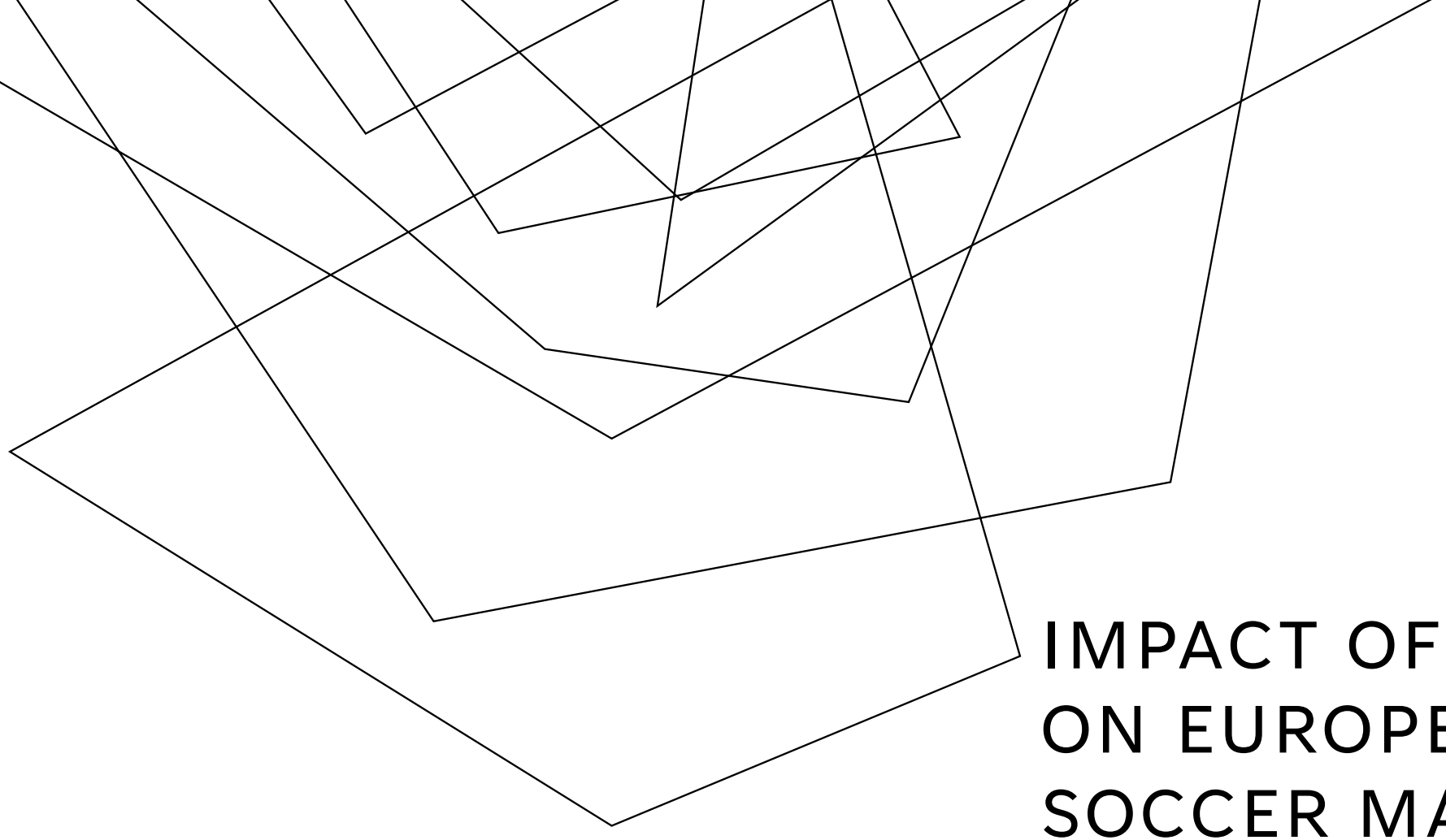


# IMPACT OF WEATHER ON EUROPEAN SOCCER MATCH OUTCOMES

Emily Wolf

IsYE 7406 – Data Mining – Spring 2023

# PROBLEM STATEMENT

The objective is to explore the impact, if any, of weather on the results of a soccer matches in the top five European soccer leagues (Bundesliga, English Premier League, La Liga, Serie A, and Ligue 1)

1. Can weather data be used to predict the outcome of a game?
    - Specifically, if there is an upset victory (the lower ranked opponent wins or draws) of if the outcome is as expected (the higher ranked opponent wins).

2. Provided that it does have an effect, are teams with certain styles of play affected differently than others?

# METHODOLOGY

1. Data Pre-Processing
    a. Match Data Set
    b. Team Data Set

2. Dimensionality Reduction
    a. PCA

3. Cluster Analysis
    a. K-Means
    b. Gaussian Mixture Model
    c. Model Comparison

4. Classification Model
    a. SVM
    b. Decision Tree
    c. Random Forest
    d. Logistic Regression

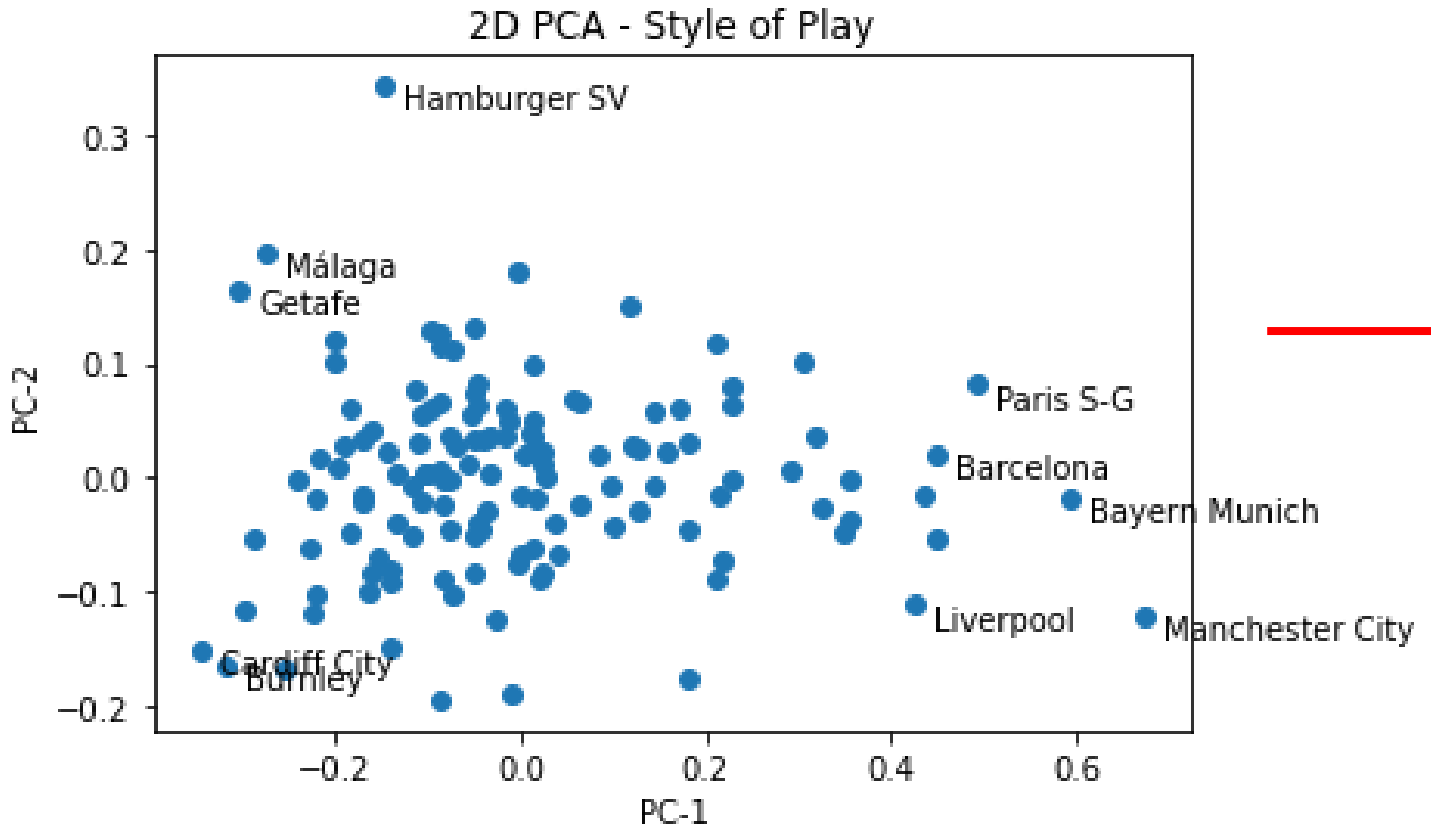
# DATA PRE-PROCESSING

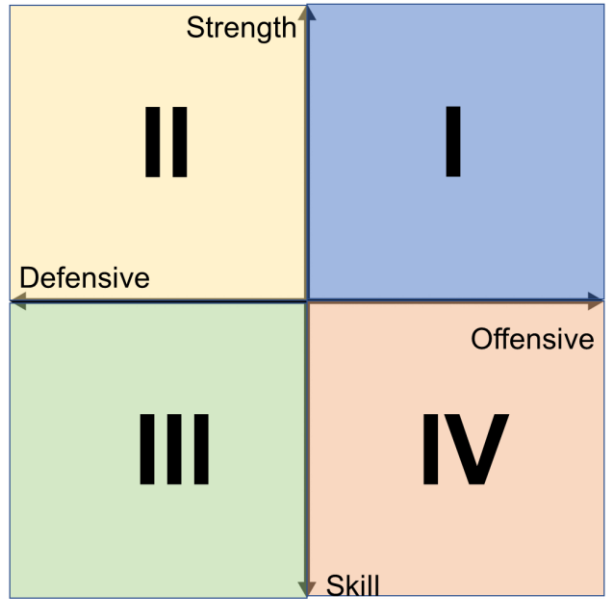

| League | Games Played |
|---|---|
| Bundesliga | 1434 |
| EPL | 1721 |
| La Liga | 1824 |
| Serie A | 1665 |
| Ligue 1 | 1903 |
| TOTAL | 8547 |

| Temperature | Relative Humidity | Dewpoint | Precipitation | Snow | Wind Speed | Wind Gust | Pressure Altitude | Adverse Condition |
|---|---|---|---|---|---|---|---|---|
| Continuous | Continuous | Continuous | Categorical | Categorical | Continuous | Continuous | Continuous | Categorical |

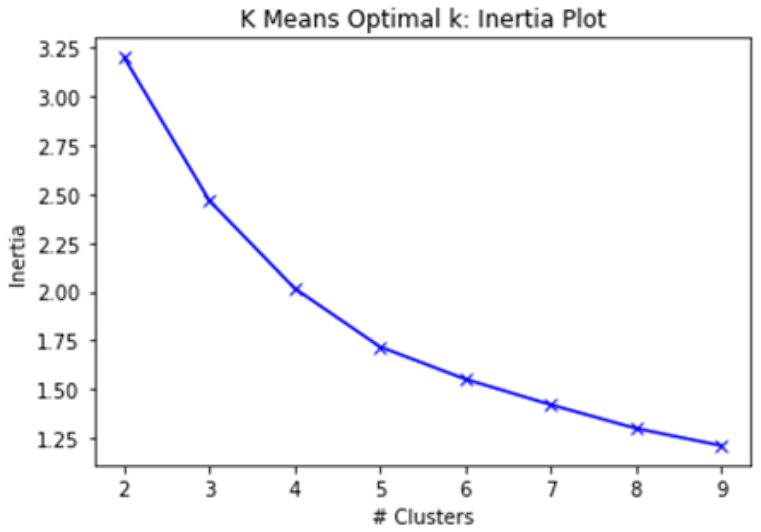
Data scraped from Fbref.com for the 2017/2018 to 2021/2022 seasons (5 total seasons).

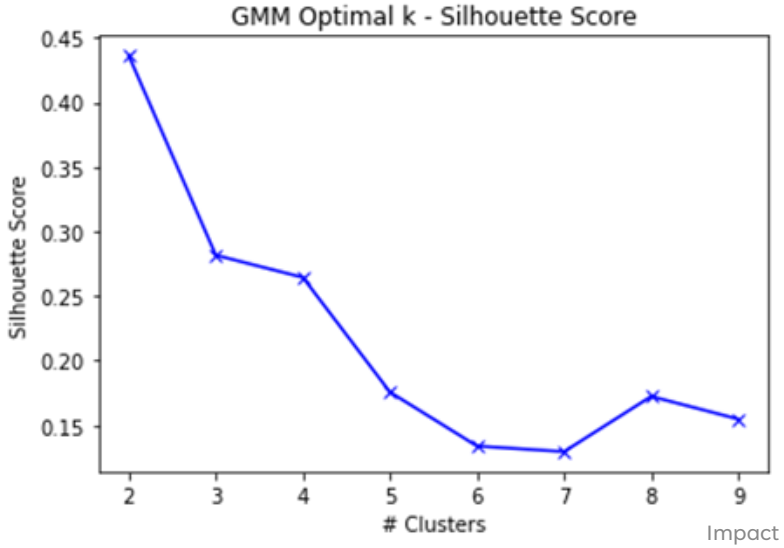
# DIMENSIONALITY REDUCTION

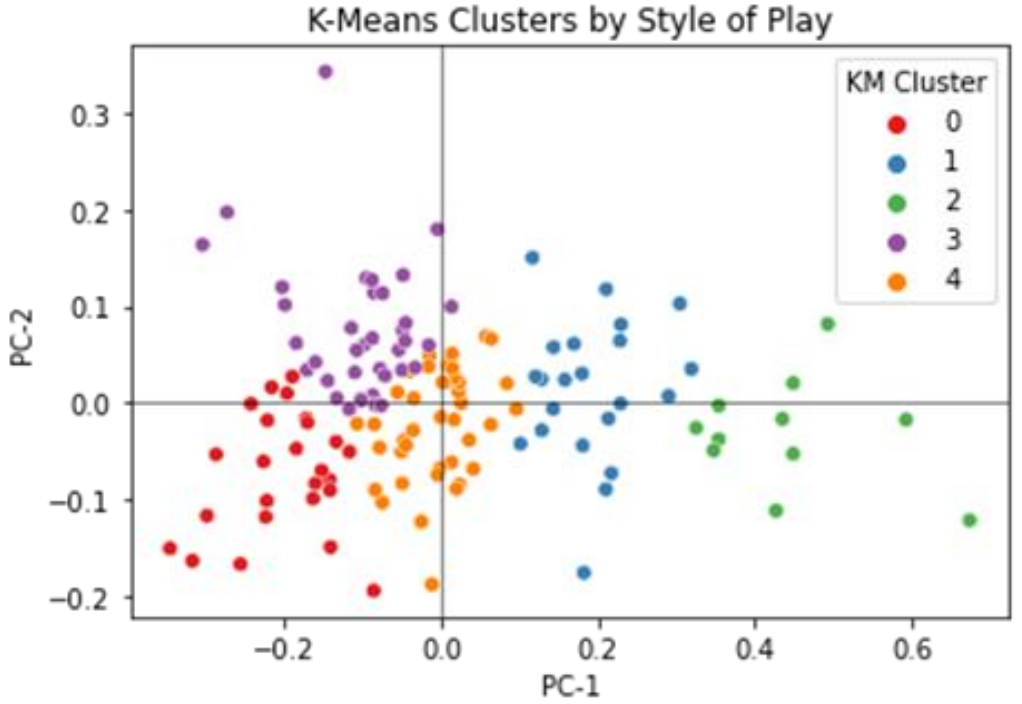


# CLUSTERING

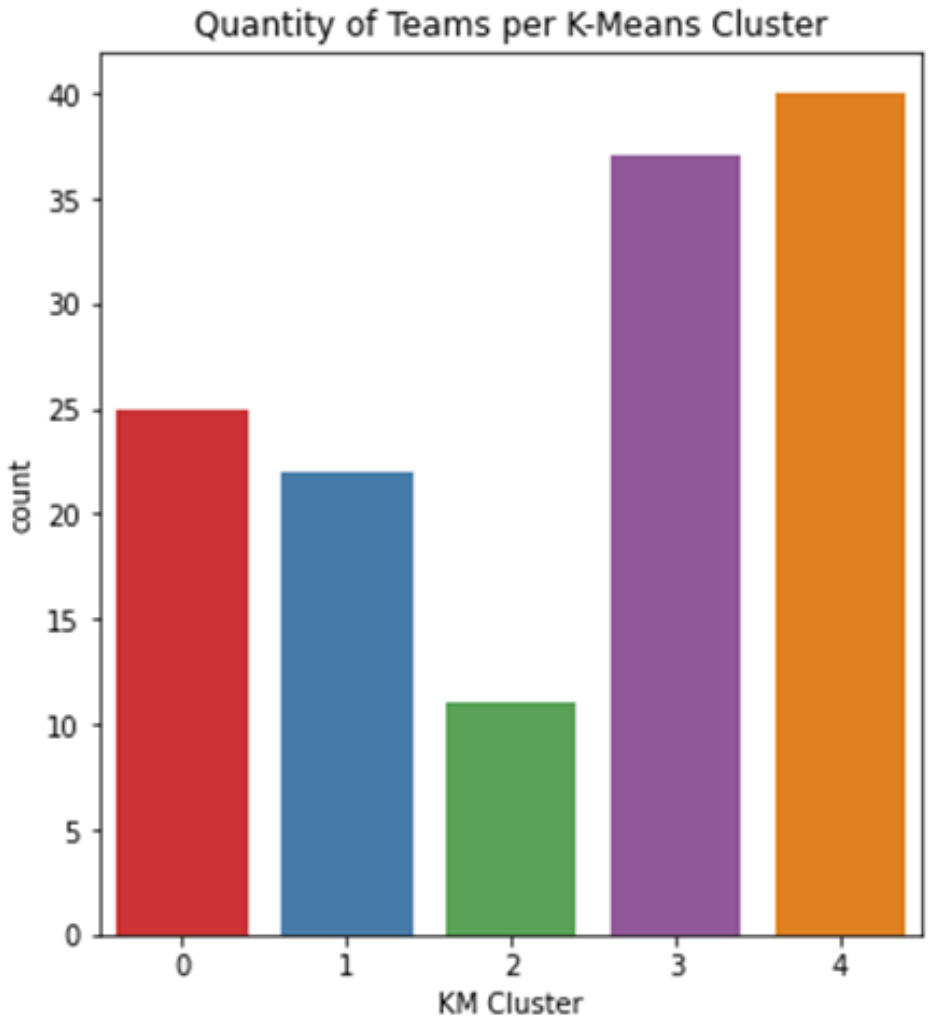


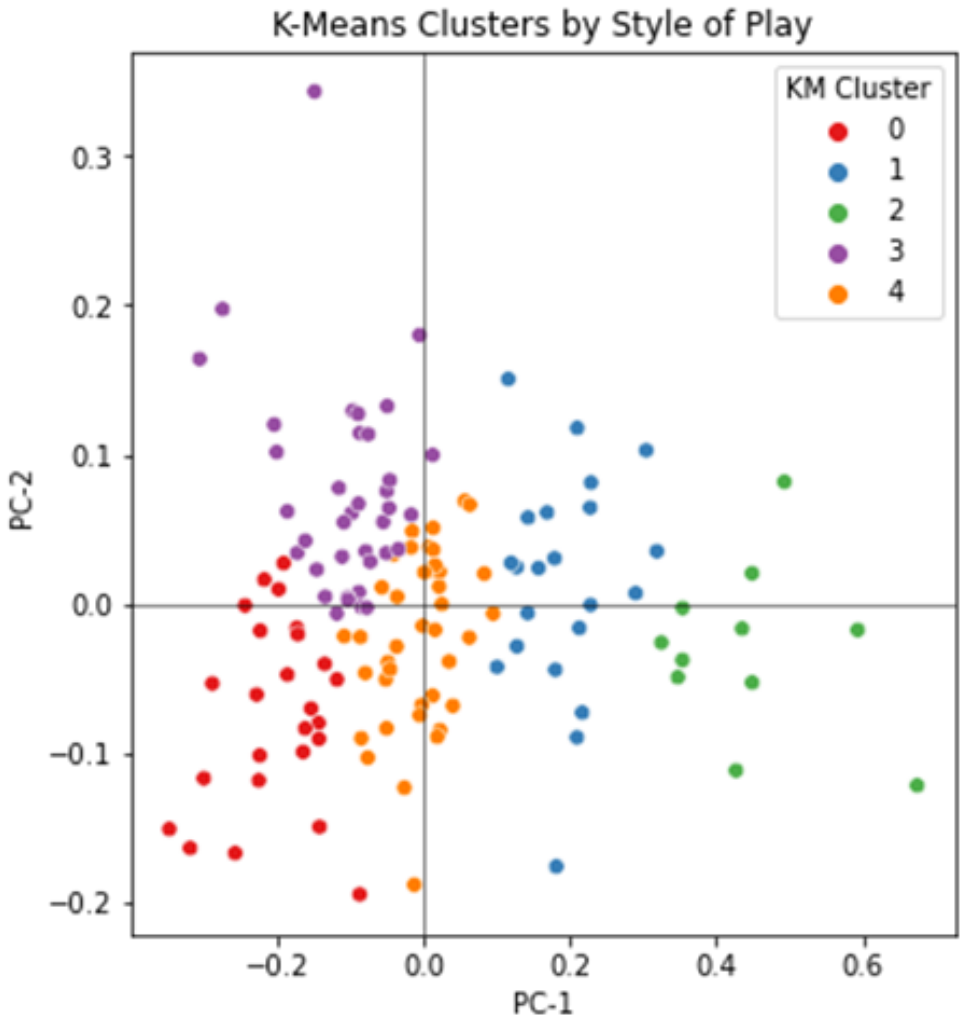




| Model | Silhouette Score |
|---|---|
| **K-Means, $k = 5$** | **0.310** |
| GMM, $k = 4$ | 0.269 |

# CLUSTERING

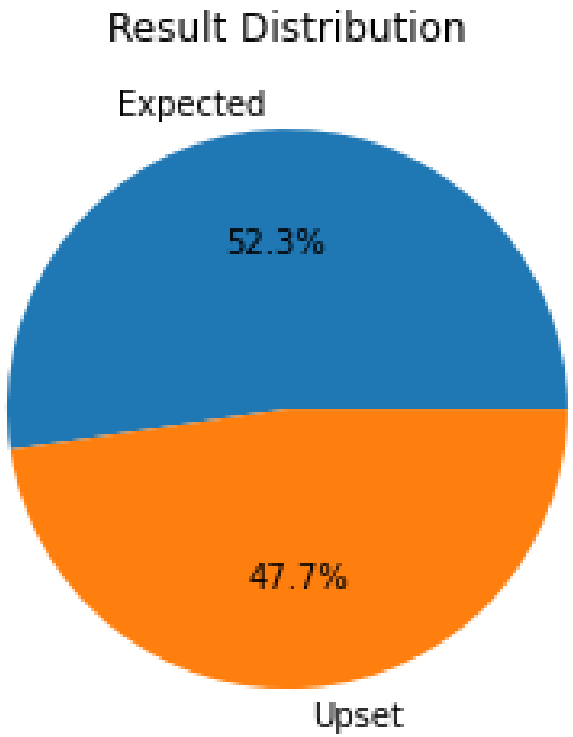

Quantity of Teams per K-Means Cluster
count
KM Cluster
0
1
2
3
4
0
5
10
15
20
25
30
35
40


K-Means Clusters by Style of Play
KM Cluster
0
1
2
3
4
PC-1
PC-2
−0.2
0.0
0.2
0.4
0.6
−0.2
−0.1
0.0
0.1
0.2
0.3

# CLUSTERING

Distribution of Results based on Cluster:

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



# CLASSIFICATION

| Model | Parameter | Search Space | Optimal Value | Accuracy |
|---|---|---|---|---|
| SVM | Kernel | Linear, Rbf, Poly, Sigmoid | Linear | 69.12 % |
| | C | 0.1, 1.0, 10, 100 | 0.1 | |
| Decision Tree | Max Features | Sqrt, Log2, Auto | Sqrt | 78.01 % |
| | Max Depth | 3, 4, 5, 6, 7 | 5 | |
| Random Forest | Max Features | Sqrt, Log2 | log2 | 76.96% |
| | Estimators | 10, 50, 100, 150, 200, 250, 300 | 10 | |
| | Min Leaf | 1,5, 10, 50, 100, 200, 500 | 50 | |
| | Max Depth | 3, 4, 5, 6, 7 | 7 | |
| Logistic Regression | Penalty | l2, l2 | L2 | 66.08 % |
| | C | 0.1, 1.0, 10, 100 | Linear | 69.12 % |

# CLASSIFICATION

| Cluster | Style of Play | Contributing Weather Conditions |
|---|---|---|
| 0 | Defensive, Skill | Presence of Precipitation/Snow, windy conditions produces Upset Win |
| 1 | Offensive, Balanced | Gusting Winds produces Upset |
| 2 | Offensive, Skill | Windy conditions and cold temperatures produce Upset |
| 3 | Defensive, Strength | Prescence of Precipitation, High Winds, and Cold weather result in Upset in their favor |
| 4 | Generally Balanced | Prescence of Precipitation, High Winds, and Cold weather result in Upset in their favor |