**Methodology Document EPL\_Challenge**

**By Daan Quaadvliet (231146)**

**Dataset Overview**

The provided datasets contain historical match, player, and team data from the English Premier League (EPL). The data was used to analyse player values for the 2014-2015 season and predict match outcomes for the 2015-2016 season.

**Datasets Used:**

* **epl\_matches\_train.csv** (2008-2015): Includes detailed match information, team formations, player lineups, match statistics (fouls, shots, possession, etc.).
* **epl\_matches\_test.csv** (2015-2016): Includes match details for the season to be predicted.
* **epl\_players.csv**: Contains player attributes (e.g., attacking work rate, defensive work rate, physical and technical abilities).
* **epl\_teams.csv**: Includes team-specific attributes (e.g., defense pressure, chance creation, passing).
* **epl\_goals.csv**: Records every goal scored from 2008-2015, including goal type and assisting player.
* **epl\_potential\_shots.csv**: Tracks shots (both on and off target) across past matches.

**Part 1: Identifying the Most and Least Valuable Players (2014-2015 Season)**

**Objective:**

Identify the **top 10 most** and **bottom 10 least** valuable players based on their contributions in the 2014-2015 season.

**Approach:**

Players were **evaluated differently** based on their **position (Goalkeeper, Defender, Midfielder, Attacker)** to ensure fair comparisons.

**Steps Followed:**

**1. Data Preprocessing**

* **Filtered** epl\_goals.csv and epl\_potential\_shots.csv for the **2014-2015 season** using match IDs.
* **Grouped data by player ID** to calculate the total number of **goals** and **close shots** for each player.
* **Merged** these statistics with epl\_players.csv to incorporate player-specific attributes.

**2. Position-Based Metrics**

Each position had a separate valuation formula based on **key attributes** that define player performance.

**For Goalkeepers:**

* **Key Metrics:**
  + Reflexes
  + Diving
  + Handling
  + Positioning
  + Jumping
  + Strength
  + Stamina
* **Final Metric Formula:**

GK Value = (Reflexes \* 0.25) + (Diving \* 0.20) + (Handling \* 0.20) +

(Positioning \* 0.15) + (Jumping \* 0.10) + (Strength \* 0.05) +

(Stamina \* 0.05)

**For Defenders:**

* **Key Metrics:**
  + Marking
  + Standing tackle
  + Sliding tackle
  + Interceptions
  + Strength
  + Aggression
  + Stamina
* **Final Metric Formula:**

DEF Value = (Marking \* 0.20) + (Standing Tackle \* 0.20) + (Sliding Tackle \* 0.15) +

(Interceptions \* 0.15) + (Strength \* 0.10) + (Aggression \* 0.10) +

(Stamina \* 0.10)

**For Midfielders:**

* **Key Metrics:**
  + Short passing
  + Long passing
  + Vision
  + Ball control
  + Dribbling
  + Positioning
  + Stamina
* **Final Metric Formula:**

MID Value = (Short Passing \* 0.20) + (Long Passing \* 0.15) + (Vision \* 0.15) +

(Ball Control \* 0.15) + (Dribbling \* 0.15) + (Positioning \* 0.10) +

(Stamina \* 0.10)

**For Attackers:**

* **Key Metrics:**
  + Finishing
  + Shot power
  + Positioning
  + Acceleration
  + Dribbling
  + Balance
  + Strength
* **Final Metric Formula:**

ATT Value = (Finishing \* 0.25) + (Shot Power \* 0.20) + (Positioning \* 0.15) +

(Acceleration \* 0.15) + (Dribbling \* 0.10) + (Balance \* 0.10) +

(Strength \* 0.05)

**3. Normalization & Ranking**

* **Normalized each metric** between 0 and 1 to ensure fair comparison.
* **Ranked players by position** and extracted the **top 10** and **bottom 10**.
* **Saved the results in** player\_list\_submission.csv.

**Part 2: Predicting Match Outcomes (2015-2016 Season)**

**Objective:**

Train a model to predict match outcomes (Win, Draw, Lose) using historical match data.

**Steps Followed:**

**1. Extracting Team Performance Metrics**

* Computed **home team statistics** from epl\_matches\_train.csv:
  + Home Win Rate
  + Home Draw Rate
  + Home Loss Rate
  + Average Goals Scored at Home
  + Average Goals Conceded at Home
* Computed **away team statistics**:
  + Away Win Rate
  + Away Draw Rate
  + Away Loss Rate
  + Average Goals Scored Away
  + Average Goals Conceded Away
* Merged these stats with epl\_matches\_train.csv for model training.

**2. Feature Engineering**

* **Extracted additional attributes** from epl\_teams.csv:
  + **Chance Creation (Passing & Shooting)**
  + **Defensive Pressure & Aggression**
  + These attributes help measure **team quality**.
* **Final Selected Features:**
  + home\_win\_rate, home\_draw\_rate, home\_loss\_rate
  + home\_goals\_scored, home\_goals\_conceded
  + away\_win\_rate, away\_draw\_rate, away\_loss\_rate
  + away\_goals\_scored, away\_goals\_conceded
  + home\_passing, home\_shooting, home\_defense\_pressure, home\_defense\_aggression
  + away\_passing, away\_shooting, away\_defense\_pressure, away\_defense\_aggression
* **Match Outcome Labeling:**
  + Win = 1
  + Draw = 0
  + Lose = -1

**3. Model Training**

* **Split training data into 80% training / 20% validation.**
* **Standardized features** using StandardScaler().
* **Trained a Random Forest Classifier (n\_estimators=150).**
* **Evaluated model performance** using:
  + **Accuracy Score**
  + **Precision, Recall, and F1-score**

**4. Match Prediction & Final Standings**

* Applied the **trained model** to the **2015-2016 test set**.
* Predicted match results and **assigned points**:
  + **Win → 3 points for home team**
  + **Draw → 1 point for both teams**
  + **Loss → 3 points for away team**
* **Computed final EPL standings** by summing points for each team.
* **Saved results in** prediction\_submission.csv.

**Final Deliverables**

1. **player\_list\_submission.csv** – List of 10 most & least valuable players.
2. **prediction\_submission.csv** – Match results for 2015-2016.
3. **Python Scripts** – Full code for both parts.
4. **This methodology document** – Explanation of metrics and procedures.

**Conclusion**

* **Player valuation was done per position** to ensure fairness.
* **Match prediction used team statistics & advanced attributes** to improve accuracy.
* The **Random Forest model achieved ~49% accuracy**, which is reasonable for match predictions given randomness in football.

This methodology ensures **clear, structured, and accurate** analysis of EPL data.