BRAC
UNIVERSITY

Inspiring Excellence

Course Name: Artificial Intelligence

Course Code: CSE422

Lab Project Report

Project Topic: Student Dropout Prediction

Table of contents

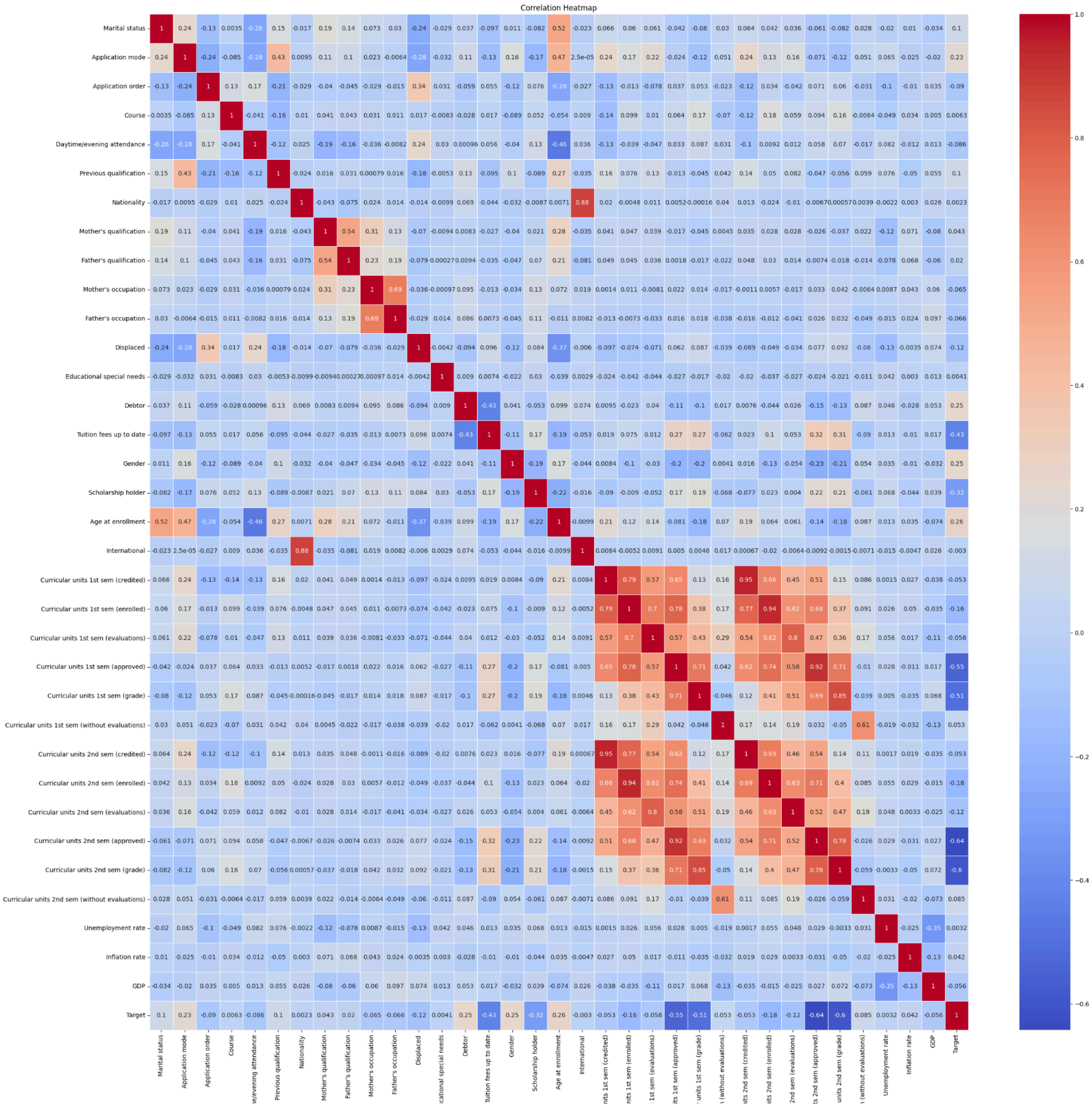
1. Introduction:

The project aims to predict whether a student will drop out or graduate with the help of a number of features. This could help educational institutions to get an estimation of which students are likely to drop out and provide them with additional support to minimize the chances of them dropping out. It could boost the graduation rate of an institution by providing data-driven insights.

2. Dataset Description:

- **Source:** Predict students' dropout and academic success
- **Number of features:** 35
- **Problem type:** Classification, as the output will be a Categorical Value, which is 0 for graduate and 1 for dropout.
- **Number of Columns:** 4424
- **Feature Types**: Quantitative and Categorical both

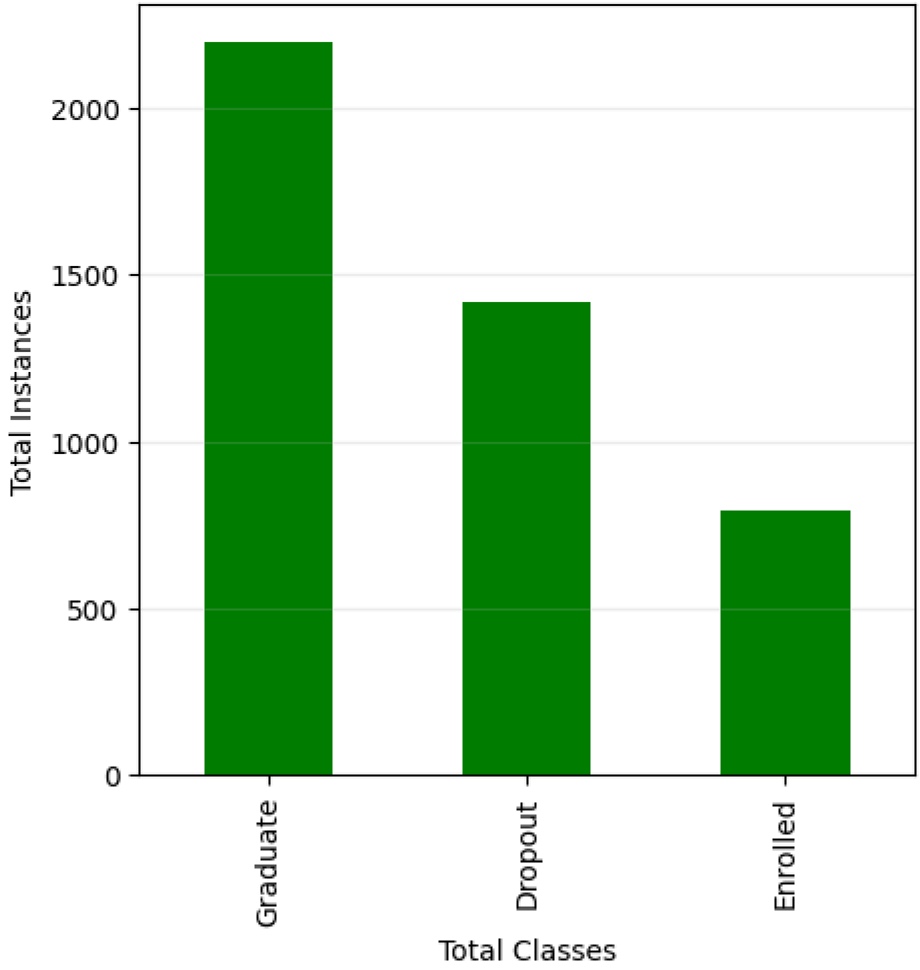
- Correlation of all features(Using Heatmap):



Correlation Heatmap

	Marital status	Application mode	Application order	Course	ne/evening attendance	Previous qualification	Nationality	Mother's qualification	Father's qualification	Mother's occupation	Father's occupation	Displaced	icational special needs	Debtor	Tuition fees up to date	Gender	Scholarship holder	Age at enrollment	International	nits 1st sem (credited)	nits 1st sem (enrolled)	1st sem (evaluations)	its 1st sem (approved)	' units 1st sem (grade)	i (without evaluations)	its 2nd sem (credited)	its 2nd sem (enrolled)	2nd sem (evaluations)	ts 2nd sem (approved)	units 2nd sem (grade)	i (without evaluations)	Unemployment rate	Inflation rate	GDP	Target
Marital status	1	0.24	-0.13	0.0035	-0.26	0.15	-0.017	0.19	0.14	0.073	0.03	-0.24	-0.029	0.037	-0.097	0.011	-0.082	0.52	-0.023	0.066	0.06	0.061	-0.042	-0.08	0.03	0.064	0.042	0.036	-0.061	-0.082	0.028	-0.02	0.01	-0.034	0.1
Application mode	0.24	1	-0.24	-0.085	-0.28	0.43	0.0095	0.11	0.1	0.023	-0.0064	-0.28	-0.032	0.11	-0.13	0.16	-0.17	0.47	2.5e-05	0.24	0.17	0.22	-0.024	-0.12	0.051	0.24	0.13	0.16	-0.071	-0.12	0.051	0.065	-0.025	-0.02	0.23
Application order	-0.13	-0.24	1	0.13	0.17	-0.21	-0.029	-0.04	-0.045	-0.029	-0.015	0.34	0.031	-0.059	0.055	-0.12	0.076	-0.28	-0.027	-0.13	-0.013	-0.078	0.037	0.053	-0.023	-0.12	0.034	-0.042	0.071	0.06	-0.031	-0.1	-0.01	0.035	-0.09
Course	0.0035	-0.085	0.13	1	-0.041	-0.16	0.01	0.041	0.043	0.031	0.011	0.017	-0.0083	-0.028	0.017	-0.089	0.052	-0.054	0.009	-0.14	0.099	0.01	0.064	0.17	-0.07	-0.12	0.18	0.059	0.094	0.16	-0.0064	-0.049	0.034	0.005	0.0063
Daytime/evening attendance	-0.26	-0.28	0.17	-0.041	1	-0.12	0.025	-0.19	-0.16	-0.036	-0.0082	0.24	0.03	0.00096	0.056	-0.04	0.13	-0.46	0.036	-0.13	-0.039	-0.047	0.033	0.087	0.031	-0.1	0.0092	0.012	0.058	0.07	-0.017	0.082	-0.012	0.013	-0.086
Previous qualification	0.15	0.43	-0.21	-0.16	-0.12	1	-0.024	0.016	0.031	0.00079	0.016	-0.18	-0.0053	0.13	-0.095	0.1	-0.089	0.27	-0.035	0.16	0.076	0.13	-0.013	-0.045	0.042	0.14	0.05	0.082	-0.047	-0.056	0.059	0.076	-0.05	0.055	0.1
Nationality	-0.017	0.0095	-0.029	0.01	0.025	-0.024	1	-0.043	-0.075	0.024	0.014	-0.014	-0.0099	0.069	-0.044	-0.032	-0.0087	0.0071	0.88	0.02	-0.0048	0.011	0.0052	-0.00016	0.04	0.013	-0.024	-0.01	-0.0067	0.00057	0.0039	-0.0022	0.003	0.026	0.0023
Mother's qualification	0.19	0.11	-0.04	0.041	-0.19	0.016	-0.043	1	0.54	0.31	0.13	-0.07	-0.0094	0.0083	-0.027	-0.04	0.021	0.28	-0.035	0.041	0.047	0.039	-0.017	-0.045	0.0045	0.035	0.028	0.028	-0.026	-0.037	0.022	-0.12	0.071	-0.08	0.043
Father's qualification	0.14	0.1	-0.045	0.043	-0.16	0.031	-0.075	0.54	1	0.23	0.19	-0.079	0.00027	0.0094	-0.035	-0.047	0.07	0.21	-0.081	0.049	0.045	0.036	0.0018	-0.017	-0.022	0.048	0.03	0.014	-0.0074	-0.018	-0.014	-0.078	0.068	-0.06	0.02
Mother's occupation	0.073	0.023	-0.029	0.031	-0.036	0.00079	0.024	0.31	0.23	1	0.69	-0.036	-0.00097	0.095	-0.013	-0.034	0.13	0.072	0.019	0.0014	0.011	-0.0081	0.022	0.014	-0.017	-0.0011	0.0057	-0.017	0.033	0.042	-0.0064	0.0087	0.043	0.06	-0.065
Father's occupation	0.03	-0.0064	-0.015	0.011	-0.0082	0.016	0.014	0.13	0.19	0.69	1	-0.029	0.014	0.086	0.0073	-0.045	0.11	-0.011	0.0082	-0.013	-0.0073	-0.033	0.016	0.018	-0.038	-0.016	-0.012	-0.041	0.026	0.032	-0.049	-0.015	0.024	0.097	-0.066
Displaced	-0.24	-0.28	0.34	0.017	0.24	-0.18	-0.014	-0.07	-0.079	-0.036	-0.029	1	-0.0042	-0.094	0.096	-0.12	0.084	-0.37	-0.006	-0.097	-0.074	-0.071	0.062	0.087	-0.039	-0.089	-0.049	-0.034	0.077	0.092	-0.06	-0.13	-0.0035	0.074	-0.12
Educational special needs	-0.029	-0.032	0.031	-0.0083	0.03	-0.0053	-0.0099	-0.0094	0.00027	-0.00097	0.014	-0.0042	1	0.009	0.0074	-0.022	0.03	-0.039	0.0029	-0.024	-0.042	-0.044	-0.027	-0.017	-0.02	-0.02	-0.037	-0.027	-0.024	-0.021	-0.011	0.042	0.003	0.013	0.0041
Debtor	0.037	0.11	-0.059	-0.028	0.00096	0.13	0.069	0.0083	0.0094	0.095	0.086	-0.094	0.009	1	-0.43	0.041	-0.053	0.099	0.074	0.0095	-0.023	0.04	-0.11	-0.1	0.017	0.0076	-0.044	0.026	-0.15	-0.13	0.087	0.046	-0.028	0.053	0.25
Tuition fees up to date	-0.097	-0.13	0.055	0.017	0.056	-0.095	-0.044	-0.027	-0.035	-0.013	0.0073	0.096	0.0074	-0.43	1	-0.11	0.17	-0.19	-0.053	0.019	0.075	0.012	0.27	0.27	-0.062	0.023	0.1	0.053	0.32	0.31	-0.09	0.013	-0.01	0.017	-0.43
Gender	0.011	0.16	-0.12	-0.089	-0.04	0.1	-0.032	-0.04	-0.047	-0.034	-0.045	-0.12	-0.022	0.041	-0.11	1	-0.19	0.17	-0.044	0.0084	-0.1	-0.03	-0.2	-0.2	0.0041	0.016	-0.13	-0.054	-0.23	-0.21	0.054	0.035	-0.01	-0.032	0.25
Scholarship holder	-0.082	-0.17	0.076	0.052	0.13	-0.089	-0.0087	0.021	0.07	0.13	0.11	0.084	0.03	-0.053	0.17	-0.19	1	-0.22	-0.016	-0.09	-0.009	-0.052	0.17	0.19	-0.068	-0.077	0.023	0.004	0.22	0.21	-0.061	0.068	-0.044	0.039	-0.32
Age at enrollment	0.52	0.47	-0.28	-0.054	-0.46	0.27	0.0071	0.28	0.21	0.072	-0.011	-0.37	-0.039	0.099	-0.19	0.17	-0.22	1	-0.0099	0.21	0.12	0.14	-0.081	-0.18	0.07	0.19	0.064	0.061	-0.14	-0.18	0.087	0.013	0.035	-0.074	0.26
International	-0.023	2.5e-05	-0.027	0.009	0.036	-0.035	0.88	-0.035	-0.081	0.019	0.0082	-0.006	0.0029	0.074	-0.053	-0.044	-0.016	-0.0099	1	0.0084	-0.0052	0.0091	0.005	0.0046	0.017	0.00067	-0.02	-0.0064	-0.0092	-0.0015	-0.0071	-0.015	-0.0047	0.026	-0.003
Curricular units 1st sem (credited)	0.066	0.24	-0.13	-0.14	-0.13	0.16	0.02	0.041	0.049	0.0014	-0.013	-0.097	-0.024	0.0095	0.019	0.0084	-0.09	0.21	0.0084	1	0.79	0.57	0.65	0.13	0.16	0.95	0.66	0.45	0.51	0.15	0.086	0.0015	0.027	-0.038	-0.053
Curricular units 1st sem (enrolled)	0.06	0.17	-0.013	0.099	-0.039	0.076	-0.0048	0.047	0.045	0.011	-0.0073	-0.074	-0.042	-0.023	0.075	-0.1	-0.009	0.12	-0.0052	0.79	1	0.7	0.78	0.38	0.17	0.77	0.94	0.62	0.68	0.37	0.091	0.026	0.05	-0.035	-0.16
Curricular units 1st sem (evaluations)	0.061	0.22	-0.078	0.01	-0.047	0.13	0.011	0.039	0.036	-0.0081	-0.033	-0.071	-0.044	0.04	0.012	-0.03	-0.052	0.14	0.0091	0.57	0.7	1	0.57	0.43	0.29	0.54	0.62	0.8	0.47	0.36	0.17	0.056	0.017	-0.11	-0.058
Curricular units 1st sem (approved)	-0.042	-0.024	0.037	0.064	0.033	-0.013	0.0052	-0.017	0.0018	0.022	0.016	0.062	-0.027	-0.11	0.27	-0.2	0.17	-0.081	0.005	0.65	0.78	0.57	1	0.71	0.042	0.62	0.74	0.58	0.92	0.71	-0.01	0.028	-0.011	0.017	-0.55
Curricular units 1st sem (grade)	-0.08	-0.12	0.053	0.17	0.087	-0.045	-0.00016	-0.045	-0.017	0.014	0.018	0.087	-0.017	-0.1	0.27	-0.2	0.19	-0.18	0.0046	0.13	0.38	0.43	0.71	1	-0.046	0.12	0.41	0.51	0.69	0.85	-0.039	0.005	-0.035	0.068	-0.51
Curricular units 1st sem (without evaluations)	0.03	0.051	-0.023	-0.07	0.031	0.042	0.04	0.0045	-0.022	-0.017	-0.038	-0.039	-0.02	0.017	-0.062	0.0041	-0.068	0.07	0.017	0.16	0.17	0.29	0.042	-0.046	1	0.17	0.14	0.19	0.032	-0.05	0.61	-0.019	-0.032	-0.13	0.053
Curricular units 2nd sem (credited)	0.064	0.24	-0.12	-0.12	-0.1	0.14	0.013	0.035	0.048	-0.0011	-0.016	-0.089	-0.02	0.0076	0.023	0.016	-0.077	0.19	0.00067	0.95	0.77	0.54	0.62	0.12	0.17	1	0.69	0.46	0.54	0.14	0.11	0.0017	0.019	-0.035	-0.053
Curricular units 2nd sem (enrolled)	0.042	0.13	0.034	0.18	0.0092	0.05	-0.024	0.028	0.03	0.0057	-0.012	-0.049	-0.037	-0.044	0.1	-0.13	0.023	0.064	-0.02	0.66	0.94	0.62	0.74	0.41	0.14	0.69	1	0.63	0.71	0.4	0.085	0.055	0.029	-0.015	-0.18
Curricular units 2nd sem (evaluations)	0.036	0.16	-0.042	0.059	0.012	0.082	-0.01	0.028	0.014	-0.017	-0.041	-0.034	-0.027	0.026	0.053	-0.054	0.004	0.061	-0.0064	0.45	0.62	0.8	0.58	0.51	0.19	0.46	0.63	1	0.52	0.47	0.19	0.048	0.0033	-0.025	-0.12
Curricular units 2nd sem (approved)	-0.061	-0.071	0.071	0.094	0.058	-0.047	-0.0067	-0.026	-0.0074	0.033	0.026	0.077	-0.024	-0.15	0.32	-0.23	0.22	-0.14	-0.0092	0.51	0.68	0.47	0.92	0.69	0.032	0.54	0.71	0.52	1	0.78	-0.026	0.029	-0.031	0.027	-0.64
Curricular units 2nd sem (grade)	-0.082	-0.12	0.06	0.16	0.07	-0.056	0.00057	-0.037	-0.018	0.042	0.032	0.092	-0.021	-0.13	0.31	-0.21	0.21	-0.18	-0.0015	0.15	0.37	0.36	0.71	0.85	-0.05	0.14	0.4	0.47	0.78	1	-0.059	-0.0033	-0.05	0.072	-0.6
Curricular units 2nd sem (without evaluations)	0.028	0.051	-0.031	-0.0064	-0.017	0.059	0.0039	0.022	-0.014	-0.0064	-0.049	-0.06	-0.011	0.087	-0.09	0.054	-0.061	0.087	-0.0071	0.086	0.091	0.17	-0.01	-0.039	0.61	0.11	0.085	0.19	-0.026	-0.059	1	0.031	-0.02	-0.073	0.085
Unemployment rate	-0.02	0.065	-0.1	-0.049	0.082	0.076	-0.0022	-0.12	-0.078	0.0087	-0.015	-0.13	0.042	0.046	0.013	0.035	0.068	0.013	-0.015	0.0015	0.026	0.056	0.028	0.005	-0.019	0.0017	0.055	0.048	0.029	-0.0033	0.031	1	-0.025	-0.35	0.0032
Inflation rate	0.01	-0.025	-0.01	0.034	-0.012	-0.05	0.003	0.071	0.068	0.043	0.024	-0.0035	0.003	-0.028	-0.01	-0.01	-0.044	0.035	-0.0047	0.027	0.05	0.017	-0.011	-0.035	-0.032	0.019	0.029	0.0033	-0.031	-0.05	-0.02	-0.025	1	-0.13	0.042
GDP	-0.034	-0.02	0.035	0.005	0.013	0.055	0.026	-0.08	-0.06	0.06	0.097	0.074	0.013	0.053	0.017	-0.032	0.039	-0.074	0.026	-0.038	-0.035	-0.11	0.017	0.068	-0.13	-0.035	-0.015	-0.025	0.027	0.072	-0.073	-0.35	-0.13	1	-0.056
Target	0.1	0.23	-0.09	0.0063	-0.086	0.1	0.0023	0.043	0.02	-0.065	-0.066	-0.12	0.0041	0.25	-0.43	0.25	-0.32	0.26	-0.003	-0.053	-0.16	-0.058	-0.55	-0.51	0.053	-0.053	-0.18	-0.12	-0.64	-0.6	0.085	0.0032	0.042	-0.056	1

- Imbalance dataset: Here, for the output feature, all the unique values do not have an equal number of instances.



3. Dataset Pre-processing

- **Modified the dataset:** We had to modify the data a bit in order to make the dataset unprocessed; this included nullifying some of the values. This is the link to the updated dataset.
- **Null-values:** There are some columns which included null values; We dropped those columns

- **Column name fix:** The “Nacionality” column name was misspelled, so fixed the spelling.
- **Mapping Categorical columns with Numerical Values:** The Columns “Country Names” & “Target” had categorical values so we replaced them with some arbitrary numerical values.
- **Dropping Unnecessary Columns:** There were some columns which we felt were a bit redundant and did not add much to the training data; therefore, we decided to drop them.
- **Dropping “Enrolled” Rows from Target Column:** The number of instances between all the three features were too imbalanced and the “Enrolled” instance was not giving us any conclusive decision that the student will drop out or not; So, we dropped the rows with “Enrolled” instances.
- **Dropping Correlated Features:** In the heat map, some features were highly correlating; so we chose to drop some of them to reduce redundancy.

4. Feature Scaling

- We scaled the columns with quantitative values from the processed dataset using the MiniMax Scaler algorithm.

5. Dataset Splitting

- 70% of the data were used as Train set and 30% of the data were used for Test set

6. Model Training and Testing

- **Naive Bayes:** We chose naive bayes for its simplicity and for the fact that it performs well with both large high-dimensional datasets as well as small datasets. It assumes conditional independence between features making the calculation simpler, therefore being able to generate responses faster.

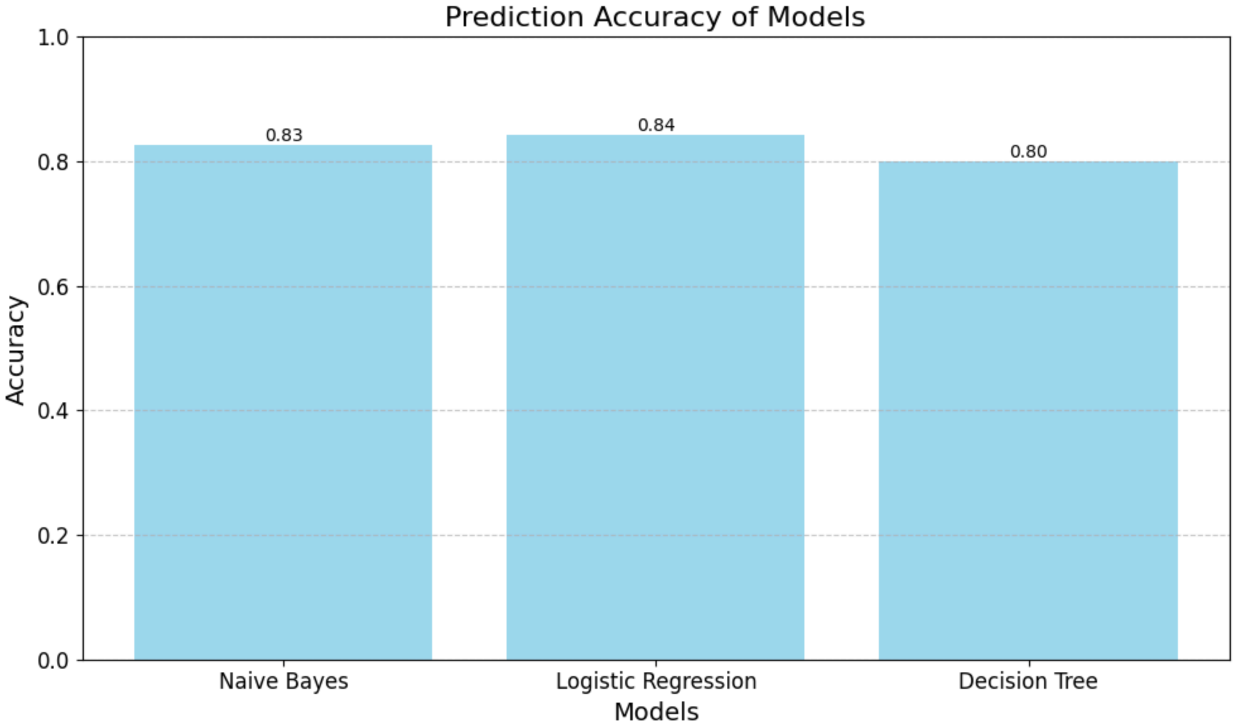
- **Logistic Regression:** Logistic Regression uses sigmoid function to map the predicted values to [0, 1] of binary classification problems. It works great with binary classification problems and it is quite easy to implement and very efficient for training. Also, it provides good accuracy and performs well for fairly simple datasets and when the dataset is linearly separable; hence, we chose to use this model.

- **Decision Tree:** We used decision tree for its insensitivity to outliers as it splits data based on feature thresholds rather than depending on distance calculations which gives a better probability to find the optimal class of the target.
  We used “class_weight = balanced” as a parameter so that it doesn't give a biased class because of target feature imbalance.

7. Comparison Analysis

- **Bar Chart showcasing prediction accuracy of models**

  We can notice that Logistic Regression gives us the most accurate prediction with 84% correct prediction, while Decision Tree gives us the worst with 80%; Naive Bayes is in the middle with 83%.

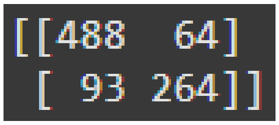


- **Precision and recall comparison**

Models	**Precision**	**Recall**
Naive Bayes	0.83	0.83
Logistic Regression	0.84	0.84
Decision Tree	0.80	0.80

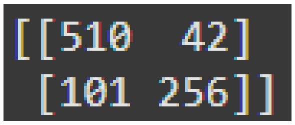
- **Confusion Matrix of each model**

Naive Bayes:



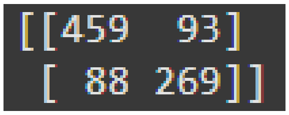
```
[[488  64]
 [ 93 264]]
```

Logistic Regression:



```
[[510  42]
 [101 256]]
```

Decision Tree:



```
[[459  93]
 [ 88 269]]
```

8. Conclusion

This project was designed to predict the chances of a student dropping out. By training the model, we tried finding an appropriate answer to the query. Although the models were inaccurate for about ~20% of the data, it can be used as a guide to predicting the likelihood of the student dropping out. Educational institutions could use something of this nature and enhance their graduation rate and help the students out greatly.