# Comparison between Decision Trees and Support-Vector Machines for Grade Prediction

## Introduction

This report will compare the decision tree and logistic regression machine learning methods within the field of learning analytics using data from the OULAD [1] to predict (by classifying) students’ final results.

## Experimental Procedure Plan

Usage of the two machine learning models need to be justified. After this, a feature set needs to be selected. Next, the dataset is split into a training and test set.

The processing and analysis of the dataset is now performed using Python with NumPy and Pandas for data manipulation, scikit-learn for model application and matplotlib for data visualisation.

Finally, model hyperparameters are tuned to try increase accuracy of the model (while minimising overfitting).

## Justification for Use of Models

A decision tree classifier generates a tree structure when training, this is primarily useful when the input and output is categorical, which in this case it mostly is (dependant on the feature selection).

A support-vector machine generates a hyperplane to separate training data in hyperspace. While it is best suited to continuous data, as explained in Data Preparation, categorical features can be mapped to meaningful continuous data.

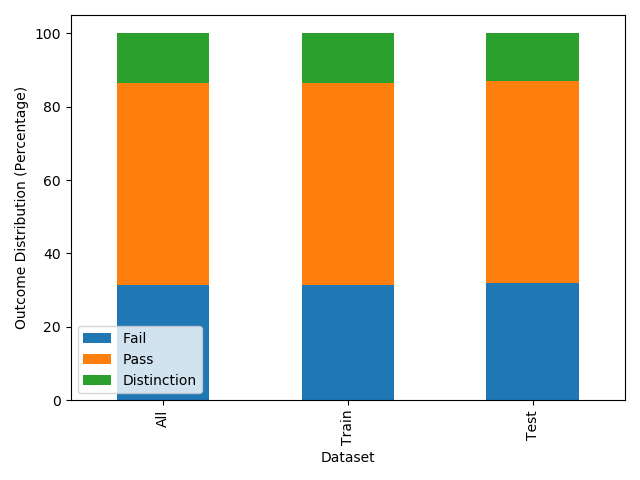
## Data Preparation

In order to use the models, the data must first be transformed into a set of input features and a corresponding set of input labels.

Since both models require class-based features to be converted into numerical inputs, decision tree may use one-hot encoding with the pandas get dummies method.

To enable both methods to be comparable, the ‘Withdrawn’ classification will not be considered, this is because it acts as a mutually exclusive entity compared to the grade classifications (meaning it cannot be assigned a numerical value using the banded system).

The second and third pie charts demonstrate that the testing and training data have less than 1% difference in label proportions.

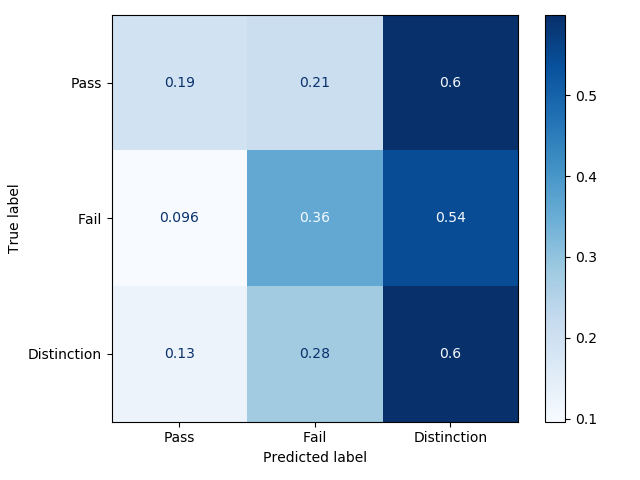


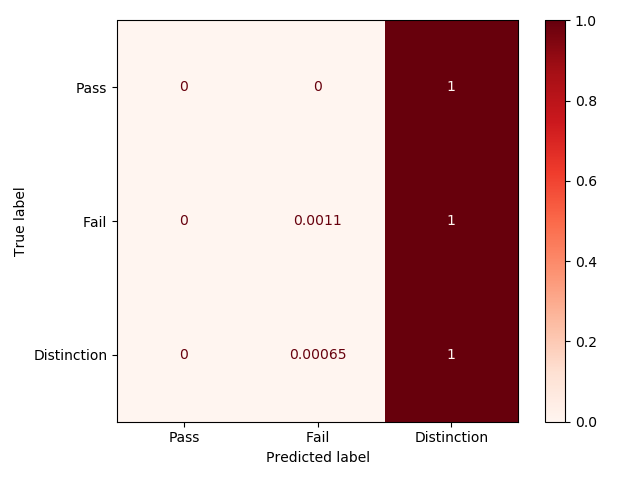
The stacked bar chart demonstrates the distribution of outcomes between the training and test datasets (of which there is less than one percent difference).

## Performance of Methods

Classification accuracy score works out the fraction of correctly classified samples when a model is used to predict labels for the test dataset.

Normalised confusion matrices visualise the fraction of instances within the sample set where a true label is output as a given prediction label. The higher the values in leading diagonal in the matrix (and the lower the values everywhere else), the greater the accuracy of the model (on the testing set).





## Conclusions

Comparisons: Training, evaluation, hyperparameter tuning [C value, OVR vs OVO, kernel], prediction

# References

|  |  |
| --- | --- |
| [1] | J. Kuzilek, M. Hlosta and Z. Zdrahal, “Open University Learning Analytics dataset,” *Scientific Data,* vol. 4, no. 1, 2017. |