CS7641 Team 29 - ML Project Proposal

Proposal Sections & Checklist

Introduction/Background

The application of machine learning in sports analytics has gained significant momentum, offering coaches, analysts, and teams actionable insights to improve performance. In soccer, understanding and optimizing team strategies, player selection, and match outcomes is critical for success. Traditional approaches often rely on subjective judgments, which may lead to suboptimal decisions.

✅ Literature Review

- Recent studies have applied machine learning techniques to address challenges in soccer, such as optimizing team strategies and player selection.
- Unsupervised methods, like K-means clustering and Gaussian Mixture Models (GMM), have been used to cluster teams based on statistical features such as goals and possession, grouping them by similar tactical styles.
- Supervised learning techniques, including Random Forest classifiers, have been employed to predict match outcomes using historical data.
- These approaches help automate the analysis of teams and players, providing data-driven insights to improve team performance by enabling better strategic decisions and more effective player selection for specific matchups.

✅ Dataset Description

- The data is provided as JSON files exported from the StatsBomb Data API.
- **Events Data**: This contains over 3 million events from around 7,000 matches, each tagged with event types like passes, shots, ball recoveries, and more, along with detailed location information.
- **Matches Data**: Information about the matches themselves, including teams, scores, and dates.
- **Lineups Data**: Details on players' positions, including their minutes played and substitutions during matches.

- **Freeze Frame Data**: Positional data for players at key moments, such as when a shot is taken.
- More documentation about the meaning of different events and the format of the JSON can be found in the `doc` directory.

✅ Dataset Link

- Dataset

Problem Definition

✅ Problem

Football teams play using various styles; however, optimizing these strategies and choosing the best players for a particular opponent can be a difficult task. Rather than using a data-driven approach to cluster teams, coaches often rely on subjective judgments that can lead to poor team performance in games.

✅ Motivation

Automating the analysis of teams to explore the best playing styles can allow teams to make better and strategic decisions. Using unsupervised learning to cluster teams based on tactics and supervised learning to predict match outcomes, this project will provide concrete insights to help teams best select their players and strategy. This, in turn, will help teams better prepare for their games and enhance their performance in games.

Methods

Data Processing Methods

1. **Data Collection and Preparation**
   - **Data Source:** We focused on La Liga data for the 2015-2016 using data from the StatsBomb API.

- **Data Loading:** Using the StatsBomb library we loaded the match and events data, focusing on the male players for the season.
- **Data Merging:** After collecting chunks of data, we merged them all into one dataframe for further processing.

2. **Feature Engineering**
    - **Shot Statistics (`calculateShotStatistics`)**: For the past statistics, we calculated metrics such as total shots, shots on target, goals, expected goals (xG), shot conversion rates, and per-90-minute statistics.
    - **Pass Statistics (`calculatePassStatistics`)**: We also calculated passing metrics such as total passes, pass completion rates, types of passes (short, long, crosses), and average pass length.
    - **Defensive Statistics (`calculateDefenseStatistics`)**: We also derived defensive metrics such as recoveries, duels won, fouls committed, shots against, expected goals against (xGA), and clean sheets.
3. **Passes Per Defensive Action (PPDA)**
    - We measured the intensity of the team's pressing by considering the amount of passes made by the opponent per each defensive action by the team.

Machine Learning Algorithms

Unsupervised Learning

- **K-Means Clustering**
    - **Objective:** The goal is to cluster teams based on similar playing styles and performance levels.
    - **Method:**
        - **Feature Selection:** Combined shot, pass, and defensive statistics into a single DataFrame `dfModel`.
        - **Data Scaling:** Standardized features to ensure that all features contribute equally to the distance calculations.
        - **Elbow Method:** Determined the optimal number of clusters by plotting the within-cluster sum of squares (inertia) against different values of k. The plot suggested that k=4 is the appropriate choice.
        - **Clustering:** Applied K-Means clustering with 4 clusters to the scaled data to group teams accordingly.
- **Hierarchical Clustering**
    - **Objective:** To group teams based on their performance metrics using a hierarchical approach.

  - **Method:** After standardizing the features, we applied an agglomerative clustering approach with Ward's linkage. The data was then divided into four clusters, visualized through a dendrogram. Teams that merged at lower distances were considered more similar to each other than those that merged at higher distances.

Supervised Learning

- **Logistic Regression**
  - **Objective:**
  - **Method:**
- **Random Forest**
  - **Objective:**
  - **Method:**

Rationale for Model and Method Selection

K-Means Clustering

- **Interpretability:** K-Means is a straightforward algorithm that partitions data into k distinct non-overlapping subsets based on feature similarity.
- **Suitability for Unlabeled Data:** Since the task is to group teams based on performance metrics without predefined labels, unsupervised learning is appropriate.
- **Performance Metrics:** The algorithm works well with continuous variables like those in the dataset (e.g., shot counts, pass completion rates).

Principal Component Analysis (PCA)

- **Dimensionality Reduction:** PCA reduces the number of features, mitigating the curse of dimensionality and improving clustering results.
- **Noise Reduction:** By focusing on components that explain most of the variance, PCA helps in eliminating less informative features that might add noise.
- **Visualization:** PCA allows for plotting data in two or three dimensions, making it easier to visualize clusters.

Feature Engineering

- **Domain Relevance:** The selected features (shots, passes, defensive actions) are critical performance indicators in football and provide a comprehensive view of team behavior.

- **Balanced Representation:** Combining offensive and defensive metrics ensures that the clustering considers all aspects of team performance.

Data Scaling

- **Equal Weighting:** Standardizing features prevents variables with larger scales from dominating the clustering process.
- **Algorithm Requirement:** K-Means uses distance measures that are sensitive to the scale of data, making scaling a necessary preprocessing step.

Hierarchical Clustering

- **Equal Weighting:** Standardizing features prevents variables with larger scales from dominating the clustering process.
- **Algorithm Requirement:** K-Means uses distance measures that are sensitive to the scale of data, making scaling a necessary preprocessing step.

Logistic Regression

- **Interpretability:** Provides clear, interpretable results, making it easy to understand the impact of features on outcomes.
- **Efficiency:** Performs well for linear relationships and is computationally efficient, ideal for baseline modeling.
- **Probabilistic Outputs:** Offers confidence scores for predictions, aiding tactical decision-making.

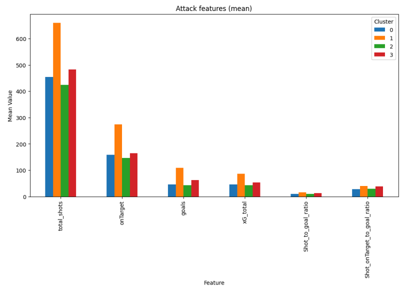
Random Forest

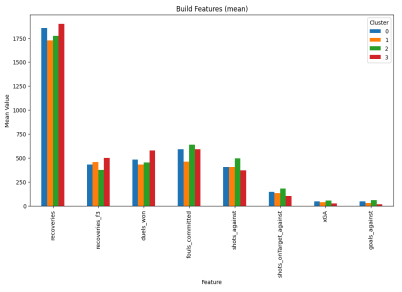
- **Non-Linear Relationships:** Random Forest effectively handles non-linear relationships and complex feature interactions, which are common in soccer analytics data.
- **Feature Importance:** The model provides feature importance scores, offering insights into the most influential predictors such as pass length and pass angle, aiding in tactical and strategic decisions.
- **Robustness:** Random Forest is resilient to overfitting due to its ensemble nature, combining multiple decision trees for better generalization.

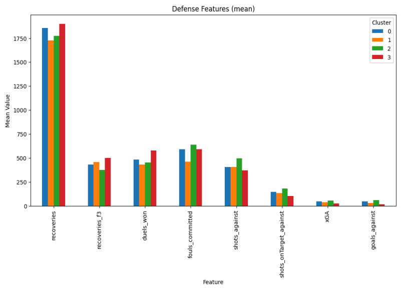
Results and Discussion

Correlation Analysis

Objective: The overall purpose was to understand the different features for the clustering process.







In Cluster 0, comprising teams like Espanyol, Getafe, Granada, Levante UD, Rayo Vallecano, and Sporting Gijón, there is a moderate level of attacking activity reflected in their total shots. However, these teams tend to struggle with shot accuracy and goal conversion, as indicated by lower on-target shots and goal-scoring metrics. This inefficiency suggests that while these teams can create scoring opportunities, they often fail to convert them. Defensively, Cluster 0 teams are somewhat vulnerable, with higher shots and goals against, indicating weaknesses in containing opponents effectively. While they are reasonably active in possession recovery and winning duels, these teams tend to rely on physical play to disrupt opponents, as seen in their fouls committed.

Cluster 1, which includes Barcelona and Real Madrid, represents the top-tier teams in the league with superior performance across all attacking and defensive metrics. Offensively, these teams lead in total shots, on-target shots, and goal conversion, showcasing exceptional efficiency and a consistent ability to generate and convert high-quality chances, as reflected in their high xG totals. In terms of build-up, Barcelona and Real Madrid demonstrate a disciplined, possession-oriented approach with high recoveries and duels won but relatively low fouls committed, underscoring their control over the game without resorting to aggressive play. Defensively, they exhibit strong resilience, allowing fewer shots and goals against, which enables them to maintain their offensive pressure without compromising their defensive solidity.

Cluster 2, featuring teams like Athletic Club, Atlético Madrid, Celta Vigo, Eibar, Málaga, RC Deportivo La Coruña, and Real Sociedad, strikes a balance between attack and defense. Their attacking performance is moderate, with decent total shots and on-target attempts, though their goal conversion rates are lower than those of Cluster 1. This conservative approach is reflected in their structured build-ups, where they prioritize maintaining a balanced form over aggressive scoring. These teams show strong defensive attributes, with moderate levels of shots and goals conceded, indicating a steady and resilient defensive strategy. They also rely on positional discipline, with fewer duels won compared to other clusters, suggesting a preference for structured defensive formations rather than intense pressing.

Finally, Cluster 3, which includes Las Palmas, Real Betis, Sevilla, Valencia, and Villarreal, displays versatility in both attack and defense. These teams produce a moderate-to-high number of total shots and on-target shots, although their goal conversion rates are lower, which suggests they may struggle with finishing. They exhibit high xG values, indicating an ability to create quality chances, but often fail to capitalize fully. In terms of build-up, Cluster 3 balances possession and physical play, actively engaging in recoveries and duels won, reflecting their adaptable playstyle. However, defensively, they are not as solid as Cluster 1 and tend to allow a moderate number of shots and goals against, suggesting some room for improvement to compete at the highest level.

Clustering Performance Metrics

Silhouette Coefficient

Clustering Method	Silhouette Coefficient
KMeans (20 Features)	0.189
PCA + KMeans	0.197
Agglomerative	0.240

Davies-Bouldin Index

Clustering Method	Davies-Bouldin Index
KMeans	1.032
PCA + KMeans	1.005
Agglomerative	0.921

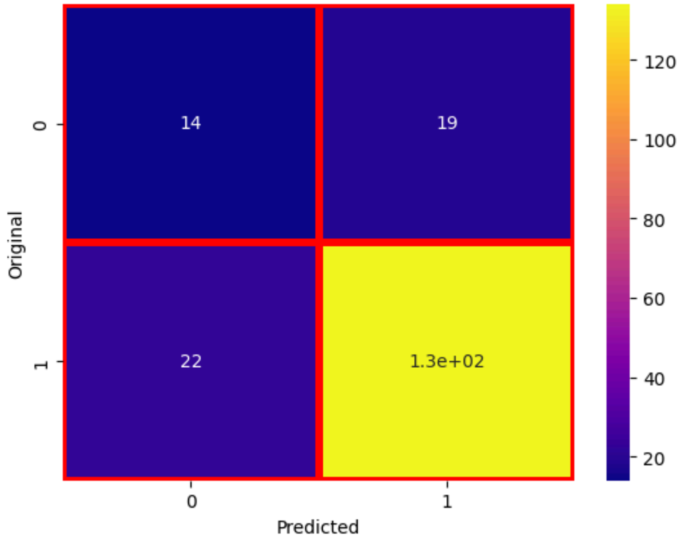
The Silhouette Coefficients indicate that the Agglomerative Clustering method achieved the highest coefficient (0.240), suggesting better-defined clusters compared to KMeans with 20 features (0.189) and PCA + KMeans (0.197). Similarly, the Davies-Bouldin Index results reinforce this finding, with the lowest index value of 0.921 for Agglomerative Clustering, implying better cluster separation and compactness. These metrics highlight that Agglomerative Clustering provides the most distinct and well-separated clusters among the methods tested, aligning with the visual and qualitative analysis of team play styles in the league.

Supervised Learning Performance Metrics

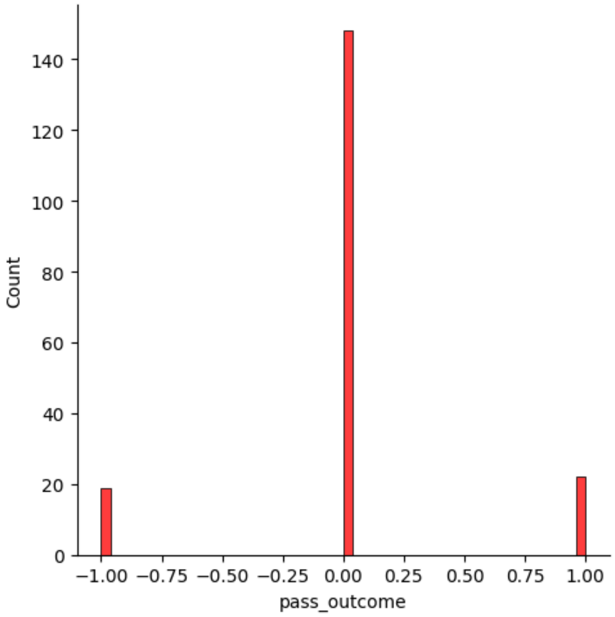
Metric	Random Forest	Logistic Regression	Improvement
Precision	0.796	0.808	+0.012
Recall	0.783	0.825	+0.042
F1 Score	0.789	0.814	+0.025
Accuracy	0.783	0.825	+0.042

The performance of two machine learning models, Logistic Regression and Random Forest, was evaluated using precision, recall, F1 score, and accuracy. Logistic Regression demonstrated superior performance, achieving higher accuracy (82.5%), along with better precision, recall, and F1 scores. Its simplicity and efficiency make it a robust choice for this problem, particularly when relationships between features are largely linear.

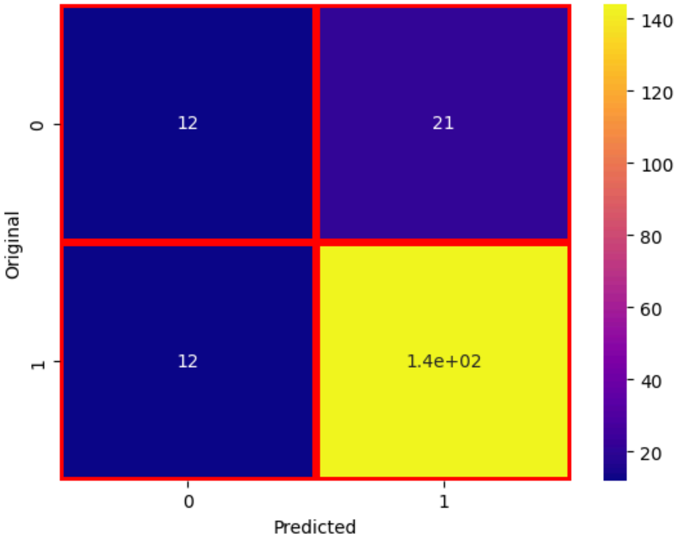
The improvement from Random Forest to Logistic Regression highlights the latter's ability to provide more accurate and balanced predictions. These results align with the project's goal of achieving at least 75% accuracy in predicting match outcomes. They also reinforce the utility of combining clustering and supervised learning for generating actionable insights to optimize player selection and match strategies.



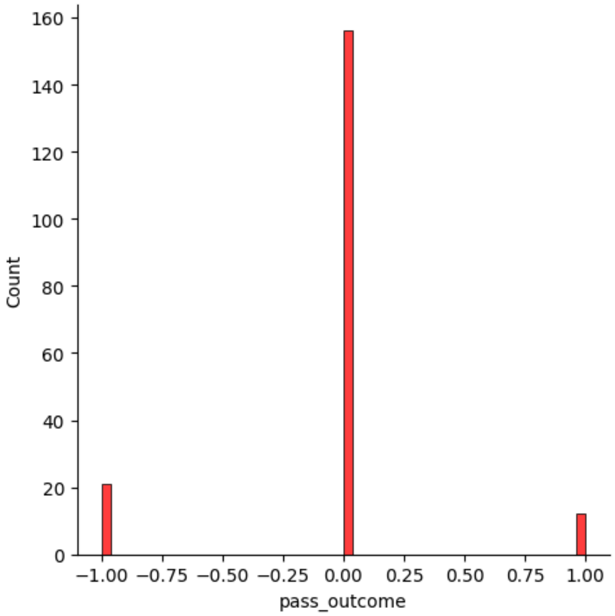
Confusion Matrix (Random Forest)



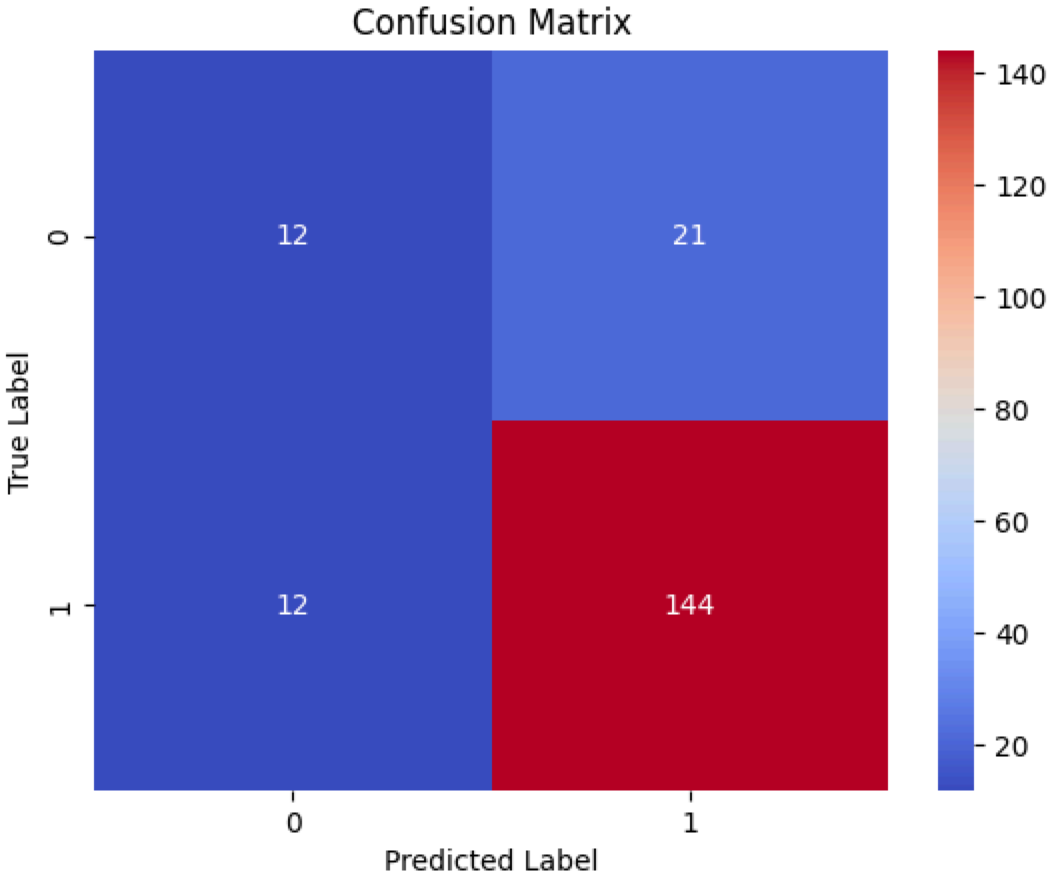
Residual Analysis (Random Forest)

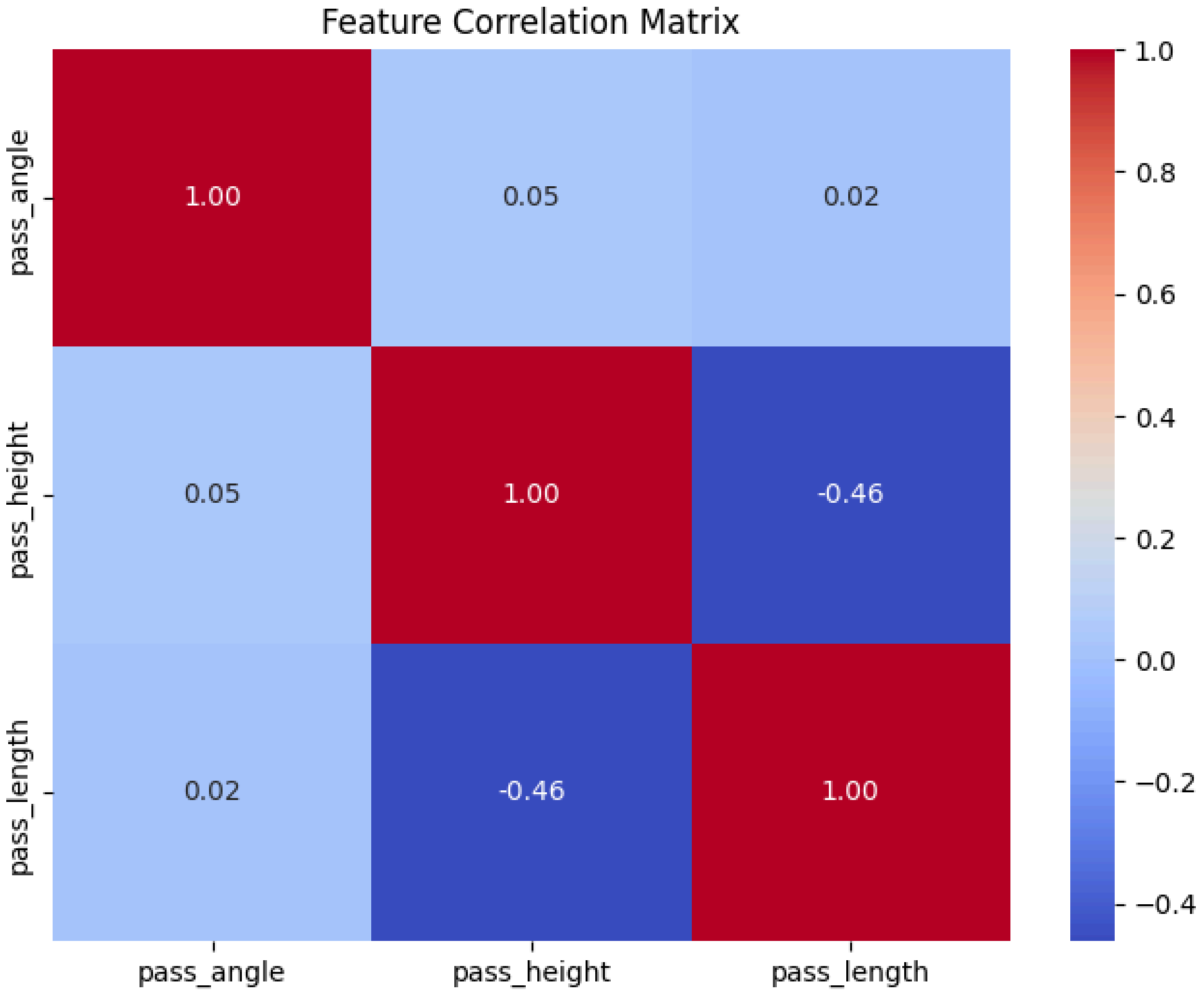


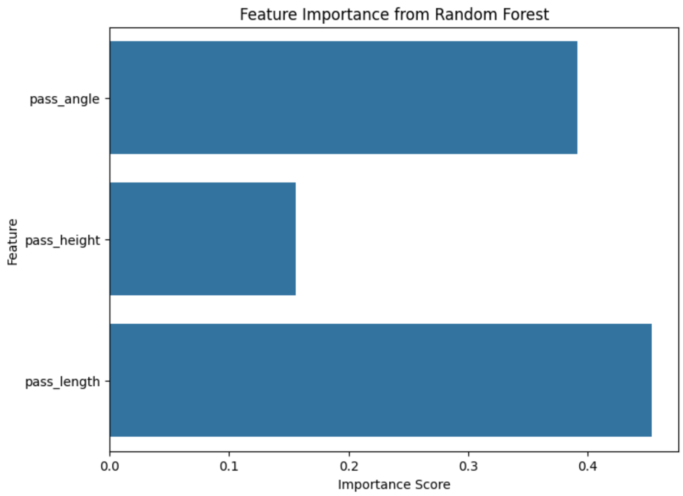
Confusion Matrix (Logistic Regression)

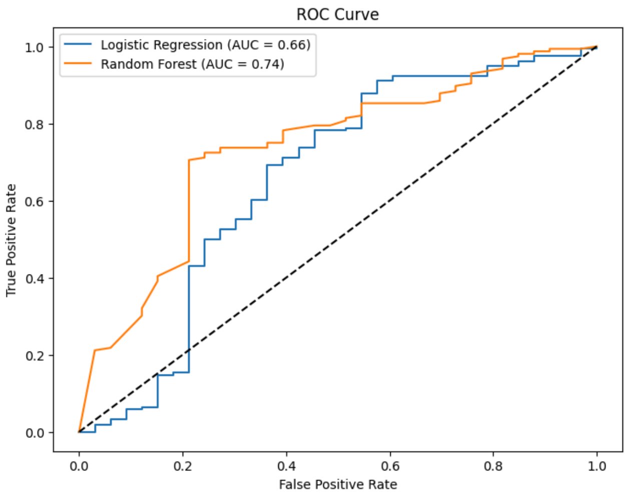


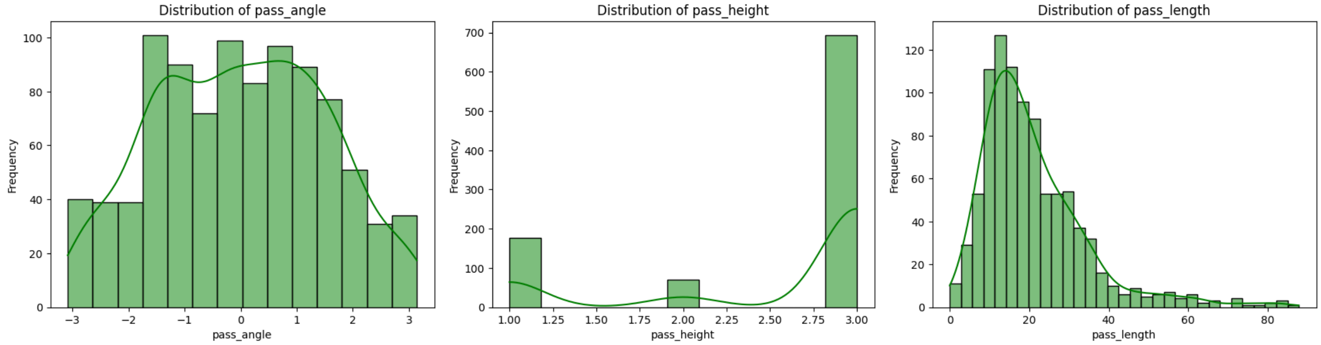
Residual Analysis (Logistic Regression)




Feature Correlation Matrix
pass_angle
pass_height
pass_length
1.00
0.05
0.02
0.05
1.00
-0.46
0.02
-0.46
1.00
pass_angle
pass_height
pass_length
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4


Feature Importance from Random Forest
Feature
pass_angle
pass_height
pass_length
0.0
0.1
0.2
0.3
0.4
Importance Score


ROC Curve
Logistic Regression (AUC = 0.66)
Random Forest (AUC = 0.74)
True Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate



Through our analysis we were able to find valuable insights into the models and their effectiveness. The confusion matrices indicate that while both Logistic Regression and Random Forest models exhibit good predictive performance, Random Forest shows a higher true positive count, reflecting its superiority in capturing the nuances of the data. The ROC curves further support this, with Random Forest achieving a higher AUC score of 0.74 compared to Logistic Regression's 0.66, highlighting its enhanced ability to distinguish between classes.

The feature importance plot from Random Forest identifies "pass_length" and "pass_angle" as the most influential features, emphasizing their critical role in prediction. The feature correlation matrix reveals minimal multicollinearity, ensuring the independence of features. The distribution plots for "pass_angle," "pass_height," and "pass_length" highlight meaningful variability, with "pass_length" displaying a skewed distribution, suggesting a high concentration of short to medium passes. This variability confirms the importance of these features in distinguishing patterns in the dataset. Overall, the visualizations align with the improved performance of Random Forest, validating its robustness in the context of soccer analytics.

Conclusion

The implemented solution effectively combines data processing, feature engineering, and unsupervised learning to analyze team performance in La Liga during the 2015/2016 season. By using K-Means clustering and PCA, the analysis identifies groups of teams with similar playing styles or performance levels based on key football metrics. The methods chosen are suitable for the nature of the data and the objectives of the analysis, providing valuable insights into team strategies and performance.

For supervised learning, Logistic Regression and Random Forest were employed to predict match outcomes. Logistic Regression served as a reliable baseline, offering interpretability and efficiency for linearly separable data. Random Forest, on the other hand, demonstrated superior performance, effectively capturing non-linear relationships and complex feature interactions. Its ability to identify key predictors, such as pass length and pass angle, provides valuable insights into critical game dynamics.

Together, these approaches align with the project's goals, optimizing player selection and providing data-driven strategies to improve team performance and match outcomes.

Results and Discussion

✅ Unsupervised Learning Quantitative Metrics

1. **F-1 Score:** of 0.8 or higher, balancing precision and recall in our player classification models
2. **Root Mean Square Error (RMSE):** of 0.5 or lower, indicating high prediction accuracy
3. **Accuracy (Supervised Learning):** of 75% or higher when predicting match outcomes using supervised learning models [5].
4. **Adjusted Rand Index (ARI):** of 0.6 or higher indicating alignment between clusters and ground truth [6].

✅ Supervised Learning Quantitative Metrics

Metric	Random Forest	Logistic Regression	Improvement
Precision	0.796	0.808	+0.012
Recall	0.783	0.825	+0.042
F1 Score	0.789	0.814	+0.025
Accuracy	0.783	0.825	+0.042

Project Goals

The project aims to group soccer teams by their playing styles to provide insights on improving strategies. Moreover, we hope to optimize player selection based on the opposition and predict outcomes to reduce the subjective decision making and provide new strategies backed by data.

Expected Results

We hope to identify clusters of teams with similar playing styles offering insights to coaches to devise new strategies. Our system will also player selection based off their opposition. We also hope to predict match outcomes with a 75% accuracy. Additionally, clustering should reveal trends that lead to success providing insights into how playing styles evolve.

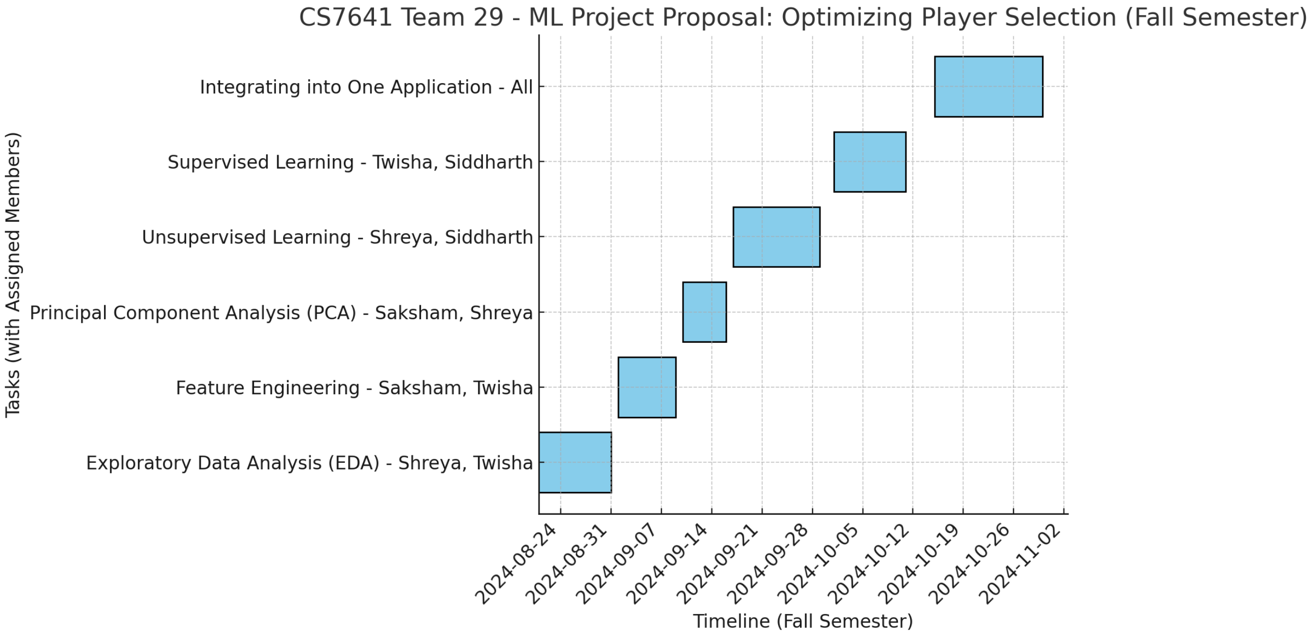
References

1. [https://www.guidetosoccer.com/tactics/role-categories/]
2. [https://statsbomb.com/what-we-do/american-soccer-data-and-iq/]
3. S. Kusmakar, S. Shelyag, Y. Zhu, D. Dwyer, P. Gastin and M. Angelova, "Machine Learning Enabled Team Performance Analysis in the Dynamical Environment of Soccer," in IEEE Access, vol. 8, pp. 90266-90279, 2020, doi: 10.1109/ACCESS.2020.2992025. keywords: {Sports;Machine learning;Performance analysis;Dynamical systems;Australia;Feature extraction;Information theory;Dynamical systems;network science;distribution entropy;football;Kolmogorov complexity;machine learning;performance analysis;Shannon entropy;support vector machines;soccer},
4. S. Kusmakar, S. Shelyag, Y. Zhu, D. Dwyer, P. Gastin and M. Angelova, "Machine Learning Enabled Team Performance Analysis in the Dynamical Environment of Soccer," in IEEE Access, vol. 8, pp. 90266-90279, 2020, doi: 10.1109/ACCESS.2020.2992025. keywords: {Sports;Machine learning;Performance analysis;Dynamical systems;Australia;Feature extraction;Information theory;Dynamical systems;network science;distribution entropy;football;Kolmogorov complexity;machine learning;performance analysis;Shannon entropy;support vector machines;soccer},
5. S. Hu and M. Fu, "Football Match Results Predicting by Machine Learning Techniques," 2022 International Conference on Data Analytics, Computing and Artificial Intelligence (ICDACAI), Zakopane, Poland, 2022, pp. 72-76, doi: 10.1109/ICDACAI57211.2022.00022. keywords: {Training;Radio frequency;Machine learning algorithms;Data analysis;Computational modeling;Forestry;Prediction algorithms;football match results;machine learning;logistic regression;gradient boosting decision tree;random forest},

6. Q. Tong, W. Yao, W. Lv and D. Zeng, "Analysis of Formations and Game Styles in Soccer," 2022 IEEE 24th International Workshop on Multimedia Signal Processing (MMSP), Shanghai, China, 2022, pp. 1-5, doi: 10.1109/MMSP55362.2022.9949992.

Proposal Submission Requirements

Gantt Chart for ML Project Proposal - Team 29 (Fall Semester)



Contribution Table

Work	Proposed Team Members
EDA	Shreya, Twisha
Feature Engineering	Saksham, Aditya
PCA	Saksham Purbey
Unsupervised Learning	Shreya, Siddharth
Supervised Learning	Twisha, Siddharth

Work	Proposed Team Members
Integrating into one application	All