Football Match Result Prediction Analysis REPORT

Analysis of Football Match Prediction Results:

This study examines how well two machine learning models predict football games' full-time results (FTR) using data from the "soccer21-22[1].csv" dataset. Random Forest Classification and Linear Regression are the models that are investigated

Data preprocessing:

The preprocessing actions listed below are carried out by the supplied code:

Label Encoding: Label Encoding is the process of converting categorical information, such as Home Team, Away Team, FTR, HTR, and Referee, into numerical values. Machine learning models that use numerical data require this.

Training and testing sets of data are separated (train-test split). The testing set (20%) is used to evaluate the models' performance on unobserved data after they have been trained on the training set (80%).

Model Selection and Evaluation:

1. The Linear Regression Method:

Predicting match results is one classification problem that linear regression is often not appropriate for. It is included here for comparison's sake, though.

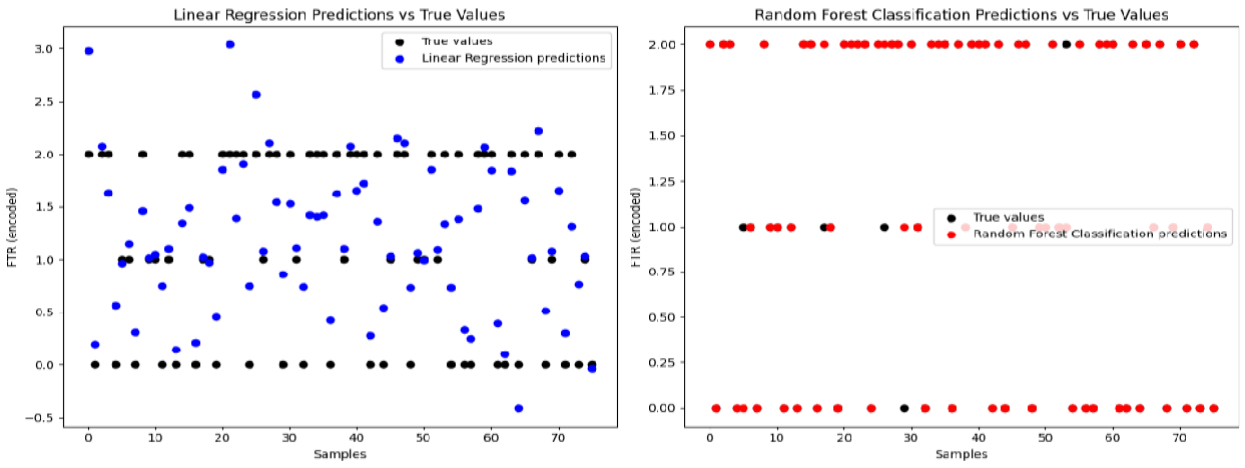
2. Random Forest Classification:

An effective approach for multi-class classification issues is Random Forest Classification. For increased accuracy, it constructs several decision trees and averages their forecasts.

The following metrics are used by the code to assess the models:

The Linear Regression:

These are the Graphs for linear regression and random forest classification:



The Mean Squared Error, or MSE, calculates the average squared difference between the actual and expected data. A lower MSE indicates better performance.
The Mean Absolute Error (MAE) is a statistical measure of the average absolute difference between the predicted and actual values. A lower MAE indicates better performance.
R-squared (R^2): Shows the proportion of the FTR volatility of the dependent variable that can be explained by the independent variables (match statistics). A greater R^2 indicates a better match.

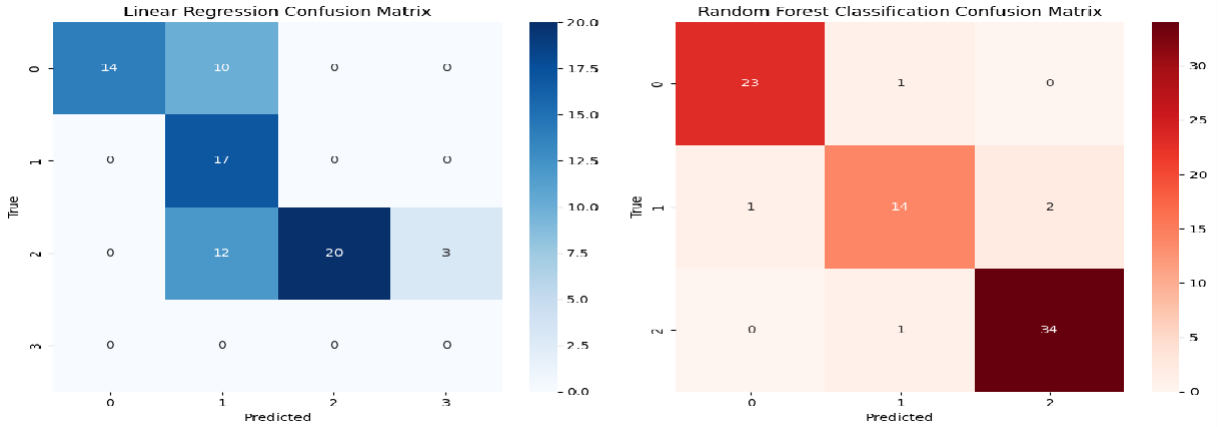
The Random Forest Categorization:

The percentage of accurately predicted FTR values is known as accuracy. Improved performance is indicated by higher accuracy.

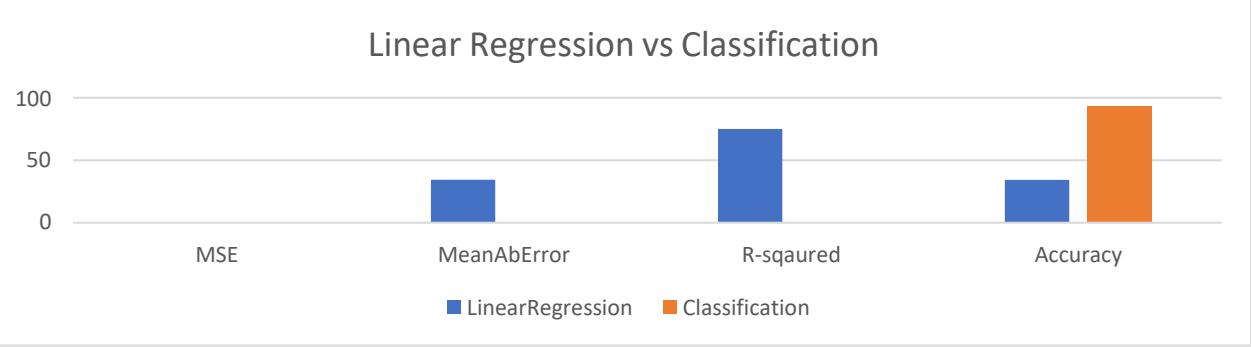
The computed values for these metrics are output by the code.

Visualisation:

The comparison graphs between linear regression and random forest classification



Bar graph:



Conclusion:

The advantages and disadvantages of various machine learning algorithms for forecasting football match outcomes are highlighted in this review. This assignment may not be best suited for Linear Regression; however, Random Forest Classification is a viable alternative.

References:

Data Preprocessing: Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

Model Selection and Evaluation:Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Science & Business Media.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). Introduction to Linear Regression Analysis. Wiley.

Visualisation:Tufte, E. R. (2001). The Visual Display of Quantitative Information. Graphics Press.

Google Colab Link:
https://colab.research.google.com/drive/1Jp2ilkU-oEbx3EMlnGnsyeXCRk1100RH?usp=sharing