ISyE 7406 – Spring 2023

Final Project

Student Name: Emily Wolf

Project Title: Impact of Weather Conditions on European Football Results

Table of Contents:

Background

Soccer is unquestionably the most popular sport across the globe, though the majority of the world's population would refer to the game as football. Despite its slow building popularity in the United States, it is estimated that there are 265 million people that play "the beautiful game" and viewership of major tournaments and championships is registered in the billions. Considered its birthplace and almost certainly its current epicenter, football has a uniquely magnetic pull within the European continent. Home to over 1000 professional club teams playing within 37 leagues, Europe has the highest grossing individual teams and collective leagues in the world. Colloquially known as the "Big Five", they represent the strongest leagues with respect to both performance and value, and consist of the English Premier League (England), La Liga (Spain), the Bundesliga (Germany), Serie A (Italy), and Ligue 1 (France).

Figure 1: Big Five European Football Leagues

Amongst sports fans, it is often said that weather is a great equalizer. Optimal conditions, such as moderate temperatures, calm winds, and a dry field, allow for a level playing field, so to speak. Comparatively, in a competition between two mismatched teams, it is thought that sub-optimal or adverse conditions stand to benefit the lower ranked team. Potentially, this can be attributed to a disruption of the superior team's style of play. Perhaps strong winds necessitate shorter passes and individual possession rather than lofty crosses and free kicks. Fields slick from rain leads to reduced traction on the ball, leading to a faster paced game. Temperature extremes could lead to an increase in errors, which are especially evident in defensive statistics. A team partial to brute force tactics might be at a greater advantage in wet or windy conditions against a more skill-focused team whose possession is disrupted by the weather. Similarly, it is possible that offensive minded teams are unable to rally the same explosive energy in exhaustingly hot or humid conditions.

Problem Statement

The objective of this project was to explore the impact of weather on the results of soccer matches in the top five European soccer leagues and to show empirically if weather is in fact an equalizer. First, it identified if weather can be used to predict the outcome of a game. Specifically, if there is an upset victory (i.e., the lower ranked opponent wins or draws) or if the result is as expected (higher ranked opponent wins). Second, provided it does have an effect, it explored if that effect was global or if certain styles of play were affected by weather conditions differently. By examining the relationship between weather conditions and performance, this project can help to inform decisions about when and where to play games and could potentially guide strategies for optimizing performance in different weather conditions.

Methodology

The main objective of this project is to predict the results of a soccer match, namely an upset or expected win. First, match and weather data were collected and pre-processed. Next, cluster analysis was performed to group teams by their style of play using the team's summary match statistic features. With these assignments, it was determined which playing style was most likely to result in each match outcome. Cluster assignments were then added as input variables and a variety of classification models were fitted to predict the outcome of the match using the weather data as the input variable. The parameters for each model were tuned to find the optimal value and the accuracy of the prediction was used as the evaluation metric. From the selected models, a decision tree was constructed that specified which playing style was more likely to result in an expected win or an upset win for each set of weather conditions.

Data Collection

Data Source

For this analysis, match statistics with corresponding meteorological data needed to be collected. For a full explanation of match statistics and weather features gathered in the dataset, refer to Appendix A (Variable Descriptions).

<u>Match Data Collection</u>

Match statistics were scraped from FBref.com, a subsidiary of Sports-Reference.com. Data was collected from the 2017/2018 to 2021/2022 season (5 total seasons) from teams within the Big Five. Data prior to the 2017/2018 season was not included because the match statistics are significantly less detailed and do not provide the level of insight required to make inferences on style of play. Scraped tables include summary, passing, pass types, defensive actions, possession, and miscellaneous statistics as well as goalkeeper statistics. All identifying player information was dropped and the totals for each team were retained for the match dataset. A partial screenshot from FBref.com is included in Figure 2 below.

Arsenal Player Stats Share & Export ▾ Glossary

Summary | Passing | Pass Types | **Defensive Actions** | Possession | Miscellaneous Stats

						Tackles					Challenges				Blocks						
Player	#	Nation	Pos	Age	Min	Tkl	TklW	Def 3rd	Mid 3rd	Att 3rd	Tkl	Att	Tkl%	Lost	Blocks	Sh	Pass	Int	Tkl+Int	Clr	Err
Folarin Balogun	26	ENG	FW	20-041	58	0	0	0	0	0	0	0		0	0	0	0	0	0	0	0
Bukayo Saka	7	ENG	LW	19-342	32	0	0	0	0	0	0	1	0.0	1	1	0	1	0	0	0	0
Martinelli	35	BRA	LW,FW	20-056	70	1	1	0	0	1	0	1	0.0	1	2	0	2	0	1	0	0
Reiss Nelson	24	ENG	RW	21-246	20	1	0	0	0	1	0	1	0.0	1	1	0	1	0	1	0	0
Nicolas Pépé	19	CIV	RW,FW	26-076	90	0	0	0	0	0	0	0		0	0	0	0	1	1	1	0
Emile Smith Rowe	10	ENG	AM	21-016	90	0	0	0	0	0	0	0		0	0	0	0	0	0	0	0
Granit Xhaka	34	SUI	DM	28-320	90	1	0	0	0	1	1	2	50.0	1	2	0	2	2	3	0	0
Albert Sambi Lokonga	23	BEL	DM	21-295	90	2	1	1	1	0	0	0		0	0	0	0	2	4	1	0
Kieran Tierney	3	SCO	LB	24-069	90	0	0	0	0	0	0	1	0.0	1	2	0	2	0	0	1	0
Pablo Marí	22	ESP	CB	27-347	90	1	0	0	1	0	1	3	33.3	2	0	0	0	1	2	7	0
Ben White	4	ENG	CB	23-309	90	1	0	0	1	0	1	1	100.0	0	0	0	0	1	2	1	0
Calum Chambers	21	ENG	RB	26-205	80	2	1	2	0	0	1	1	100.0	0	1	0	1	0	2	3	0
Nuno Tavares	20	POR	RB	21-199	10	0	0	0	0	0	0	0		0	1	0	1	0	0	1	0
Bernd Leno	1	GER	GK	29-162	90	0	0	0	0	0	0	0		0	0	0	0	0	0	0	0
14 Players					**990**	**9**	**3**	**3**	**3**	**3**	**4**	**11**	**36.4**	**7**	**10**	**0**	**10**	**7**	**16**	**15**	**0**

Arsenal Goalkeeper Stats Share & Export ▾ Glossary

				Shot Stopping					Launched			Passes				Goal Kicks			Crosses			Sweeper	
Player	Nation	Age	Min	SoTA	GA	Saves	Save%	PSxG	Cmp	Att	Cmp%	Att	Thr	Launch%	AvgLen	Att	Launch%	AvgLen	Opp	Stp	Stp%	#OPA	AvgDist
Bernd Leno	GER	29-162	90	3	2	1	33.3	1.0	2	8	25.0	27	3	14.8	27.4	8	50.0	41.5	7	1	14.3	0	6.0

Figure 2: FBRef.com Team Statistics Table

Meteorological Data Collection

Meteorological data was gathered from Meteostat's Python library. It is a simple API for accessing historical open weather and climate data from weather stations around the globe. Using the date, time, and venue location from each of the matches, the Meteostat library provides hourly weather data to include temperature, relative humidity, precipitation, and wind speeds. A screenshot of the hourly weather output in Python is provided in Figure 3.

Out[90]:

time	temp	dwpt	rhum	prcp	snow	wdir	wspd	wpgt	pres	tsun	coco
2021-08-14 12:00:00	21.0	14.2	65.3	0.0	NaN	220.0	18.2	29.6	1020.3	NaN	4.0
2021-08-14 13:00:00	21.2	14.3	65.1	0.0	NaN	230.0	19.5	29.6	1019.8	NaN	3.0
2021-08-14 14:00:00	22.5	15.0	62.7	0.0	NaN	250.0	17.7	29.9	1019.3	NaN	3.0

Figure 3: Meteostat Hourly Weather

Weather forecasts were collected for the match's start time plus two hours to account for 90 minutes of play plus halftime and any additional stoppage time. The maximum values were kept as part of the dataset to capture the worst-case conditions. The exception to this rule was temperature: for temperatures below 10 degrees Celsius, the lowest temperature value was kept.

Data Pre-Processing

Data Cleaning

The dataset was verified for completeness and accuracy. There were some matches that had a final score but did not include match statistics. These cases were normally due to cancellation or forfeiture and were thus excluded from the dataset. For a small percentage of games, Meteostat was unable to provide any meteorological data (i.e., all weather features were null) so those matches were also eliminated from the dataset.

For meteorological data with only some null features, domain knowledge allowed for the replacement of missing data. Null entries for precipitation, snow, and total sunshine features were replaced by zero (mm and minutes). For the wind gust feature, the prevailing wind speed replaced empty entries. For the remaining weather features, any rows with null values were dropped from the dataset.

Feature Selection

Given that the match data set was scraped from multiple, detailed statistics tables, over 150 match statistic features (home and away) were pruned for redundancy (e.g., number of fouls listed twice) or irrelevance (e.g., average length of a goal kick). Domain knowledge was used to justify feature relevance to determining style of play. From the gathered data, 125 admin features, match features (for both home, away teams), and weather features were selected for analysis.

Feature Engineering

New features were added to determine if the result of a match was as expected or an upset. Using the Home Goals and Away Goals features, the match winner was determined (Home Team, Away Team, or Draw), and the winner awarded three points or both teams awarded one point for a tie. The points per game were calculated for each team and then used as a metric for the relative strength for that team (3 PPG is best, 0 PPG is worst). Finally, the home and away team strength feature was compared for each match. If the lower strength team won or tied, the match was deemed an upset; otherwise, the result was reported as expected. An additional feature was created to capture the home field advantage (HFA) which was recorded if the home team won the match.

As another step of the feature engineering, categorical attributes were decomposed and created for simplicity. The Match Result and HFA feature were converted to categorical variables. In addition, the Precipitation, Snow, and Weather condition code features were all converted to categorical variables, with an explanation of features listed in Table 1 below.

Feature	**Precipitation**	**Snow**	**Weather Condition**
Attributes	None (0 mm)	None (0 mm)	Fair (Code 1 – 4)
	Light (0 – 2.5 mm)	Light (0 – 1 mm)	
	Moderate (2.5 – 4 mm)	Moderate (1 – 2.5 mm)	Adverse (Codes 5 – 27)

	Heavy (> 4 mm)	Heavy (>2.5 mm)

Table 1: Categorical Variables

Finally, two interaction terms were created to capture the relationship between some of the weather features. The first was the product of temperature and relative humidity, the second was the product of wind speed and precipitation, and the third was the product of wind speed and absolute difference of wind gust and prevailing wind.

<u>Data Transformation, Splitting</u>

A minimum/maximum scaler was used to standardize the data from 0 to 1 to account for the differences in the magnitude of the various features. Finally, the data set was split into training and testing sets, with an 80-20% split.

<u>Processed Data Set</u>

The final match dataset consisted of a total of 8547 rows (matches) and 136 columns (features). Table 2 provides the breakdown of games and teams per league.

	Bundesliga	**EPL**	**La Liga**	**Ligue 1**	**Serie A**	**Total**
Teams	25	28	28	26	28	135
Matches	1434	1721	1824	1903	1665	8547

Table 2: Data Set Summary

Evaluation and Results

Dimensionality Reduction

For the purposes of defining a team's style of play, a separate data set was constructed, hereafter referred to as the team data set, where the match statistic features were aggregated by team across all seasons. The features for each team were normalized by the number of games played to account for relegation (two worst performing teams forced down to second-tier league) and promotion (replacement from second league), and finally standardized from 0 to 1.

Initially, Principal Component Analysis (PCA) was used to reduce the dimensionality of the team data set from 56 to 2. However, the two principal components could explain less than 56% of the variation in the data and failed to visualize the data along the intended axes. An alternative method that employed domain knowledge was used instead for dimensionality reduction.

Team features were assigned to one of four groups: Offense, Defense, Skill, or Strength. While overlap may exist between these groups, feature assignment was absolute and non-repeated. A score was calculated for each group by summing each feature and dividing by the number of features in the group. To account for those features where a lower value was "better" (e.g., goals against for defense), those values were made negative. Table 3 provides some of the features used in each group. Refer to Appendix B for the complete list of feature assignments.

Axis	**Offense**	**Defense**	**Skill**	**Strength**

Match Statistic Feature	Goals For	Goals Against	Possession	Total Tackles
	Shots on Target For	Defensive Error	Passes Completed	Yellow/Red Cards
	Corner Kicks	Interceptions	Touches	Fouls Committed

Table 3: Feature Group Assignment Examples

Using the scores obtained for Offense, Defense, Skill, and Strength, each team was projected in four dimensions for use in the clustering analysis. From the correlation matrix provided in Figure 4, skill and offense are strongly correlated as well as defense and strength.

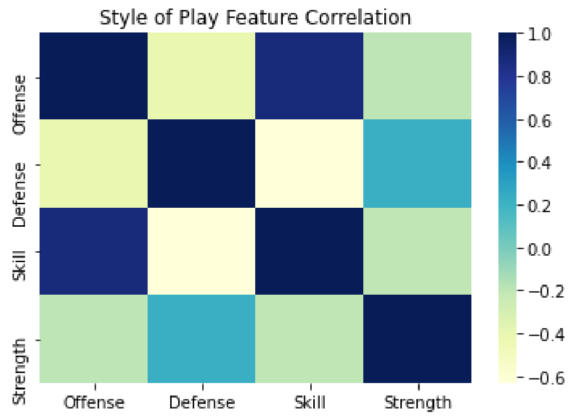


Figure 4: Style of Play Correlation Matrix

Using the reduced data set, PCA was applied in order to reduce the dimensionality to two principal components for visualization purposes. Those two components were able to explain 90% of the variation in the data and as such is a much better fit for the data set. Figure 5 depicts each team's style of play projected into two dimensions.

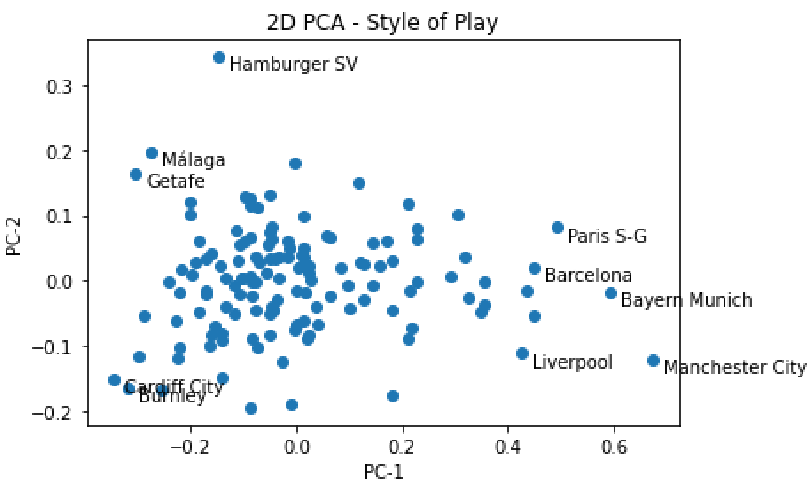


Figure 5: PCA Style of Play

	PC-1	PC-2
Offense	0.741	0.147
Defense	0.184	-0.144
Skill	0.635	0.043
Strength	0.113	-0.978

Table 4: PCA Components

Each team's style of play can be surmised from their relative position on each axis. Teams such as Manchester City, Bayern Munich, Liverpool are the strongest performing teams offensively.

Alternatively, Cardiff City, Burnley, and Hamburger SV are the three strongest teams defensively. PC-1 can most simply be interpreted as the Defense – Offense axis.

Comparatively, Manchester City, Barcelona, and Paris S-G are the teams that most favor skill and possession tactics. On the opposite end of the spectrum, Hamburger SV, Malaga, and Getafe prefer tactics of brute force. To that end, PC-2 is the Strength – Skill axis.

Clustering

The intent of clustering is to categorize each team's style of play along two axes, Offense-Defense and Strength-Skill. As shown in Figure 6, those teams in Quadrant I would be more offensive in nature, relying on skill and possession while those in Quadrant III rely more on their physical strength and aggressiveness to assume a more defensive approach.

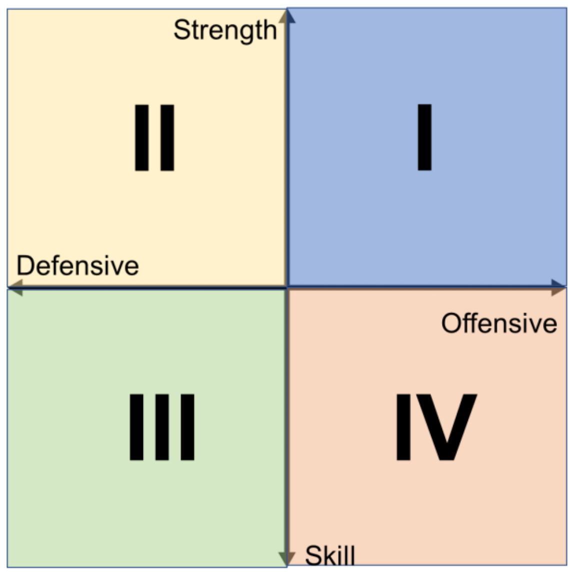


Figure 6: Style of Play Qualities

Two clustering algorithms were implemented to group similar playing styles: K-Means and Gaussian Mixture Model (GMM). K-Means partitions data into k clusters based on their similarity, iteratively assigning points to the nearest cluster center, and then recalculating the centroid of each cluster. Comparatively, GMM is a statistical model that assumes the underlying distribution is a combination of multiple Gaussian distributions.

Compared to GMM, K-Means has the benefit of simplicity, computational efficiency, and ease of interpretation. However, GMM can capture complex data distributions better than K-Means, particularly when the data has non-linear or non-spherical clusters.

<u>K-Means</u>

An inertia plot was used to determine the optimal number of k clusters to form for the K-Means algorithm. Otherwise known as the 'Elbow Method', the number of clusters was determined by visually assessing the bend in the plot of inertia values at various k's. From Figure 7, the bend appears to occur where $k = 5$. Figure 8 illustrates the cluster assignment for each team.

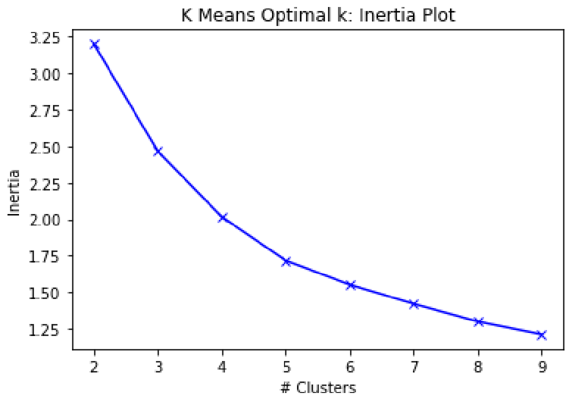


Figure 7: K-Means Inertia Plot

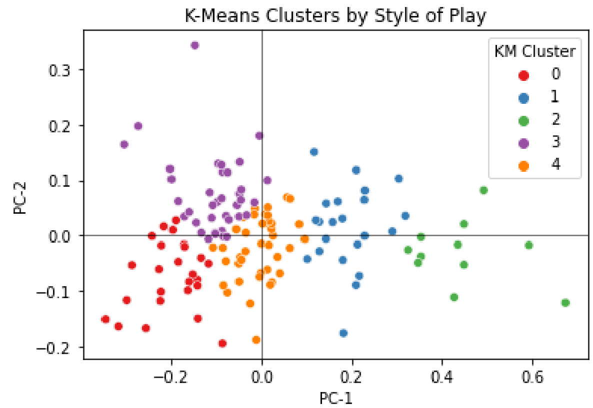


Figure 8: K-Means Clustering

<u>GMM</u>

For the GMM algorithm, the number of clusters was determined using a combination of plots of the silhouette score and BIC. The silhouette score is a measure of the intra-cluster similarity compared to other clusters so the optimal k occurs where the score is maximized using the fewest clusters. Comparatively, the BIC criterion gives an estimation for the model's ability to predict the data given. By evaluating the gradient, it can more clearly be seen that the optimal number of clusters occurs when the gradient is constant.

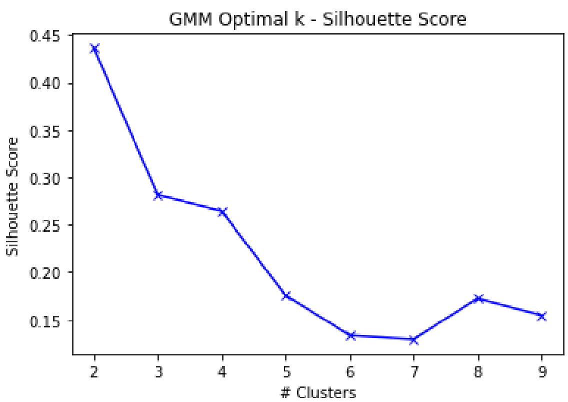


Figure 9: Silhouette Plot

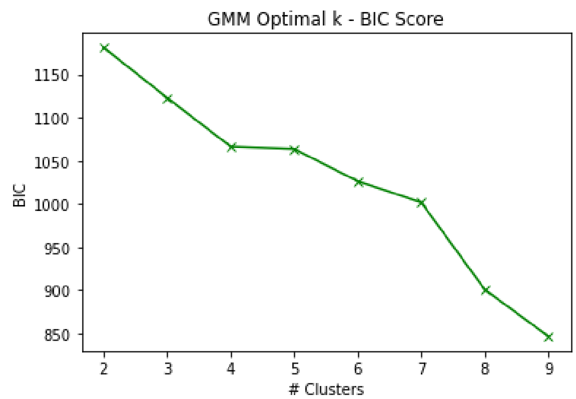


Figure 10: BIC Plot

Using $k = 4$ clusters, Figure 11 illustrates the distribution of teams per GMM cluster.

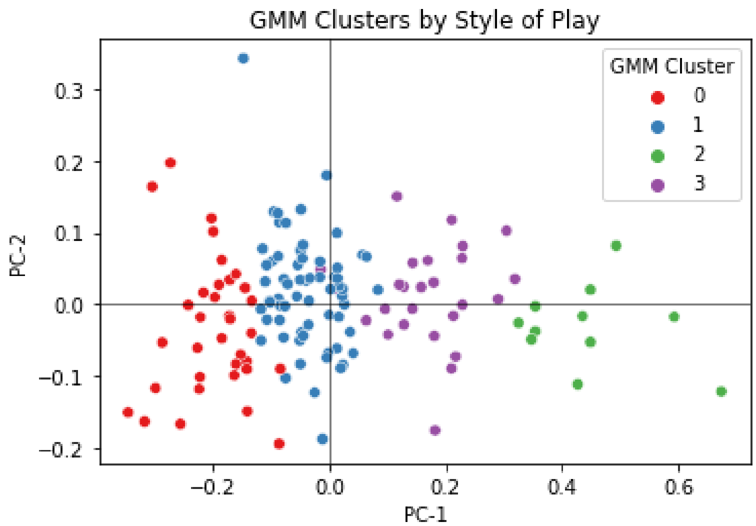


Figure 11: GMM Clustering

<u>Model Comparison</u>

The performance of the two clustering algorithms were compared using the silhouette score at the optimal number of clusters. Table 5 outlines each model's score, with K-Means performing slightly better than GMM.

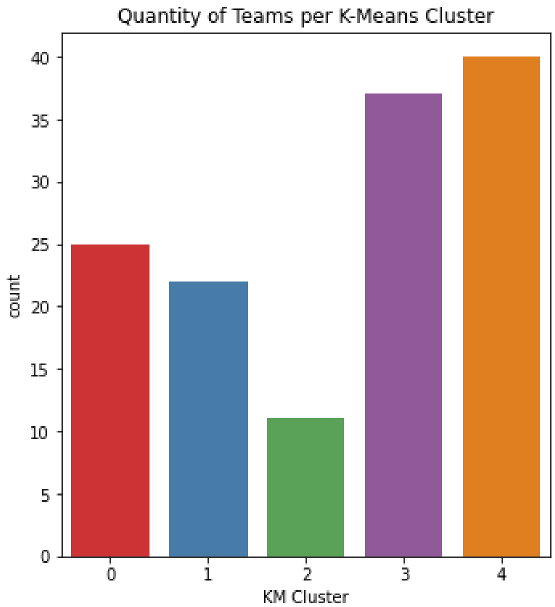
	K-Means $k = 5$	**GMM** $k = 4$
Silhouette Score	0.310	0.269

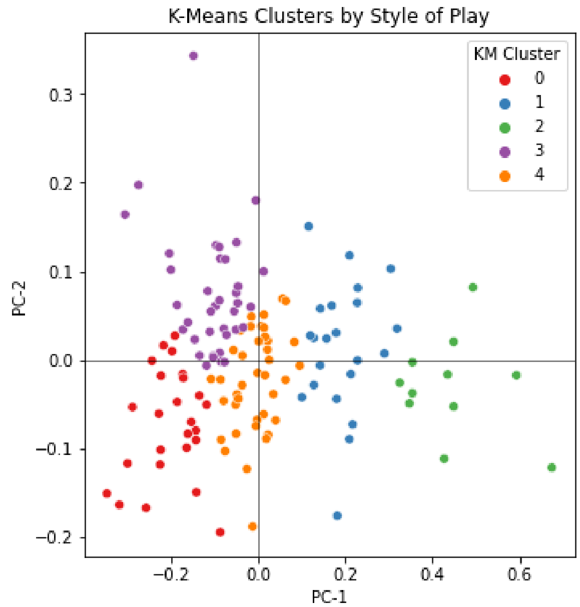
Table 5: Clustering Model Comparison

Given the silhouette score, computational efficiency, and the delineations drawn between the more defensive teams, K-Means was chosen as the clustering model of choice with 5 clusters.

<u>Cluster Qualities</u>

With K-Means as the chosen clustering model, Figures 12 and 13 provide a detailed distribution of each cluster and Table 6 specifies the predominant qualities in each.





Figures 12, 13: K-Means Cluster Distribution

Cluster	Distribution	Qualities
0	19%	Defensive, Skill
1	16%	Offensive, Balanced
2	8%	Offensive, Skill
3	27%	Defensive, Strength
4	30%	Generally Balanced

Table 6: K-Means Cluster Style of Play Qualities

The cluster assignments were added to the original match data set as categorical variables. In the pie plot and bar plot in Figures 14 and 15, expected results occur just over 52% of the time and upsets are surprisingly more common. This distribution holds true over each season.

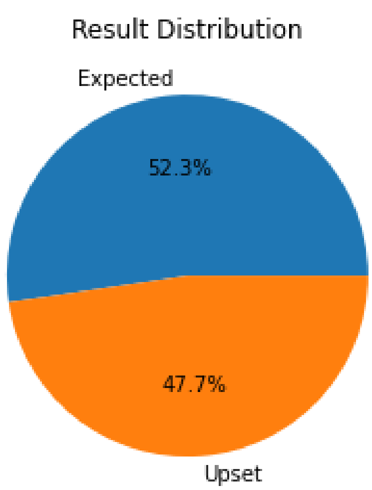


Figure 14: Results Distribution (Average)

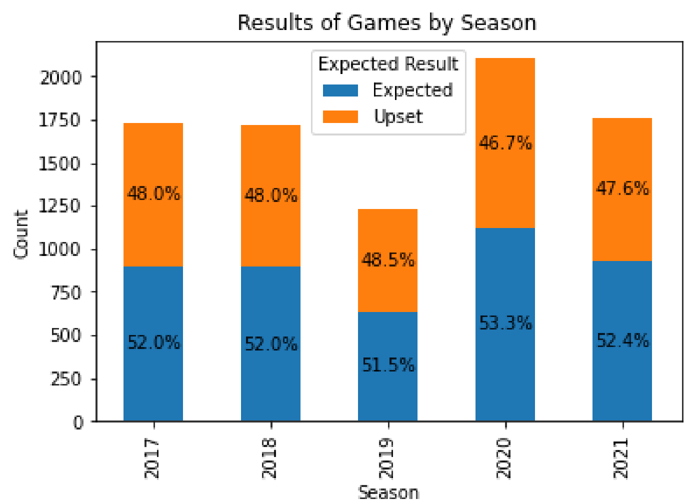


Figure 15: Results Distribution (by Season)

For each of the four clusters of playing styles, the percentage of expected results and upset wins was calculated in order to compare against the average. From Table 7, it can be shown that Clusters 3 and 4 are more likely than average to result in an upset result. Comparatively, Clusters 0, 1, and especially Cluster 2 are more impervious to variations in the expected result.

Cluster	**0**	**1**	**2**	**3**	**4**
Expected	54%	53%	66%	49%	49%
Upset	46%	47%	34%	51%	51%

Table 7: Playing Style Results

Classification

The classification models chosen for this bivariate problem included SVM, Decision Tree, Random Forest, and Logistic Regression. SVM aims to find the optimal hyperplane that separates two classes of data. The algorithm works well with high-dimensional datasets and can handle linear and non-linear boundaries through different kernel functions. Comparatively, Decision Tree is much easier to visualize and consequently interpret given that its output is a tree-like structure where the branches represent the decision rules for its leaf predictions. Random Forest was included because it is generally more robust and accurate than a single decision tree since it reduces overfitting by combining multiple decision trees. Finally, Logistic Regression models the probability of a binary outcome based on the relationship between the independent variables and the dependent variable. It has the benefit that it can be regularized to prevent overfitting, which is especially important in a high-dimensional dataset.

Together, GridSearchCV and cross-validation were used to tune and evaluate the performance of each model, and to select the best combination of hyperparameters that produces the highest accuracy. Table 8 reports the optimal values of each hyperparameter and the performance of each of the models.

Model	**Parameter**	**Search Space**	**Optimal**	**Accuracy**
SVM	Kernel	Linear, Rbf, Poly	Linear	69.12 %
	C	0.1, 1, 10, 100	0.1	
Decision Tree	Max Features	Sqrt, Log2, None	Sqrt	78.01 %
	Max Depth	3,4,5,6,7	5	
Random Forest	Max Features	Sqrt, Log2, Auto, None	log2	76.96%
	Estimators	10,50,100,150,200,250,300	10	
	Min Leaf	1,10,50,100	50	
	Max Depth	3,4,5,6,7	7	
Logistic Regression	Penalty	l1, l2	L2	66.08 %
	C	0.1, 1, 10, 100	10	

Table 8: Classification Model Performance (Accuracy)

The accuracy of each model is in the same general range, with Decision Tree reporting the best performance at 78% accuracy and closely followed by Random Forest. Given the similarity in the reported accuracy of the models, additional evaluation metrics were considered for selection. Table 9 provides the precision, recall, F1-score, and ROC-AUC for each model using the optimal hyperparameters found in Table 9.

Model	Precision	Recall	F1-Score	ROC-AUC
SVM	0.677	0.687	0.682	0.691
Decision Tree	0.796	0.732	0.762	0.778
Random Forest	0.741	0.752	0.746	0.754
Logistic Regression	0.597	0.608	0.603	0.613

Table 9: Classification Model Performance (Additional Measures)

From these results, Decision Tree was selected as the classification model of choice since its accuracy, precision, and F1-score were superior to every other model.

Decision Tree Interpretation

In addition to outperforming the other models in terms of their evaluation metrics, Decision Tree has the benefit of easy visualization and interpretation. Using the hyperparameters tuned in the previous section, a decision was fit to the data set to determine which playing style is more likely to result in an upset win under certain weather conditions. The full-size decision tree is provided in Appendix C.

Interpreting the model, it was found that matches with Cluster 2 teams (Offensive, Skill) are more likely to result in an upset in the presence of windy and/or cold conditions (less than 8 degrees Celsius). Intuitively, this is a reasonable conclusion since long or airborne passes are more challenging in such conditions. For Cluster 1 teams (Offensive, Balanced), wind was against the primary disruptor which led to an upset. On the opposite end of the spectrum, Cluster 0 teams (Defensive, Skill) were susceptible to rain and gusting wind conditions. With rainfall making the field slick, a greater number of errors are likely to occur with devastating consequences for the defense. For Clusters 3 (Defensive, Strength) and 4 (Generally Balanced), they did not enter the Decision Tree until a depth of 8 at which point, overfitting is a concern. Given that these clusters are more likely to produce an upset victory, it can be inferred that those weakness of the previous clusters would serve as their strengths.

Conclusion

In this project, it was found that weather does influence the outcome of a match, specifically an expected result or an upset. Armed with weather forecasts and knowledge of a team's playing style, we can predict with 75% accuracy if the results of a match are as expected. Weather particularly disrupts the game play for offensive, skill-based team with respect to wind, rain, and temperature.

Appendix

Appendix A : Variable Descriptions

Variable	Description
League	-
Date	-
Time	-
Location	-
Latitude	-
Longitude	-
('Performance', 'Gls')	Home / Away Goals
('Performance', 'Ast')	Home / Away Assists
('Performance', 'PK')	Home / Away PK
('Performance', 'PKatt')	Home / Away PK Attempt
('Performance', 'Sh')	Home / Away Shots
('Performance', 'SoT')	Home / Away Shots on Target
('Performance', 'CrdY')	Home / Away Yellow Card
('Performance', 'CrdR')	Home / Away Red Card
('SCA', 'SCA')	Home / Away Shot Creating Action
('SCA', 'GCA')	Home / Away Goal Creating Action
('Passes', 'PrgP')	Home / Away Penetrating Pass
('Carries', 'Carries')	Home / Away Carries
('Carries', 'PrgC')	Home / Away Penetrating Carries
('Total', 'Cmp')	Home / Away Passes Complete
('Total', 'Att')	Home / Away Passes Attempted
('Total', 'Cmp%')	Home / Away Pass Complete %
('Total', 'TotDist')	Home / Away Total Distance Passed
('Total', 'PrgDist')	Home / Away Progressive Distance Passed
('Short', 'Cmp%')	Home / Away Short Pass Completion %
('Medium', 'Cmp%')	Home / Away Medium Pass Completion %
('Long', 'Cmp%')	Home / Away Long Pass Completion %
('Pass Types', 'Sw')	Home / Away Switch Pass
('Pass Types', 'Crs')	Home / Away Cross Pass
('Pass Types', 'CK')	Home / Away Corners
('Tackles', 'Tkl')	Home / Away Tackles
('Tackles', 'TklW')	Home / Away Tackles Won
('Challenges', 'Tkl%')	Home / Away % Dribblers Tackled
('Blocks', 'Blocks')	Home / Away Total Blocks
('Blocks', 'Sh')	Home / Away Shots Blocked
('Blocks', 'Pass')	Home / Away Passes Blocked

('Unnamed: 18_level_0', 'Int')	Home / Away Defensive Interceptions
('Unnamed: 20_level_0', 'Clr')	Home / Away Defensive Clearance
('Unnamed: 21_level_0', 'Err')	Home / Away Defensive Error
('Touches', 'Touches')	Home / Away Touches
('Take-Ons', 'Att')	Home / Away Take On Attempts
('Take-Ons', 'Succ')	Home / Away Take On Success
('Take-Ons', 'Succ%')	Home / Away Take On Success %
('Carries', 'Carries')	Home / Away Carries
('Carries', 'TotDist')	Home / Away Total Distance Carried
('Carries', 'Mis')	Home / Away Miscontrol
('Carries', 'Dis')	Home / Away Dispossed
('Performance', 'Fls')	Home / Away Fouls
('Performance', 'Fld')	Home / Away Fouls Drawn
('Performance', 'Off')	Home / Away Offsides
('Performance', 'PKwon')	Home / Away PK Won
('Performance', 'PKcon')	Home / Away PK Conceded
('Performance', 'OG')	Home / Away Own Goal
('Performance', 'Recov')	Home / Away Recovery
('Aerial Duels', 'Won')	Home / Away Aerial Won
('Aerial Duels', 'Lost')	Home / Away Aerial Lost
('Aerial Duels', 'Won%')	Home / Away Aerial Won %
('Shot Stopping', 'SoTA')	Home / Away Shots on Target Against
('Shot Stopping', 'GA')	Home / Away Goals Against
('Shot Stopping', 'Saves')	Home / Away Saves
Home / Away Team	Home / Away Team
temp	The air temperature in °C
dwpt	The dew point in °C
rhum	The relative humidity in percent (%)
prcp	The one hour precipitation total in mm
snow	The snow depth in mm
wdir	The average wind direction in degrees (°)
wspd	The average wind speed in km/h
wpgt	The peak wind gust in km/h
pres	The average sea-level air pressure in hPa
tsun	The one hour sunshine total in minutes (m)
coco	The weather condition code

Code	**Weather Condition**
1	Clear

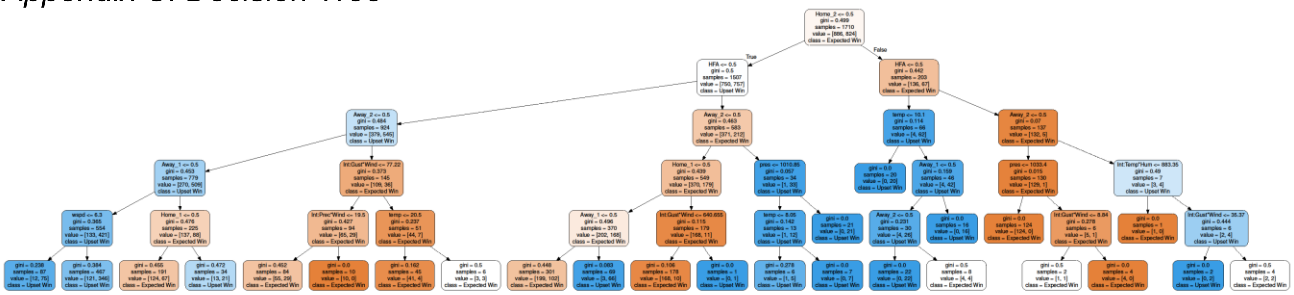
2	Fair
3	Cloudy
4	Overcast
5	Fog
6	Freezing Fog
7	Light Rain
8	Rain
9	Heavy Rain
10	Freezing Rain
11	Heavy Freezing Rain
12	Sleet
13	Heavy Sleet
14	Light Snowfall
15	Snowfall
16	Heavy Snowfall
17	Rain Shower
18	Heavy Rain Shower
19	Sleet Shower
20	Heavy Sleet Shower
21	Snow Shower
22	Heavy Snow Shower
23	Lightning
24	Hail
25	Thunderstorm
26	Heavy Thunderstorm
27	Storm

Appendix B: Style of Play Axis Features

Offense	**Defense**	**Strength**	**Skill**
Goals For	Total Blocks	Tackles	Carries
Shots For	Shots Blocked	Tackles Won	Passes Complete
Shots on Target	Passes Blocked	% Dribblers Tackled	Passes Attempted
Shot Creating Action	Defensive Interceptions	Yellow Card	Total Distance Passed
Goal Creating Action	Defensive Clearance	Red Card	Progressive Distance Passed
Penetrating Pass	Defensive Error	Fouls	Short Pass %
Penetrating Carries	Own Goal	PKs Conceded	Medium Pass %
Corners	Recovery		Long Pass %
Offside	Aerial Won		Touches
	Aerial Lost		Take on Success
	Shots on Target Against		Total Distance Carried

	Goals Against		Miscontrol
	Saves		Dispossessed
			Fouls Drawn
			PK Won

Appendix C: Decision Tree



References

[1] URL: https://dev.meteostat.net/

[2] URL: https://fbref.com/en/comps/Big5/Big-5-European-Leagues-Stats

[3] URL: https://dreampuf.github.io/GraphvizOnline/