Final Project Report
Machine Learning & Deep Learning

Project Title:
Football Match Prediction (FMP)

Prepared by:
MOHAMED SAID (N01654326)

AIGC 5002 – SEPTEMBER 2023
Humber College

1. INTRODUCTION

1.1 Background

Football's widespread popularity and the complexity of match outcomes make predictions challenging yet crucial for strategic and entertainment purposes. The shift from traditional expert-based forecasts to data-driven approaches has emphasized the need for accurate, analytical predictions in sports. This project addresses this need by offering data-backed insights for strategic planning.

1.2 Problem Statement

The Football Match Prediction project focuses on developing a model to forecast football match outcomes and identify key statistical metrics influencing these outcomes. It involves analyzing historical data to find patterns crucial for predictions. This prediction includes identifying whether it will be a home win, away win, or a draw as well as highlight the most impactful game performance metrics (stats).

1.3 Objectives:

- To Predict the Outcome of a football Match
- To Get Key Statistical metrics which significantly determine the outcome of a match.

1.4 Significance:

The importance of the Football Match Prediction project in the context of machine learning (ML) and deep learning (DL) lies in its potential to showcase the capability of these technologies in accurately analyzing and predicting complex, real-world outcomes. It demonstrates how ML/DL can handle vast datasets to uncover hidden patterns and insights, contributing to strategic decision-making in sports. Although I employed one of the foundational methods to address the project goals, sports science is a rapidly growing field that will significantly contribute to advancements in ML/DL.

1.5 Impact:

Teams, coaches, analysts, and football fans will benefit from improved strategic decision-making and enhanced understanding of game dynamics through data-driven insights provided by this project's predictive modeling.

2. METHODOLOGY

2.1 Data Source:

Data for all English Premier League matches which was collected via web scraping from the official Premier League website and uploaded to Kaggle by P. FREITAS[i]. There are 4,070 matches, from 2010 to 2021, with 113 stats features for each match and team performance throughout the season at the time of each match.

2.2 Data Preprocessing:

I. Analyzed and understood all the dataset columns. Selected 24 most common stats based on industry knowledge.
II. Created target class ['Result_Class'] from full time result score feature.
III. Checked for class imbalance.
IV. Converted class label from string values to integer as ML deals with INT better.
V. Used Pearson correlation analysis to select most impactful features on class label. 5 positive correlation (home team features) and 5 negative correlation (away team features).
VI. Final features scaled using MinMaxScalar.
VII. Detected using IQR and dealt with outliers by removing them.
VIII. Dataset split: Training 80% and Testing 20%

2.3 Model Selection:

Logistic regression with One-vs-Rest (OvR) was chosen for its simplicity and efficacy in multiclass problems, like predicting football match outcomes. Additionally, research, notably the "Machine Learning Models Reveal Key Performance Metrics of Football Players to Win Matches in Qatar Stars League"[ii] paper, demonstrated logistic regression's superior performance in similar predictive tasks, further validating its selection for this project.

2.4 Training Process:

HPO Training selected parameters:

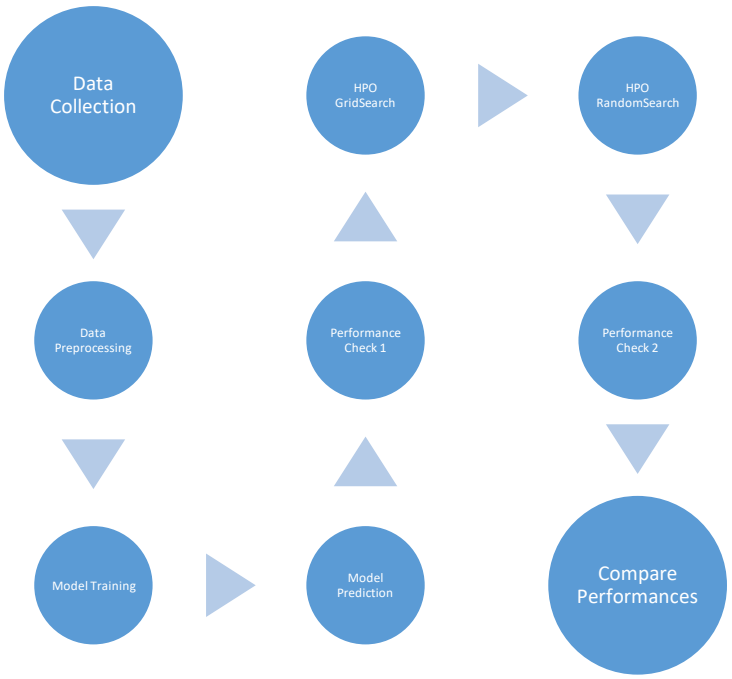
C = Represents the inverse of regularization strength in logistic regression.

Solver = Different solvers can be more effective for different types of datasets and may have different computational efficiency.

Penalty = Determines the type of regularization applied to the model. L1 (Lasso regularization) or L2 (Ridge regularization).

Max_iter = Set the maximum iteration to 1000 for model to converge.

3. IMPLEMENTATION




```
# Initialize logistic regression model using one-versus-rest method
logreg = LogisticRegression(multi_class='ovr', random_state=42)
logreg.fit(X_train, y_train) # Train the model
y_pred = logreg.predict(X_test) # Predict on the test set
```

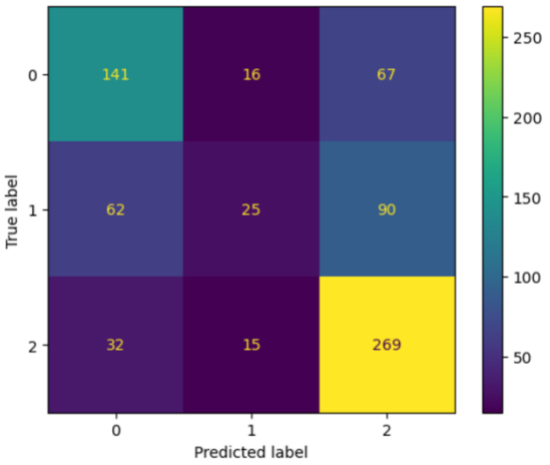
```
%%time # To display the time it took for cell code to run. To time how long HPO took.
# The parameter grid for RandomizedSearchCV
param_dist = {
    'C': np.logspace(-4, 4, 20),  # Regularization strength
    'solver': ['newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga'],  # Algorithms to use in the optimization
    'max_iter': [1000]
}
# RandomizedSearchCV with Logistic Regression model
random_search = RandomizedSearchCV(LogisticRegression(multi_class='ovr', random_state=0), param_distributions=param_dist,
                                   n_iter=10, cv=5, scoring='accuracy', random_state=0)
# Fitting to the data
random_search.fit(X_train, y_train)
# Best parameters and best score
best_params_random = random_search.best_params_
best_score_random = random_search.best_score_
```

4. RESULTS AND DISCUSSIONS

The initial model was evaluated based on accuracy, precision, recall, and F1-score. The accuracy achieved was approximately 60.9%, with varying performance across the different classes. The model showed a better ability to predict certain outcomes (Class 2) compared to others (Class 1).

	precision	recall	f1-score
0	0.603448	0.625000	0.614035
1	0.510638	0.135593	0.214286
2	0.623288	0.863924	0.724138
accuracy	0.609484	0.609484	0.609484

Following hyperparameter optimization—adjusting C, solver, penalty, and max_iter through both Grid and Randomized SearchCV—the model's accuracy improved by 1.8%, reaching 62.83%.



New Accuracy after HPO: 62.83%

The final model's performance, post hyperparameter tuning, indicated modest accuracy in predicting soccer match outcomes. The findings suggest that while logistic regression provided a baseline, the complexity of predicting soccer match outcomes might require more sophisticated models or feature engineering.

5. ETHICAL

No customer or player personal data were collected; all data utilized were from public records.

The 'team name' column was omitted to prevent bias, though future studies may consider including it to evaluate its influence on model accuracy.

6. CONCLUSION

The project demonstrated the feasibility of using logistic regression for predicting soccer match outcomes. Future work could explore alternative algorithms, such as Random Forests or Neural Networks, and incorporate additional features like team form, weather conditions, or player statistics to enhance prediction accuracy.

CITATION

[i] **Dataset Source:** All Premier League matches from 2010 to 2021 https://www.kaggle.com/datasets/pablohfreitas/all-premier-league-matches-20102021

[ii] **Researched Paper:** Machine Learning Models Reveal Key Performance Metrics of Football Players to Win Matches in Qatar Stars League https://ieeexplore.ieee.org/document/9261335