Identify the categories with low frequency
  low_freq_cats <- names(counts[counts < min_freq])

# Only combine categories if at least 2 categories have frequency less than min_freq
  if (length(low_freq_cats) >= 2) {

# Combine the low frequency categories into a single "Other" category represented as -1
    students.df.cleaned[, col] <- ifelse(students.df.cleaned[, col] %in% low_freq_cats, -1, students.df.cleaned[, col])

# Identify the categories that were combined
    combined_cats_dict[[col]] <- paste(low_freq_cats, collapse = ", ")
    other_percent <- length(students.df.cleaned[students.df.cleaned[, col] == -1, col]) / nrow(students.df) * 100
    cat_msg <- paste("For", col, "variable, -1 (Other) constitutes", other_percent, "% & combines categories", combined_ca
ts_dict[[col]])
    print(cat_msg)
    }
  }
}
```

Performance Metrics for Decision Tree Model: dt_model_optimal

Accuracy Score

In [58]:

```
y_pred = dt_model_optimal.predict(x_test)
```

In [59]:

```
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
```

```
Accuracy: 0.9077134986225895
```

Precision Score

In [60]:

```
print("Precsion Score:" , metrics.precision_score(y_test, y_pred , average = None))
```

```
Precsion Score: [0.88389513 0.92156863]
```

F1 Score

In [61]:

```
from sklearn.metrics import f1_score
print("F1 Score:" , f1_score(y_test , y_pred , average = None))
```

```
F1 Score: [0.87569573 0.92661555]
```

Recall Score

In [62]:

```
from sklearn.metrics import recall_score
print("Recall Score:" , recall_score(y_test , y_pred , average = None))
```

```
Recall Score: [0.86764706 0.93171806]
```

Confusion Matrix

In [64]:

```
metrics.confusion_matrix(y_test  ,  y_pred)
```

Out[64]:

```
array([[236,  36],
       [ 31, 423]], dtype=int64)
```

In [65]:

```
confusion_matrix = metrics.confusion_matrix(y_test, y_pred)
cm_display = metrics.ConfusionMatrixDisplay(confusion_matrix = confusion_matrix, display_labels = [False, True])

cm_display.plot()
plt.show()
```


Classification Report

In [66]:

```
print(classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.88      0.87      0.88       272
           1       0.92      0.93      0.93       454

    accuracy                           0.91       726
   macro avg       0.90      0.90      0.90       726
weighted avg       0.91      0.91      0.91       726
```

Cohen Kappa Score

In [67]:

```
print("cohen_kappa_score:" , cohen_kappa_score(y_test , y_pred))
```

```
cohen_kappa_score: 0.8023213284240813
```

Receiver Operating Characteristic (ROC) Curve

In [79]:

```
fpr, tpr, thresholds = roc_curve(y_test, y_pred)
roc_auc = auc(fpr, tpr)
print("False Positive Rates:" , fpr)
print("True Positive Rates:" , tpr)
print("Threshold:",threshold)
```

```
False Positive Rates: [0.         0.13235294 1.        ]
```

```
True Positive Rates: [0.         0.93171806 1.        ]
Threshold: [2 1 0]
```

In [75]:

```
plt.plot(fpr, tpr, label='ROC curve (area = %0.2f)' % roc_auc)
plt.plot([0, 1], [0, 1], 'k--', label='Random guessing')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate (1 - Specificity)')
plt.ylabel('True Positive Rate (Sensitivity)')
plt.title('Receiver Operating Characteristic')
plt.legend(loc="lower right")
plt.show()
```

Important Features in the Decision Tree Model

In [78]:

```
importances = dt_model_optimal.feature_importances_

# Sort feature importances in descending order
indices = np.argsort(importances)[::-1]

# Get the names of the top 10 most important features
top_features = x_train.columns[indices][:10]

# Plot feature importances
plt.figure(figsize=(10, 5))
plt.title("Feature importances")
plt.bar(range(10), importances[indices][:10])
plt.xticks(range(10), top_features, rotation=90)
plt.show()
```