Student Number: 254899

1. Introduction

The data provided come from photos which are whether memorable or not. Each sample in the training data has a predict label, 1 for memorable, 0 for not memorable, and a confidence label. The classification process can be divided into three steps: data processing on training data, training of classifier and prediction of the test data. In this project, several classifiers are tested and used to perform the binary classification task.

2. Methods

2.1. Pre-processing

Feature selection can be considered as the process of selecting a subset of relevant features for use in model construction Therefore it enhances the predictor's prediction performance [1]. In the data provided for training, there are labels for confidence. In order to train the model more accurately, samples with low confidence level are dropped (0.66 is considered as threshold). Next, features are classified whether they are extracted from CNN or GIST and the mean of features importance is calculated. Based on the calculations, the importance of CNN features is 0.000244 and GIST features is 0.001957. According to the results, GIST features are relatively more important compared to CNN features.

Rescaling is one of the most significant steps in pre-processing. When digital input variables are scaled to a standard range, it can be observed that various machine learning algorithms perform better. One of the many rescaling methods is Min Max Scaler, which is calculated by dividing all values by the maximum value encountered or by subtracting the minimum value, then dividing by the range between the maximum and minimum values [4]. As stated in the assignment brief, it is possible that some training sets might have some missing data. In order to overcome the issue, data imputation approach is used. Data imputation approximates the value of the column based on the existing values and replaces the missing ones with the estimated statistics accurately and quickly. In order to implement the data imputation, SciKit-Learn's Simple Imputer method with the main strategy is used.

Finally, as the final stage of pre-processing, dimensionality reduction is performed. by concentrating on the most relevant variables, dimensionality reduction can generate a more interpretable representation of the desired notion [1]. Principle Component Analysis is one of the most well-known dimensionality reduction methods which can be considered as projection method.

2.2. Selecting, Training, and Testing the Model

By using Repeated Stratified K-Fold (n times repetition of Stratified K-Fold with separate randomization in each iteration), dataset is taken and splitted into a training and test datasets.

Next, the most popular 7 approaches with spot-checking algorithms (Naive Bayes, Logistic Regression, Linear Discriminant Analysis, Random Forest Classifier, Support Vector Machine, Multi-layer Perceptron Classifier, K-Nearest Neighbors) are evaluated for the training dataset and then the best-performing algorithms are chosen and applied algorithm tuning for each to improve the performance of the model. Due to its ease while applying, covering the whole space, and not being based on randomness; pipeline and grid search is used [2].

After, the best classifier to predict the test dataset is used and output the prediction results.

3. Results

Based on the experimental results (see Figure 1) and the assumptions of mine, I decided to move on with Naive Bayes Classifier, Logistic Regression Classifier and Multi-layer Perceptron Classifier.

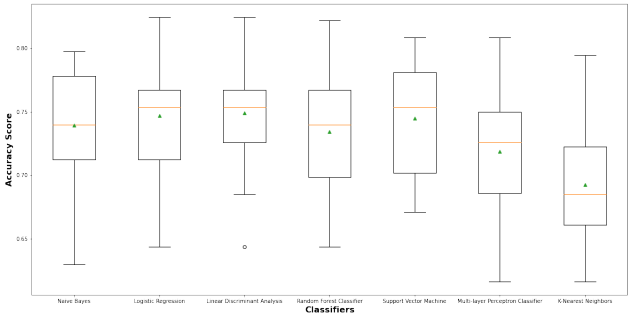

Figure 1: Model Results

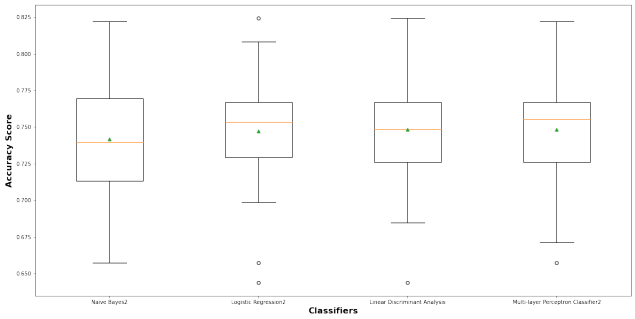

Figure 2: Selected Model Results

For the further stages, since it performed well on the first run, it does not make assumptions about the distribution of classes in the feature space, and performs better regardless of the distribution of the data is normal or non-normal [3], I decided to focus especially on logistic regression.

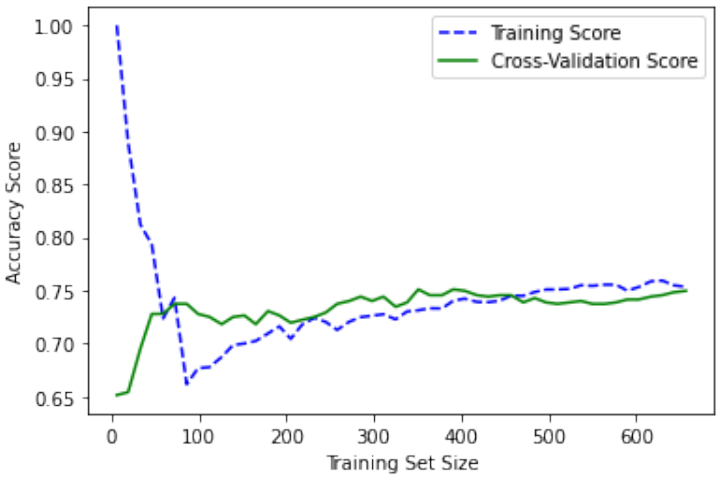

Figure 3: Learning Curves of the Final Logistic Regression Model

At the end, the final logistic regression model's accuracy score was 0.747334.

4. Discussion

Performance of the model can be improved by training it with better sampled and collected data. In order to avoid low-confidence related issues, there might be better suggestions than dropping them.

References

[1] I. Guyon and A. Elisseeff. An introduction to feature extraction. *Feature Extraction*, page 1–25. 1

[2] M. J. Kochnderfer and T. A. Wheeler. *Algorithms for optimization*. The MIT Press, 2019. 1

[3] C.-Y. Liong and S.-F. Foo. Comparison of linear discriminant analysis and logistic regression for data classification. *AIP Conference Proceedings*, 2013. 2

[4] I. H. Witten, E. Frank, and M. A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. Elsevier, 2011. 1