**Paper Template for COMP30027 Report**

**Anonymous**

1. **Introduction**

IMDB ratings are votes (from 0 to 4) cast by IMDB users based on every released title in the database. IMDB ratings can act as strong predictors as to how a movie will be received by viewers, hence, by focusing on IMDB rating as a target, it can give insight to the performance of the movie.

IMDB ratings can be predicted by using several machine learning models. This report analyses different models and data pre-processing techniques to predict the IMDB rating of movies based on various features of the movie including (actor 1 name, director name, etc) given in a dataset.

This report explores the Gaussian Naïve Bayes Model, K-Nearest Neighbour Model, SVM Model and Neural Networks. These models have been implemented individually along with pre-processing techniques involving one-hot encoding and Normalisation of numerical data to evaluate model performance based on the percentage accuracy between the predicted the expected values of the IMDB ratings.

Training and testing of the accuracy of the models was achieved through the employment of the train-test split method, in which 80% of the training data was used to train the model, and 20% was used to test the model. In addition to this, the accuracy of the data was also examined by uploading the predicted IMDB ratings onto Kaggle, in which the percentage accuracy of the result was thereby evaluated.

1. **Methodology**

**2.1 General Pre-processing**

To perform the classification, feature encoding utilised to convert columns of string value to floats. Columns such as (‘language’, ‘country’, ‘content\_rating’, ‘movie\_title’, ‘actor\_3\_name’) underwent one hot encoding to change into numerical values. In addition, a correlation matrix was employed for feature selection. Features with a correlation of above 60% were removed to reduce the dimensionality of the dataset and thereby reducing the space and runtime. Removing closely correlated features also reduces noise in the dataset, resulting in higher predicted accuracy of the IMDB ratings.

* + 1. **Gaussian Naïve Bayes**

The Gaussian Naïve Bayes Model (GNB) utilises probabilistic classification based on the Bayes Theorem for predicting labels in the dataset. It is generally useful for continuous data and assumes independence between features of the dataset.

For the dataset, standardisation was selected over min-max scaling as GNB assumes its features to be normally distributed. Standardisation is more useful when the data is Gaussian, hence when using a Gaussian Naïve Bayes Model, standardisation should theoretically improve its accuracy.

The Gaussian Naïve Bayes Model was selected due to its suitability to handle continuous data such as (duration, budget etc).

* + 1. **K-Nearest Neighbour**

K-Nearest Neighbour operates by finding the K-Nearest Neighbours to a given datapoint and uses their class to predict the class or value of the new datapoint.

For the KNN model, the standardised dataset was used to train the model. Models such as KNN are most affected by the range of features present. Due to the large quantity of data, noise can be easily introduced as outliers in the dataset, hence standardisation was employed as pre-processing to train the model.

The KNN model was chosen for the dataset due to its non-parametric nature. Features such as movie ratings are useful for this, as the relationship between the features and the label (ratings) might not follow a specific statistical distribution.

* + 1. **SVM (Support Vector Machines)**

Support Vector Machines works by mapping data to a high dimensional feature space so data points can be categorised.

For SVM both Min-Max scaling and standardisation was used as pre-processing techniques to train the dataset. The kernel was set to linear to increase the computational efficiency as the final dataset after preprocessing had 12637 columns, hence resulting in faster training time.

SVM was chosen due to its effectivity for high dimensioned datasets. The training dataset includes numerous features such as textual data (plot keywords), categorical data (genres) and numerical data (budget, Facebook likes). SVM can handle this high dimensionality effectively, making it a suitable model for predicting the IMDB ratings based on the large quantity of features.

**2.1.4 MLP (Multilayer Perceptron)**

MLP are artificial neural networks that combines several neurons organised into 3 layers. MLP learns to map the input features to movie ratings through the iterative process of forward propagation, loss computation and backpropagation.

For the MLP classifier, hidden layer size was set to (100, 50) so that the model has two hidden layers (first containing 100 neurons, 2nd containing 50 neurons). This prevents both overfitting and underfitting of the dataset. In addition, increasing the number of neurons in the hidden layers increases the capacity of the model to learn from the data, however as capacity of the model increases, more intricate representations of the features of the input dataset is learnt, which can potentially result in overfitting of the data and inaccurate IMDB predictions.

1. **Results** 
   * 1. **Result for Gaussian Naïve Bayes**

From the training data, the standardised dataset that has been split into 80% training and 20% yielded an accuracy of 63.4%.

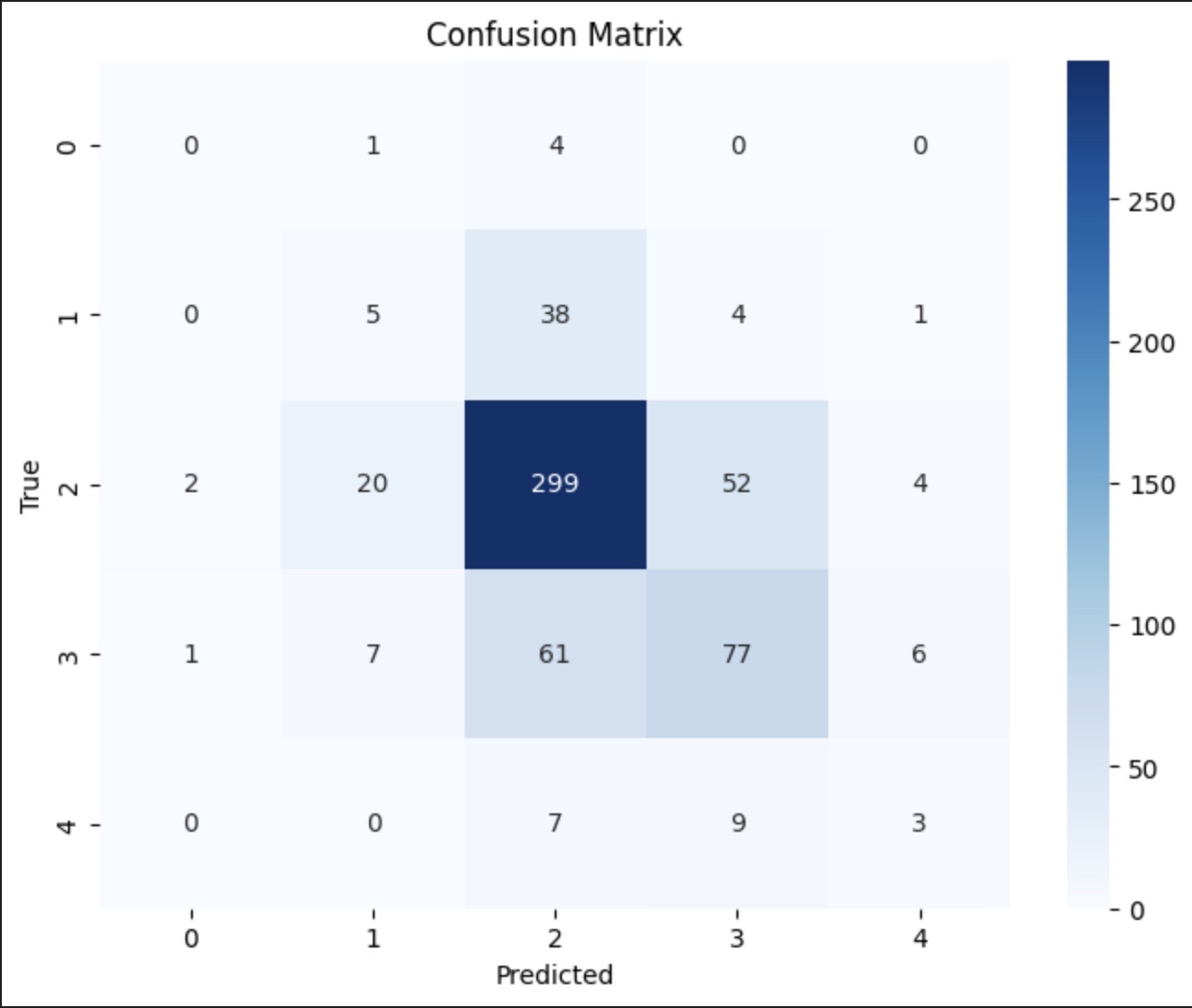


Figure 1

Figure 1 shows the confusion matrix generated for the Gaussian Naïve Bayes Model when training on the standardised train test split data. The axis labels (0-4) showcase the IMDB ratings predicted. From Figure 1, the diagonals show that an IMDB rating of 2 has the most accurate classification, followed by an IMDB rating of 3. Ratings of (0, 1, 4) had low accuracy.

However, when looking at the false positives and negatives of the IMDB ratings of 2, it is significantly higher than those of other ratings. Therefore, a high accuracy for a rating of 2 does not imply that the model performs well on 2-star ratings.

A graph of a bar chart

Description automatically generated

Figure 2

Figure 2 shows a histogram of the counts of the correctly classified labels with the counts of the occurrence of the actual labels in the dataset.

From the histogram, IMDB ratings of 2 has the highest occurrence in the actual IMDB ratings

* + 1. **Result for K Nearest Neighbour**

From the training data, the standardised dataset that has been split into 80% training and 20% yielded an accuracy of 62.7%.

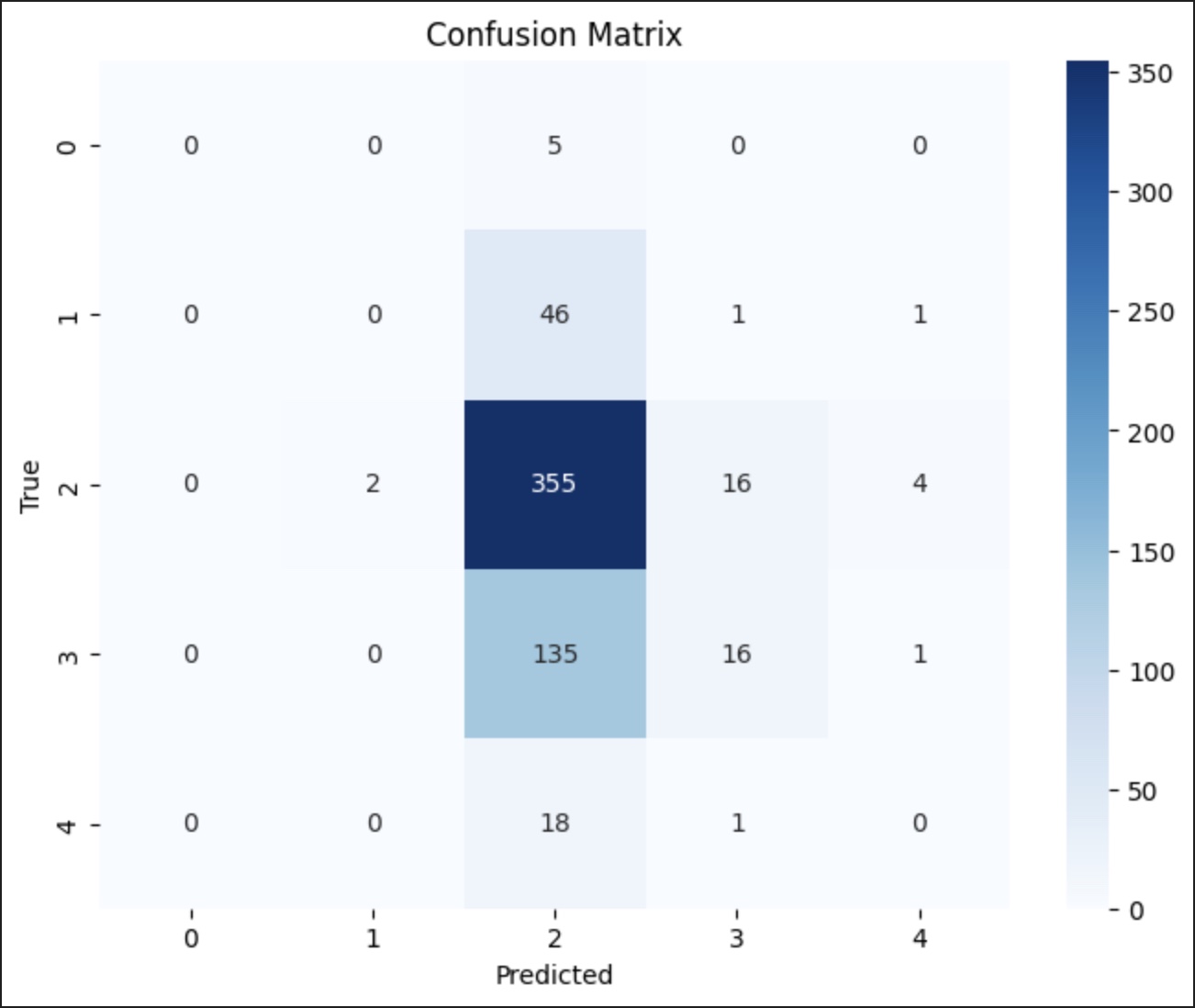


Figure 3

Figure 3 depicts the confusion matrix generated for the accuracy for IMDB rating predicts when using the KNN classifier on the standardised training dataset.

From the confusion matrix, it is evident that the IMDB rating of 2 showed the highest

A graph of a bar chart

Description automatically generated with medium confidence

Figure 4

* + 1. **Result for Support Vector Machines**

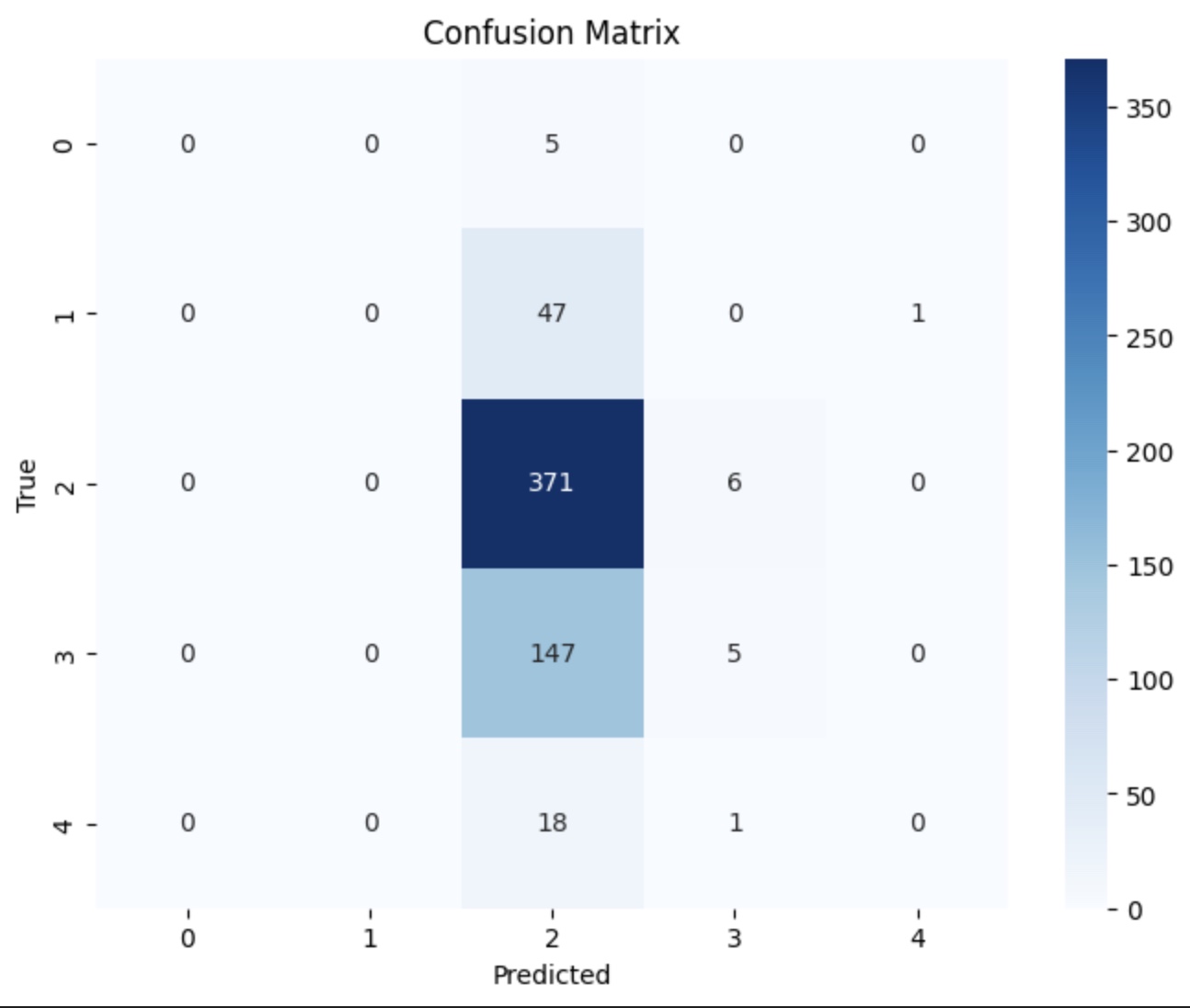
****

Figure 5

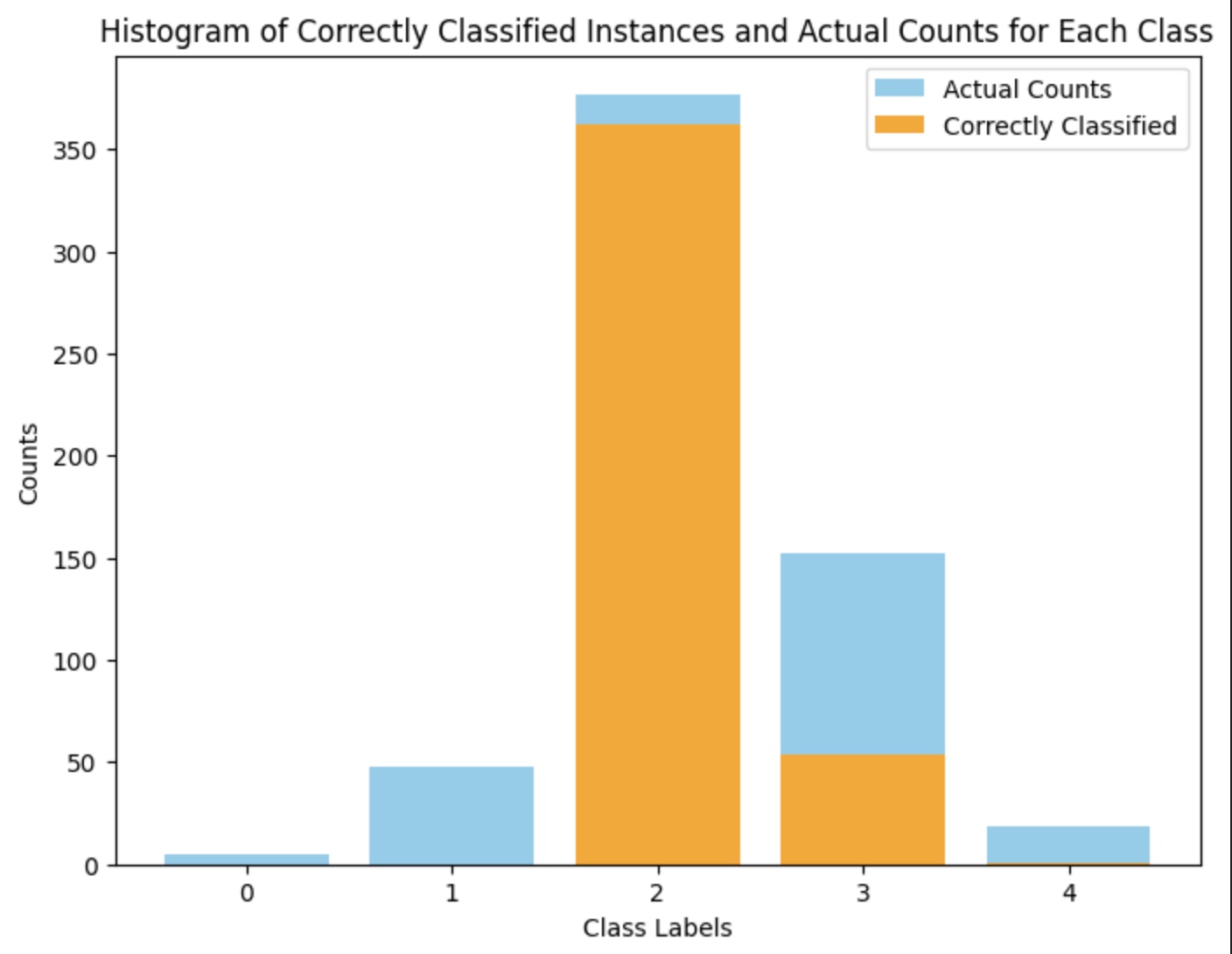
****

Figure 6

From the training data, the standardised dataset that has been split into 80% training and 20% yielded an accuracy of 69.4%.

* + 1. **Result for Multilayer Perceptron**

From the training data, the standardised dataset that has been split into 80% training and 20% had t

1. **Discussion and critical Analysis**

1. **Conclusion**
2. **References**