A red square with white letters

Description automatically generated with medium confidence

**Prepared For**

Summer School

**Prepared By**

LSE ID: 202275952

**Module**

Machine Learning in Practice (ME315)

**Submission Date: 7th August 2023**

Football is life

Contents

[Dataset Description 1](#_Toc139557135)

[Exploratory Analysis 3](#_Toc139557136)

[Overview of the League 3](#_Toc139557137)

[How different is the style of play? 3](#_Toc139557138)

[Unsupervised Learning 5](#_Toc139557139)

[Hierarchical Clustering 5](#_Toc139557140)

[K-Means Clustering 6](#_Toc139557141)

[Supervised Learning 7](#_Toc139557142)

[Classification 7](#_Toc139557143)

[Cleaning up process 8](#_Toc139557144)

[Correlation of Significant Variables 8](#_Toc139557145)

[Splitting the Data 9](#_Toc139557146)

[Comparing LDA, QDA and Logistic Regression models 9](#_Toc139557147)

[Conclusion 11](#_Toc139557148)

[References 12](#_Toc139557149)

# Dataset Description

A screenshot of a computer

Description automatically generated with low confidenceA screenshot of a computer

Description automatically generated with low confidence

The dataset utilized in this project comprises two distinct sets of data: 'ginf.csv' on the left and 'events.csv' on the right, both summarized in a consolidated table. This comprehensive table includes column names, concise variable descriptions, and data types. It offers a detailed perspective on 9,074 games, encompassing a total of 941,009 events from the five most prominent European football leagues: England, Spain, Germany, Italy, and France known as Premier League, La Liga, Bundesliga, Serie A and Ligue 1 respectively. The dataset spans from the 2011/2012 season to the 2016/2017 season, as of January 25, 2017.  
  
Cleaning up process

To streamline the information, we will consolidate all the essential details into a single table using a left join. This join operation merges the rows from the left data frame (i.e., "events") with the corresponding rows from the right data frame ("ginf"), based on a shared column or variable (in this case, "id\_odsp").

Additionally, the league column in the events table is currently represented nominally, which can be difficult and challenging to interpret during exploratory analysis. To enhance clarity, we will transform the league column into the respective league names, namely 'Bundesliga', 'Premier League', 'Ligue 1', 'Serie A', and 'La Liga'.

Lastly, we will convert certain categorical data into factors using the built-in R function 'as.factor()'. This conversion ensures that these variables are correctly recognized as categorical when performing analysis or generating plots.

## Exploratory Analysis

### Overview of the League

Figure

A picture containing text, font, screenshot, receipt

Description automatically generatedFigure 1 provides a visual representation of key statistics, including the total number of goals, matches played, and the goals-to-match ratio. Upon analyzing this data, it becomes apparent that the Premier League stands out as the most fiercely competitive and intense league in comparison to the others. However, it is important to note that this conclusion is merely an initial observation based on the information presented in the figure.

A picture containing text, screenshot, diagram, plot

Description automatically generated

Figure

The visualization of Figure 1 indicates a correlation between the number of matches and the number of goals. It is evident that the Bundesliga (Germany) appears to have the fewest games played, resulting in the lowest total number of goals among the five leagues. However, the Bundesliga boasts the highest ratio of 2.88 goals per match, surpassing the other leagues. This suggests that while there may be fewer overall goals, each individual match in the Bundesliga tends to have a higher goal-scoring intensity compared to the other leagues.

### How different is the style of play?

A picture containing text, screenshot, diagram, font

Description automatically generated  
Contrary to the assumption based on Figure 1, Figure 3 reveals an intriguing insight. It demonstrates that the Premier League has the lowest number of cards shown among the five leagues, indicating less risky challenges and a less aggressive style of play. This challenges the notion of the Premier League being the most competitive and intense solely based on the data from Figure 1. It highlights the need to consider multiple factors and perspectives when evaluating the level of competition in different leagues.

Figure

A picture containing screenshot, text, colorfulness, line

Description automatically generated  
A picture containing text, screenshot, diagram, font

Description automatically generatedFigure 4 provides a detailed breakdown of the cards per match, revealing interesting insights about teams within each league. La Liga, Serie A, and Bundesliga teams display more aggressive playstyles, resulting in a higher number of cards. In contrast, the Premier League and Ligue 1 teams demonstrate comparable card frequencies. This analysis delves deeper into the playing dynamics and disciplinary patterns of teams in each league.

Figure

Figure

Figure 5 highlights the top four teams from each league that hit the crossbar. This occurrence signifies that these teams took shots that came agonizingly close to scoring, missing the goal by only a few centimeters. It indicates that these teams are proactive in taking shots and seizing opportunities but have been unlucky in converting them into goals. The data underscores their offensive mindset and willingness to take chances in pursuit of scoring.

Figure

A screenshot of a graph

Description automatically generated with low confidence   
  
  
Figure 6 examines the scoring preferences of the top five teams from each league. It reveals how teams utilize different methods, such as scoring with the right foot, left foot, or head, to find the back of the net. This data provides insights into the diverse play styles and goal-scoring strategies employed by these teams.

A picture containing text, screenshot, font, parallel

Description automatically generated

A picture containing text, screenshot, font, parallel

Description automatically generatedFigure 7 focuses on goals scored in the first 15 minutes, while Figure 8 highlights goals scored in the last 15 minutes of matches. These statistics emphasize that early or late scoring is not merely a matter of luck but rather a significant indicator of a team's play style. The timing of goals provides valuable insights into how teams approach matches and their ability to score early or late.   
  
Using the data from Figure 2 to Figure 8, we will employ these factors to cluster teams with similar play styles, aggressiveness, and patterns in their gameplay.

Figure 7 & 8

# Unsupervised Learning

## Hierarchical Clustering

#### Comparing Single & Complete Linkage

A picture containing diagram, text, line, parallel

Description automatically generatedOur objective in employing unsupervised learning is to cluster teams based on their similar play styles and patterns. By applying unsupervised learning techniques to the data, we aim to group teams that exhibit comparable characteristics, allowing us to gain a deeper understanding of the underlying patterns and playing styles within the dataset.

A picture containing diagram, sketch, drawing, technical drawing

Description automatically generatedBased on the Single Linkage clustering results shown in the figure above, it is evident that the clustering accuracy is inadequate. Notably, 'Barcelona' is placed in Cluster 2, while 'Real Madrid' is assigned to Cluster 3. The remaining teams are grouped together in Cluster 1. This outcome suggests that the Single Linkage method is ineffective in accurately clustering teams with similar characteristics.

A picture containing text, font, letter, document

Description automatically generatedIn contrast to the Single Linkage clustering, the Complete Linkage clustering method has demonstrated greater accuracy in producing more compact and spherical clusters. Compact clusters are desirable as they indicate minimized within-cluster variation, resulting in improved separation between clusters. The Complete Linkage clustering approach has effectively generated clusters that better capture the similarities and patterns within the data, leading to more accurate and reliable results.  
  
  
  
In the Complete Linkage figure, we observe that Cluster 3 (indicated by the red box) comprises exceptionally strong teams. It is logical that these teams are clustered together as they dominate their respective leagues. This clustering reveals that these teams share similarities in play style, including their scoring methods, level of aggression, timing of goals, and willingness to take chances.

#### Conclusion on Hierarchical Clustering Complete Linkage demonstrates better results compared to Single Linkage. Single Linkage uses an iterative process to merge nearest neighbors, while Complete Linkage considers the maximum distance between points in different clusters. Complete Linkage produces more compact and spherical clusters, whereas Single Linkage tends to generate long chains and elongated clusters. This difference may be attributed to Single Linkage being more affected by outliers, as it relies on the shortest distance between points. Outliers far from most data points can lead to the merging of distant clusters and the formation of elongated chains. Therefore, for datasets with compact and well-separated data points without significant outliers, Single Linkage can be effective. However, in the case of this dataset, Complete Linkage proves to be the more appropriate method, as it generates more accurate and compact clusters.

K-Means Clustering   
The K-means clustering method utilizes the parameter K, which represents the desired number of clusters, and relies on Euclidean distances to create these clusters. The algorithm follows the following steps to cluster the data points:

1. Randomly initializes K centroids within the data set.
2. Calculates the Euclidean distance between each data point and the centroids. Each data point is then assigned to the cluster with the closest centroid.
3. Relocates each centroid to the center of its respective cluster.
4. A graph of a number of clusters

   Description automatically generated with low confidenceRepeats steps 2 and 3 until either the maximum number of iterations is reached or the changes in the centroids become insignificant.

To determine the optimal number of clusters, we employ the within-group sum of squares (WSS) plot. Utilizing the elbow method, we observe a plot where the number of clusters is plotted against the corresponding WSS. By analyzing the WSS plot, we identify the point where the decrease in WSS begins to slow down significantly. This point, often resembling an elbow, indicates the optimal number of clusters. In this case, the plot suggests that the optimum number of clusters is 4 and beyond this, reduction in WSS is less substantial.

Figure 9

A picture containing text, diagram, screenshot, design

Description automatically generatedIn Figure 10, the K-means clustering results are visualized. After extracting the teams belonging to Cluster 2 (indicated by the green color), we can examine a specific subset of teams. These teams include "AC Ajaccio," "Lyon," "AS Nancy Lorraine," "Juventus," "Granada," "Sampdoria," "Sassuolo," and "Sunderland."

Figure 10

Upon closer examination, it becomes evident that apart from "Lyon" and "Juventus," the other teams are not widely recognized for their aggressiveness or high quality. This finding suggests that the clusters may not accurately represent the playstyles of the teams, which was the main objective of the clustering analysis. Therefore, based on this observation, it can be concluded that the K-Means clusters may not provide an accurate representation of the desired playstyle categorization.

#### Comparing K-Means and Hierarchical Clustering

It is evident that Hierarchical Clustering with Complete Linkage outperformed K-means Clustering. The advantage of Complete Linkage can be attributed to its effectiveness in handling noisy data and outliers compared to K-means Clustering.  
  
Complete Linkage clustering considers the maximum distance between points when merging clusters, which reduces the likelihood of merging distant clusters due to the presence of outliers. This characteristic makes Complete Linkage more suitable for creating robust clusters that are less influenced by noise or outlying data points.  
  
Considering the characteristics of the dataset at hand, it can be concluded that Hierarchical Clustering with Complete Linkage is the most appropriate clustering method. It provides more reliable and accurate results, considering the presence of noise or outliers in the data.

# Supervised Learning

## Classification

In the realm of supervised learning, our focus will be on classification models applied to the football information dataset. Specifically, we will utilize three classification models: Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), and Logistic Regression. The primary objective is to predict the "is\_goal" variable, using carefully selected predictor variables. From there, we will find out the best classification model with the best prediction and the least error. To determine the predictor variables, we will begin by conducting a correlation matrix analysis. Additionally, a heatmap visualization will be generated to highlight the significant predictors. These selected variables will then be utilized in our models to ascertain their ability to predict goal scoring.

### Cleaning up process

As part of the data cleaning process, we will address the columns that contain categorical information such as "event\_type" and "location." Currently, these columns are represented numerically, where each numeric value corresponds to a specific category. For example, the value 1 in the "event\_type" column represents "Attempt."

To facilitate data analysis and the implementation of supervised learning models, we will convert these categorical variables into binary dummy variables. This process involves creating new columns for each category, where a value of 0 indicates that the event did occur, and a value of 1 indicates that it did not occur. By transforming the data in this manner, it will be easier to interpret and utilize the data in subsequent analysis and modeling stages.

### A picture containing text, screenshot, diagram, line Description automatically generatedCorrelation of Significant Variables

After conducting the correlation analysis with the "is\_goal" variable, we have identified the most significant predictors that are likely to lead to goals. The heatmap visualization in Figure 11 illustrates this relationship.

The selected predictors that demonstrate a notable association with goal scoring are as follows:

1. Attempt: This represents the act of taking a shot at the goal.
2. Key\_Pass: It denotes the last key pass made before a player attempts a shot.
3. Centre\_box: This indicates that the shot was taken from the center box area near the goal.
4. Penalty\_spot: It signifies a shot taken from the penalty spot.
5. Close\_range: This refers to shots taken in close proximity to the goal, typically within a range of around 15 meters.

Figure 11

While it is important to note that scoring goals in football is a challenging task, the correlation between these predictors and goal scoring, although not extremely high, is comparatively stronger than other variables. These selected variables have demonstrated a meaningful relationship with the outcome of interest and will be instrumental in our subsequent analysis and modeling.

### Splitting the Data

Firstly, we split the data into training and testing set using a 70:30 ration. We also set.seed(123) to make sure reproducibility of the random sampling.

### Comparing LDA, QDA and Logistic Regression models

In our analysis, we have employed three classification models: Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), and Logistic Regression. These models are designed to classify observations into one of two classes, namely "goal" or "not goal."

During the training phase, the models learn the patterns and relationships within the predictor variables. Their objective is to effectively separate the observations into these distinct classes based on the values of the predictors. Once the models are trained, they can be utilized to predict the class labels for the test set.

To evaluate the performance of these models, we calculate the misclassification error rate. This rate quantifies how well the models accurately classify the observations in the test set. It is determined by comparing the predicted class labels with the actual class labels and recording the results in a logical vector. The mean of this logical vector is then computed, providing the misclassification rate.

A lower misclassification rate indicates a better performance in predicting the correct class labels. This metric serves as a measure of how accurately the models classify observations into the appropriate goal or not goal categories.

A picture containing text, receipt, font, white

Description automatically generated

After calculating all the misclassification values of the three models, we have summarized them into the table in Figure 12. We can see that Logistic Regression performed the best with the lowest error rate.

Figure 12

#### Further dive into Logistic Regression

#### A picture containing text, screenshot, diagram, line Description automatically generated

Figure 13

In Figure 13, we present a Receiver Operating Characteristic (ROC) Curve, which offers a visualization of the performance of the previously conducted Logistic Regression model. The ROC curve displays the relationship between the true positive rate (sensitivity) on the y-axis and the false positive rate (specificity) on the x-axis.

Each point on the ROC curve corresponds to a different threshold used for classification. The curve showcases the model's performance across a range of various thresholds, allowing us to observe how the sensitivity and specificity trade off as the threshold changes.

Additionally, the Area under the Curve (AUC) is a summary measure that assesses the overall performance of the classifier based on the ROC curve. The AUC value ranges from 0 to 1, where a value closer to 1 indicates better discrimination between the classes. The AUC provides an indication of how well the model can distinguish between positive and negative instances, with higher values representing stronger discriminatory power.

A picture containing text, font, typography, white

Description automatically generated

Figure 14

In Figure 14, the Logistic Regression model demonstrates a high AUC value of 0.935, indicating its strong performance in effectively distinguishing between instances with goals and no goals. A higher AUC value suggests a better ability to discriminate between the two classes.

The F1 score, displayed as 0.87 in Figure 14, is a metric that combines both precision and recall, providing an overall measure of the classifier's accuracy. This score indicates that the Logistic Regression model achieves a good balance between precision (the ability to correctly classify positive instances) and recall (the ability to minimize false negatives and positives), resulting in accurate predictions.

Additionally, the Time Elapsed of 0.0098 seconds indicates that the Logistic Regression model can make predictions efficiently, implying its computational speed and responsiveness in providing classification results.

# Conclusion

This report summarizes the analysis of football data using machine learning methods. The study explored both unsupervised and supervised learning methods to gain insights into team play styles, goal-scoring patterns, and the prediction of goal outcomes.

Through hierarchical clustering, the report discovered that Complete Linkage produced more accurate and compact clusters, providing a better representation of team similarities compared to Single Linkage. This robust clustering method allowed the identification of teams with similar play styles and patterns, enabling a deeper understanding of their gameplay strategies.

The supervised learning phase focused on classification models, including Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), and Logistic Regression. By selecting significant predictor variables through correlation analysis, the models were trained to classify goal-scoring outcomes. The evaluation of the models' performance revealed the strength of Logistic Regression in accurately distinguishing between instances with goals and no goals.

Overall, this report demonstrates how machine learning can extract valuable insights from football data. By employing clustering and classification techniques, we can uncover hidden patterns, identify play styles, and accurately predict goal outcomes, which can be useful in strategic decision-making for teams.

# References

Dataset: <https://www.kaggle.com/datasets/secareanualin/football-events?datasetId=712&sortBy=voteCount>