Student Number: 231012

1. Introduction

The task required for this binary class experiment involves the training of data to assess for ability to remember its contents. The solution to be presented involves a method of machine learning to correctly label a confidence rating of either 1 (for total confidence/ability to remember), 0.66 for partial ability or 0 not memorable. This report will discuss in part, pre-processing techniques, methods/models needed and finally the data obtained from classification techniques implemented.

2. Methodology

Due to the nature of the project, there are a set number of methods that could be used and implemented for this task. The most optimal solution is one in which the level of accuracy retrieved, is at its highest. The first of these methods is the usage of a Naïve Bayes (NB) Classifier, which utilizes the Bayes' Theorem to statistically train/build models to predict on probability [1]. In addition to this, other models like the use of a multi-layer perceptron [MLP], could prove to be invaluable. These networks integrate a method of generating an output through methods like backpropagation. Whilst a simplistic concept, the idea of generating an output by working backwards throughout a network in a nonlinear approach has proven to be particularly successful [2].

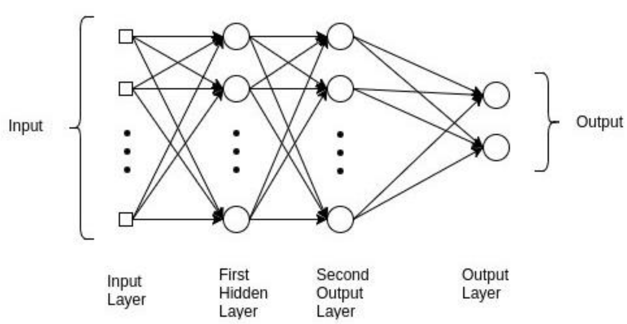

Figure 1: An example of a Multilayer Perceptron [3]

As seen in figure 1 as the name suggests, multiple hidden and output layers comprise of the metaphorical "meat" of the MLP. To be able to train this MLP, supervised training would have to occur, whereby forward propagation would take place, and then go backwards to assess if the end output resolves back into put. Overall, MLP structures are known to be especially effective at producing high accuracy ratings. Finally, the use of a Logistic Regression (LR) Modeling tool will play a big role in this task. LR modeling is a form of statistical analysis that can create predictions from a trained dataset. This is achieved through the usage of functions like that of sigmoidal relationships. This refers to an activation function that assists with non-linear values and makes a judgement on what to pass as an output [4]. LR modeling usually begins by defining, non-linear (but not restricted to it), boundaries which are then connected to a probability within the classifier. Furthermore, there are other certain events within a LR model that need to be accounted for, for example, Cross – Entropy (CE). CE within LR models, tends to represent the loss function, generally as a form of declaring the difference between level of uncertainty that a model has, and its realistic probability score. The general rule of thumb with LR modeling, is to reduce the CE function as much as possible to achieve the highest accuracy ratings.

3. Pre-Processing

Pre-processing refers to a state of normalization or data preparation for a set of procedures. This the state of this task, pre-processing has been used to investigate which sections of the training back can be procured to retrieve as optimal accuracy ratings as possible. To begin, an aspect of attempting to gain best results possible, is by removing points of weakness or delicacy. Since low confident data is rated from X>0.66, it would be ideal to remove anything that classifies as being X, thus providing a strong set of accurate predictions for a classifier to learn from. Once this process is complete, comparisons between GIST and CNN extraction methods need to occur. Implementing data from both these features tools, a level of importance indicator was retrieved.

```
0.00024420026
0.0019569471
```

Figure 2: GIST (Lower) and CNN (Top) Importance ratings.

Figure 2 shows that the GIST feature extractions are relatively more of significance and would be beneficial to use for greater accuracy, since there was better outlook on object recognition and scene-based classification [5].

The training set used within this project, also had instances of missing information, as commonly seen by "NaN".

To combat this dilemma, all these points of Anomaly had to be addressed through data imputation. This technique is best suited for these scenarios whereby missing information must be replaced using a substitute value approximate to pre-existing data. This results in a complete and concise data matrix to create accurate predicts with, because of SimpleInputer, a package by scikit.

Finally, the most important part is using the newly formatted data for the purpose of training and learning. Since currently there is no total confidence in understanding which model is superior in this task, a set of training data will be implemented to test 5 different sets of data.

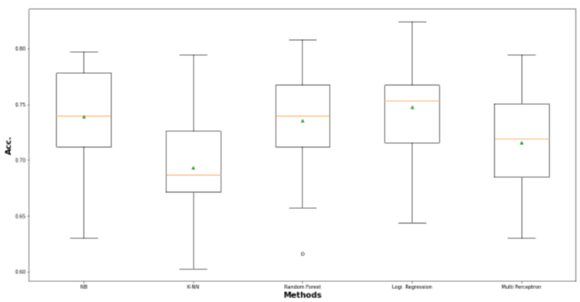

Figure 3: (From L to R) NB, K-NN, Random Forest & LR

```
NB : 0.739152 (0.044049)
K-NN : 0.693533 (0.044018)
Random Forest : 0.735505 (0.046478)
Logi. Regression : 0.747797 (0.040121)
Multi Perceptron : 0.715864 (0.045373)
```

Figure 4: Accuracy Values of classifiers

After the 5 different tests on different classifier methods, it is quite apparent that the LR Classification (as seen in figure 3 & 4) has a better accuracy score of 0.747797, with a better overall average. This was a result that was to be expected in some degree, as LR is a statistical analysis tool with a strong emphasis on memorization of objects/scenes, with strong cases of working effectively in cases of binary values. This means that for proper implementation, the LR will be the main subject of attempting to create a set of accurate predictions.

4. Results

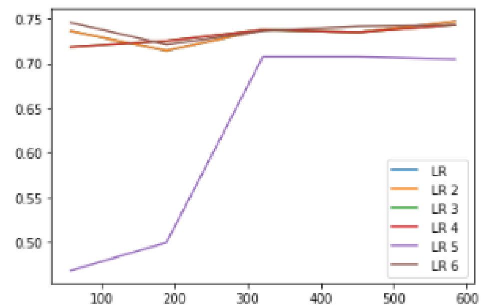

Figure 5: LR Learning Curve

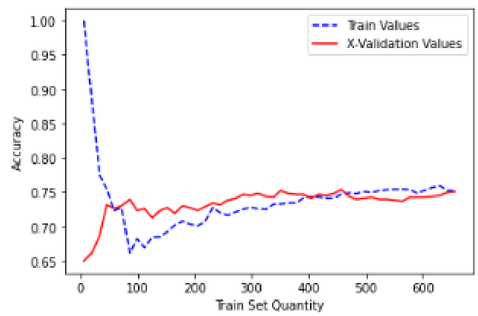

Figure 6: Learning curve of accuracy vs data-size

5. Discussion

Based on the data collected from the results, the LR worked quite effectively when it came to accurate predictions. Figure 5 shows that highest confidence rating to be around 0.750987, When examining figure 6, this is only more apparent as cross-validation averages overall accuracy just above the 0.75 mark. To potentially increase these values in a future experiment, it would be ideal to have a pre-existing dataset of values, instead of having to rely on external tools like data imputation. Furthermore, by further increasing the dataset, it may mean that the LR model would have a better amount of information collected to be tested against and hence, performs a lot more precisely.

References

[1] Ranganathan, S., Gribskov, M., Nakai, K. and Schönbach, C., 2019. *Encyclopedia of bioinformatics and computational biology*. 1st ed. pp.403-404.

[2] Chauvin, Y. and Rumelhart, D., 2009. *Back propagation*. New York: Psychology Press, pp.1-7.

[3] Peixoto, F., 2020. *A Simple Overview of Multilayer Perceptron (MLP) Deep Learning*. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2020/12/mlp-multilayer-perceptron-simple-overview/

[4] Kumawat, D., 2021. *Introduction to Logistic Regression - Sigmoid Function, Code Explanation | Analytics Steps*. [online] Analyticssteps.com. Available at: https://www.analyticssteps.com/blogs/introduction-logistic-regression-sigmoid-function-code-explanation

[5] J. Yin, H. Li and X. Jia, "Crater Detection Based on Gist Features," in *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, no. 1, pp. 23-29, Jan. 2015