Machine Learning - Assignment 2

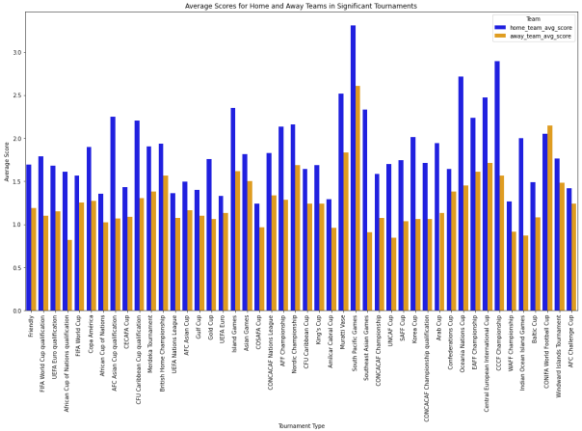
International football results Data insights
and Home team Winning prediction

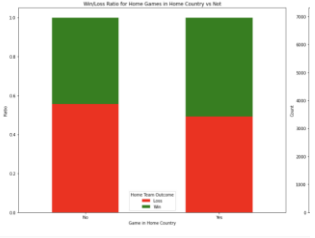
Amit Malik | **Nir Manor**

Data Exploration & Visualization

- **Impact of Location:** Home teams are more likely to win when playing in their own country, though the effect is minor.

- Across most tournaments, home teams generally have higher average scores compared to away teams. This indicates a significant home advantage in many of the "big" tournaments. We Can also notice there are specific tournaments types in which Home-Team as noticeable advantage





Home country Games

Games Not in Home Country:

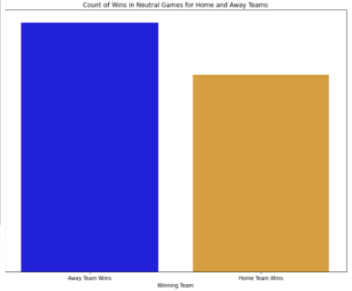
Home teams have a lower win ratio and Tends to **lose more**

Games in Home Country:

Home teams have a lower loss ratio and tend to **win more** by a bit

Insight:

Home teams are more likely to win when playing in their own country. However, The effect seems Minor



Neutral Games

Home Team Wins:

Home teams win **less often in neutral games**.

Away Team Wins:

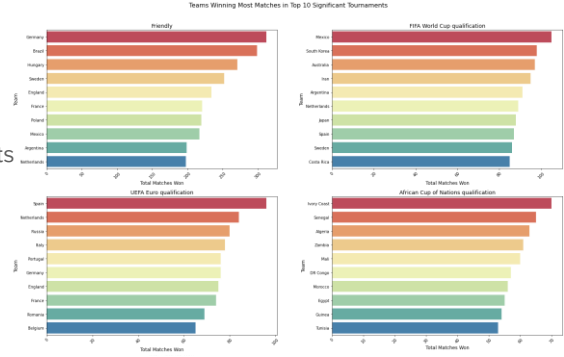
Away teams win **more often in neutral games.**

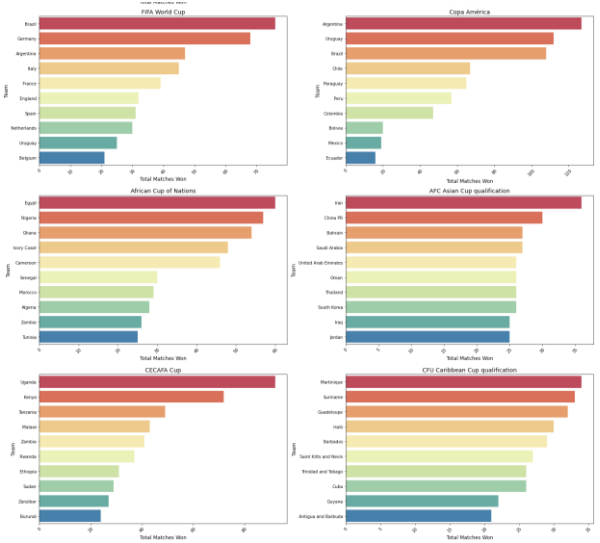
Insight:

In neutral games, away teams tend to win more frequently than home teams

- **Countries who Scored the most at Top 10 significant tournaments, includes teams that Appear in Multiple Tournaments as Top scorers - which indicates they strength**

Germany: Appears in 4 tournaments
Brazil: Appears in 4 tournaments
Argentina: Appears in 3 tournaments
Netherlands: Appears in 3 tournaments
Mexico: Appears in 3 tournaments
Spain: Appears in 3 tournaments
England: Appears in 3 tournaments
France: Appears in 3 tournaments
Italy: Appears in 3 tournaments
Sweden: Appears in 3 tournaments





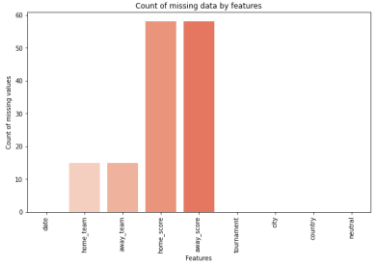
Data Preprocessing

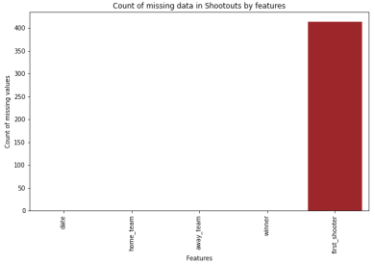
- **Imputation and Drops:**
  results due to high Missing values in 'home_score', 'away_score', 'home_team', and 'away_team' were addressed using imputation and dropping rows.
  We discovered that the data contains matches that **has not occurred yet**
  shootouts - **'first_shooter'** feature column was dropped as it seems irrelavant ~ 400 missing values from ~ 600 data rows!
  GoalScorers - Filling **'scorer'** missing values with the player who scored the most

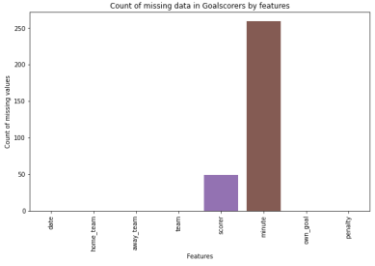
Filling **'minute'** missing values with the average of scoring minute for that team

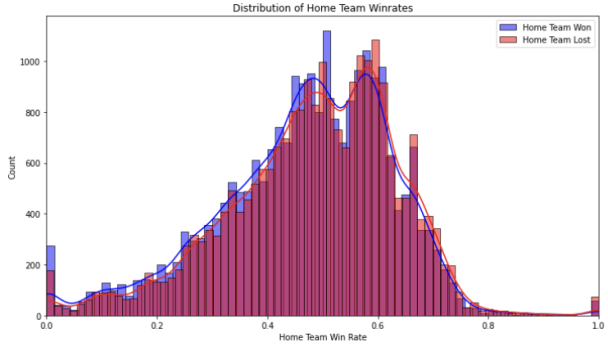
- **Transformation:** Boolean **'neutral'** feature transformed to 0/1

- **Feature Engineering:** features such as 'Home team won', 'Home team win rate', 'Away team win rate', 'Home team average goals', and 'Away team average goals' were created.
  We also created new features from our own:
  - Head-to-Head Win Ratio: Historical win ratio of the home team against the away team.
  - Head-to-Head Goal Difference: Historical goal difference between the home and away team in previous encounters
  - **Home Team Recent Performance**: Average points (win = 3, draw = 1, loss = 0) obtained by the home team in the last 7 matches.
  - **Away Team Recent Performance:** Average points obtained by the away team in the last 7 matches.
  -**Tournament Experience:** each team exp. In each tournament
  -Total Teams Goals (up to the match)
  - ELO rating

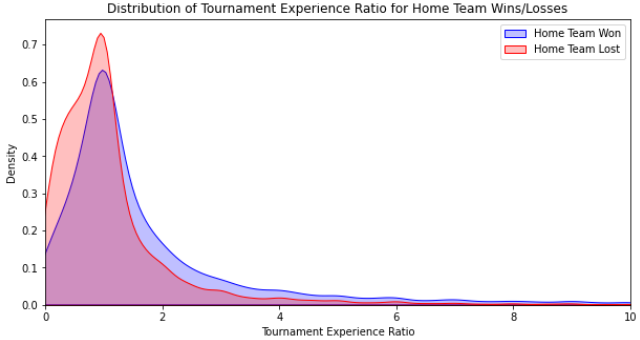






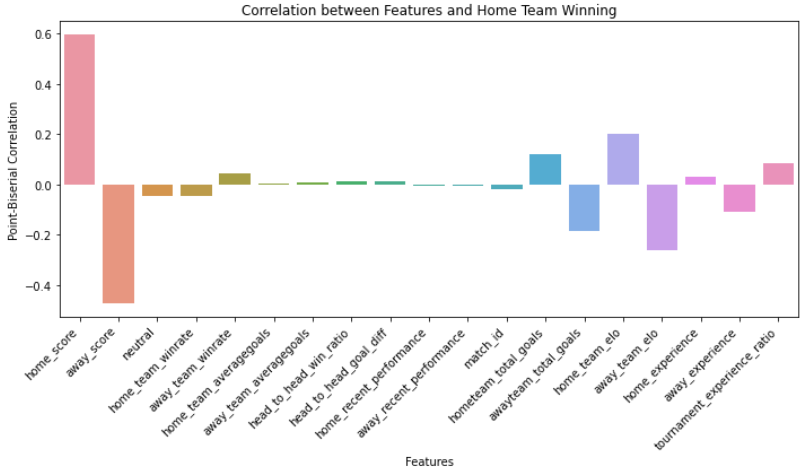


We were surprised to see that Winrate had such a small impact as it grows



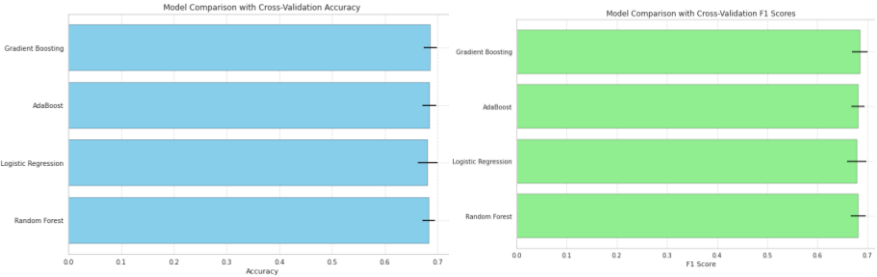
We can see that as the tournament exp. Ratio (home team/away team exp.) **in each tournament affected home team winning while was >1**

Home Team Winning Prediction

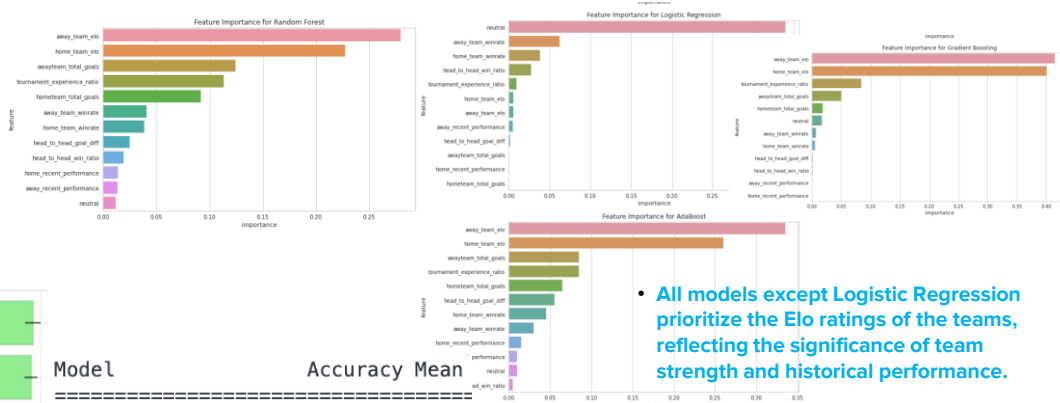


- Importance of Visualization:
  Identifying Relationships: Correlation plots reveal the strength and direction of the relationship between variables, guiding us in refining our model by focusing on the most influential features.
  Interesting Findings - Low Correlation Surprises (probably dur to poor data quality)
  Average Goals: Contrary to our expectations, features like the average goals scored by teams have a minimal correlation with the home team's chances of winning. This suggests that merely scoring more goals on average is not a strong predictor of winning at home.
  Head-to-Head Win Ratio: Similarly, the historical head-to-head win ratio between teams shows surprisingly low correlation with the home team's winning probability. This indicates that past match outcomes between the same teams do not significantly influence the result of a new match.
  Home Team's Recent Performance: Surprisingly, the recent performance of the home team also exhibits a low correlation with the home team winning. This suggests that even if the home team has been performing well in recent matches, it doesn't strongly predict their chances of winning the current match.

- **Model Selection:** 4 machine learning models were selected
  Random Forest, Logistic Regression, AdaBoost, and Gradient Boosting
- **Parameter Tuning:** Hyperparameters for each model were tuned to optimize performance.
- **Evaluation:** Models were evaluated using accuracy, precision, and recall metrics.



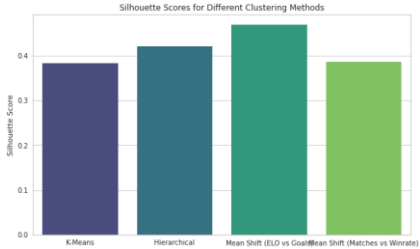
Model	Accuracy Mean
Random Forest	0.6832
Logistic Regression	0.6814
AdaBoost	0.6845
Gradient Boosting	0.6867

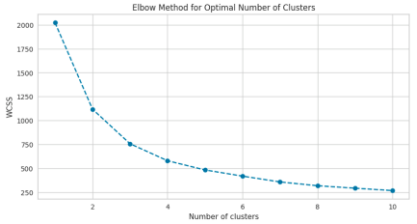


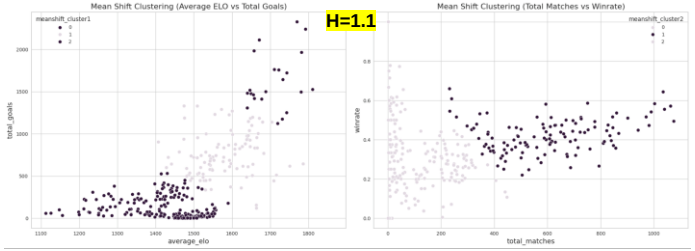
- **All models except Logistic Regression prioritize the Elo ratings of the teams, reflecting the significance of team strength and historical performance.**
- Total goals scored by the teams and their experience in tournaments are also common important features for the ensemble methods (Random Forest, AdaBoost, Gradient Boosting).

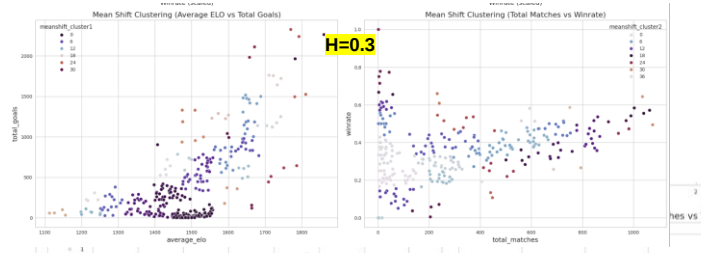
Clustering Teams

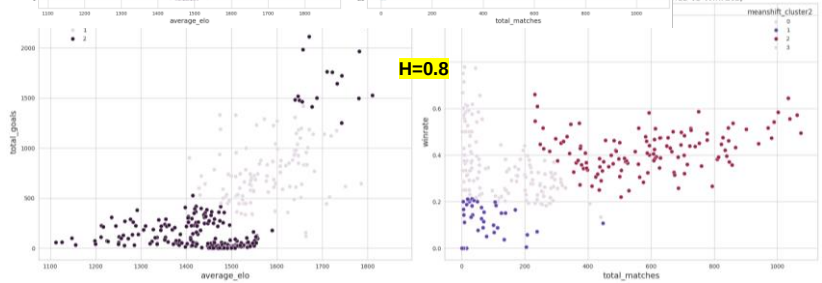
- **Algorithms Used:**
  K-Means , SVM , Hierarchical clustering , Mean Shift
- **Mean Shift Clustering appears to be the best ,**
  however it took a lot of parameter's tuning (bandwidth) and feature selection until we have reached to optimal clustering





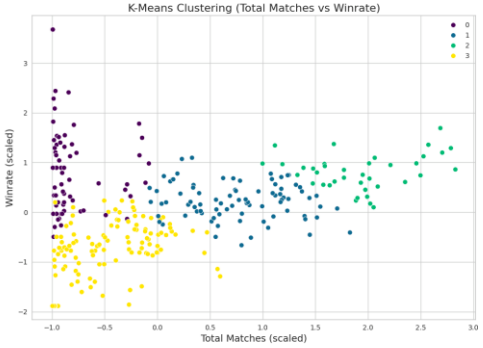


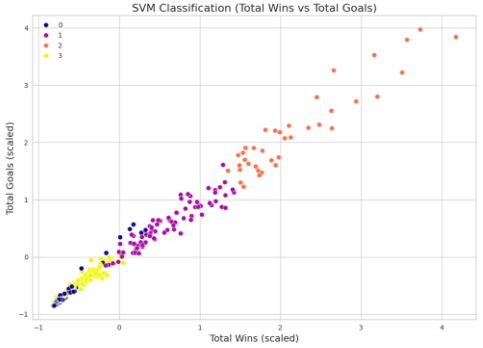


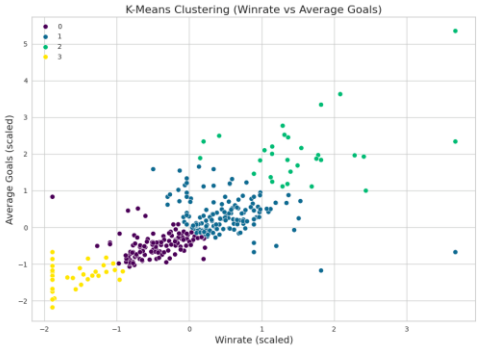


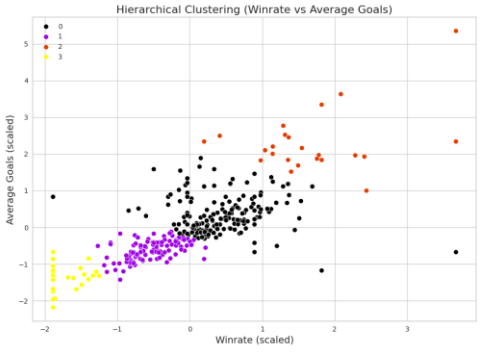
Clustering Different Methods & Features Visualizations

Choosing the correct features also took an important part at the rest of the methods







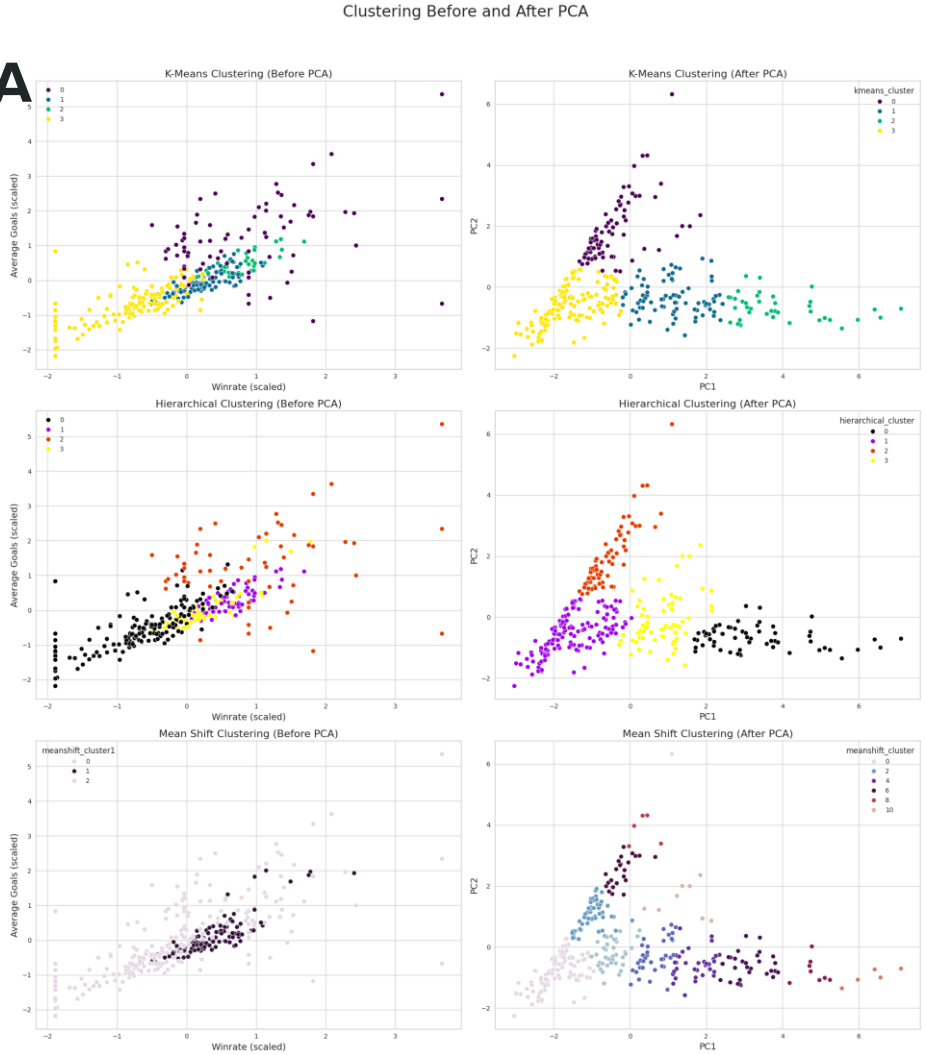


Dimensionality Reduction with PCA

- **Principal Component Analysis (PCA):** PCA was applied to reduce the dimensions of the teams' data, capturing the majority of variance with fewer components.

To estimate the quality of the clusters, the Silhouette Score was used. This method measures how similar each point is to its own cluster compared to other clusters. High silhouette scores indicate well-defined clusters with significant intra-cluster similarity and inter-cluster dissimilarity.
Silhouette Scores for Clustering Methods:
•K-Means Clustering Silhouette Score (Before PCA): 0.383
•K-Means Clustering Silhouette Score (After PCA): 0.474
•Hierarchical Clustering Silhouette Score (Before PCA): 0.420
•Hierarchical Clustering Silhouette Score (After PCA): 0.439
•Mean Shift Clustering Silhouette Score (Before PCA): 0.385
•Mean Shift Clustering Silhouette Score (After PCA): 0.384

Conclusion
Applying PCA before clustering helps in reducing the dimensionality of the data, removing noise and redundancy, and improving the cluster quality. The principal components capture the majority of the variance in the data, leading to more compact and well-separated clusters. This comparison highlights the benefits of using PCA as a preprocessing step for clustering algorithms. The improved silhouette scores for K-Means and Hierarchical Clustering after PCA demonstrate the effectiveness of this approach.



Exploring Players' Performance

We created Players Dataset and Added Features such as:

home_away_goals_ratio: The ratio of home goals to away goals, calculated as (home_goals + 1) / (away_goals + 1) to handle with 0's

best_tournament: The tournament where the player scored the most goals.

winning_contribution: The number of matches in which the player's goals contributed to the team's win.

average_goal_minute: The average minute during the game when the player scored their goals.

team_dependence: how much goals the player has scored from all team's goals

opponent_strength: The total number of goals scored by the opponent team in all matches, used as a proxy for the team's strength.

match_importance: A synthetic value indicating the importance of the match. Higher values are assigned to later stages of tournaments (e.g., 3 for finals, 2 for semifinals, 1 for earlier stages).

weighted_goal: A weighted value of goals scored by the player, calculated as the product of match_importance and opponent_strength for each goal scored.

Question:
Can we predict the winning contribution of a player based on their performance metrics?

We used **Random Forest Regression** model with cross-validation. Here are the key findings and reflections on the results:
Mean Squared Error (**MSE**): **0.9139175097778707**
R-squared (R^2): **0.9279234861997144**
The high R-squared values indicate that the model explains a substantial portion of the variance in the winning contribution. The low MSE values suggest that the predictions are close to the actual values, demonstrating the accuracy of the model.

This insight can be valuable for team managers and analysts in evaluating player impact