1. INTRODUCTION

This is a summarized report of the Data Mining coursework for the production of a pipeline capable of predicting newly discovered epitopes contained within the proteins of the Trypanosoma Cruzi virus. The dataset provides feature, info, and class columns where we work on the feature columns in reference to the info-cluster column to predict the epitopes.

2. EDA AND DATA PRE-PROCESSING

Exploratory Data Analysis(EDA)

EDA is an essential requirement for the initiation of a Data Mining project to guide the following procedures in the project. This analysis provides a summary of a dataset's characteristics including its components' structure, data types, and patterns [1]. Here, it has been conducted by observing the columns, their datatypes, missing values, class balance, and visualization.

The dataset was found to have *12402* observations and *300* variables with datatypes of integers and floats. As it was mentioned that info columns would have negative impact on further processes, feature columns were isolated from them while keeping the info_cluster for grouping. This grouping in the EDA aided in balancing the positive and negative values in the classes.

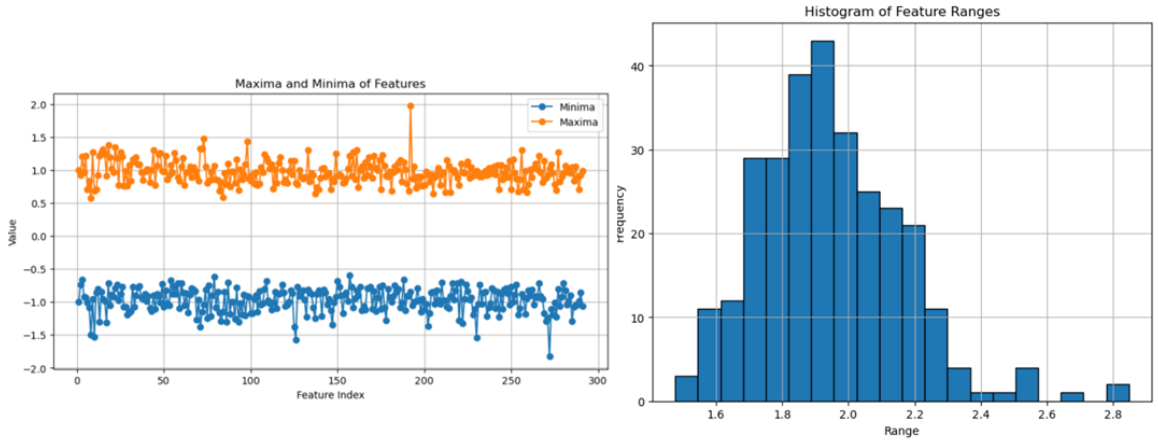

Figure 1 Line and histogram plot for the maxima and minima of the dataset

By evaluating the diagrams above, the maxima and minima values of the dataset were found to be *2* and almost *-2*, and roughly visualize the presence of outliers.

Following the plotting, the observations and features detected with large number of missing values were removed as they would indicate poor quality in data and effect the model's performance.

Data Pre-Processing

Taking plenty of processing time with cleaning, normalization, feature selection, and more, DPP ensures that the resulting dataset is clean and adaptable to the following algorithms of learning models and predictions [2].

- Group Splitting
  To start the pre-processing, the dataset was split into Train and Test datasets using the GroupShuffleSplit method to distribute them with sizes of 80% and 20% respectively on the basis of the info_cluster column. This allowed the train-data to be used in training the model and the test-data for the model evaluation.

- Outlier detection and handling
  Isolation Forest(IF) was used for the outlier detection as its unsupervised learning abilities enables scarce unlabelled data to be identified as outliers. Other than this algorithm, Angle Based Outlier Detection(ABOD) was ruled out due to its inability to produce adequate detection. *415* outliers were detected by the algorithm and removed using the predicted mask from the algorithm.

- Scaling of the Dataset
  With Min-Max Scaler, the dataset was scaled and transformed to an overall specific range to improve its interpretability and ensure that all the features have equal contribution in the analysing and modelling process. The simplicity and compatibility of this scaler allows the production of stable results and less influence from the outliers.

3. FEATURE REDUCTION

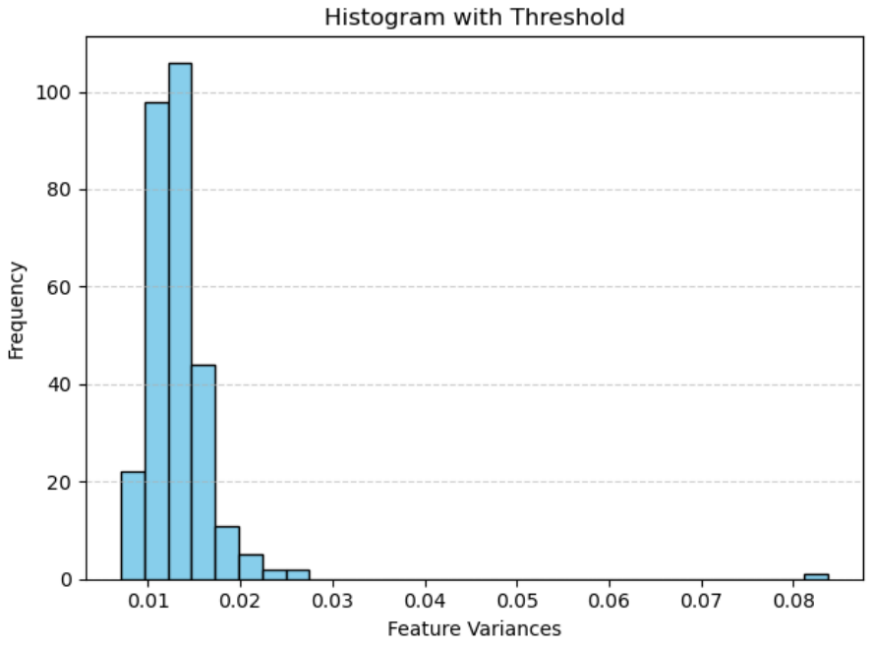

Figure 2 Histogram for Threshold Variance

Here, the threshold value, determined by plotting a histogram with the variances obtained from the scaled dataset, was implemented in the Variance Threshold method. The features selected from the scaled dataset with this method are further used for modelling. While it could not promote the full predictive power of the dataset as it does not properly consider the importance of predicting target variables, this method is compatible for the dataset with large quantity of features.

4. MODELLING AND ASSESSMENT

Modelling

In terms of model selection, a model is chosen for a classification or a regression problem based on the best performance exhibited by it during fitting and evaluation to address a problem [3]. For this coursework, decision tree and logistic regression were the two models trained where each of their accuracy score, balanced accuracy score, and classification reports were calculated.

SMOTE was introduced to create resampled data for the oversampling of the previous data. The resampled data were used to run the models again and the scores obtained here were compared with those from before the SMOTE.

Model Assessment

The assessment of the model begins with Hyperparameter Tuning which is the process of obtaining the best parameters and scores out for the chosen model.

Depending on the models' performance metrics obtained from the hyperparameter tuning presented below,

Balanced accuracy score for,

Decision Tree: *0.8515007294234826*

Logistic Regression: *0.8084950999283775*

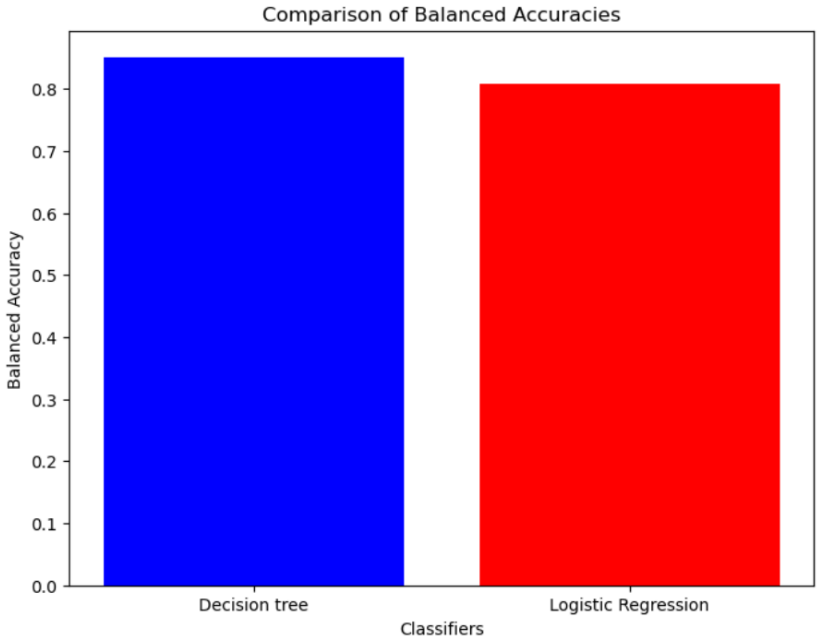

Figure 3 Comparison Plotting for Balanced Accuracies

Decision Tree was found to be the best model for prediction due to the higher balanced accuracy. Furthermore, the best parameters found for the Decision Tree model are as follows:

```
Best parameters for decision_tree: {'max_depth': 3, 'min_samples_split': 2}
```

Figure 4 Best Parameters for Best Model

To further assess the chosen model, pipeline is created with the preprocessing steps and the parameters determined from the tuning. This pipeline is fitted to the resampled data with the selected features.

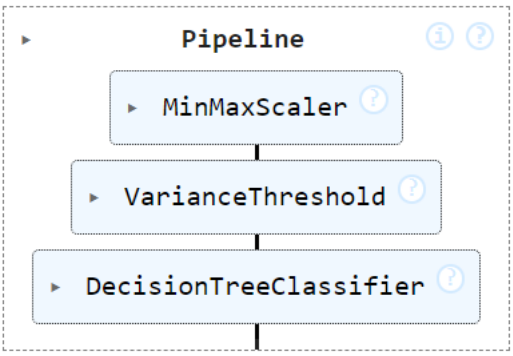

Figure 5 Pipeline for Decision Tree

Finally, the pipeline was used to predict the classes for the holdout_data containing unseen data. This prediction was put into a csv file as required by the coursework.

5. CONCLUSION

The purpose of this coursework was to predict the classes that were never seen before to determine the epitopes in the virus. Following EDA, pre-processing, and model related steps by training using the provided dataset, the coursework led to the creation of a pipeline used to predict unique classes for the holdout dataset. From this coursework, several models were trained to determine the best one that was finalized to produce the best predictions.

Bibliography

[1] M. Komorowski, D. C. Marshall, J. D. Salciccioli and Y. Crutain, “Exploratory Data Analysis,” in *Secondary Analysis of Electronic Health Records*, Springer, 2016.

[2] S. Garcia, J. Luengo and F. Herrera, Data Preprocessing, Springer, 2015.

[3] J. Brownlee, “A Gentle Introduction to Model Selection for Machine Learning,” 26 September 2019. [Online]. Available: https://machinelearningmastery.com/a-gentle-introduction-to-model-selection-for-machine-learning/. [Accessed April 2024].