BASIC INFORMATION

Title of Project: STUDENT DROPOUT PREDICTION USING MACHINE LEARNING MODELS

Student Name: Asif Iqbal Khan

Enrollment Number: 00819011622 (LE08)

Batch: AIML B1

Email ID: asifiqbalk99@gmail.com

Contact Number: 7017546946

Google Drive Link: https://drive.google.com/drive/folders/10jObn6BHFCwS3mpF_qiGWpNaAHw57msW?usp=sharing

Google Website Link: https://studentdropoutprediction.blogspot.com/2023/06/student-dropout-prediction-using-ml.html

YouTube Video Link: https://youtu.be/TrGEx338Ets

STUDENT DROPOUT PREDICTION USING MACHINE LEARNING MODELS

Abstract: Higher education institutions record a significant amount of data about their students,representing a considerable potential to generate information, knowledge, and monitoring. Both school dropout and educational failure in higher education are an obstacle to economic growth,employment, competitiveness, and productivity, directly impacting the lives of students and their families, higher education institutions, and society as a whole. The dataset described here results from the aggregation of information from different disjointed data sources and includes demographic,socioeconomic, macroeconomic, and academic data on enrollment and academic performance at the end of the first and second semesters. The dataset is used to build machine learning models for predicting academic performance and dropout that provides information to the tutoring team with an estimate of the risk of dropout and failure. The dataset is useful for researchers who want to conduct comparative studies on student academic performance and also for training in the machine learning area.

Keywords: academic performance; machine learning in education; educational data mining; learning management system; prediction

1. Introduction:

Academic success in higher education is vital for jobs, social justice, and economic growth. Dropout represents the most problematic issue that higher education institutions must address to improve their success. There is no universally accepted definition of dropout. The proportion of students who dropout varies between different studies depending on how dropout is defined, the data source, and the calculation methods. Frequently, dropout is analyzed in the research literature based on the timing of the dropout (early vs. late). Due to differences in reporting, it is not possible to compare dropout rates across institutions. In this work, we define dropouts from a micro-perspective,where field and institution changes are considered dropouts independently of the timing these occur. This approach leads to much higher dropout rates than the macro-perspective, which considers only students who leave the higher education system without a degree.

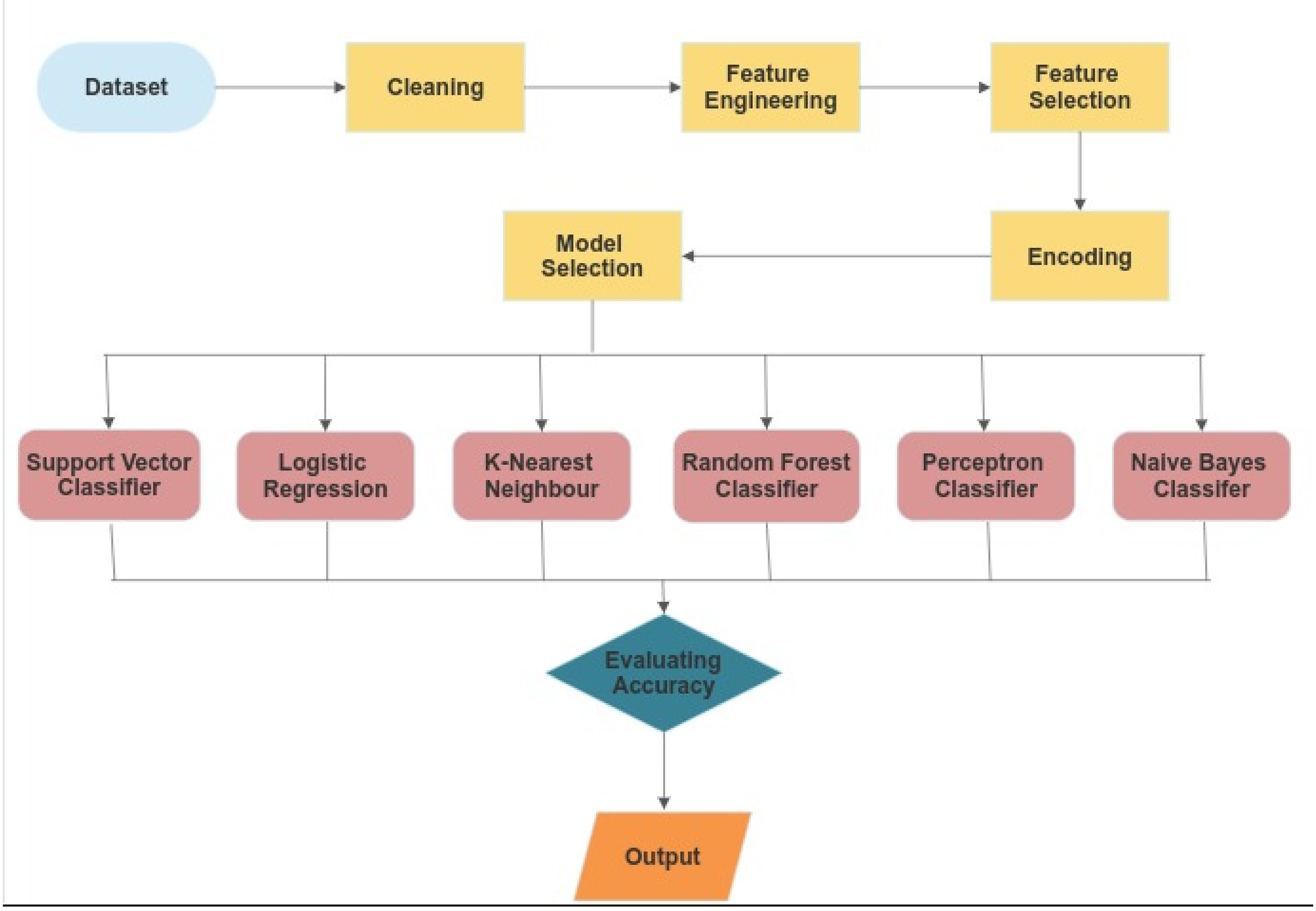
According to the independent report for the European Commission, too many students drop out before the end of their higher education courses [4]. Even in the most successful country (Denmark), only around 80% of students complete their studies, while in Italy, this rate is only 46%. This report highlights key factors that lead students to drop out, with the major cause being socioeconomic conditions.

In recent years, early prediction of student outcomes has attracted increasing research interest. However, despite the research interest and the considerable amount of data that the universities generate, there is a need to collect more and better administrative data, including dropout and transfer reasons.

This descriptor presents a dataset created from a higher education institution (acquired from several disjoint databases) related to students enrolled in different undergraduate degrees, such as agronomy, design, education, nursing, journalism, management, social service, and technologies. The dataset includes information known at the time of student enrollment (academic path, demographics, and macroeconomics and socioeconomic factors) and the students' academic performance at the end of the first and second semesters. The data are used to build classification models to predict student dropout and academic success. The problem is formulated as a three-category classification task (dropout, enrolled, and graduate) at the end of the normal duration of the course. These classification models are part of a Learning Analytic tool that includes predictive analyses which provide information to the tutoring team at our higher education institution with an estimate of the risk of dropout and failure. With this information, the tutoring team provides more accurate help to students.

The dataset contained 4424 records with 35 attributes, where each record represents an individual student and can be used for benchmarking the performance of different algorithms for solving the same type of problem and for training in the machine learning area.

2. Proposed Methodology



2.1 Dataset

The dataset includes demographic data, socioeconomic and macroeconomic data, data at the time of student enrollment, and data at the end of the first and second semesters. The data sources used consist of internal and external data from the institution and include data from (i) the Academic Management System (AMS) of the institution, (ii) the Support System for the Teaching Activity of the institution (developed internally and called PAE), (iii) the annual data from the General Directorate of Higher Education (DGES) regarding admission through the National Competition for Access to Higher Education (CNAES), and (iv) the Contemporary Portugal Database (PORDATA) regarding macroeconomic data.

The data refer to records of students enrolled between the academic years 2008/2009 (after the application of the Bologna Process to higher education in Europe) to 2018/2019. These include data from 17 undergraduate degrees from different fields of knowledge, such as agronomy, design, education, nursing, journalism, management, social service, and technologies. The final dataset is available as a comma-separated values (CSV) file encoded as UTF8 and consists of 4424 records with 35 attributes and contains no missing values

Table 1 describes each attribute used in the dataset grouped by class: demographic,socioeconomic, macroeconomic, academic data at enrollment, and academic data at the end of the first and second semesters. Appendix A contains the descriptions of possible values for the attributes, and the URL referenced in the Supplementary Material contains more detailed information.

Table 1. Attributes used grouped by class of attribute.

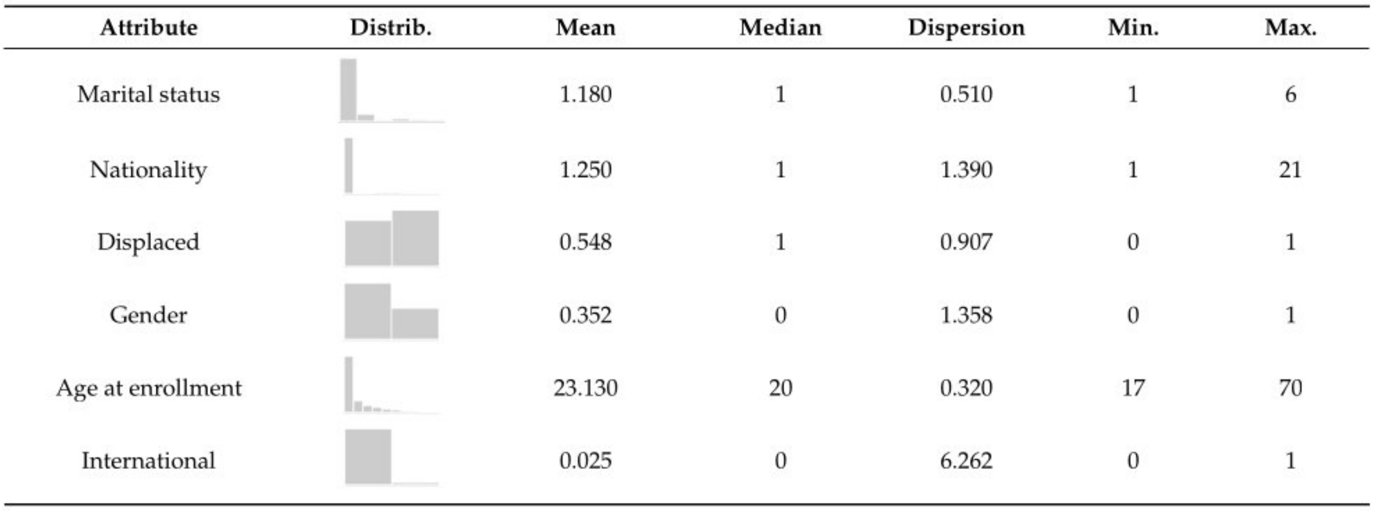
Class of Attribute	Attribute	Type
Demographic data	Marital status	Numeric/discrete
	Nationality	Numeric/discrete
	Displaced	Numeric/binary
	Gender	Numeric/binary
	Age at enrollment	Numeric/discrete
	International	Numeric/binary
Socioeconomic data	Mother’s qualification	Numeric/discrete
	Father’s qualification	Numeric/discrete
	Mother’s occupation	Numeric/discrete
	Father’s occupation	Numeric/discrete
	Educational special needs	Numeric/binary
	Debtor	Numeric/binary
	Tuition fees up to date	Numeric/binary
	Scholarship holder	Numeric/binary
Macroeconomic data	Unemployment rate	Numeric/continuous
	Inflation rate	Numeric/continuous
	GDP	Numeric/continuous
Academic data at enrollment	Application mode	Numeric/discrete
	Application order	Numeric/ordinal
	Course	Numeric/discrete
	Daytime/evening attendance	Numeric/binary
	Previous qualification	Numeric/discrete
Academic data at the end of 1st semester	Curricular units 1st sem (credited)	Numeric/discrete
	Curricular units 1st sem (enrolled)	Numeric/discrete
	Curricular units 1st sem (evaluations)	Numeric/discrete
	Curricular units 1st sem (approved)	Numeric/discrete
	Curricular units 1st sem (grade)	Numeric/continuous
	Curricular units 1st sem (without evaluations)	Numeric/discrete
Academic data at the end of 2nd semester	Curricular units 2nd sem (credited)	Numeric/discrete
	Curricular units 2nd sem (enrolled)	Numeric/discrete
	Curricular units 2nd sem (evaluations)	Numeric/discrete
	Curricular units 2nd sem (approved)	Numeric/discrete
	Curricular units 2nd sem (grade)	Numeric/continuous
	Curricular units 2nd sem (without evaluations)	Numeric/discrete
Target	Target	Categorical

2.2 Data Analysis

We performed a brief exploratory data analysis in Python 3 using the Pandas library version 2.0.2, the Scikit-learn library version 1.2.2, and the Seaborn library version 0.12.2 for visualizations.

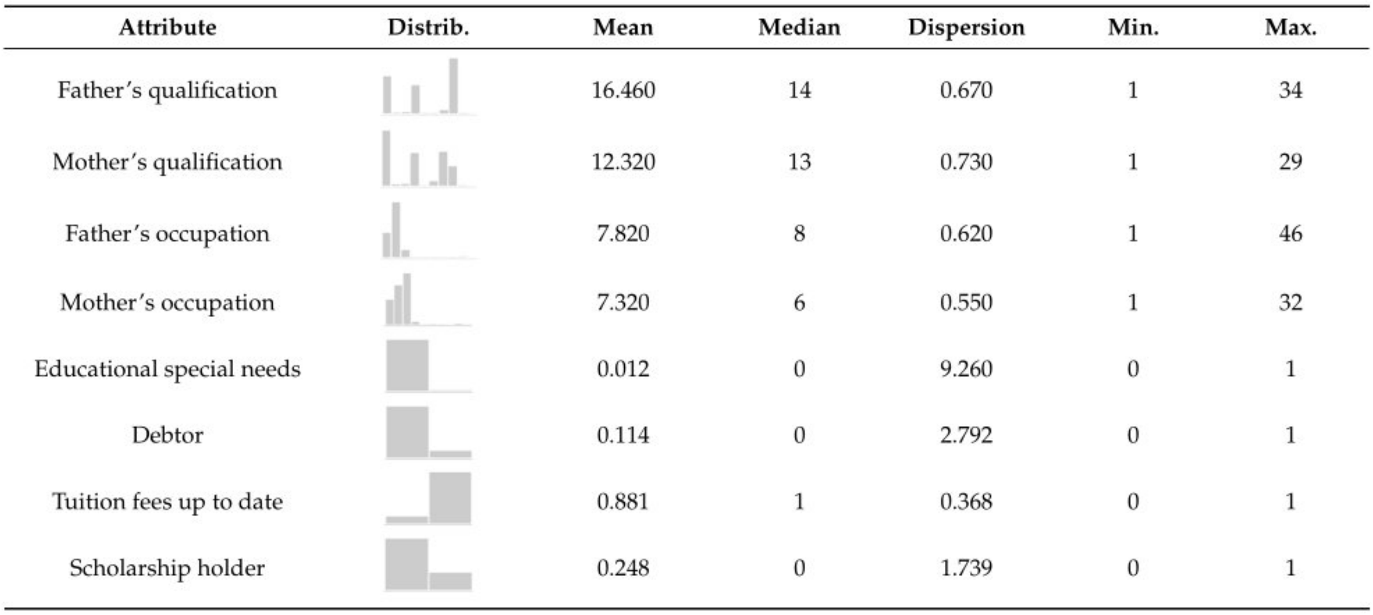
Tables 2.1–2.6 contain basic statistics about all the attributes. These tables include a histogram of attribute values, the central tendency of each attribute value (mode for categorical attributes and mean for Numeric attributes), the median of each attributes values, the dispersion of the attributes values, and the minimum and maximum value for numerical attributes only.

Table 2.1 Basic Statistics information about demographic data



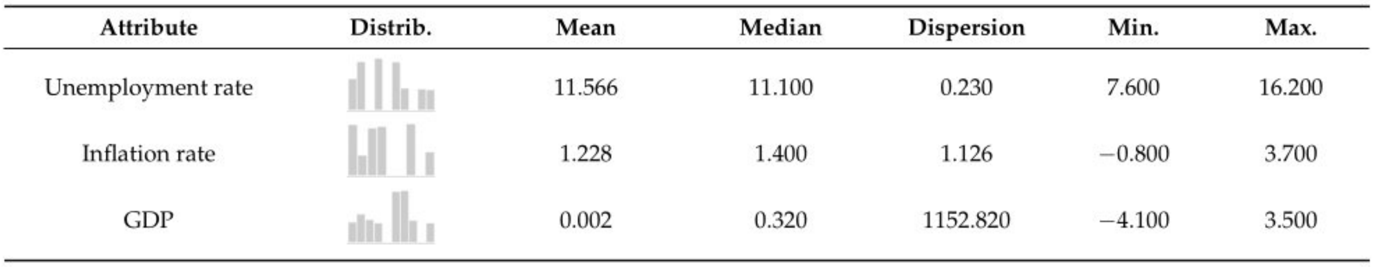
Attribute	Distrib.	Mean	Median	Dispersion	Min.	Max.
Marital status		1.180	1	0.510	1	6
Nationality		1.250	1	1.390	1	21
Displaced		0.548	1	0.907	0	1
Gender		0.352	0	1.358	0	1
Age at enrollment		23.130	20	0.320	17	70
International		0.025	0	6.262	0	1

Table 2.2. Basic statistics information about Socioeconomic data.



Attribute	Distrib.	Mean	Median	Dispersion	Min.	Max.
Father′s qualification		16.460	14	0.670	1	34
Mother′s qualification		12.320	13	0.730	1	29
Father′s occupation		7.820	8	0.620	1	46
Mother′s occupation		7.320	6	0.550	1	32
Educational special needs		0.012	0	9.260	0	1
Debtor		0.114	0	2.792	0	1
Tuition fees up to date		0.881	1	0.368	0	1
Scholarship holder		0.248	0	1.739	0	1

Table2.3. Basic statistics information about macroeconomic data



Attribute	Distrib.	Mean	Median	Dispersion	Min.	Max.
Unemployment rate		11.566	11.100	0.230	7.600	16.200
Inflation rate		1.228	1.400	1.126	−0.800	3.700
GDP		0.002	0.320	1152.820	−4.100	3.500

Table 2.4. Basic statistics information about academic data at enrollment

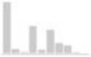
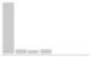
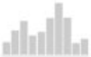
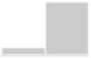
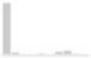
Attribute	Distrib.	Mean	Median	Dispersion	Min.	Max.
Application mode		6.890	8	0.770	1	18
Application order		1.730	1	0.760	1	9
Course		9.900	10	0.440	1	17
Daytime/evening attendance		0.891	1	0.350	0	1
Previous qualification		2.530	1	1.570	1	17

Table 2.5. . Basic statistics information about academic data at the end of the first semester

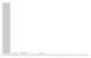
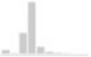
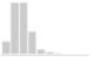
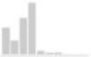
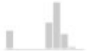
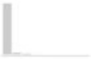
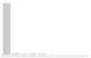
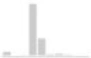
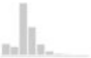
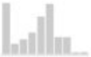
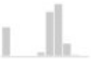
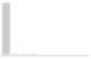
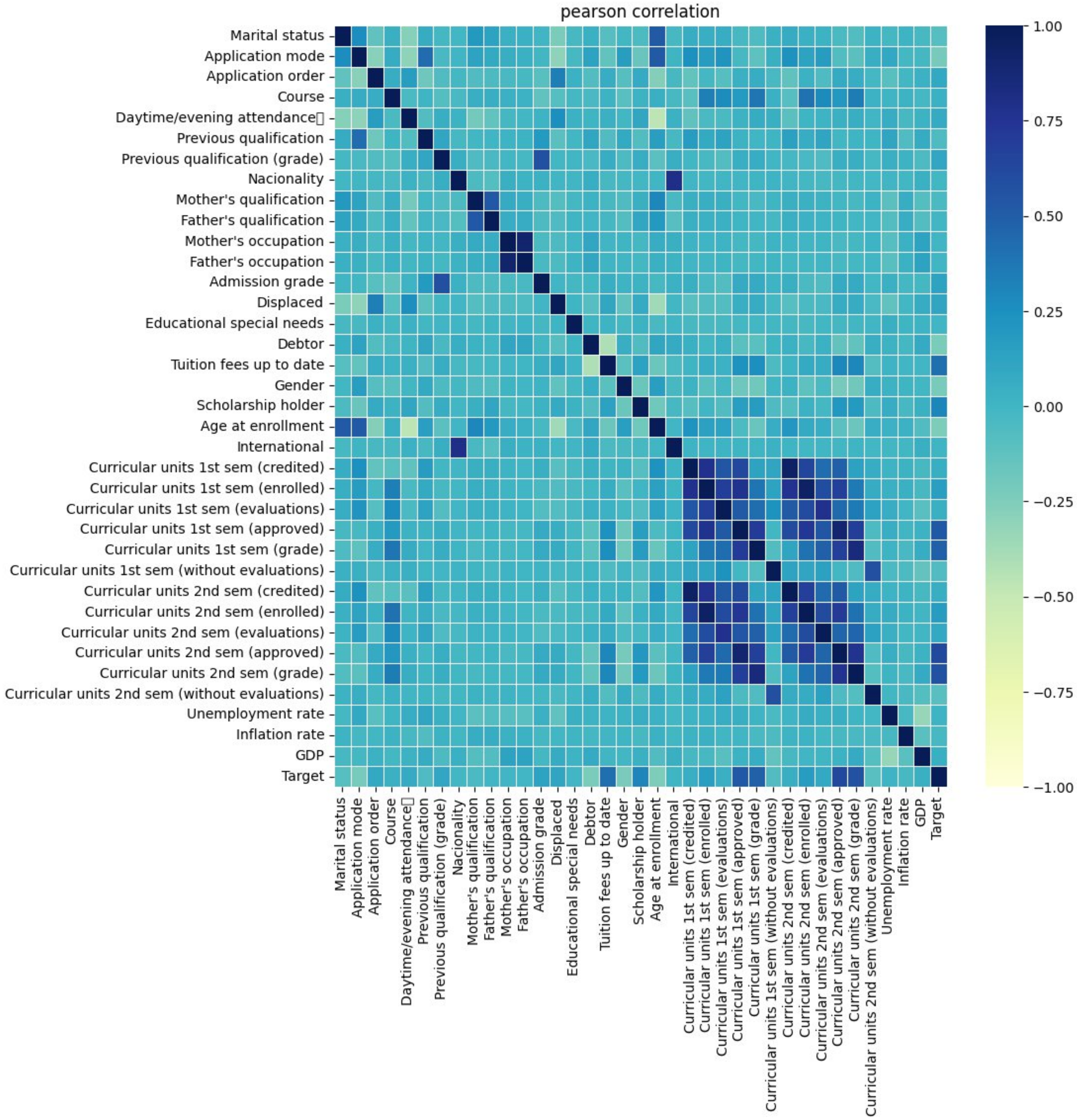
Attribute	Distrib.	Mean	Median	Dispersion	Min.	Max.
Curricular units 1st sem (credited)		0.710	0	3.320	0	20
Curricular units 1st sem (enrolled)		6.270	6	0.400	0	26
Curricular units 1st sem (evaluations)		8.300	8	0.500	0	45
Curricular units 1st sem (approved)		4.710	5	0.660	0	26
Curricular units 1st sem (grade)		10.641	12.286	0.455	0.000	18.875
Curricular units 1st sem (without evaluations)		0.140	0	5.020	0	12

Table 2.6.Basic statistics information about academic data at the end of the second semester

Attribute	Distrib.	Mean	Median	Dispersion	Min.	Max.
Curricular units 2nd sem (credited)		0.540	0	3.540	0	19
Curricular units 2nd sem (enrolled)		6.230	6	0.350	0	23
Curricular units 2nd sem (evaluations)		8.060	8	0.490	0	33
Curricular units 2nd sem (approved)		4.440	5	0.680	0	20
Curricular units 2nd sem (grade)		10.230	12.200	0.509	0.000	18.571
Curricular units 2nd sem (without evaluations)		0.150	0	5.010	0	12

The analysis of the heatmap, using the Pearson correlation coefficient, shows that there are some pairs of features having high correlation coefficients, which increases multi-collinearity in the dataset. The collinearity is strongest within the same group of features, but we can also find higher values of correlation between groups.



2.3 Data Preprocessing

Before training the machine learning models, several preprocessing steps were applied to the dataset:

2.3.1 Handling Missing Values:

Missing values in the dataset were identified and treated appropriately. Different strategies such as mean imputation, mode imputation, or removal of instances were employed based on the nature and significance of the missing values. This particular dataset not contain any missing values so, Handling missing values are not needed.

2.3.2 Feature Selection:

To enhance model performance and reduce computational complexity, feature selection techniques like correlation analysis, information gain, or stepwise regression were used to select the most relevant attributes for training the models. In this particular dataset all the attributes either have significant corelation with the target or among each other as in section 2.2 data analysis so, all the features are taken for the model training.

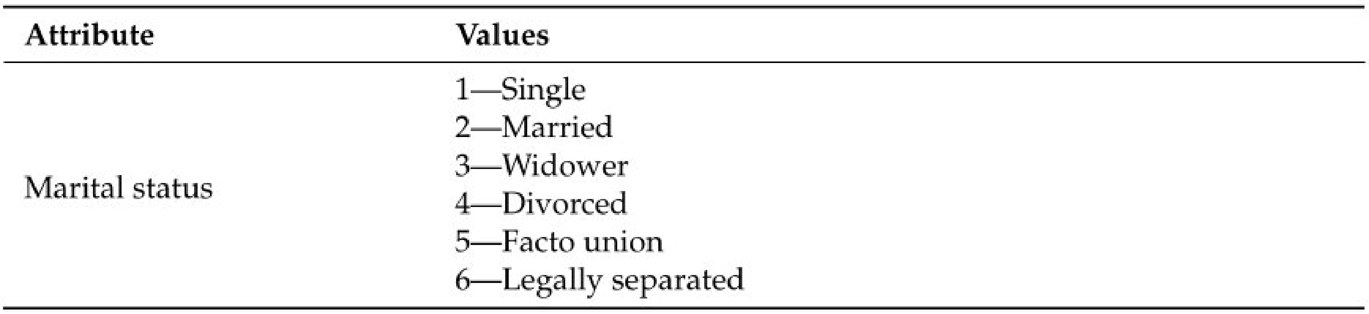
The target column contain three classes Graduate, Dropout and enrolled. As per our assertion we are predicting whether a student will dropout or not so, the number of "Enrolled" student is irrelevant because it won’t give us any new information as all the Graduate and Dropout are also enrolled. We only need to know whether a student graduated or dropedout. So, we are dropping the "Enrolled" values and going forward with "Graduate" & "Dropout" values.

2.3.3 Encoding:

Categorical variables are encoded into numerical representations to enable machine learning algorithms to process them effectively. Techniques like one-hot encoding or label encoding are employed. The Target column which contain Text values are encoded using the label encoder in the scikit learn library which randomly assign numerical values to all the unique values.

All the other attributes are already encoded in the data set as per the following table:

Table A1. Marital status values.



Attribute	Values
Marital status	1—Single 2—Married 3—Widower 4—Divorced 5—Facto union 6—Legally separated

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)

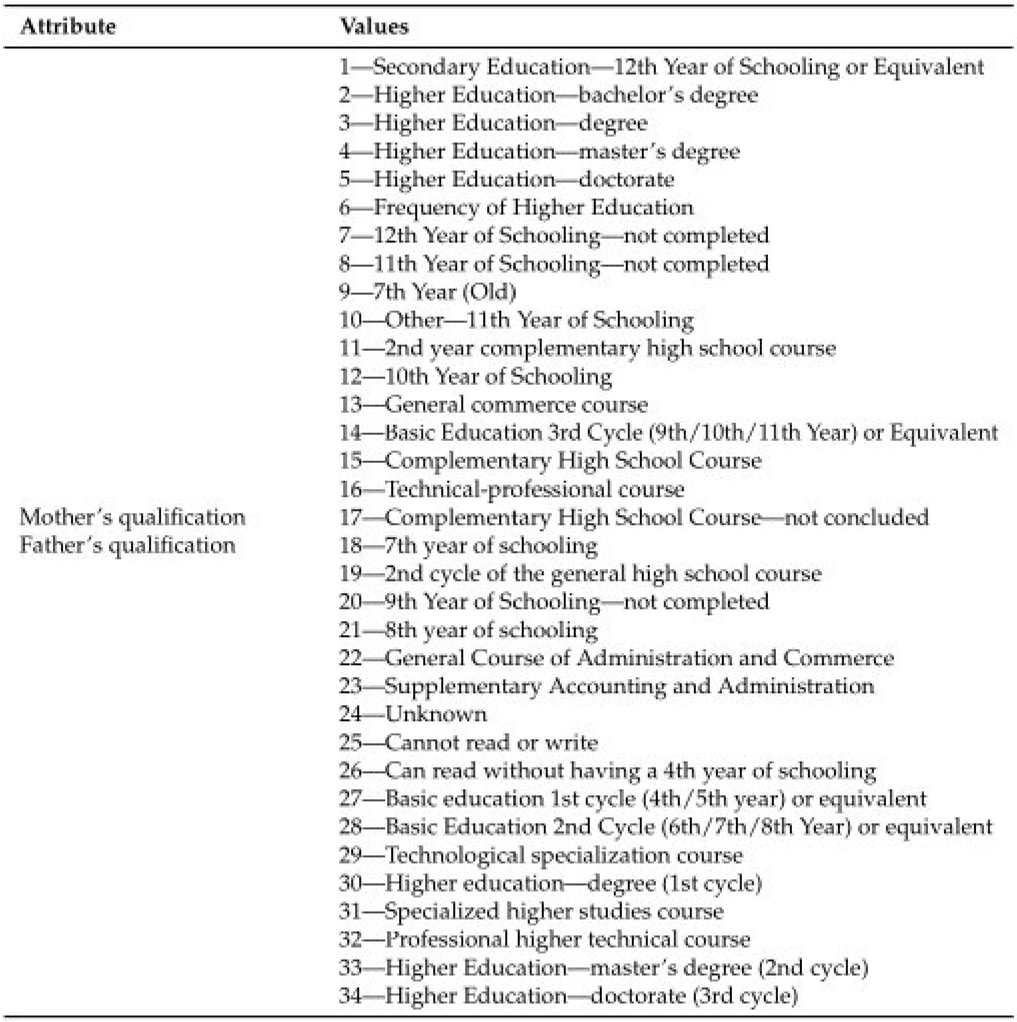
Table A4. Course values.

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 qualification 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 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 Educational special needs Debtor Tuition fees up to date Scholarship holder International	1—yes 0—no

2.3.4 Feature Scaling:

To ensure that all features have a similar scale, feature scaling techniques such as normalization or standardization are applied. In this dataset we apply standard scaler of the Sklearn library to scale the features. It Standardize features by removing the mean and scaling to unit variance.

The standard score of a sample x is calculated as:

z = (x - u) / s

where u is the mean of the training samples, and s is the standard deviation of the training sample.

2.3.5 Data Splitting:

The pre-processed dataset is divided into training and testing sets, with 80% of the data allocated for training the models and 20% for evaluating their performance. Splitting is done using the train_test_split function of the sklearn.

2.4 Model Training and evaluation

Six ML models, such as Naive-Bayes classifier, logistic regression, KNN Classifier, random forests, support vector machines (SVM), and Perceptron, are trained using the training dataset. Each model is trained with the goal of accurately predicting dropout or academic success.

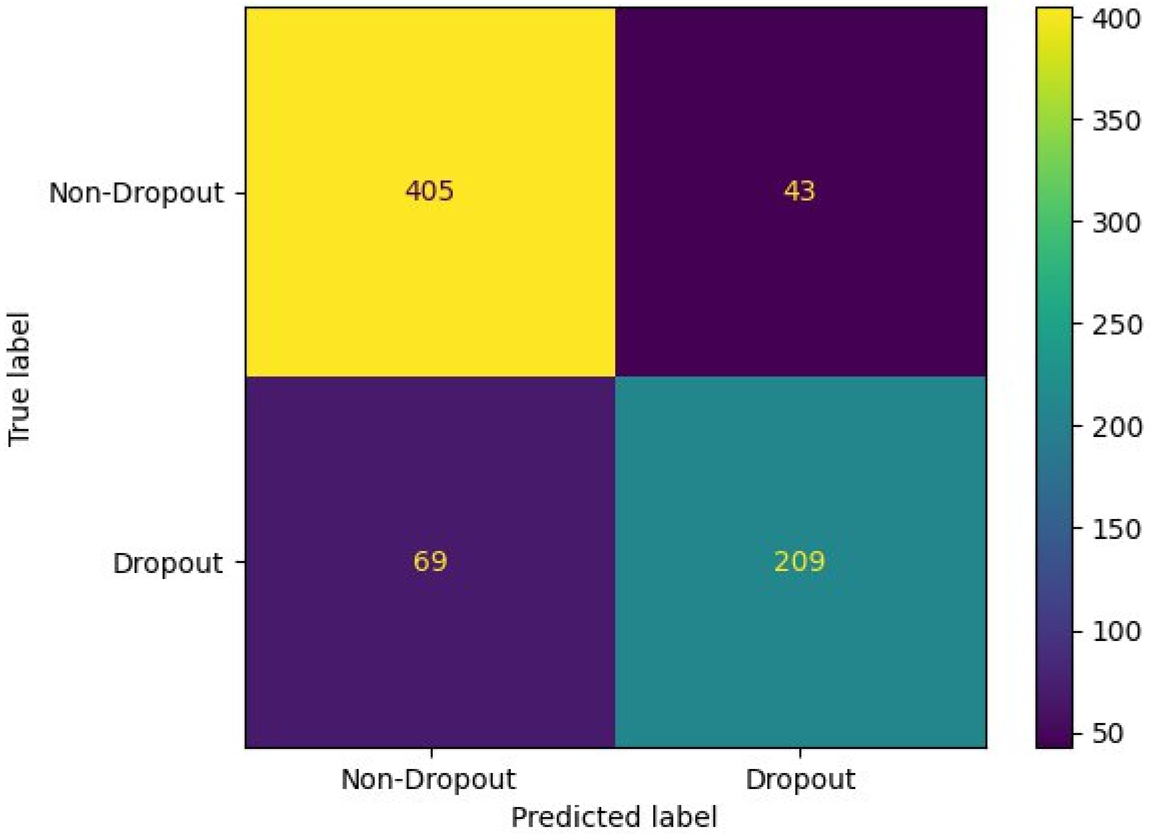
The trained models are evaluated using appropriate evaluation metrics, such as accuracy, precision, recall, and F1-score, to assess their performance and compare their effectiveness in predicting student outcomes.

2.4.1 Naive Bayes

Naive Bayes is a popular machine learning algorithm used for classification tasks. It is based on Bayes' theorem and assumes that features are conditionally independent given the class labels. Naive Bayes is particularly useful when dealing with large feature spaces and relatively small training datasets. NaiveBayes is used using the Sklearn library.

The Evaluation of model is as below:

```
******************************************************
                Classification Report
******************************************************
              precision    recall  f1-score   support

           0       0.85      0.90      0.88       448
           1       0.83      0.75      0.79       278

    accuracy                           0.85       726
   macro avg       0.84      0.83      0.83       726
weighted avg       0.84      0.85      0.84       726

******************************************************
```

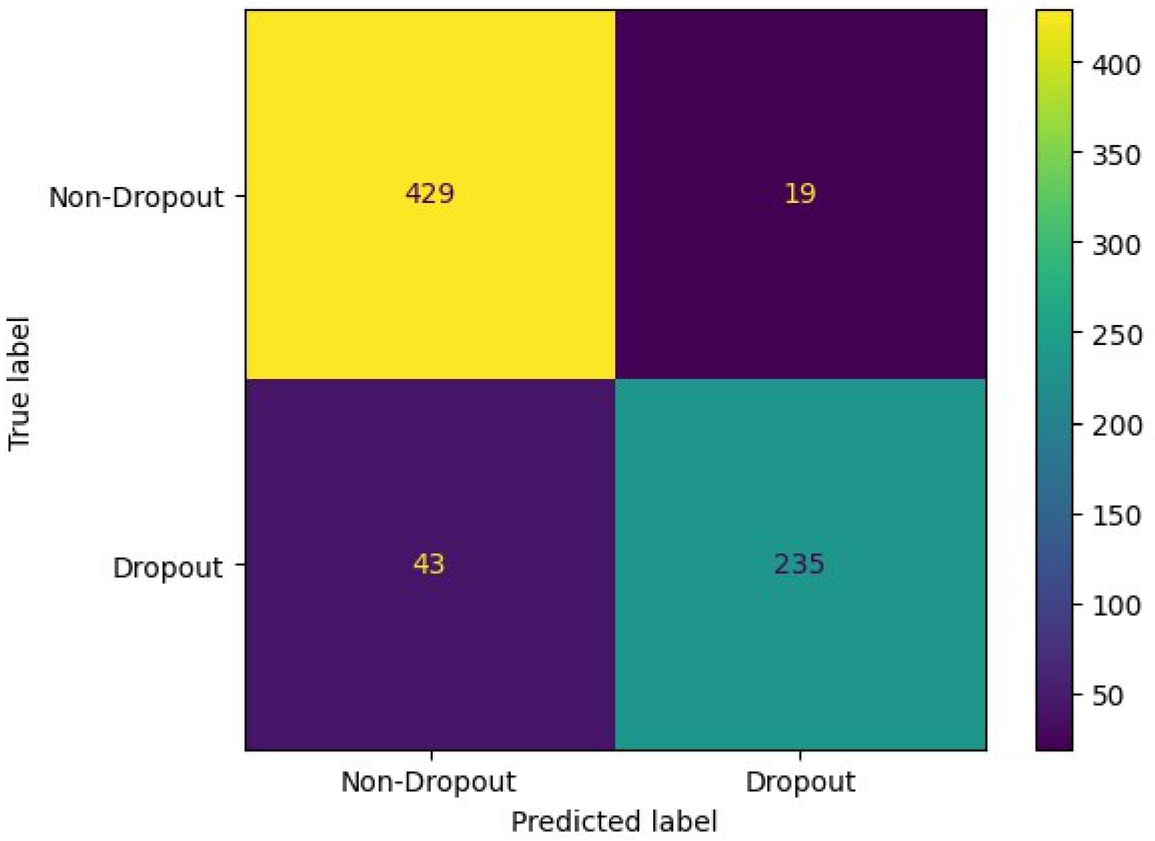


2.4.2 Logistic Regression

Logistic regression is a widely used machine learning algorithm for binary classification problems. It is a type of regression analysis that models the relationship between a set of independent variables (features) and a binary dependent variable (the target variable) using the logistic function. Since it is a binary classification problem logistic regression would be best for that.

The Evaluation of model is as below:

```
*****************************************************
               Classification Report
*****************************************************
             precision    recall  f1-score   support

           0       0.91      0.96      0.93       448
           1       0.93      0.85      0.88       278

    accuracy                           0.91       726
   macro avg       0.92      0.90      0.91       726
weighted avg       0.92      0.91      0.91       726

*****************************************************
```

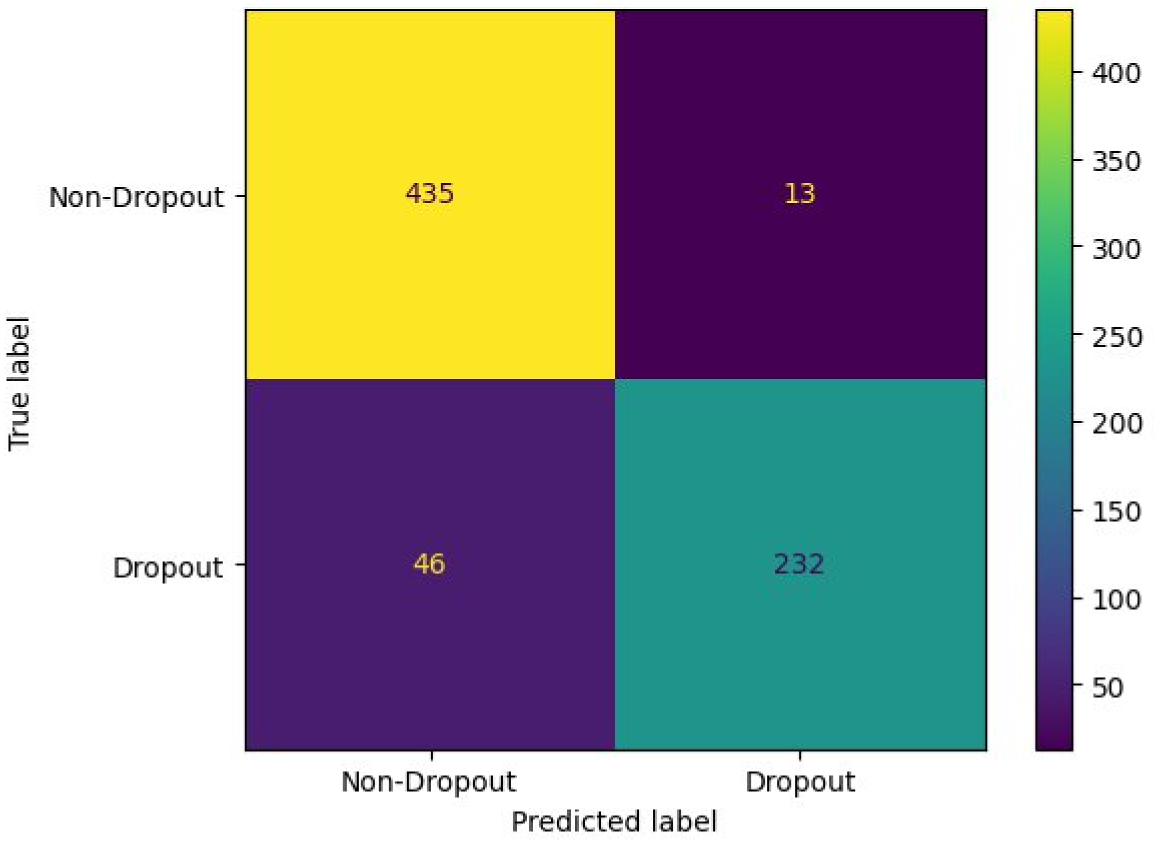


2.4.3 Random Forest

Random Forest is a popular machine learning algorithm that is commonly used for both classification and regression tasks. It is an ensemble learning method that combines multiple decision trees to make predictions. In this dataset we apply random Forest with following hyperparameters Random Forest n_estimators=500,criterion='entropy'.

The Evaluation of model is as below:

```
*******************************************************
                Classification Report
*******************************************************
              precision    recall  f1-score   support

           0       0.90      0.97      0.94       448
           1       0.95      0.83      0.89       278

    accuracy                           0.92       726
   macro avg       0.93      0.90      0.91       726
weighted avg       0.92      0.92      0.92       726

*******************************************************
```

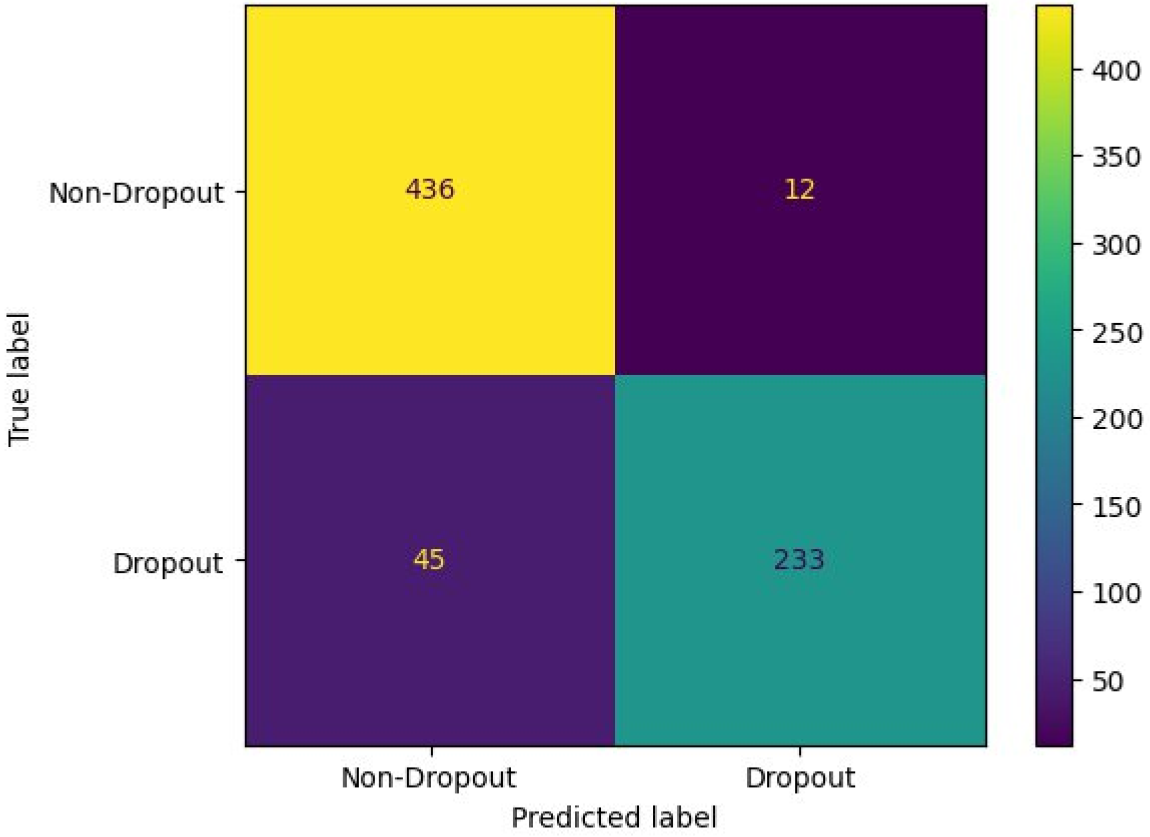


<u>2.4.4 Support vector Classifier</u>

Support Vector Classifier (SVC), also known as Support Vector Machine (SVM), is a popular machine learning algorithm used for both binary and multi-class classification tasks. It is particularly effective when dealing with complex decision boundaries and datasets with high-dimensional feature spaces. Here we apply the SVC with the following hyper parameters C=0.1,kernel='linear'.

The Evaluation of model is as below:

```
******************************************************
                Classification Report
******************************************************
              precision    recall  f1-score   support

           0       0.91      0.97      0.94       448
           1       0.95      0.84      0.89       278

    accuracy                           0.92       726
   macro avg       0.93      0.91      0.91       726
weighted avg       0.92      0.92      0.92       726

******************************************************
```

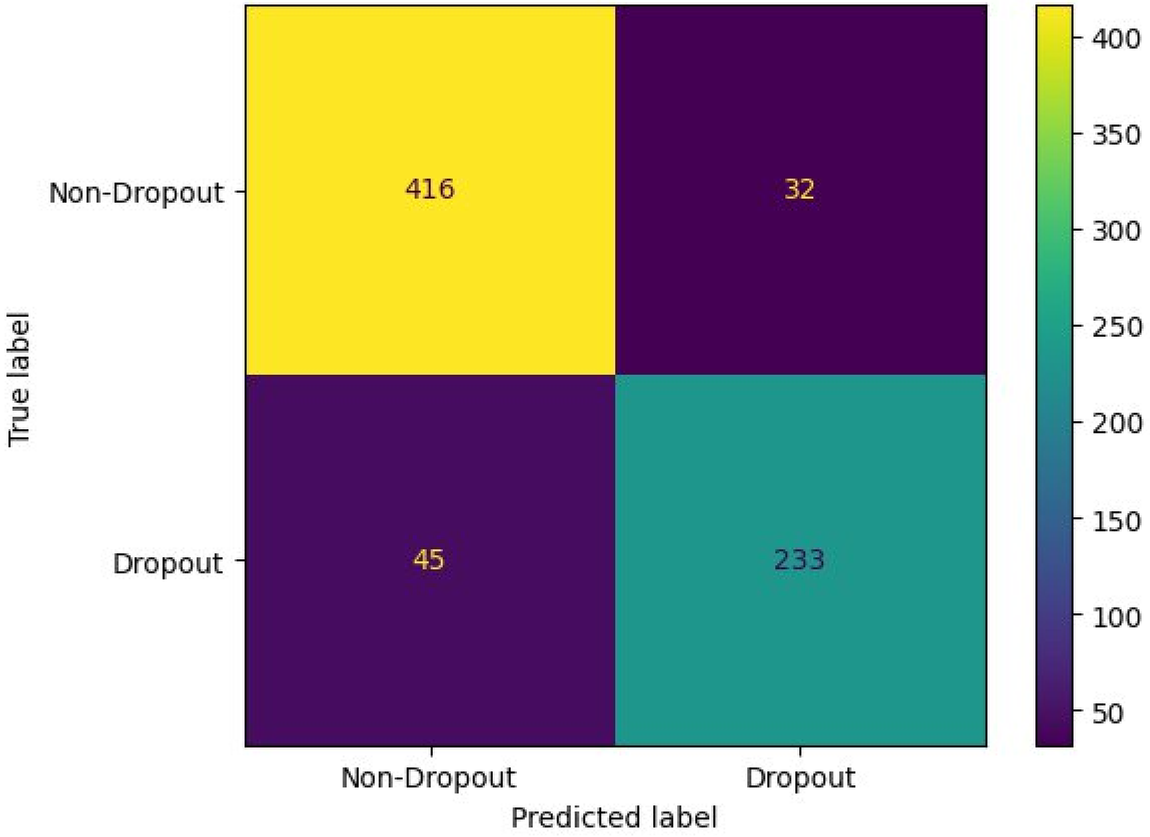


2.4.5 Perceptron

The Perceptron classifier is a linear binary classification algorithm that is used when dealing with linearly separable datasets. It is one of the earliest and simplest forms of artificial neural networks and is typically used for solving binary classification problems. Here perceptron in trained with the following hyper parameters.

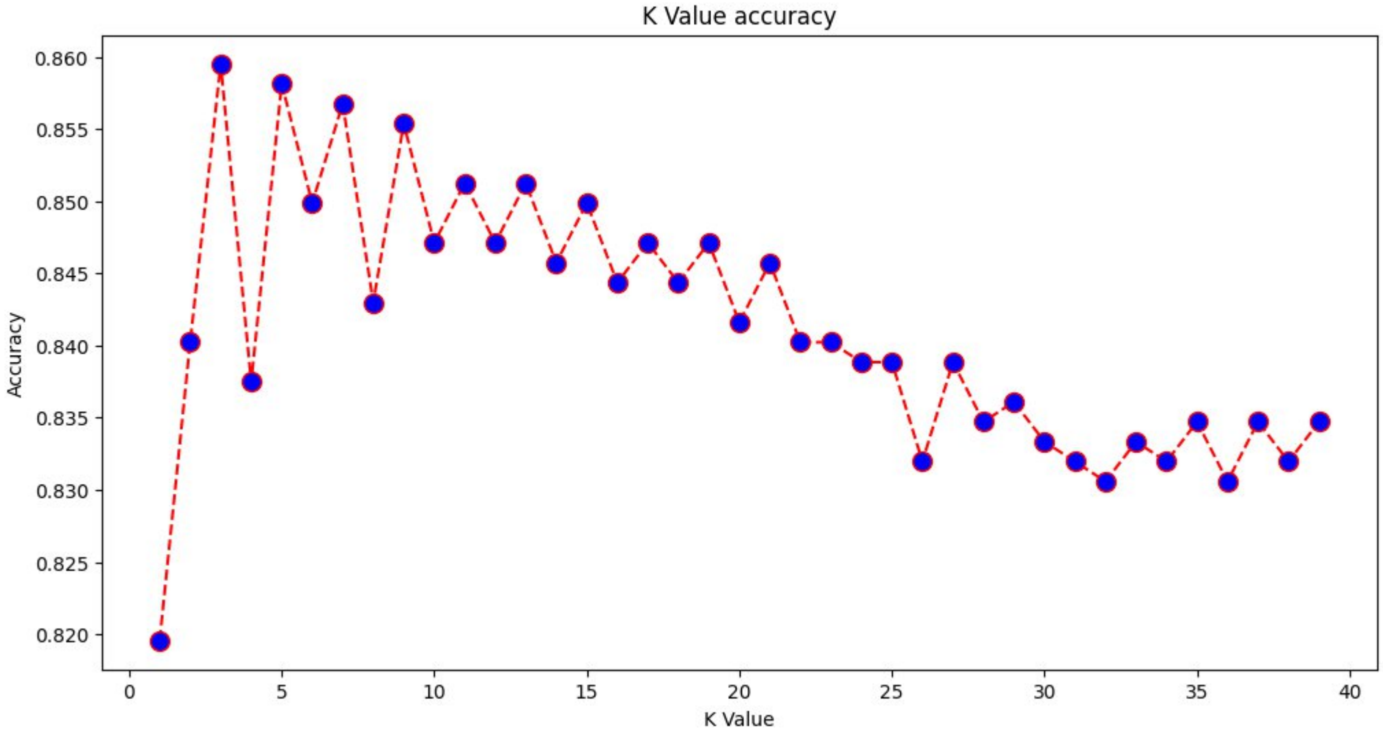
The Evaluation of model is as below:

```
******************************************************
                Classification Report
******************************************************
              precision    recall  f1-score   support

           0       0.90      0.93      0.92       448
           1       0.88      0.84      0.86       278

    accuracy                           0.89       726
   macro avg       0.89      0.88      0.89       726
weighted avg       0.89      0.89      0.89       726

******************************************************
```



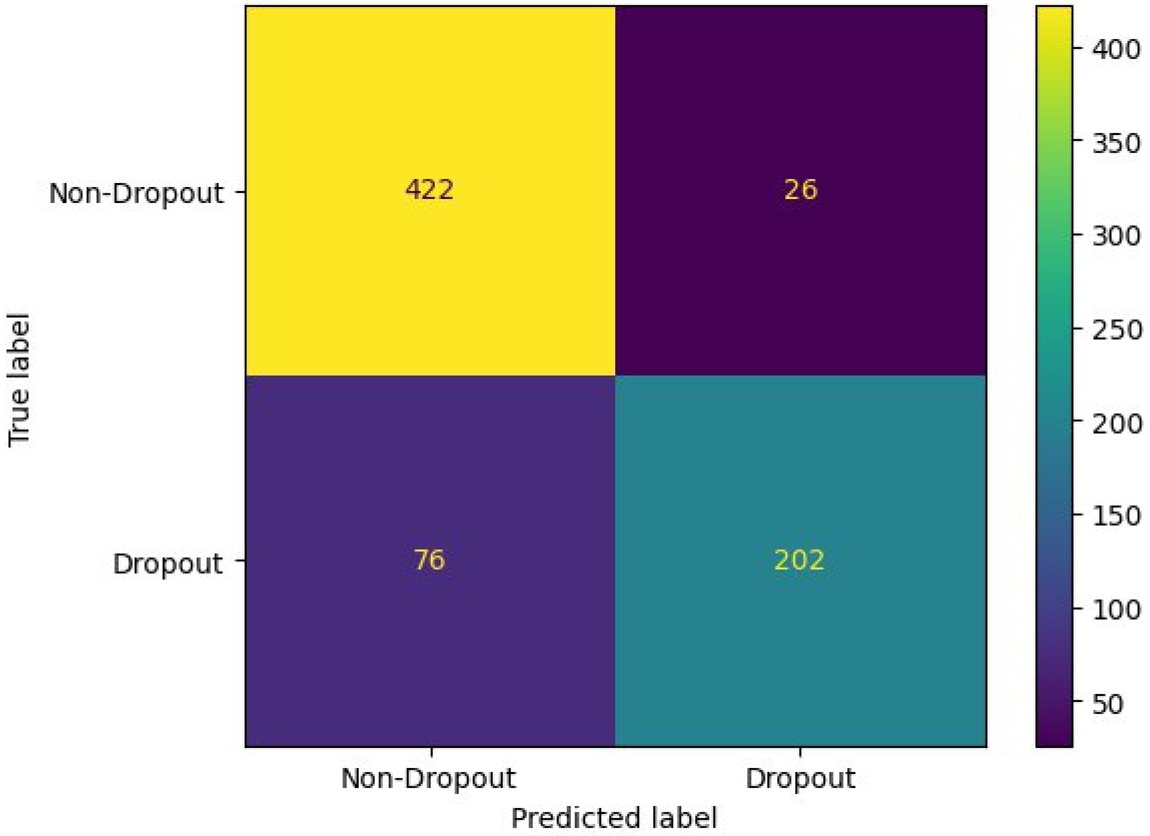
2.5.6 KNN Classifier

The K-Nearest Neighbors (KNN) classifier is a versatile machine learning algorithm used for both classification and regression tasks. It is a non-parametric algorithm that makes predictions based on the similarity of input instances to its neighboring data points. We use elbow method to find the optimal value of the K in the KNN which turn out to be 3 so, we trained the model with the hyper parameter n_neighbors=3 .

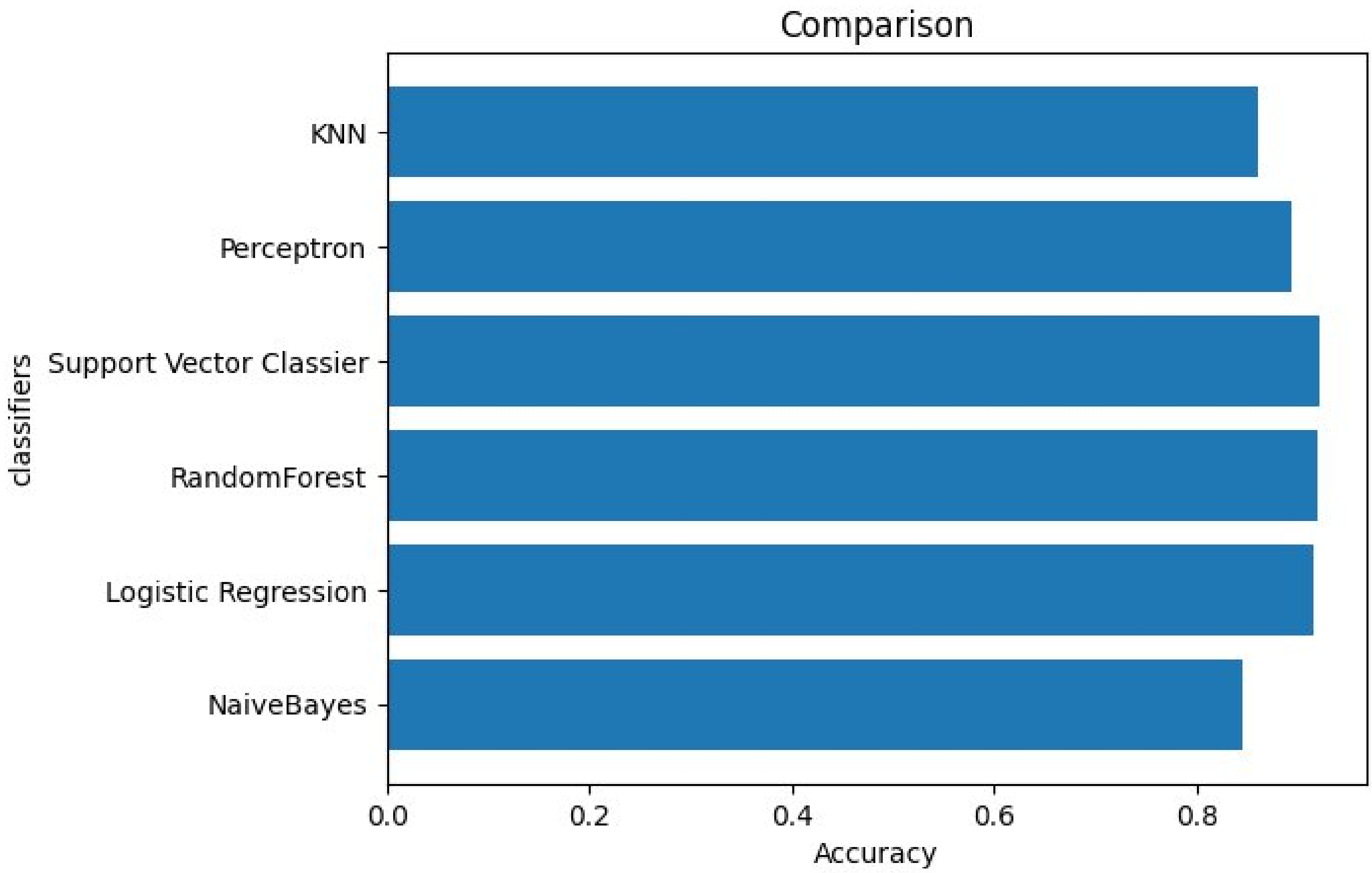


The Evaluation of model is as below:

```
******************************************************
                Classification Report
******************************************************
              precision    recall  f1-score   support

           0       0.85      0.94      0.89       448
           1       0.89      0.73      0.80       278

    accuracy                           0.86       726
   macro avg       0.87      0.83      0.85       726
weighted avg       0.86      0.86      0.86       726

******************************************************
```



2.5.7 Comparison between models



As per the above comparison graph we get the information that all the selected classifiers performs well but among the all SVC, Random Forest and Logistic Regression performs the best.

3. Result and Discussion

During the evaluation phase, the performance of the trained machine learning models for predicting students' dropout and academic success is assessed. The evaluation metrics used include accuracy, precision, recall, and F1 score, which provide insights into the models' effectiveness in classification tasks.

The results obtained from the evaluation reveal the performance of each model on the given dataset. The models are compared based on their metrics, allowing for the identification of the most suitable model for the task at hand. For example, the logistic regression model may demonstrate higher accuracy but lower recall, indicating that it correctly predicts non-dropout instances but struggles with identifying actual dropout cases. On the other hand, the decision tree model may show a higher recall but lower precision, suggesting that it captures more actual dropout cases but also produces more false positives.

Furthermore, the implications of the models' performance in real-world scenarios are discussed. The potential benefits of accurately predicting dropout and academic success are highlighted, emphasizing the importance of early intervention and targeted support for at-risk students. The limitations of the models and areas for improvement are also addressed, paving the way for future work in the field.

4. Conclusion and Future Work

In conclusion, this project successfully develops and evaluates machine learning models for predicting students' dropout and academic success using the provided dataset. The conclusions drawn from the results highlight the potential applications of machine learning in the education domain. By effectively predicting dropout and academic success, educational institutions can implement proactive measures to support students and improve their overall outcomes.

In terms of future work, several avenues can be explored to enhance the accuracy and robustness of the predictive models. Firstly, additional relevant features can be incorporated into the dataset to capture a more comprehensive understanding of students' backgrounds and experiences. This could include factors such as socio-economic status, extracurricular activities, or family support systems.

Secondly, experimenting with different machine learning algorithms or ensemble methods can be undertaken to identify the most optimal approach for predicting dropout and academic success. Techniques like gradient boosting, deep learning, or recurrent neural networks may yield better performance or uncover hidden patterns in the data.

Lastly, integrating real-time data and implementing a feedback loop system would enable continuous model refinement and adaptation to evolving student dynamics. This would provide more up-to-date predictions and allow for timely interventions.

By addressing these areas of improvement, the accuracy and applicability of the predictive models can be further enhanced, leading to more effective strategies for preventing dropout and promoting academic success in educational settings.

5. References

1. https://link.springer.com/chapter/10.1007/978-3-030-52237-7_11
2. http://dspace.nm-aist.ac.tz/handle/20.500.12479/71
3. https://onlinelibrary.wiley.com/doi/abs/10.1111/exsy.12135
4. https://www.mecs-press.net/ijeme/ijeme-v7-n2/IJEME-V7-N2-2.pdf
5. https://dl.acm.org/doi/abs/10.1145/3388792
6. https://ieeexplore.ieee.org/abstract/document/8363340
7. https://www.sciencedirect.com/science/article/abs/pii/S0360131509001249
8. https://www.tandfonline.com/doi/abs/10.1080/10494820.2020.1802300
9. https://ieeexplore.ieee.org/abstract/document/8398887
10. https://journals.sagepub.com/doi/abs/10.1177/0735633118757015?journalCode=jeca
11. https://www.mdpi.com/2076-3417/9/15/3093
12. https://link.springer.com/chapter/10.1007/978-3-319-27000-5_12
13. https://www.sciencedirect.com/science/article/abs/pii/S0167923623000155
14. https://books.google.co.in/books?hl=en&lr=&id=USGLCgAAQBAJ&oi=fnd&pg=PA319&dq=student+dropout+prediction&ots=FtkekmJSTH&sig=ExjooPHfEkc_AaoRs38TSfyr_1U&redir_esc=y#v=onepage&q=student%20dropout%20prediction&f=false
15. https://www.mdpi.com/2079-9292/10/14/1701