# Project Title: Predicting English Premier League Match Outcomes Using Machine Learning

## 1. Introduction

### Why Predict Match Outcomes?

The outcome of football matches, particularly in highly competitive leagues like the English Premier League (EPL), is of great interest to a wide audience including teams, coaches, analysts, and sports bettors. Predicting match outcomes can provide tactical insights for coaches, help analysts identify key patterns and trends, and assist in strategic decision-making. Furthermore, accurate predictions can offer an edge in the betting markets, making it a valuable tool for investors and traders.

The challenge in predicting football matches lies in the complex interplay of factors such as team strategies, formations, individual player performances, historical form, and even psychological factors like home crowd support. Machine learning models offer a unique approach to dealing with such complexity by identifying hidden patterns in historical data, helping make more accurate and data-driven predictions.

### Importance of This Approach

This project takes a focused approach by predicting three possible outcomes (home win, away win, or draw) rather than dealing with the complexities of each team’s individual performance. The use of historical data, including formations, set pieces, and audience impact, combined with machine learning models like logistic regression and KNN, allows for a comprehensive and systematic analysis of match outcomes. The insights derived from these models could be extended beyond match predictions to tactical decision-making and performance analysis.

# 2. Data Collection and Features

## 2.1 Data Source

FBref.com: EPL match data for the last two seasons was gathered from this website. The dataset includes detailed match events, player statistics, team formations, and more.

## 2.2 Data Features

Match Metadata:

* Home and Away Teams
* Match Date
* Outcome (Home Win, Away Win, Draw)

Formations:

* Commonly used team formations (e.g., 4-3-3, 4-4-2, 3-5-2) and their impact on match results.

Set Pieces:

* Corners, free kicks, penalties, and their relationship to goals scored.

Match Statistics:

* Shots attempted, shots on target, possession, tackles, and fouls.

Audience Impact:

* Number of spectators for home wins, draws, and losses.

Derived Features:

- Recent team form based on the last 10 matches.

- Head-to-head historical performance.

- Win percentage for home teams based on recent performances.

# 3. Data Preprocessing

## 3.1 Cleaning and Handling Missing Data

- Drop irrelevant columns.

- Handle missing values using appropriate imputation techniques.

## 3.2 Feature Engineering

Recent Form Calculation:

* For each team, calculate recent form by assigning a score to their last 10 matches (win = 3, draw = 1, loss = 0).

A graph with red and green dots

Description automatically generated

Win Percentage:

* Calculate the win percentage for home teams over their last 10 home matches.

A graph of red and green dots

Description automatically generated

Audience Normalization:

* Normalize audience data to account for outliers in attendance (e.g., matches played during COVID-19 restrictions).

## 3.3 Data Split

- Split the data into 80% training and 20% test sets.

- Perform stratified k-fold cross-validation (10 folds) to ensure balanced class representation across folds.

# 4. Exploratory Data Analysis (EDA)

## 4.1 Formation Analysis

Objective:

* Determine if certain formations contribute to a higher likelihood of winning.

Methodology:

* Calculate win rates for each formation.
* Conduct a chi-square test of independence to evaluate if formations significantly affect match outcomes.

Findings:

* Formation 4-3-3 has the highest win rate.
* The Chi-square test resulted in a p-value of 0.000004117, indicating a significant relationship between formation and match outcome.

A graph with numbers and a green line

Description automatically generated

## 4.2 Audience Impact Analysis

Objective:

* Investigate the impact of crowd size on match outcomes.

Methodology:

* Compare average audience sizes for matches where home teams win, lose, or draw.

Findings:

* Home wins average around 55,000 spectators.
* The draws have an average of 33,000 spectators.
* Losses average around 30,000 spectators.

A graph showing a number of people

Description automatically generated with medium confidence

## 4.3 Correlation Analysis

Objective:

* Examine which matches statistics most influence goals scored.

Methodology:

* Correlate goals scored with variables such as shots attempted, shots on target, corners, free kicks, and penalties.

Findings:

* Shots attempted and shots on target show the highest correlation with goals scored.
* Corners also make a meaningful contribution to goal-scoring opportunities.

A graph of a graph with blue squares

Description automatically generated with medium confidence

# 5. Model Development

## 5.1 Logistic Regression

Model Objective:

* Classify matches into home win, away win, or draw based on input features.

Cross-Validation:

* 10-fold cross-validation used for model training.

Performance Metrics:

Precision:

* Home win: 61%
* Away win: 56%
* Draw: 0%

Recall:

* Home win: 87%
* Away win: 56%
* Draw: 0%

F1-Score: Home win 72%, Away win 56%, Draw 0%

Overall Accuracy: 60%

A graph with blue lines

Description automatically generated

## 5.2 K-Nearest Neighbors (KNN)

Model Objective:

* Classify match outcomes based on the similarity to nearby data points.

Model Details:

* Used 9 nearest neighbors.
* Euclidean distance metric for neighbor selection.

Performance Metrics:

* Home win accuracy: 18/22 (~82%)
* Away win accuracy: 17/24 (~71%)
* Draw accuracy: 3/13 (~23%)

A graph with a line

Description automatically generated

# 6. Model Evaluation

Comparison of Models:

* Logistic regression performed reasonably well for home and away wins but struggled to classify draws.
* KNN showed better performance for home and away wins but also faced difficulties with draw predictions.

Evaluation Challenges:

* Class imbalance: Draws are underrepresented in the dataset, leading to lower predictive power for this class.
* Further model tuning is needed to improve draw prediction accuracy.

# 7. Conclusion

Conclusion:

* Team formations and audience size significantly affect match outcomes.
* Logistic regression is a simple yet effective model for home and away win predictions, but it performs poorly for draws.
* KNN can be improved by exploring more sophisticated distance metrics and neighbor selection strategies.