TASK 3

Introduction

Classification is a technique used to forecast either the objectives or categories of a data point and it is a supervised machine learning. Three classification models would be employed in this study to determine from a dataset whether a player would play in the NBA for more or less than five years. The models include Gaussian Naive Bayes, Logistic Regression, and Neural Networks. Target 5Yrs, a target variable in the dataset, has a binary classification with two classes because it falls between 0 and 1 according to the provided data. The multi-layer perception classifier (MPL classifier), which is frequently used for binary classification, was the neural network classifier that was utilized to generate the model. In the Python file Task3.py, the dataset for this task would be computed using machine learning.

Keynotes

- Accuracy of Prediction: The frequency with which the model forecasts the target variable.
- F1 Score: A precision and recall weighted average with values ranging from 0 to 1. The better the performance, the closer the F1 score is to one.

STEP1

The dataset comprises a total of 1340 inputs and 21 features. There are 11 missing values . As shown in Task3.pyfile, the missing values were substituted with their mean. The least and greatest number of games played was found to be 11 and 82, respectively.

To determine how each feature related to the target variable (Target 5yrs), a correlation matrix was applied to all of the features. It was discovered that the target variable and the total number of Games played have a high correlation. Based on the target variable and the number of games played, the dataset was grouped and recorded. It was found that 831 players had careers longer than or equal to five years, whereas 509 players had careers shorter than or equal to five years.

STEP2

The variable for input and output was described, reshaped, and scaled. The dataset was divided into training and testing for the models, using 1/3 to test the data and 2/3 to train the data. The dataset was fitted and predicted using the three model techniques of Logistic Regression, Gaussian Naive Bayes, and Neural Network (MPL Classifier) utilizing a single variable (features). The scatter plot on matplotlib was used to depict the models' predictions, as can be seen in Figures 1, 2, and 3, respectively.

Where:

X= Games played

Y=Target_5years

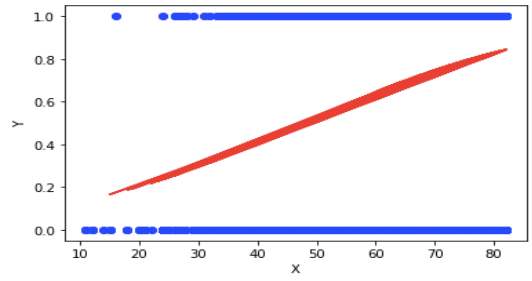


Figure 1: Logistic Regression Visualization

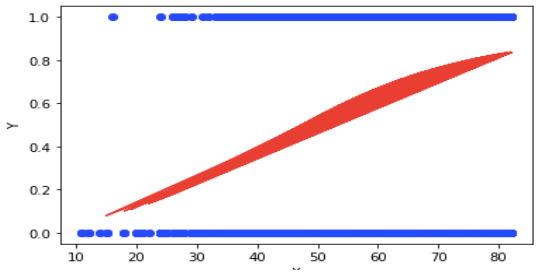


Figure 2: Guassian Naïve Bayes Visualization

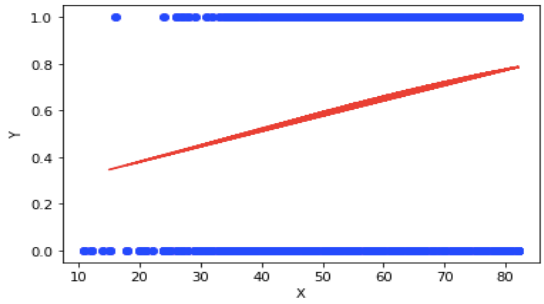


Figure3: Neural Network Visualization

Using One feature the accuracy score was 68.2%,68.2%, and 65.5% for logistics regression, GaussianNB and Neural Networks respectively. The three models showed fair performance because, in classification analysis, the closer the model accuracy value is to 1, the better the model performance.

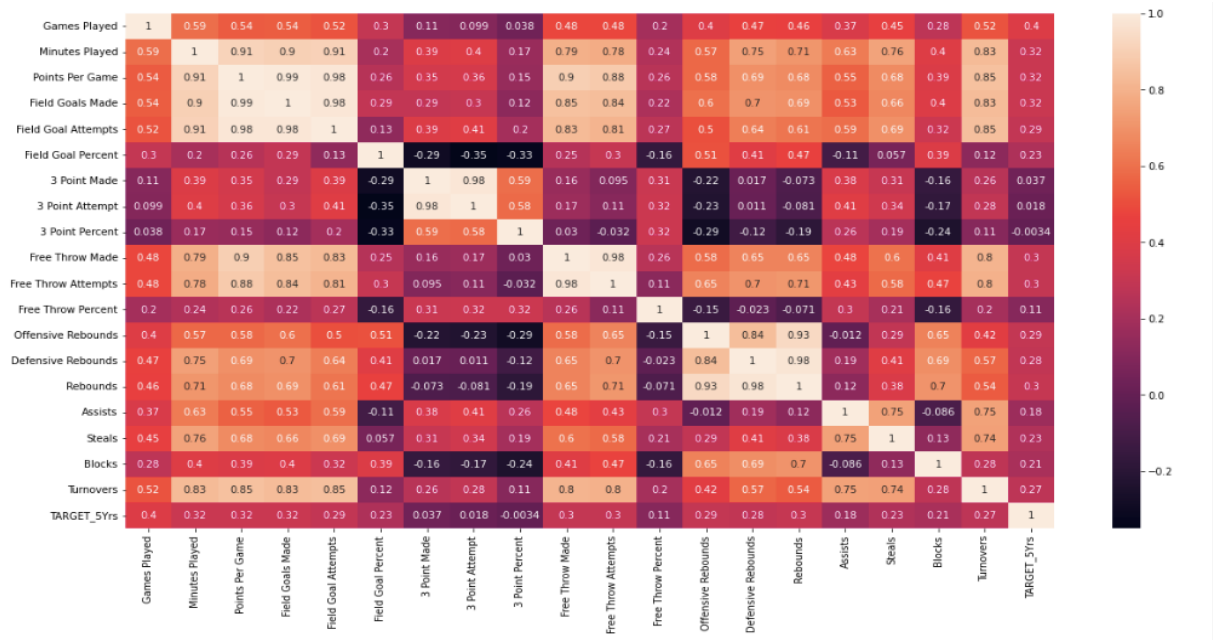
Using three features the accuracy score was 69%,70%, and 66% for logistics regression, GaussianNB and Neural Networks respectively. This result shows a slight improvement in all models.

Using all features the accuracy score was 71.6%,67.34%, and 72.04% for logistics regression, GaussianNB and Neural Networks respectively. The neural network showed a better

performance using all the features while GaussianNB shows a lesser performance using all features.

More features should be employed in the model's training in order to achieve improvement. The crucial elements required to develop the model can be chosen with the use of feature selection.

The correlation between the predictor and the target variable was determined using the correlation matrix depicted in the heatmap below. The correlation between the predictor and the target variable is strongest when it is near to 1. Based on this, it was found that the predictors and target variables had a poor association, with values less than 0.5.



Conclusion

The best model that fits the data set is the Neural Network. Showing a 72.04% chance of the player staying greater or equal to 5years using all features.