Topic: Predictive factors linked with student dropout or completion in higher education

1. Data description and research question

1.1 Dataset description

The dataset offers a comprehensive view of students enrolled in various undergraduate degrees at a higher education institution. It comprises several categories of variables to help analyze and understand student dropout and completion rates better. We can group these variables into the following categories:

1. Demographic Information:
   - Marital status (Categorical)
   - Gender (Categorical)
   - Age at enrollment (Numerical)
   - Nationality (Categorical)
   - International (Categorical)
   - Displaced (Categorical)

2. Application and Enrollment Details:
   - Application mode (Categorical)
   - Application order (Numerical)
   - Course (Categorical)
   - Daytime/evening attendance (Categorical)
   - Previous qualification (Categorical)
   - Scholarship holder (Categorical)

3. Parental Information:
   - Mother's qualification (Categorical)
   - Father's qualification (Categorical)
   - Mother's occupation (Categorical)
   - Father's occupation (Categorical)

4. Student Financial and Special Needs Status:
   - Debtor (Categorical)
   - Tuition fees up to date (Categorical)
   - Educational special needs (Categorical)

5. Academic Performance:
   - Curricular units 1st sem (credited) (Numerical)
   - Curricular units 1st sem (enrolled) (Numerical)
   - Curricular units 1st sem (evaluations) (Numerical)
   - Curricular units 1st sem (approved) (Numerical)

By categorizing the variables in this manner, we can gain insights into the various aspects of students' lives, such as their demographic backgrounds, academic performance, financial statuses, and family circumstances. Analyzing these factors together can help us better understand and predict the reasons for student dropout and completion rates.

1.2 Research question

The research question we aim to answer is: What are the predictive factors linked with student dropout or completion in higher education?

2. Data preparation and cleaning

In this section, we detail the process of data cleaning and preparation, ensuring that the dataset is suitable for further analysis and modeling. We follow a step-by-step approach to clean, transform, and validate the dataset.

2.1 Initial Data Exploration and Renaming Columns

First, we explore the dataset by examining its dimensions, variable types, and first ten rows. We notice that there are 35 variables and 4,424 rows of data. Most variables are either integers or numerals, except for the 'Target' variable, which is a character. To improve readability, we rename columns and create a new dataframe called 'student_new_df'.

2.2 Checking Data Quality

Next, we check the dataset for missing values using the skim() function. We find that there are 109 missing values in the 'Target' column. To further analyze the dataset, we use the describe() function to check for missing values, mean, median, mode, highest, lowest, and distinct values.

2.3 Data Validation

To ensure that our data does not contain any out-of-range values, we apply the validator() function with a set of rules for each variable. After validating the data, we confront it with the 'student_data_rules' and check the summary of the validation.

2.4 Handling Missing Values

We discover that around 2.5% of the values in the 'Target' column are missing. To avoid introducing bias into our analysis, we impute the missing values by assigning an equal proportion of the three categories (Enrolled, Graduate, Dropout). We then revalidate the dataset to ensure that there are no more missing values.

2.5 Removing Duplicates

we check the dataset for duplicate rows and remove them if any are found. After deleting duplicates, we recheck the dimensions and duplication status of the dataset to ensure its cleanliness.

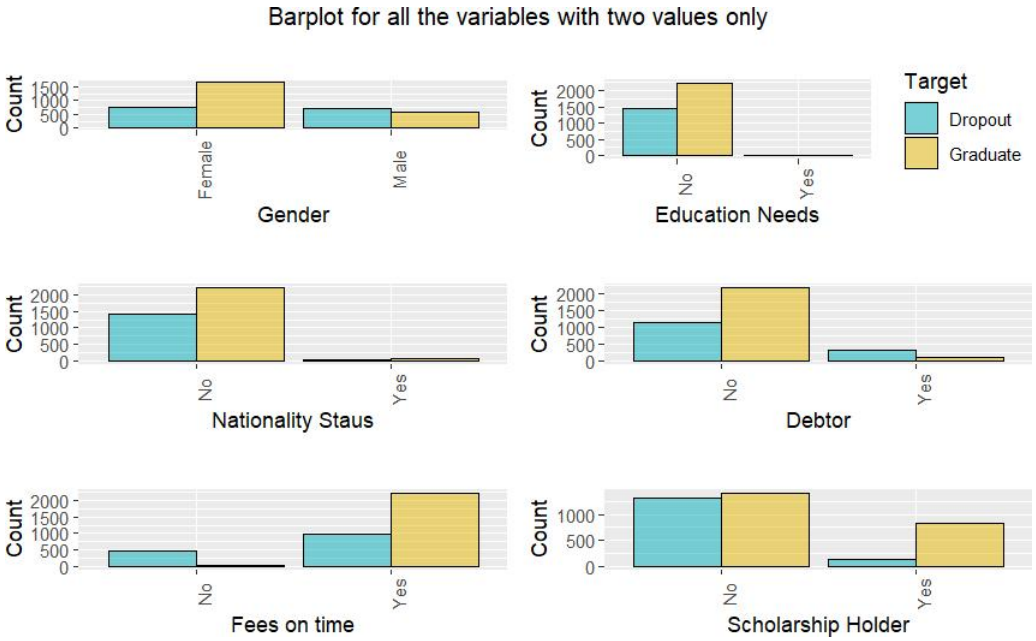
2.6 Removing Unnecessary values

Lastly, in our target variable we have “Enrolled” column which is not addressing our research question. Only “Dropout” and “Graduated” are the values that will be helpful to answer our research question.

3. Exploratory data analysis

3.1 Visualizing Dataset with respect to our Target variable:

The following barplots depict the relationship between various categorical features with two values and the target variable (dropout or graduate):



Gender: There is a notable difference in the graduation rates between male and female students. Female students graduate at a higher rate than male students. This could be attributed to various factors, such as motivation, discipline, or support systems in place. Further analysis could be conducted to understand the reasons behind this disparity and develop strategies to improve graduation rates for both genders.

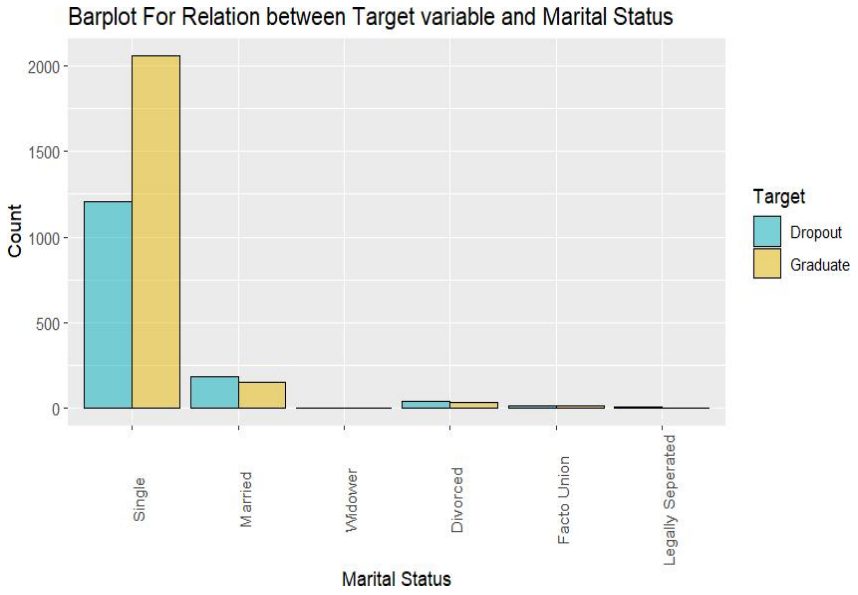
Nationality: The majority of students in the dataset are domestic students. This might indicate that the institution has a strong local presence and caters primarily to the local population. However, it may also suggest that there is room for improvement in attracting international students and enhancing the institution's global reputation.

Educational special needs: The analysis suggests that having educational special needs does not significantly impact the target variable (dropout or graduate). This might indicate that the support system and accommodations provided by the institution are effective in addressing the needs of students with educational special needs.

Debtor status and Tuition fees up to date: Students who have paid their tuition fees on time and those not in debt tend to graduate at a higher rate. This could imply that financial stability plays a role in students' ability to complete their studies. Providing more financial aid or support programs to students facing financial difficulties could potentially improve graduation rates.

Scholarship holder: Scholarship recipients are more likely to complete their studies compared to non-recipients. This could be due to the additional financial support or increased motivation to excel academically. The institution could explore expanding scholarship programs to benefit more students and boost overall graduation rates.

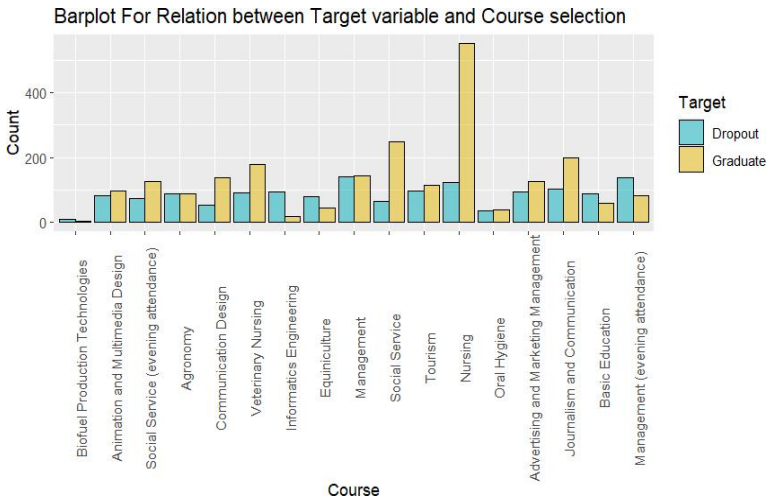
3.2 Marital status vs Target Variable:



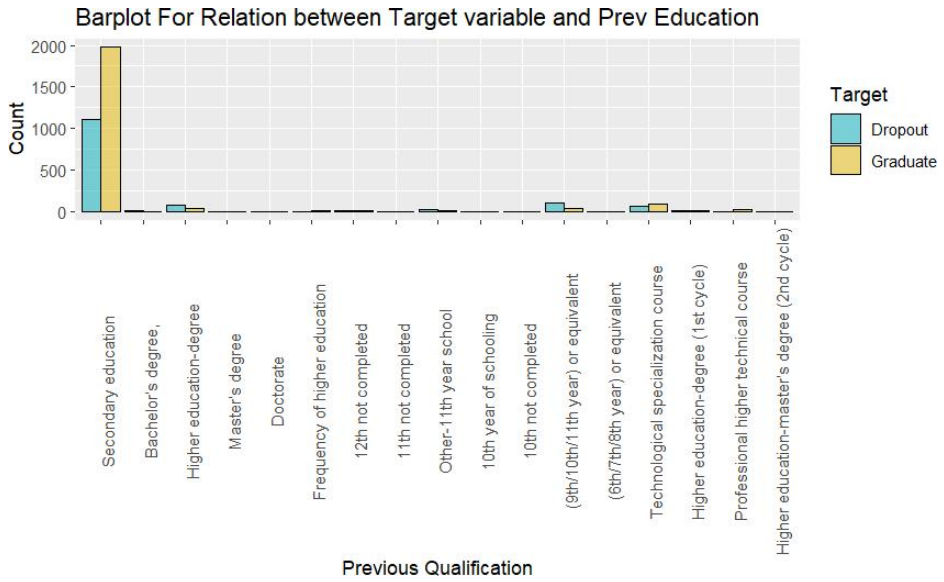
Single students have a higher graduation rate compared to other marital statuses. This might be due to fewer family responsibilities and distractions, allowing them to focus more on their studies. Further research could investigate the challenges faced by students with different marital statuses and identify support mechanisms to help them succeed academically.

3.3 Course selection vs Target:

The nursing course has the highest graduation rate, indicating that students enrolled in this program may be more committed or better supported to complete their studies. Analyzing specific elements contributing to the success of nursing students could provide insights for enhancing other programs and improving overall graduation rates.

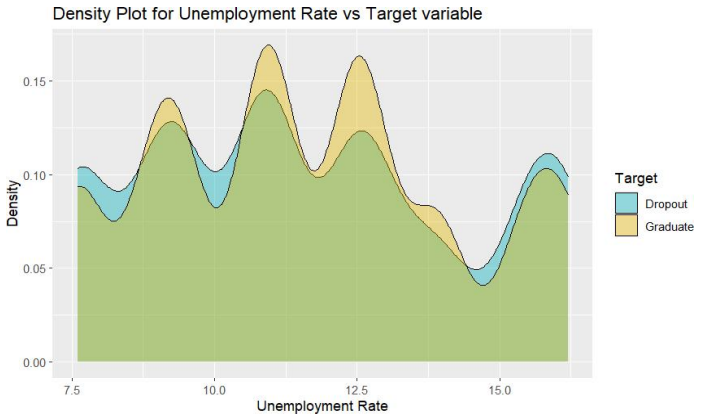


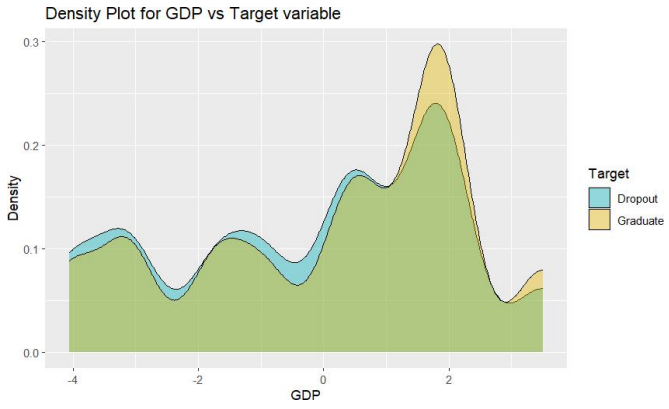
3.4 Previous qualification vs Target:



Students who have completed their secondary education tend to graduate at a higher rate. This may suggest that a strong academic foundation is essential for success in higher education. The institution could consider implementing more robust support systems for students entering higher education with weaker academic backgrounds, to improve their chances of graduation.

3.5 MacroEconomic data vs Target:

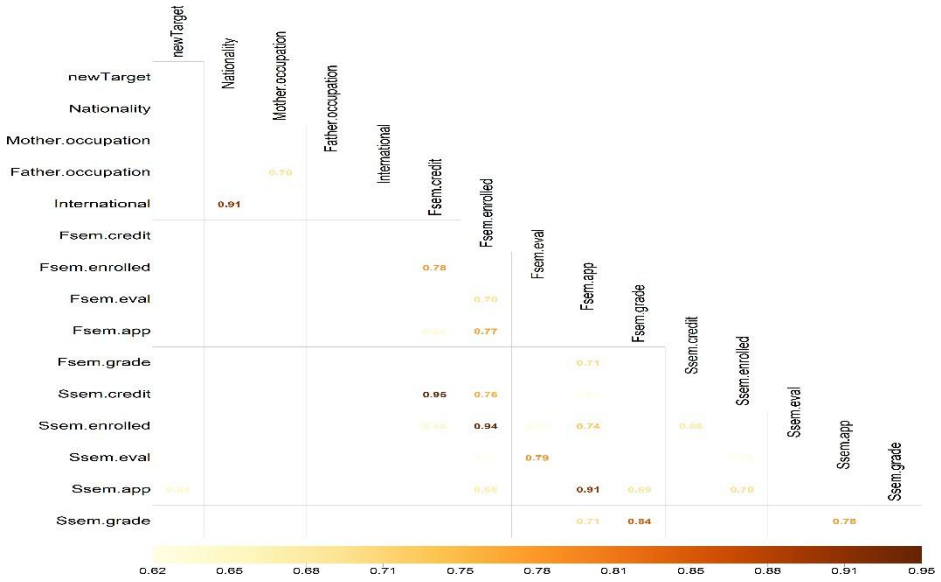




While the macroeconomic factors (GDP, unemployment rate, and inflation rate) do not show significant impact on the students' passing rates in this analysis, it would be essential not to disregard their potential long-term effects on higher education institutions and students' decision-making regarding education.

3.6 Dimensionality Reduction:

3.6.1 Correlation:



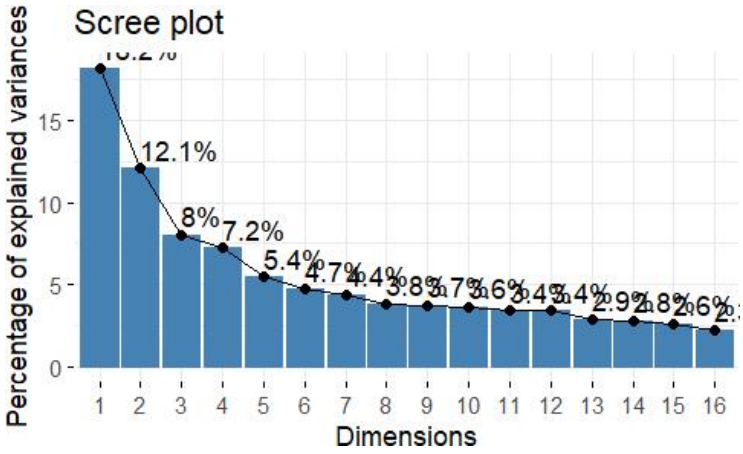
In this section, the purpose is to explore the correlations between different variables in the dataset and identify any strong correlations that might cause multicollinearity or redundancy. Multicollinearity occurs when two or more variables are highly correlated, making it difficult to determine the individual influence of each variable on the target variable. By removing variables with high correlations, we can simplify the model and make it more interpretable. The threshold for identifying significant correlations is set at an absolute value greater than 0.6, which focuses on strong relationships between variables. The visualization of the correlation matrix using the corrplot package enables a clear representation of the relationships between different variables in the dataset. The analysis identifies multicollinearity issues among some of the variables, which might negatively impact the reliability and stability of a model if not addressed. In this case, seven columns (International, Fsem.enrolled, Fsem.app, Fsem.credit, Fsem.eval, Fsem.grade, and Ssem.app) are removed from the dataset to mitigate multicollinearity issues.

By addressing multicollinearity and focusing on significant correlations, the resulting dataset is more robust and better suited for building accurate predictive models.

3.6.2 Principle Component Analysis (PCA):

This section is aiming to reduce the dimensionality of the data while retaining as much variance as possible. The insights from the PCA steps are as follows:

The PCA is computed using the princomp function with the cor = TRUE argument, ensuring that each variable is scaled to have a mean of 0 and a standard deviation of 1 before PCA calculation.



The eigenvalues, which represent the variance captured by each Principal Component (PC), are calculated and visualized using a scree plot. This plot indicates the percentage of variance explained by each PC (e.g., PC1 - 12%, PC2 - 22%, and so on).

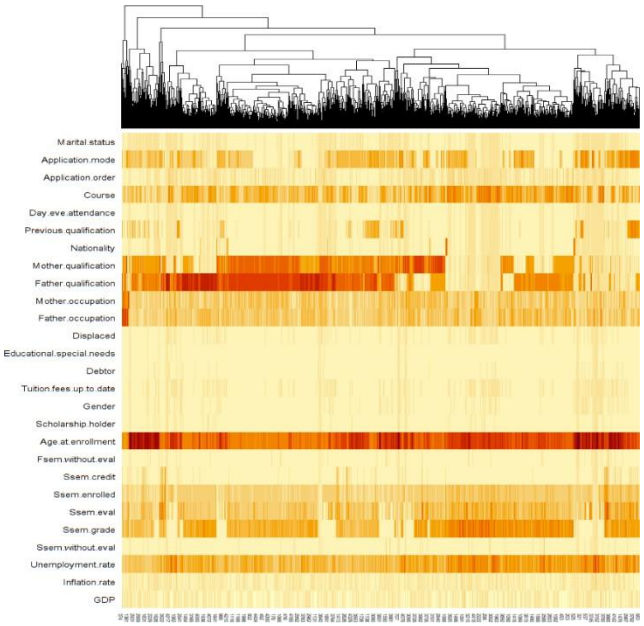
The PCA visualization includes a graph of individuals, where those with similar profiles are grouped together, and a graph of variables, where positively correlated variables point to the same side of the plot, and negatively correlated variables point to opposite sides.
A biplot combines both the individual and variable plots, providing an overview of the relationships between individuals and variables. In this case, the biplot is quite messy due to the number of data points, but row numbers close to each other on the plot have similar data patterns. Cumulative variance percentages are assessed to determine how much of the dataset's variability can be explained by a given number of dimensions (e.g., 30% by 3 dimensions, 41% by 5 dimensions, and 81% by 16 dimensions).

As the optimal number of variables to remove is unclear, the analysis will proceed with hierarchical clustering to further investigate the dataset's structure.

3.7 Clustering

3.7.1 Hierarchical clustering:

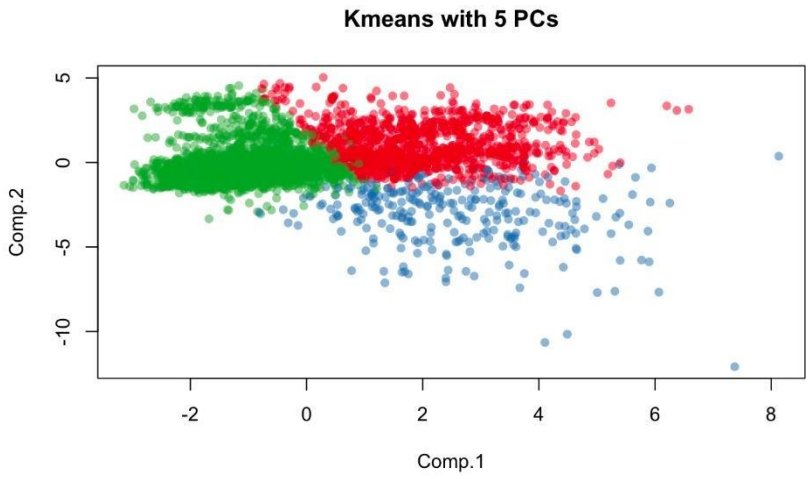
In this section, hierarchical clustering is performed on the PCA scores obtained from the PCA analysis. The shows the clustering results for 3PCA, 5PCA, and 16PCA scores. The dendrogram plots show that there are no clear clusters in the data. It's possible that the data does not contain any meaningful clusters, and all points are more or less similar to each other. In this case, a hierarchical clustering algorithm may assign all points to a single cluster.



3.7.2 K-means Clustering

K-means clustering is another unsupervised machine learning technique that groups similar observations together. The code performs Kmeans clustering on the original dataset, using all 27 variables and 3 clusters. The resulting cluster assignments are stored in a vector. The code combines the cluster assignments with the original dataset, computes the mean values of each variable within each cluster using the aggregate function, and plots two variables with color-coded points for each cluster.

Based on the exploratory data analysis (EDA) performed on the student dataset, we have gained some important insights. We started with the data cleaning process by removing missing values, duplicates and irrelevant columns. The data had a mix of categorical and numerical variables. We transformed the categorical variables into numeric values by using label encoding.



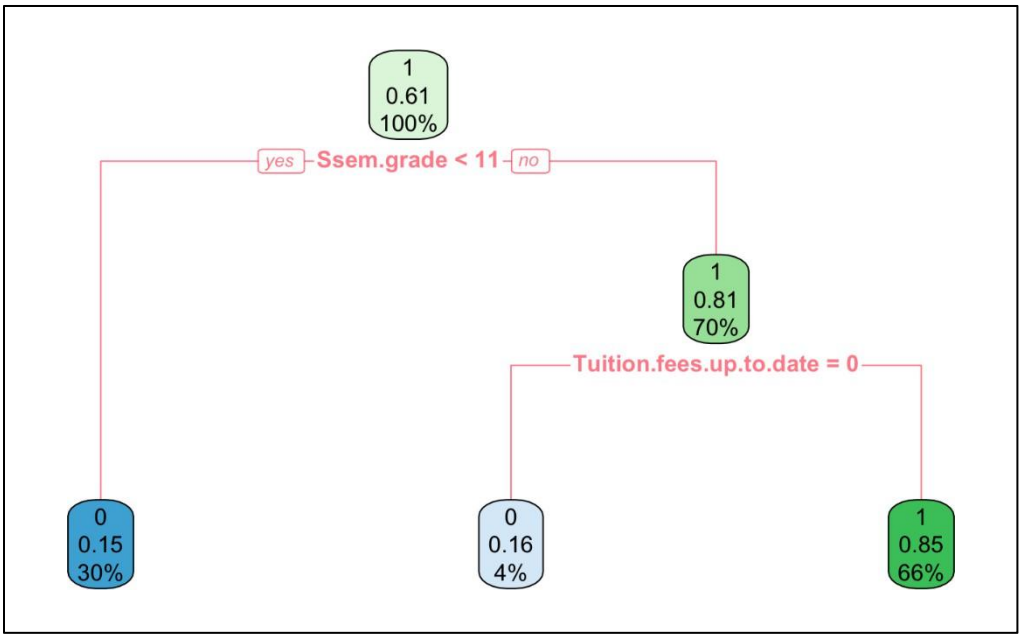
We then explored the data statistically using a correlation matrix to identify the relationship between the target variable and other variables. We removed some variables with high collinearity, leaving us with 28 variables for further analysis. We then performed principal component analysis (PCA) to reduce the dimensions and visualize the data in 2D and 3D plots.

However, the hierarchical clustering and k-means clustering did not produce clear and distinct clusters, which could indicate that the data does not contain any meaningful clusters, and all points are more or less similar to each other.

In conclusion, we were unable to reduce the dimension of the dataset through PCA as it did not produce better results compared to the original dataset. Therefore, we will proceed with all 28 variables for further analysis.

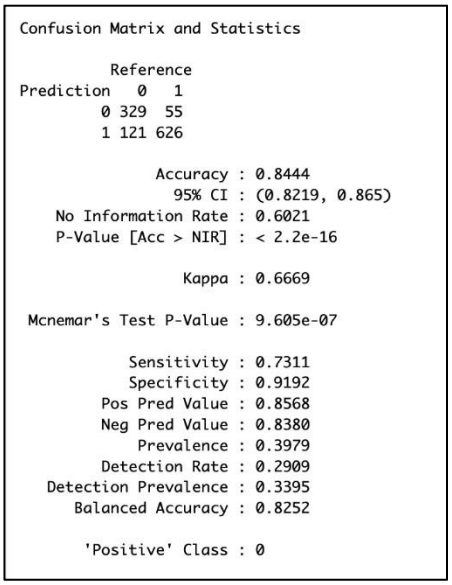
4. Machine learning prediction

For this section I have applied decision tree algorithm.



Accuracy: It measures the proportion of correct predictions among the total number of predictions made. In this case, the accuracy is 0.8444, which means the decision tree model is correct 84.44% of the time.

Kappa: The Kappa statistic measures the agreement between the predicted and actual values, accounting for the possibility of agreement occurring by chance. The Kappa value of 0.6669 suggests a moderate level of agreement between the predicted and actual values.



Sensitivity (Recall or True Positive Rate): Sensitivity measures the proportion of true positive instances that were correctly identified by the model. In this case, the sensitivity is 0.7311, which means that the decision tree model correctly identifies 73.11% of the positive instances.

Specificity: Specificity measures the proportion of true negative instances that were correctly identified by
the model. In this case, the specificity is 0.9192, which means that the decision tree model correctly identifies 91.92% of the negative instances.

Positive Predictive Value (Precision): Precision measures the proportion of true positive instances among the instances that were predicted as positive by the model. In this case, the positive predictive value is 0.8568, which means that 85.68% of the instances predicted as positive are actually positive.

Negative Predictive Value: It measures the proportion of true negative instances among the instances that were predicted as negative by the model. In this case, the negative predictive value is 0.8380, which means that 83.80% of the instances predicted as negative are actually negative.

Balanced Accuracy: Balanced accuracy is the average of sensitivity and specificity. In this case, the balanced accuracy is 0.8252, which is a more balanced metric for evaluating the model's performance, especially when dealing with imbalanced datasets.

Overall, the decision tree model has performed well on this dataset with an accuracy of 84.44%. However, there is room for improvement, particularly in terms of sensitivity (identifying true positives). You may consider fine-tuning the model or trying other algorithms to achieve better performance.

	Actual Positive	Actual Negative
Predicted Positive	329	55
Predicted Negative	121	626

From these table, it can be clearly said that the algorithm is good at predicting the negative cases whereas the model needs more improvement to predict positive cases. This bias could be possible because of imbalance in the dataset.

5. High Performance Computational implementation

As due to large dataset, the time require to train the model became high and due to lack of computational infrastructure we are not able to train the model on the larger dataset. To make the algorithm predict the values accurately, it is best to train the model with as large data as possible. To address this issue, the high-performance computation was invented which uses the distributed and parallel computing technique to train the model on the larger dataset. Therefore, in this section we are implementing the decision tree algorithm with and without HPC.

```
assembler = VectorAssembler(inputCols=feature_columns, outputCol="features")

# Use StringIndexer to convert categorical variables into numerical indices
indexers = [StringIndexer(inputCol=column, outputCol=column+"_index") for column in feature_columns]

# Apply the Decision Tree Classifier
dt = DecisionTreeClassifier(labelCol="newTarget", featuresCol="features")

# Create a pipeline with the indexers, assembler, and decision tree
pipeline = Pipeline(stages=indexers + [assembler, dt])

# Split the data into training and testing sets
train_data, test_data = data.randomSplit([0.7, 0.3])

# Train the model
start_time = time.time()
model = pipeline.fit(train_data)
end_time = time.time()


training_time = end_time - start_time
print(f"Time taken to train the Decision Tree model: {training_time:.2f} seconds")


# Make predictions on the test data
predictions = model.transform(test_data)
```

```
Time taken to train the Decision Tree model: 36.26 seconds
```

By introducing the Pyspark in training the model. It took only 36.26 seconds to train without loosing accuracy.

6. Performance evaluation and comparison of methods

The team applied different classification algorithms based on that they shared some results of their model performance. This section will compare the results and based on precision, recall, accuracy and F1 score.
The models included are: Logistic Regression, Random Forest (RF), K-Nearest Neighbors (KNN), Neural Network, Support Vector Machines with Linear Kernel (SVM-LK), and Decision Tree (from the previous discussion).
Here's a summary of the results for each model:

1. Logistic Regression:
    - Precision: 0.835
    - Recall: 0.928
    - Accuracy: 0.846
    - F1 Score: 0.87
2. Random Forest (RF):
    - Precision: 0.821
    - Recall: 0.949
    - Accuracy: 0.844
    - F1 Score: 0.88
3. K-Nearest Neighbors (KNN):
    - Precision: 0.81
    - Recall: 0.9778
    - Accuracy: 0.846
    - F1 Score: 0.8866
4. Neural Network:
    - Precision: 0.28
    - Recall: 0.13
    - Accuracy: 0.35
    - F1 Score: 0.47
5. Support Vector Machines with Linear Kernel (SVM-LK):
    - Accuracy: 0.9055

- Kappa: 0.7971
- C parameter: 1 (held constant)

6. Decision Tree:
    - Precision: 0.8568
    - Recall: 0.7311
    - Accuracy: 0.8444
    - Balanced Accuracy: 0.8252

Based on the comparison of the performance metrics, we can make the following observations:

- The Neural Network model has the poorest performance among the models, with low precision, recall, and accuracy.
- The SVM-LK model has the highest accuracy (0.9055) but only reports accuracy and kappa as performance metrics. It is essential to consider other metrics as well for a comprehensive understanding of the model's performance.
- KNN has the highest F1 Score (0.8866), which is a balanced metric considering both precision and recall. It also has the highest recall (0.9778), meaning it is excellent at identifying positive instances.
- The Logistic Regression and Random Forest models have similar performance, with slightly higher F1 Score for Random Forest.

In conclusion, the KNN model and SVM-LK model seem to have the best performance among the presented models, based on the provided metrics. However, it is essential to consider the context and the specific use case when choosing a model. For example, if minimizing false negatives is more critical, then the KNN model with its high recall would be more suitable. It is also worth noting that, depending on the dataset and problem, the performance of the models can vary, so cross-validation and further fine-tuning may be necessary to select the most appropriate model for your use case.

7. Discussion of the findings

This study aimed to answer the research question: What are the predictive factors linked with student dropout or completion in higher education? To address this question, we conducted an in-depth analysis of the student dataset, which included data cleaning, exploratory data analysis (EDA), dimensionality reduction, and machine learning prediction.

EDA revealed several factors that are associated with student dropout or completion rates. The analysis found that female students graduate at a higher rate than male students, suggesting that gender may play a role in student outcomes. The majority of students in the dataset are domestic, which might indicate that the institution primarily caters to the local population. Students with educational special needs did not show a significant difference in their graduation rates compared to other students, which may suggest that the institution's support system is effective in addressing these students' needs. Students who were debt-free and had paid their tuition fees on time were more likely to graduate, indicating the importance of financial stability in student outcomes. Additionally, scholarship holders were more likely to complete their studies, suggesting that financial support and motivation could be influential factors in student success.

PCA and hierarchical clustering were employed to reduce the dataset's dimensionality and potentially reveal underlying patterns. However, neither method produced clear and distinct clusters, indicating that the data points are more or less similar to each other. Thus, the analysis proceeded with all 28 variables for further analysis.

A decision tree algorithm was applied to predict student dropout or completion rates based on the given dataset. The model achieved an accuracy of 84.44%, a Kappa value of 0.6669, a sensitivity of 73.11%, a specificity of 91.92%, and a positive predictive value of 85.68%. These results indicate that the decision tree model provides a moderately accurate prediction of student outcomes in higher education, considering the dataset at hand.

8. Conclusion:

In conclusion, our analysis identified several factors that are linked with student dropout or completion in higher education, including gender, nationality, educational special needs, financial stability, and scholarship status. While the decision tree model provided moderately accurate predictions, it is crucial to note that these findings are specific to the dataset and institution studied. Further research could investigate these factors across various institutions and contexts to generalize the findings and develop targeted strategies for improving student outcomes in higher education.

It is also essential to consider other factors that may not be present in the current dataset but may play a role in predicting student outcomes, such as psychological factors, socio-economic background, and institutional support systems. Additionally, future research could explore the use of other machine learning algorithms or ensemble methods to enhance the prediction accuracy and provide better insights into the factors influencing student dropout and completion rates in higher education.