

## **2.? Background: Logistic Regression**

Demonstrate and describe the mechanics of the selected machine learning techniques.

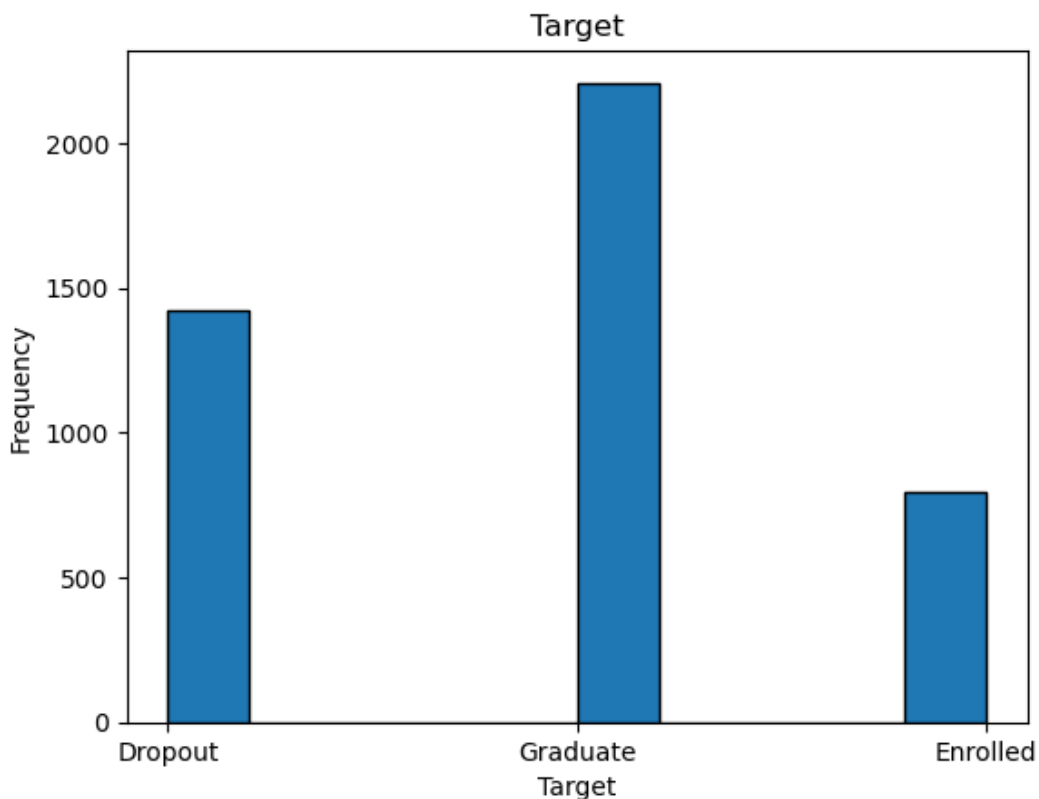
### 3. Data Preparation

#### 3.1. Exploratory Analysis

The initial dataset consisted of 4424 records with 36 independent variables, with no missing values. The dataset was processed using the *pandas* library in Python.

- **Note: Should I include means of variance tables despite not all variables being in the final MI dataset?**

Each student entry has one of 3 Target variables: Graduated, Dropout, or Enrolled (where the student took another three years to complete the course). The distribution of the Target variable is shown in Figure 1, which shows that it is imbalanced, where the Graduated Target makes up 50% of the data records. This was addressed using balancing **What is SMOTE?** The SMOTE implementation from *imbalanced-learn* in the *sci-kit learn* library was used.



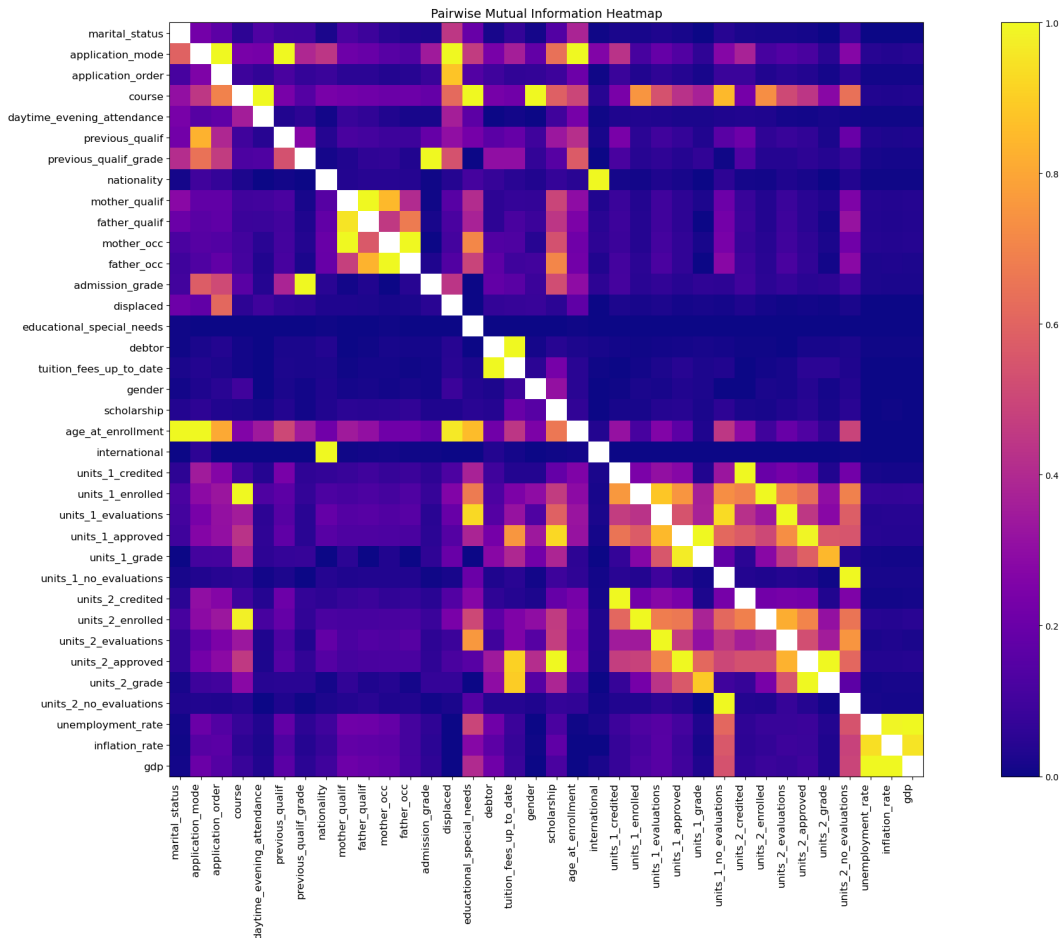
**Figure 1:** Count of Target variable.

#### 3.2. Feature Engineering

Mutual Information (MI) is a filter-based feature selection method which measures how mutually dependent two variables are [1]. This method was used to filter out redundant variables, and was opted for since MI can be used for both discrete and continuous variables. The MI score with target was computed for each variable using *sci-kit learn*. Each feature is assigned a scoring value, with the resulting features being organised in descending order based on the scores and are assigned rankings for the features [1]. A limitation of this method is that

while it measures a feature's importance by its correlation with the target, it assumes independence of features from each other [2]. Therefore, choosing the highest scoring variables from MI can lead to redundant features can be selected.

To address this, a pairwise MI matrix was computed. The values were normalised to a range between 0 and 1, presented in Figure 2. Redundant variables were removed by checking each MI pairwise score according to a set threshold (0.99) and removing the variable with the lower MI score with the target. The final selected predictors resulted in 17 independent variables.



**Figure 2:** Heatmap of pairwise Mutual Information scores.

Principal Component Analysis (PCA) was then performed on the dataset for dimensionality reduction [3]. **Explain PCA briefly.** The final resulting dataset consisted of 8 independent variables.

## 4.? Logistic Regression

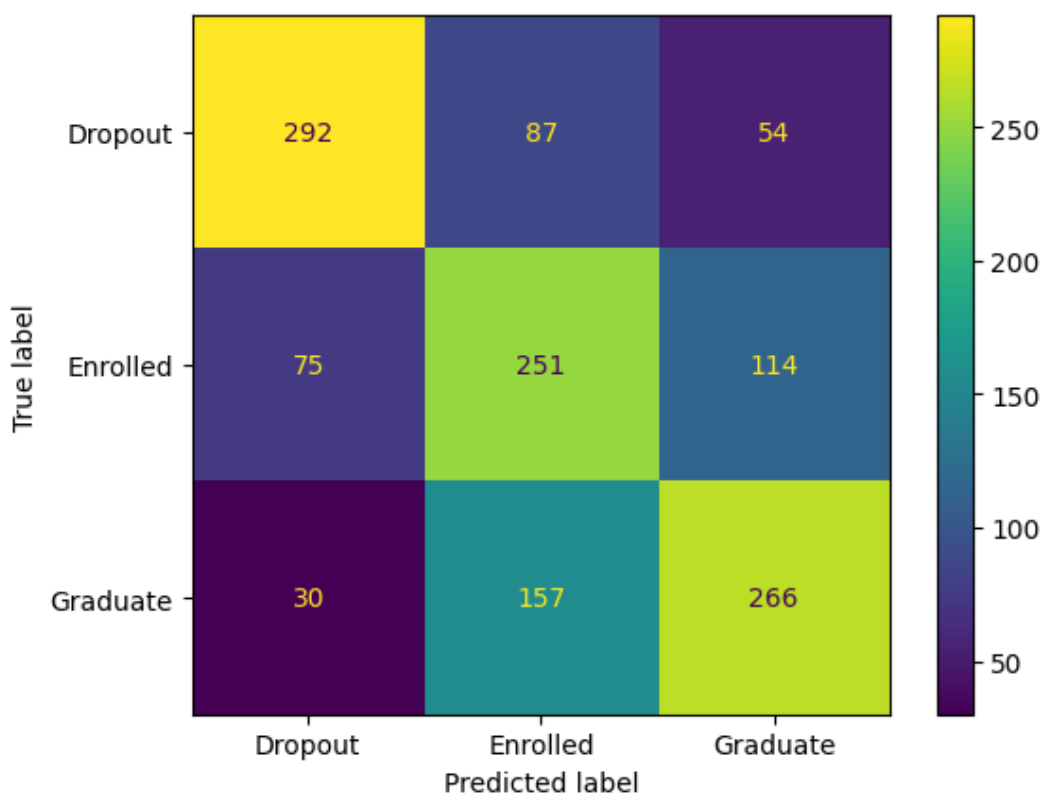
The multi-class logistic regression model was implemented using the *sci-kit learn* library in Python. The dataset was split into training, validation, and test sets using a 70/20/10 split to ensure that the model was trained on a representative sample and tested on unseen data. Feature scaling was applied using **StandardScaler?** to ensure that the coefficients of the logistic regression were appropriately scaled.

The LogisticRegression class was utilised with the 'multinomial' option for handling multiple classes. **Solver lbfgs**. Model evaluation was performed on the test set using metrics the accuracy, precision, recall, F1-score metrics, and a confusion matrix to assess classification performance. The *scikit-learn* `classification_report` and `confusion_matrix` functions were used for this.

Table 1 shows the results of the classification report and Figure 3 shows the confusion matrix. The overall model accuracy is 61%. Results show that the Dropout target has the highest Precision and F1-Scores, indicating that the model has the best results for predicting this class over others. The model struggles most in predicting the Enrolled target, with an F1-score of 0.54.

Target	Precision	Recall	F1-Score
Graduate	0.61	0.59	0.60
Dropout	0.74	0.67	0.70
Enrolled	0.51	0.57	0.54

**Table 1:** Classification report for Logistic Regression model.



**Figure 3:** Confusion Matrix of the Logistic Regression model.

## References

- [1] B. Venkatesh and J. Anuradha, "A review of feature selection and its methods," *Cybernetics and Information Technologies*, vol. 19, no. 1, pp. 3–26, 2019. DOI: 10.2478/cait-2019-0001. [Online]. Available: <https://doi.org/10.2478/cait-2019-0001>.
- [2] J. Li, K. Cheng, S. Wang, *et al.*, "Feature selection: A data perspective," vol. 50, no. 6, Dec. 2017, ISSN: 0360-0300. DOI: 10.1145/3136625. [Online]. Available: <https://doi.org/10.1145/3136625>.
- [3] F. L. Gewers, G. R. Ferreira, H. F. D. Arruda, *et al.*, "Principal component analysis: A natural approach to data exploration," *ACM Comput. Surv.*, vol. 54, no. 4, May 2021, ISSN: 0360-0300. DOI: 10.1145/3447755. [Online]. Available: <https://doi-org.ejournals.um.edu.mt/10.1145/3447755>.