REPORT
INTELLIGENT DATA ANALYSIS
SEC 701

Project Topic:

Predict Students' Dropout and Academic Success using Classification Models

LECTURER - ASST. PROF. DR. PRAPAPORN TECHAANGKOON

Linn Htet Aung

671615515

Linnhtetaung_1@cmu.ac.th

Predictions of Student Dropout or success using Classification Models

Abstract: Higher Education institutes collect a vast amount of student data to closely monitor student's performance to provide necessary support. However, not all students in universities and colleges graduate within the academic period. Some students drop out of school. Student dropouts pose a significant challenge, contributing to higher unemployment rates and affecting not only the individuals but also their families and the economy. This dataset contains various variables of demographic, socioeconomic conditions, and academic data. This is used to build machine learning: various classification models to predict the students at risk of dropping out or achieving academic success. The models are evaluated using accuracy, recall, precision, and F1 score to decide the best model. The results will help educational institutions implement policy and plans to reduce dropout rates and support student success. The findings are expected to contribute to better academic planning and policy development aimed at improving student retention and outcomes.

Introduction

The economy is evolving rapidly in the technology sector and different sectors, and the demand for highly skilled labor is increasing. For greater productivity and efficiency, higher education is paramount in shaping the workforce. Those who are equipped with advanced skills and knowledge are called skilled labor. They are important in economic development. However, higher institutions student dropout rates pose one of the significant challenges to the economy.

There are a lot of students who drop out of university around the world, around 40% of college students in the United States drop out before completing their degrees (Vardishvili, 2024). This high dropout rate poses a critical concern, not only for educational institutions but for society and the economy. College dropouts often face considerable disadvantages in the labor market. On average, they earn $21,000 less per year compared to their counterparts with college degrees, translating into 35% lower annual earnings (ThinkImpact, 2021). The lack of a degree limits the access to higher-paying jobs, which are often reserved for skilled professionals in fields such as healthcare and education.

Research shows that individuals with only a high school diploma have a 12.7% higher chance of living in poverty, compared to the 4.8% poverty rate for bachelor's degree holders (EDI, 2021). This income gap contributes to broader societal inequalities and impedes economic progress. College dropouts are also found to have lower levels of financial literacy, which can further exacerbate their financial challenges (Research.com, 2021).

This project aims to predict which students are at risk of dropping out by using machine learning techniques. By analyzing a combination of demographic, socioeconomic, and academic data, the models developed in this study will provide insights that can help universities implement targeted interventions and support strategies. The goal is to reduce dropout rates and support students in successfully completing their education, thereby contributing to a more skilled and economically productive workforce.

This project leverages data from the UCI Machine Learning Repository's ***"Predict Students' Dropout and Academic Success"*** *https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success*.

The dataset encompasses various demographic, socioeconomic, and academic variables that help build predictive models to identify students at risk of dropping out. These variables include student grades, course enrollments, financial aid information, parental backgrounds, and other socioeconomic factors.

This project focuses solely on predicting dropouts to help higher education institutions implement effective strategies to reduce dropout rates. These strategies could include providing scholarships, offering financial incentives, and delivering targeted academic support to students who are identified as high-risk. The machine learning models will be evaluated based on their accuracy, recall, precision, and F1 score, with the goal of offering educational institutions actionable insights to enhance student retention and success.

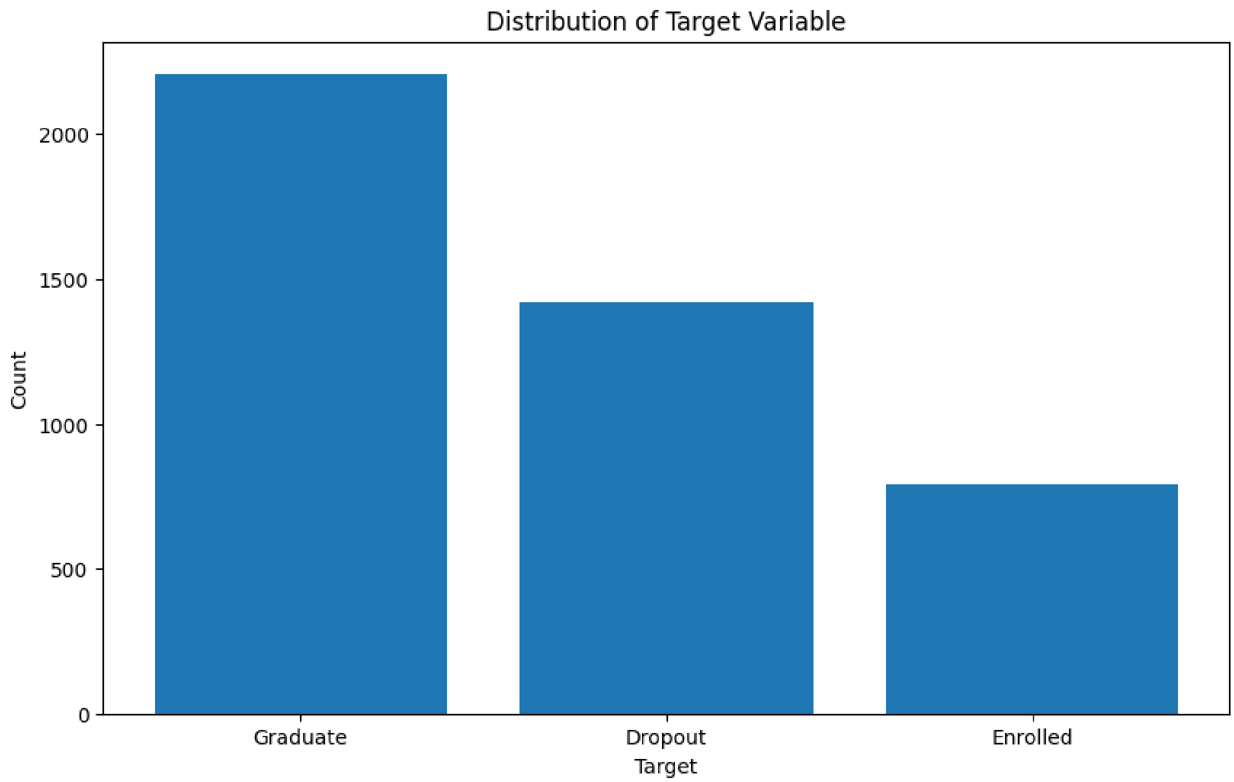
Method

The dataset contains 36 variables from student grades, course enrollments, financial aid information, parental backgrounds, and other socioeconomic factors. The target values are categorical, therefore; we will use classification models to predict the outcomes. The classification methods (KNN, Navie Bayes, Logistics Regression and Decision Tree) are used to train and evaluated. Before proceeding to the model training method, dataset must be checked if there are any missing value and duplications. However, this dataset has no missing value and duplications. The steps of handling missing value are not needed.

Dataset information.

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4424 entries, 0 to 4423
Data columns (total 37 columns):
 #   Column                                          Non-Null Count  Dtype  
---  ------                                          --------------  -----  
 0   Marital status                                  4424 non-null   int64  
 1   Application mode                                4424 non-null   int64  
 2   Application order                               4424 non-null   int64  
 3   Course                                          4424 non-null   int64  
 4   Daytime/evening attendance                      4424 non-null   int64  
 5   Previous qualification                          4424 non-null   int64  
 6   Previous qualification (grade)                  4424 non-null   float64
 7   Nacionality                                     4424 non-null   int64  
 8   Mother's qualification                          4424 non-null   int64  
 9   Father's qualification                          4424 non-null   int64  
 10  Mother's occupation                             4424 non-null   int64  
 11  Father's occupation                             4424 non-null   int64  
 12  Admission grade                                 4424 non-null   float64
 13  Displaced                                       4424 non-null   int64  
 14  Educational special needs                       4424 non-null   int64  
 15  Debtor                                          4424 non-null   int64  
 16  Tuition fees up to date                         4424 non-null   int64  
 17  Gender                                          4424 non-null   int64  
 18  Scholarship holder                              4424 non-null   int64  
 19  Age at enrollment                               4424 non-null   int64  
 20  International                                   4424 non-null   int64  
 21  Curricular units 1st sem (credited)             4424 non-null   int64  
 22  Curricular units 1st sem (enrolled)             4424 non-null   int64  
 23  Curricular units 1st sem (evaluations)          4424 non-null   int64  
 24  Curricular units 1st sem (approved)             4424 non-null   int64  
 25  Curricular units 1st sem (grade)                4424 non-null   float64
 26  Curricular units 1st sem (without evaluations)  4424 non-null   int64  
 27  Curricular units 2nd sem (credited)             4424 non-null   int64  
 28  Curricular units 2nd sem (enrolled)             4424 non-null   int64  
 29  Curricular units 2nd sem (evaluations)          4424 non-null   int64  
 30  Curricular units 2nd sem (approved)             4424 non-null   int64  
 31  Curricular units 2nd sem (grade)                4424 non-null   float64
 32  Curricular units 2nd sem (without evaluations)  4424 non-null   int64  
 33  Unemployment rate                               4424 non-null   float64
 34  Inflation rate                                  4424 non-null   float64
 35  GDP                                             4424 non-null   float64
 36  Target                                          4424 non-null   object 
dtypes: float64(7), int64(29), object(1)
memory usage: 1.2+ MB
```

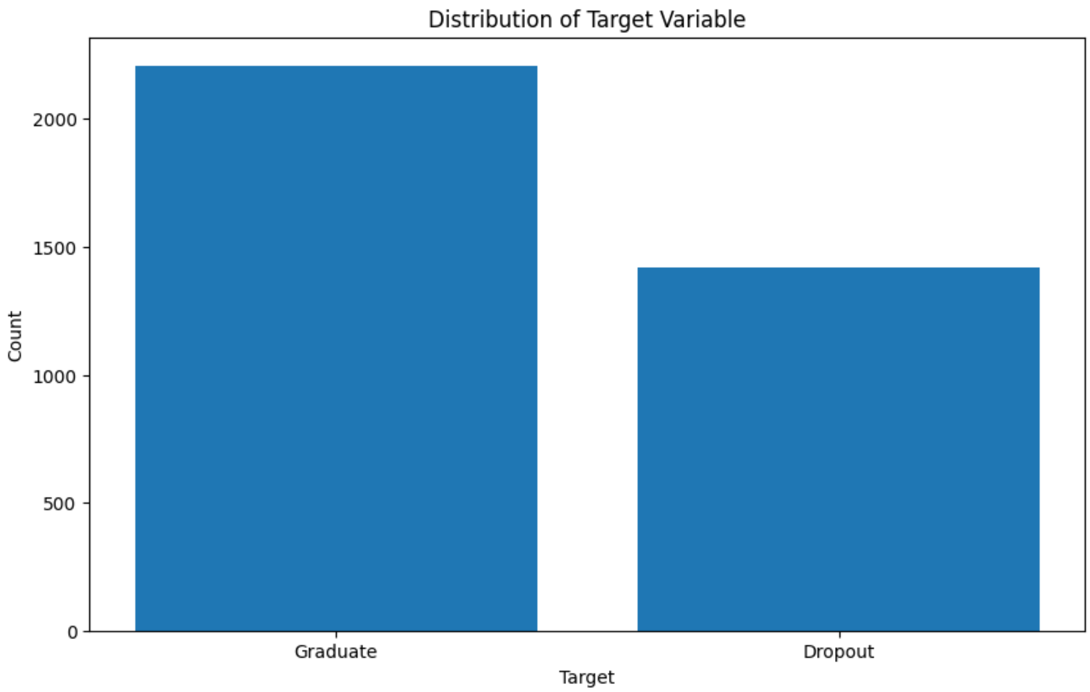


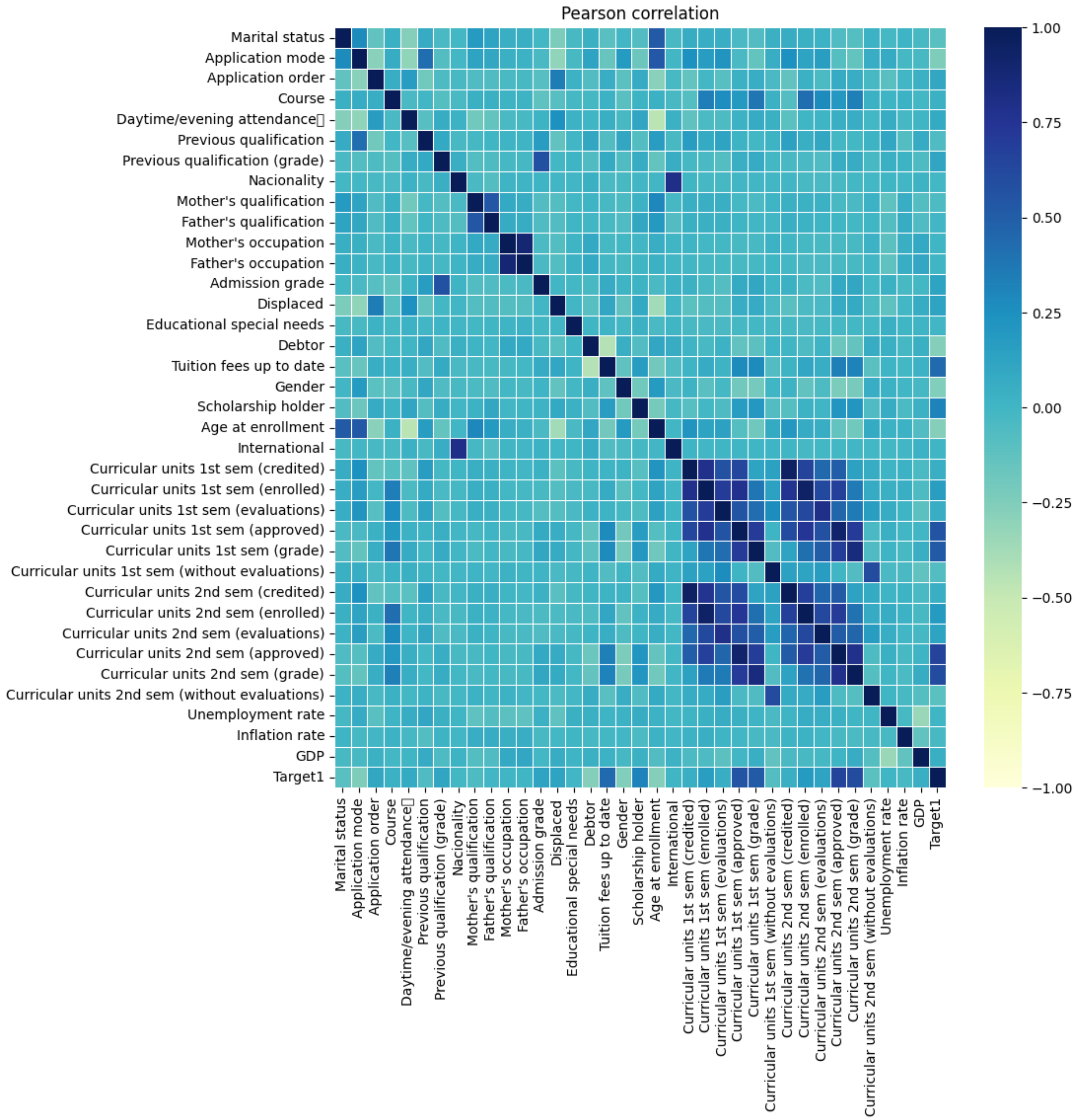

```
[84] data['Target'].value_counts()
```

Target	count
Graduate	2209
Dropout	1421
Enrolled	794

dtype: int64

The unique value of Target column is being generated by using .value_count(). There are 2209 Graduate,1421 Dropout. However, the model is to find whether students dropout or success. Therefore, enrolled value is dropped.



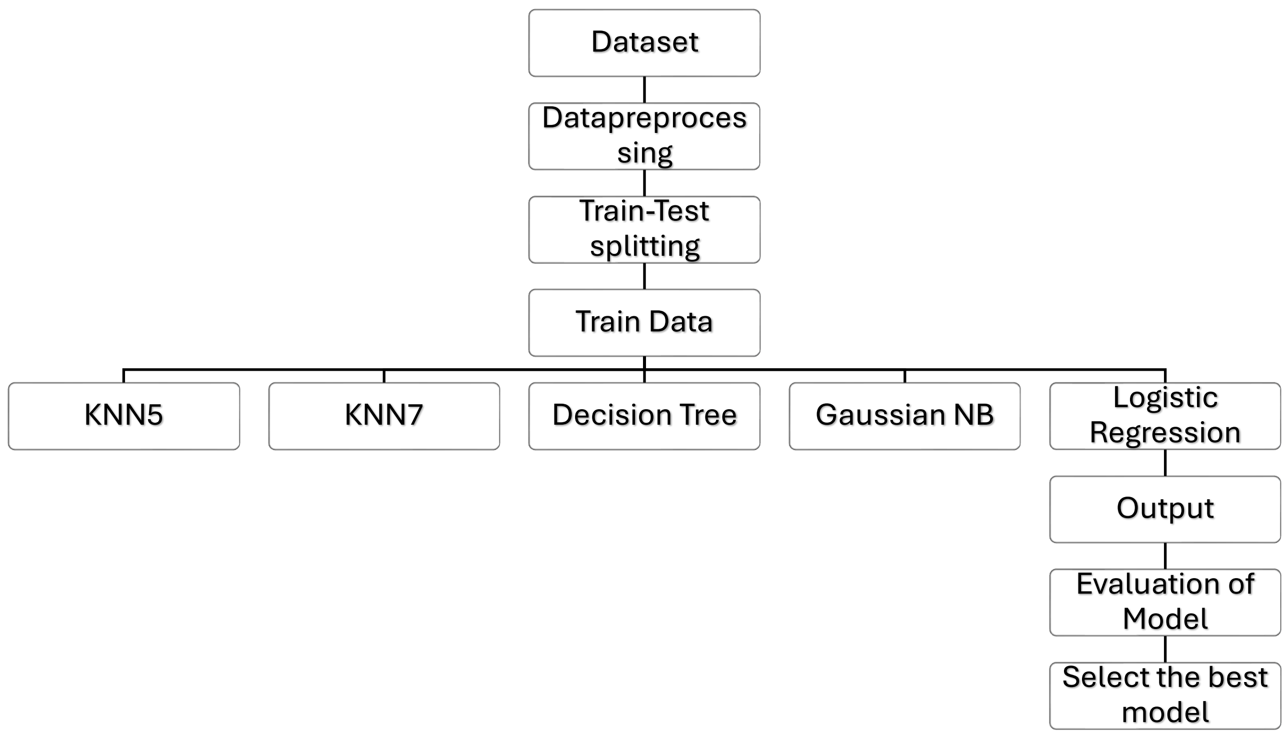

Pearson correlation
Marital status
Application mode
Application order
Course
Daytime/evening attendance
Previous qualification
Previous qualification (grade)
Nacionality
Mother's qualification
Father's qualification
Mother's occupation
Father's occupation
Admission grade
Displaced
Educational special needs
Debtor
Tuition fees up to date
Gender
Scholarship holder
Age at enrollment
International
Curricular units 1st sem (credited)
Curricular units 1st sem (enrolled)
Curricular units 1st sem (evaluations)
Curricular units 1st sem (approved)
Curricular units 1st sem (grade)
Curricular units 1st sem (without evaluations)
Curricular units 2nd sem (credited)
Curricular units 2nd sem (enrolled)
Curricular units 2nd sem (evaluations)
Curricular units 2nd sem (approved)
Curricular units 2nd sem (grade)
Curricular units 2nd sem (without evaluations)
Unemployment rate
Inflation rate
GDP
Target1
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

After that, I made some scaling for the data for smooth algorithm by using standaization. The data is standardized using StandardScaler () from scikit-learn. This preprocessing ensures that features are on the same scale, which improves the performance and convergence of machine learning models.

```
from sklearn.preprocessing import LabelEncoder, StandardScaler
X = data[["Tuition fees up to date",
          "Curricular units 1st sem (approved)",
          "Curricular units 1st sem (grade)",
          "Curricular units 2nd sem (approved)",
          "Curricular units 2nd sem (grade)"]].values
print(X)
X = StandardScaler().fit_transform(X)
X
```

```
[[ 1.          0.          0.          0.          0.        ]
 [ 0.          6.         14.          6.         13.66666667]
 [ 0.          0.          0.          0.          0.        ]
 ...
 [ 1.          7.         14.9125      1.         13.5       ]
 [ 1.          5.         13.8         5.         12.        ]
 [ 1.          6.         11.66666667  6.         13.        ]]
array([[ 0.39316683, -1.48003375, -2.08322431, -1.42901395, -1.83108537],
       [-2.54344959,  0.37330582,  0.68521698,  0.46855487,  0.66238282],
       [-2.54344959, -1.48003375, -2.08322431, -1.42901395, -1.83108537],
       ...,
       [ 0.39316683,  0.68219574,  0.86566003, -1.11275248,  0.63197467],
       [ 0.39316683,  0.06441589,  0.64566782,  0.1522934 ,  0.35830134],
       [ 0.39316683,  0.37330582,  0.2238101 ,  0.46855487,  0.54075023]])
```

Steps

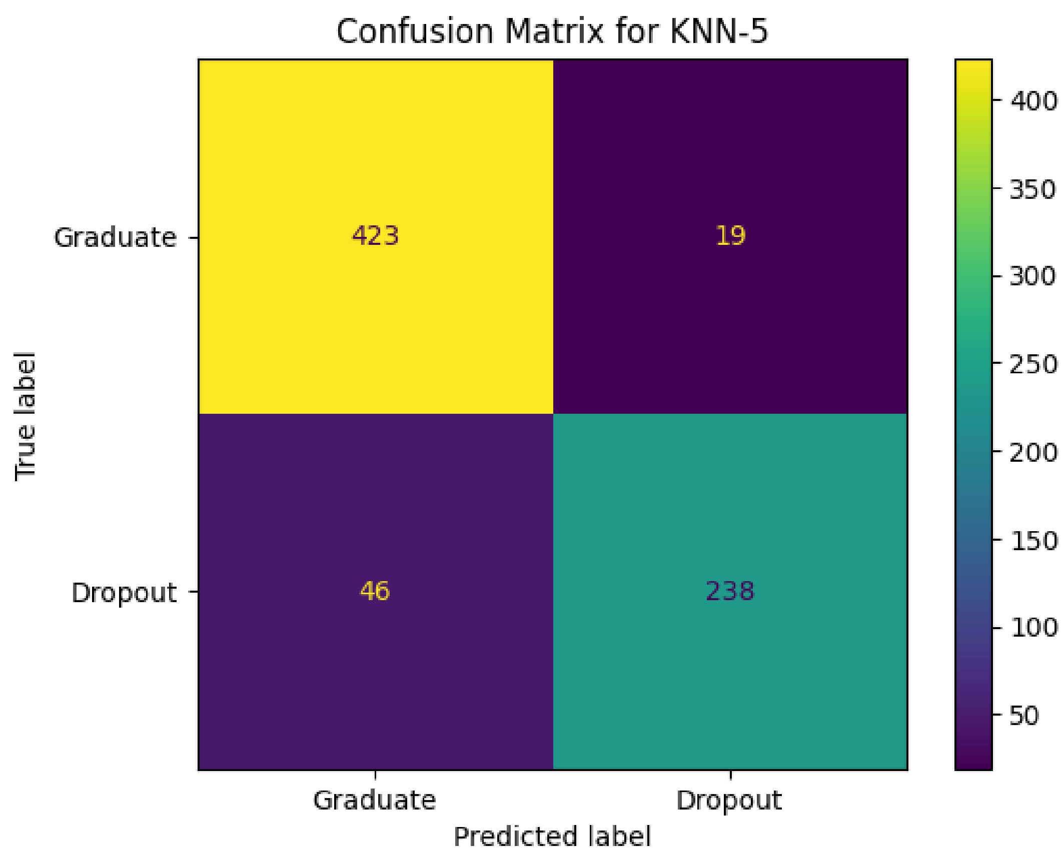

Dataset
Datapreprocessing
Train-Test splitting
Train Data
KNN5
KNN7
Decision Tree
Gaussian NB
Logistic Regression
Output
Evaluation of Model
Select the best model

Model Training and Evaluation

Model	Accuracy	Precision	Recall	F1-Score
KNN-5	0.910468	0.92607	0.838028	0.879852
KNN-7	0.915978	0.930502	0.848592	0.887661
Tree	0.867769	0.815436	0.855634	0.835052
GaussianNB	0.867769	0.894958	0.75	0.816092
LogReg	0.926997	0.913978	0.897887	0.905861

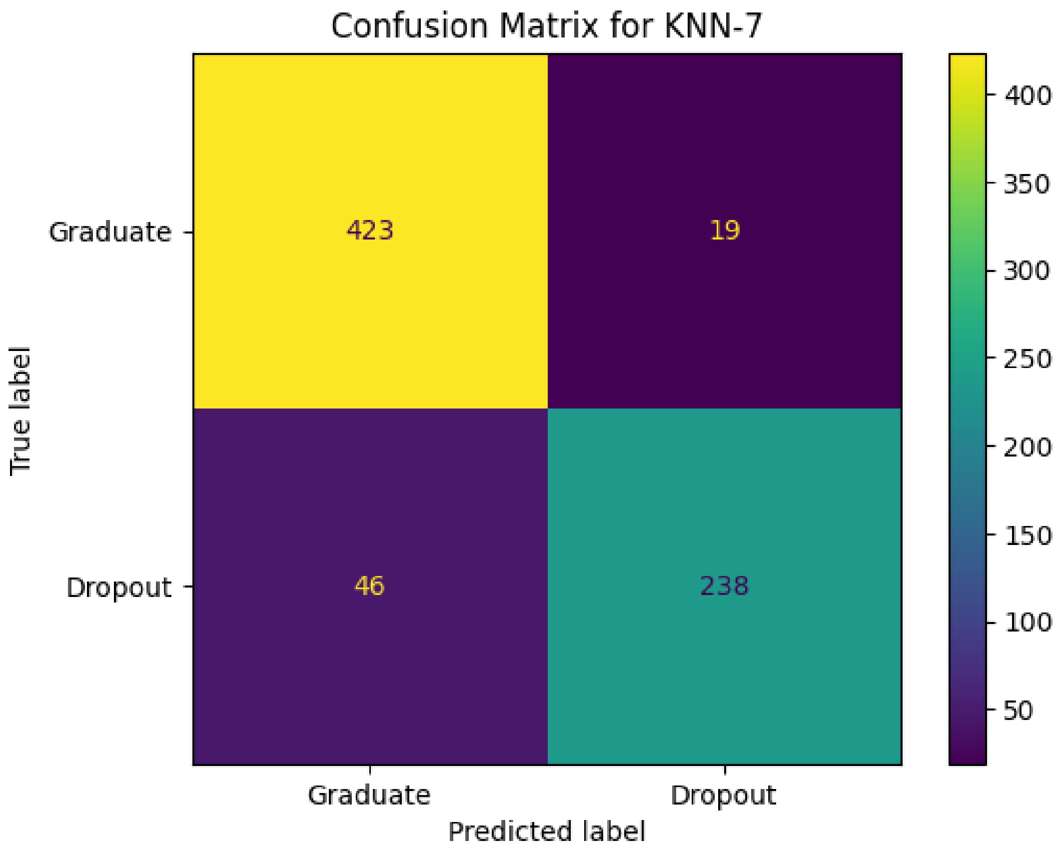
KNN-5

- **Accuracy: 91.05% – The model makes correct predictions about 91% of the time.**
- **Precision: 92.61% – When predicting positives, it is correct 92.61% of the time.**
- **Recall: 83.80% – It captures 83.80% of actual positives, slightly lower than precision.**
- **F1-Score: 87.99% – This metric indicates a solid balance between precision and recall.**



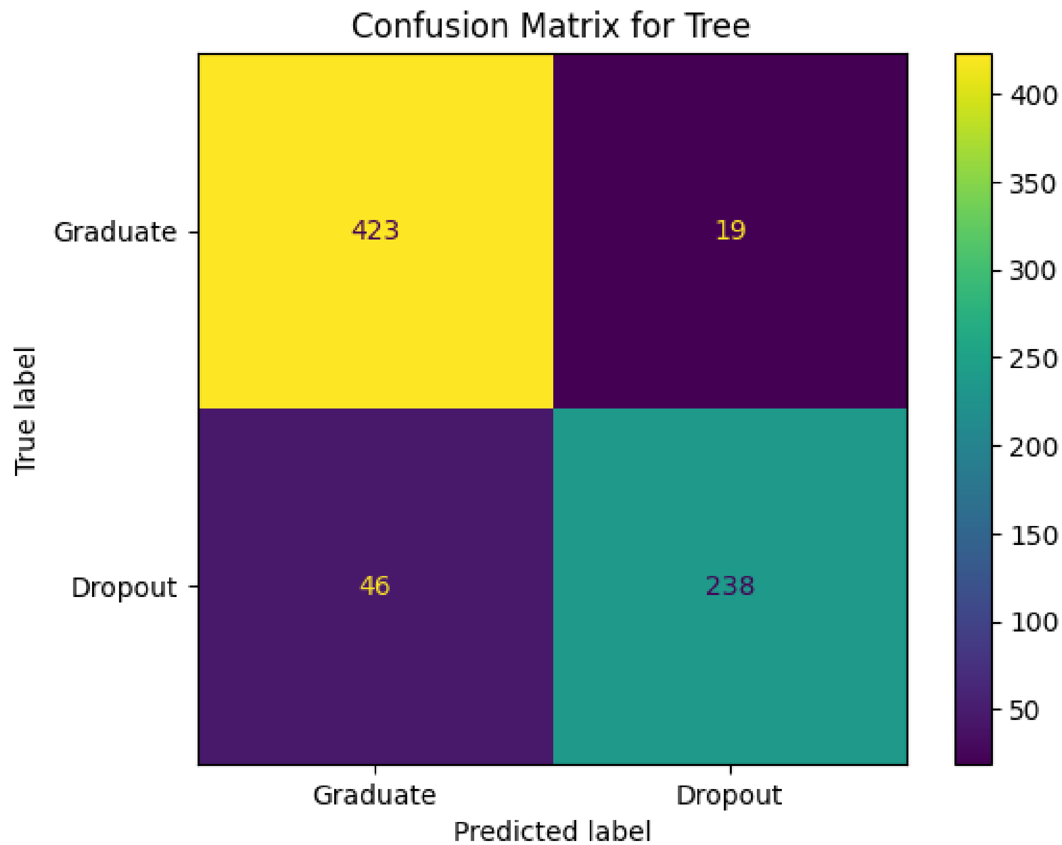
KNN-7

- **Accuracy: 91.60% – The model performs better, with an accuracy of 91.60%.**
- **Precision: 93.05% – It has higher precision compared to KNN-5, making correct positive predictions 93.05% of the time.**
- **Recall: 84.86% – It captures 84.86% of the actual positives, an improvement over KNN-5.**
- **F1-Score: 88.77% – The F1-Score reflects the best overall balance between precision and recall among the models.**



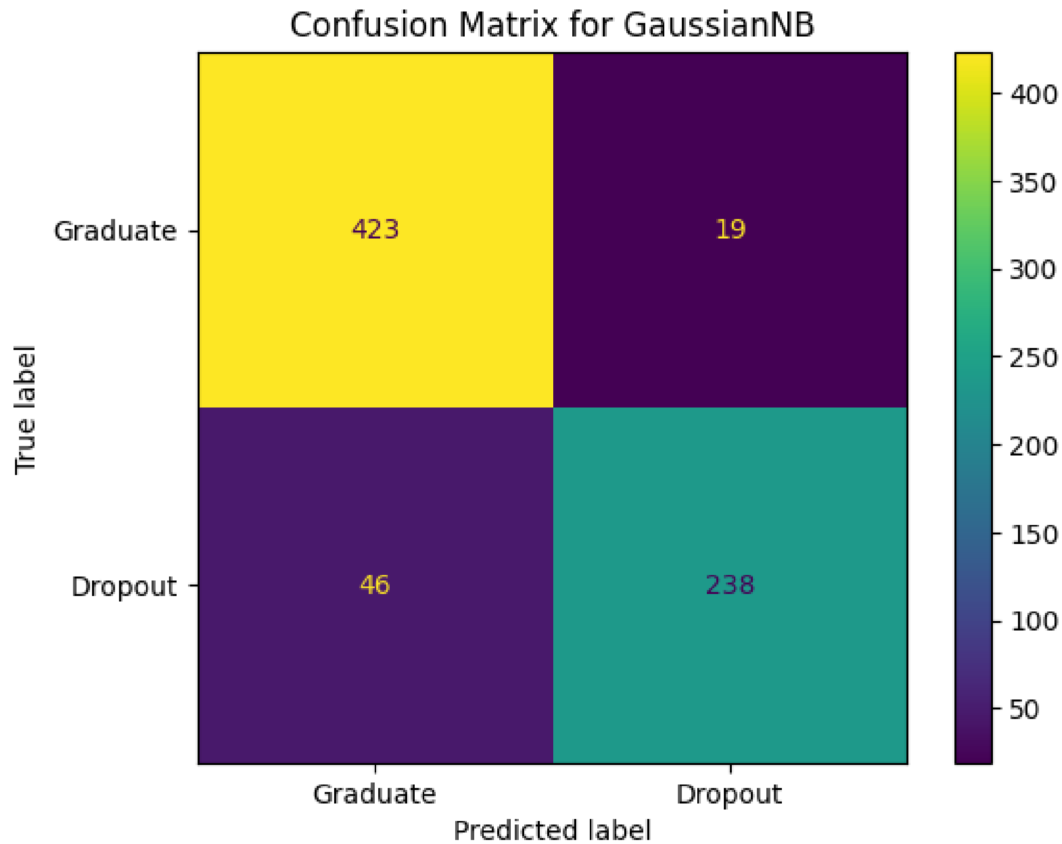
Decision Tree

- **Accuracy: 86.78% – This model is less accurate than the KNN models.**
- **Precision: 81.54% – It is more precise in predicting positives, though not as reliable as KNN or Logistic Regression.**
- **Recall: 85.56% – It captures more positives than KNN-5, but less than KNN-7 and Logistic Regression.**
- **F1-Score: 83.51% – A lower F1-Score suggests that the model is not as balanced.**



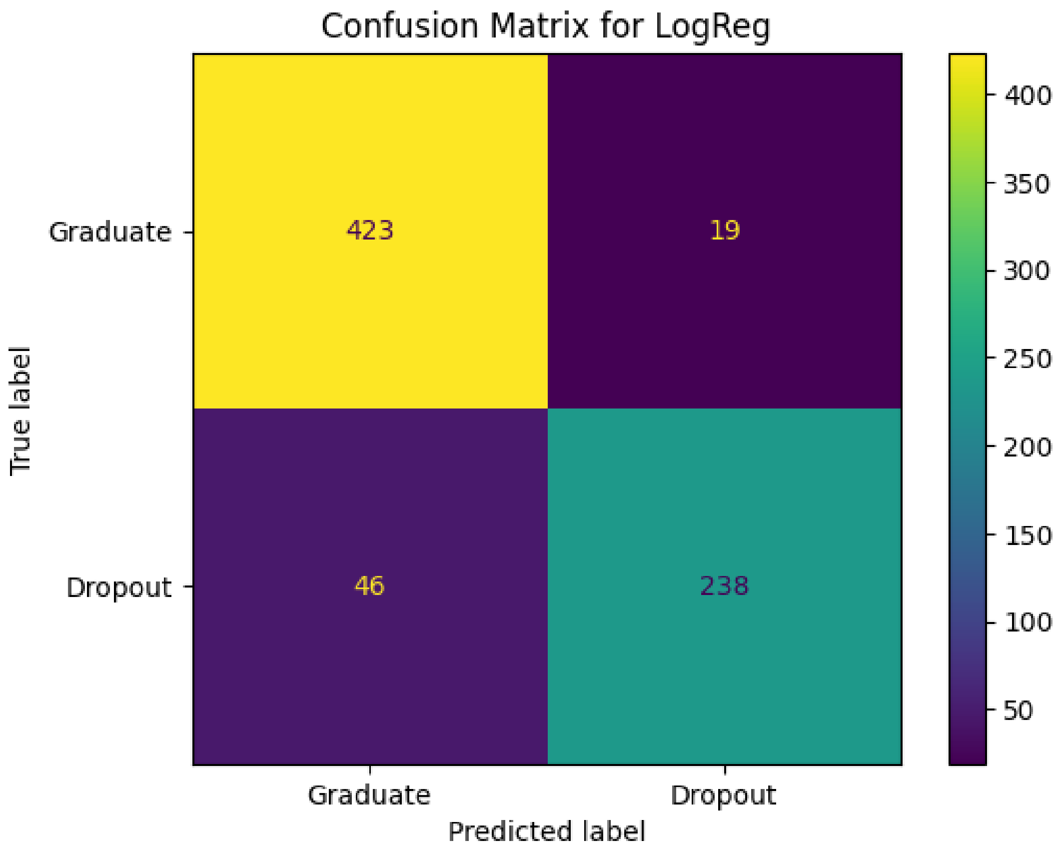
Gaussian Naive Bayes

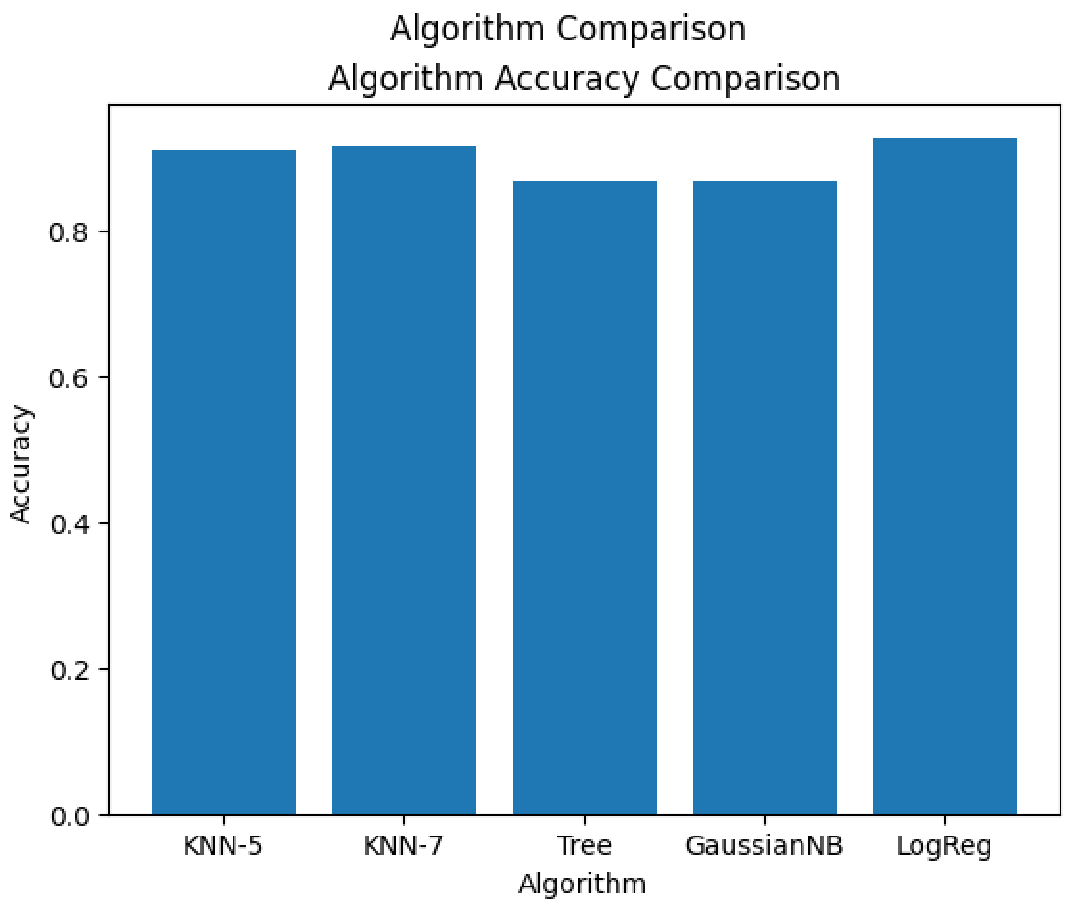
- **Accuracy: 86.78% – Accuracy matches that of the Decision Tree.**
- **Precision: 89.50% – The model is highly precise but still falls behind the KNN models.**
- **Recall: 75.00% – It captures fewer positives compared to the other models.**
- **F1-Score: 81.61% – While highly precise, it is less effective at capturing positives.**



Logistic Regression

- **Accuracy: 92.70% – This model offers the best accuracy overall.**
- **Precision: 91.40% – It makes correct positive predictions 91.40% of the time, closely matching KNN-7.**
- **Recall: 89.79% – It captures a high percentage of positives, second only to KNN-7.**
- **F1-Score: 90.59% – This F1-Score indicates the best balance of all models between precision and recall.**




Algorithm Comparison
Algorithm Accuracy Comparison
Accuracy
0.0
0.2
0.4
0.6
0.8
KNN-5
KNN-7
Tree
GaussianNB
LogReg
Algorithm

https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success.

https://research.com/universities-colleges/college-dropout-rates#1

https://educationdata.org/college-dropout-rates

https://www.researchgate.net/publication/314077615_DECREASING_SCHOOL_DROPOUT_RATE_AS_A_FACTOR_OF_ECONOMIC_GROWTH_AND_SOCIAL_EMPOWERMENT_THEORETICAL_INSIGHTS