“The art of getting sacked”

A *research by Jahnvi Rameshbhai Patel & Yusuf Ejaz*

Abstract

In this study, we investigate the factors that may influence whether or not a football (soccer) manager gets sacked. To achieve this, we collected data on various performance indicators, such as league position, number of wins, losses, draws, goals scored and goals conceded, along with managerial information such as the number of games overseen, winning, and losing streaks, and the date when the manager was sacked.

The study also looks at recent examples of managerial dismissals in professional football, including those involving Thomas Tuchel at Chelsea, Antonio Conte at Tottenham, and Liverpool's decision to retain their manager despite a recent dip in form. The time frame, we are taking into account is the previous six seasons, which allows for a more comprehensive analysis of trends and patterns and their consequences that may emerge.

Our study utilizes different software tools such as Microsoft Excel and RapidMiner to analyze the data and employ machine learning algorithms. We apply classification and clustering techniques to identify patterns in the data and determine which factors are most strongly associated with the likelihood of a manager being sacked. Our results suggest that the number of games managed, league position, and winning and losing streaks are the most significant factors, while sentiment analysis may provide additional insights into the decision-making processes of club management.

By providing insights into the factors that contribute to managerial dismissals and their impact on club performance, the study offers a valuable contribution to the world of professional football.

1 | Introduction

The world of professional football is no stranger to the phenomenon of managerial dismissals, with poor team performance often being the most common reason cited for sacking a manager. It is often seen as a last resort by club management when results are poor, or when a manager is deemed to be underperforming. However, the decision to sack a manager can be a complex one, influenced by a range of factors, including the team's performance on the field, the manager's tactics, and the club's financial situation. In recent years, there has been growing interest in understanding the factors that influence whether or not a manager gets sacked. In this study, we aim to investigate this issue using data analysis techniques. We use a range of data sources, including manager and club names, league position, number of games managed, winning, and losing streaks, goals conceded and scored, and sentiment analysis, to identify the factors that are most strongly associated with the likelihood of a manager being sacked. By doing so, we hope to contribute to the ongoing debate around the role of club management in football, and the factors that influence their decision-making processes.

2 | Literature review

Both research papers address the question of when a football manager in the English Premier League should be sacked, but they differ in the data used and methodology employed. Research paper by Chris Hope uses data from the earlier seasons of 1995-2001 to develop a simple model based on management science techniques based on team position, recent performance, and the manager's length of tenure. In contrast, our research paper uses more recent data from the seasons of 2017-2023 and employs machine learning algorithms to predict whether or not a manager will be sacked by featuring league position, number of games managed, winning and losing streaks, goals scored and conceded, and the date of the sacking. While both papers have

the same goal, they offer different approaches and insights due to the differences in their data sources and methods.

3 | Data

3.1 | Collection Process

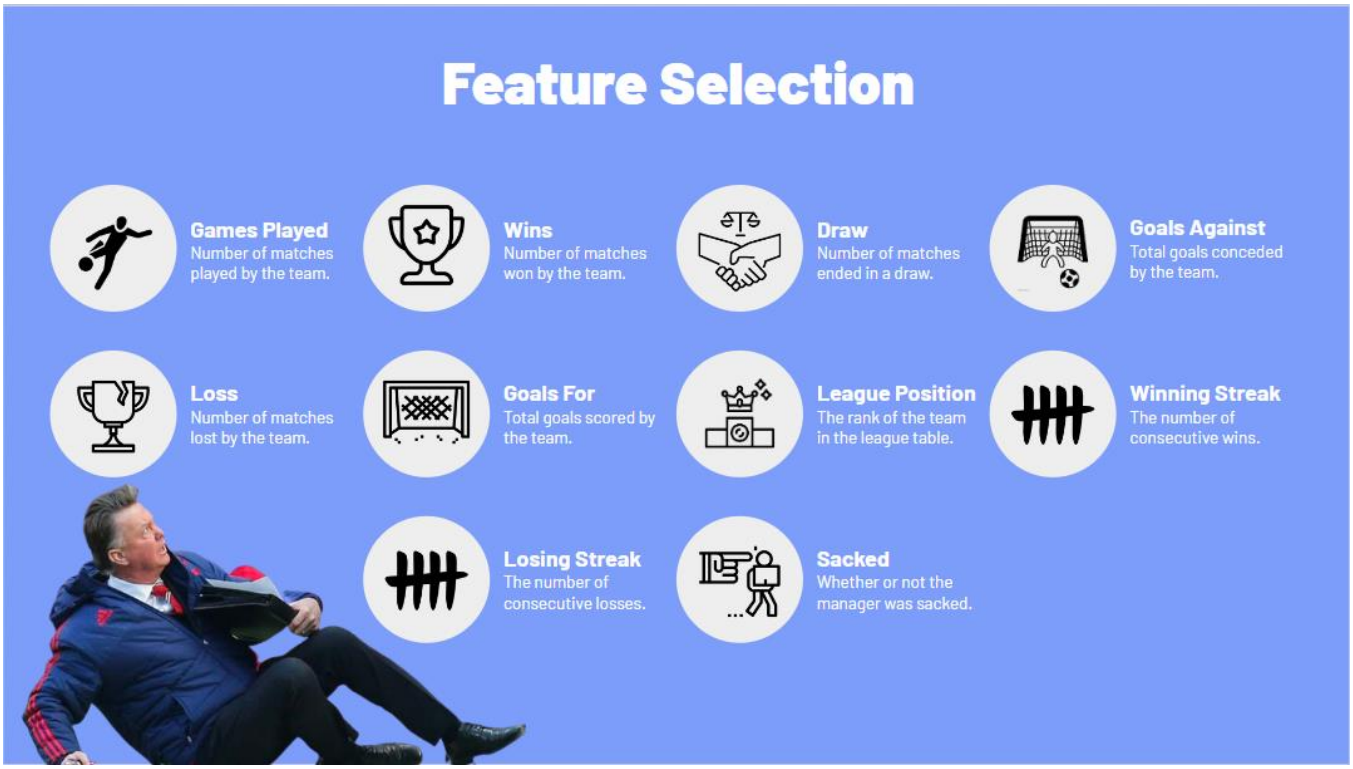
The data was gathered from several third-party websites that host football databases. The websites used include Sky Sports, 11v11, and TWTD. These sources were chosen as they provide comprehensive and reliable information on football clubs, managers, and their performance. By using multiple sources, we were able to gather a broad range of data, which can provide a more accurate representation of the situation. This data was then used to analyze the factors that contribute to a manager's departure from a football club.

3.2 Feature selection

We have used a feature selection method called "weight by information gain" in RapidMiner, which allowed us to select the most informative features for our prediction task, which can help to improve the overall quality and reliability of our analysis.

attribute	weight ↓
Wins	0.633
Goal Scored	0.555
Games Incharge	0.547
Winning Streak	0.364
League Position	0.302
Draw	0.267
Goals Conceded	0.262
Loss	0.094
Losing Streak	0.089

- **Manager & club names:** Identifying the manager and the club is essential to contextualize the data and understand the specifics of the situation.
- **Season:** The season is significant because the performance of the team may vary from one season to another. It provides a broader perspective on how the club has performed over time.
- **League position:** The league position is crucial because it indicates the success of the team. If the team is performing poorly, there may be pressure on the manager to improve results.
- **Number of games the manager oversaw:** The number of games the manager oversaw gives an idea of the duration of the manager's tenure. It helps in analyzing how the team's performance changed over time.
- **Winning & losing streak:** The winning and losing streaks provide information on the team's consistency and form. It can help to identify the causes of a sudden decline in performance.
- **Number of winnings, loss, draw:** The number of winnings, losses, and draws provide an overview of the team's performance. It can help in analyzing the overall trend of the team's performance.
- **Goals conceded, goals scored:** The number of goals conceded and scored provide insights into the team's attacking and defensive capabilities.



- **Whether or not the manager was sacked:** Knowing whether the manager was sacked or not is essential in determining the outcome of the situation. It can provide insights into the club's expectations and the reasons behind the manager's departure.
- **Date when the manager got sacked:** The date when the manager got sacked provides a timeline of events. It helps to identify any patterns or trends in the club's decisions regarding the manager's tenure.

4 | Methodology

In our study, we utilized different software tools for various stages of our analysis. Microsoft Excel was used for data preparation, correlation analysis and charting purposes, while RapidMiner was employed for machine learning algorithms. These software tools allowed us to perform our analyses effectively and efficiently, facilitating the interpretation of results and aiding in the achievement of our research goals.

4.1 | Data Cleaning Techniques

In order to prepare the data for analysis, several data cleaning techniques were applied to the dataset.

Removing duplicates was not a significant issue in our dataset as our focus was on the managers rather than the clubs they worked for. However, we still ensured that any duplicate records were removed to prevent any unnecessary biases or inaccuracies in our analysis.

Data scraping was used to extract relevant data from third-party websites such as Sky Sports, 11v11, and TWTD. This process allowed us to collect a comprehensive dataset that was suitable for our research needs.

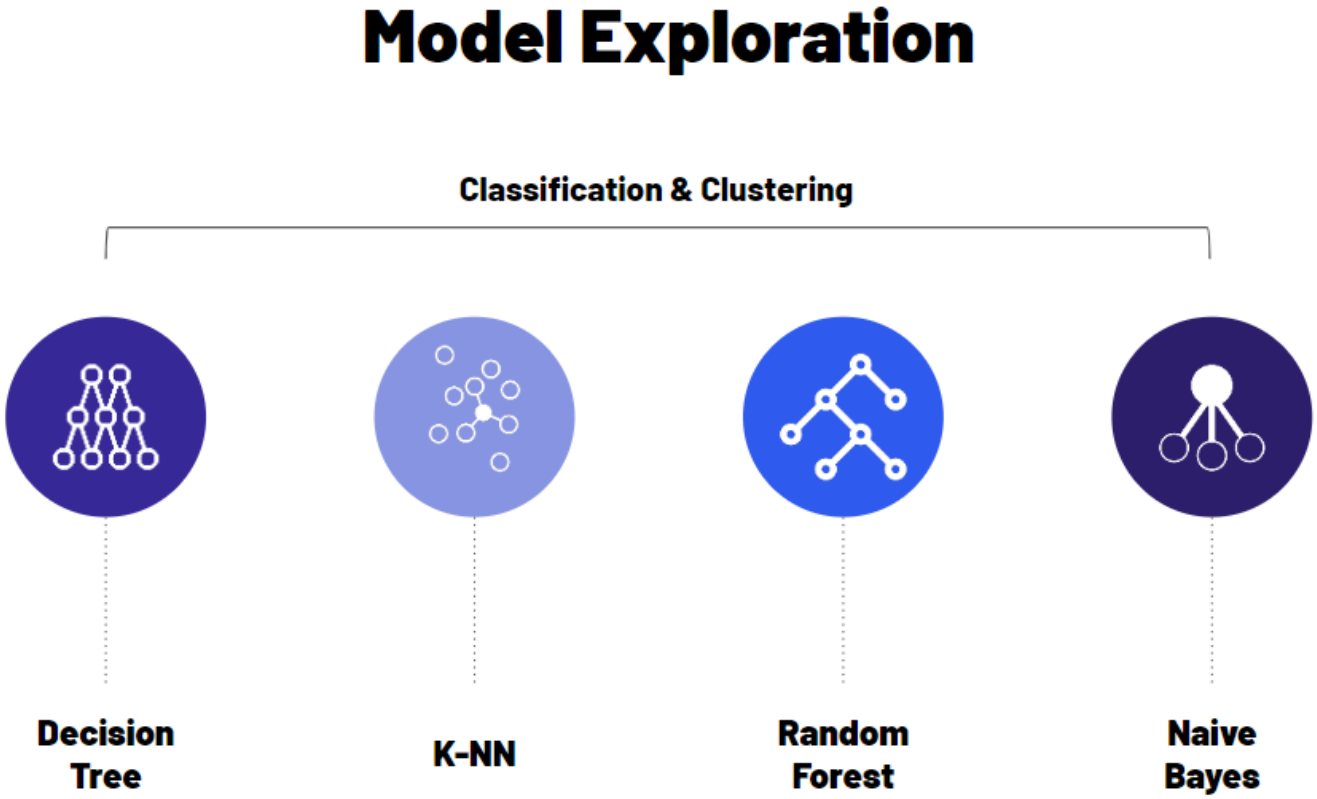
Handling missing values and categorical variables was necessary to ensure that the data was complete and consistent. We used appropriate techniques such as mean imputation, mode imputation, and one-hot encoding to handle these variables.

Standardizing data was also performed to bring all the variables to the same scale. This is necessary when using clustering algorithms, as they can be sensitive to differences in scale between variables.

Finally, *correcting typos* was necessary to ensure the accuracy of the data. Any typos or errors in the data were manually corrected to avoid any potential inaccuracies in our analysis.

4.2 | Algorithms & Statistical Methods

Initially, we used JMP and regression analysis to test our data. However, we found that these methods were not appropriate for our research question, which was to predict whether a manager would be sacked or not. Regression analysis is more suitable for predicting continuous numerical outputs, such as the exact timing of when a manager will be sacked.



Therefore, we decided to use machine learning classification and clustering algorithms to predict whether a manager would be sacked or not. These algorithms are designed to handle categorical or binary data and are well-suited for classification tasks.

For classification, we used algorithms such as Naive Bayes, Decision Tree, Random Forest, and K-Nearest Neighbors (K-NN). These algorithms are able to classify data into different categories based on patterns in the data.

For clustering, we used K-Means clustering, which is a common unsupervised learning algorithm that groups similar data points together based on their similarity. This allowed us to group managers based on their performance metrics and identify any patterns or trends that may be associated with being sacked.

4.3 | Limitations or Assumptions of our approach

One of the limitations of our approach is that it is based on certain assumptions about the data. We only considered data that we believed was most suitable for predictions, but there may be other attributes that could have been included in our dataset. For instance, we did not include sentiments in our analysis, but this attribute could be influential in determining whether a manager will be sacked or not.

Another limitation is the size of our dataset. During the initial testing phase, we only considered data from the last 3 seasons, but we found that the results were overfit. Therefore, we expanded our analysis to include the last 6 seasons, which covers a total of 77 managers. However, even with this expanded dataset, it may still be considered small compared to other studies in the field.

5 | Results

The Results section presents the outcomes of the study in terms of classification and clustering accuracy. To provide a clear visual representation of the accuracy, a funnel chart was used to display the classification accuracy for the decision tree, random forest, KNN, and Naive Bayes models. Additionally, a clustered bar chart was used to illustrate the clustering accuracy of the K-means and hierarchical clustering algorithms. The following subsections provide detailed information on the accuracy of each model and algorithm.

Decision tree: A decision tree is a tree-like model of decisions and their possible consequences. The model is constructed by recursively splitting the data based on the most informative feature at each node. The accuracy of the decision tree model is 80.18%.

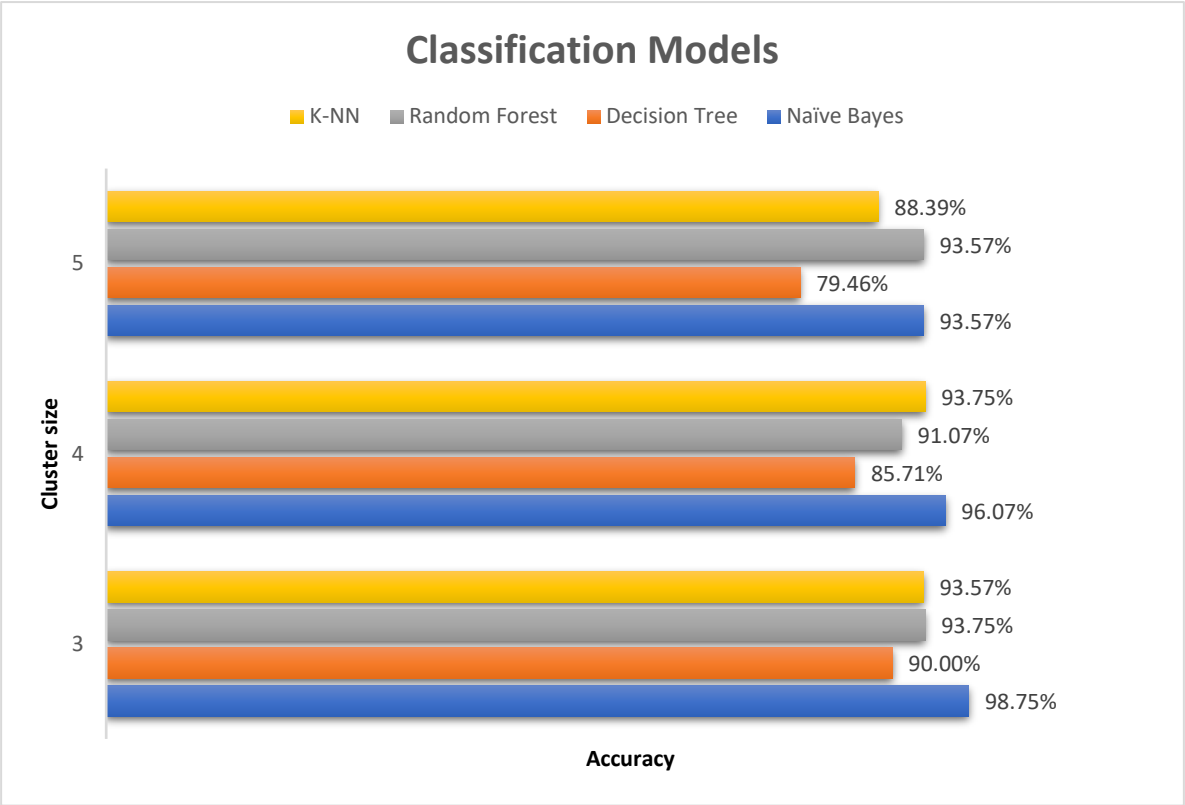
Random Forest: Random Forest is a type of ensemble learning method for classification, regression, and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. The accuracy of the random forest model is 90.89%.

KNN: The k-nearest neighbors algorithm is a non-parametric classification method that classifies a new data point based on the majority class among its k-nearest neighbors in the training data. The accuracy of the KNN model is 88.21%.

Naive Bayes: Naive Bayes is a probabilistic algorithm that uses Bayes' theorem to calculate the probability of a given data point belonging to a certain class. It assumes that all the features are conditionally independent of each other, given the class label. The accuracy of the Naive Bayes model is 92.32%.

Based on the accuracy values, we can see that the Naive Bayes model has the highest accuracy of all the models, followed by the random forest and KNN models. Therefore, we can select the Naive Bayes model for predicting whether or not a manager will be sacked in football.

Each of the four classification models (Random Forest, Decision Tree, KNN, and Naive Bayes) was evaluated using clustering accuracy with different values of k.



Random Forest had the highest accuracy for all values of k, ranging from 91.07% to 93.75%. Decision Tree had a lower accuracy compared to Random Forest, ranging from 79.46% to 90.00%. KNN had accuracy ranging from 88.39% to 93.75%, with the highest accuracy for k=3 and k=4. Naive Bayes had the highest accuracy among all models, ranging from 93.57% to 98.75%.

Based on the results, Naive Bayes seems to be the most suitable model for predicting whether or not a manager will be sacked in football, as it had the highest clustering accuracy for all values of k.

6 | Conclusion & Findings

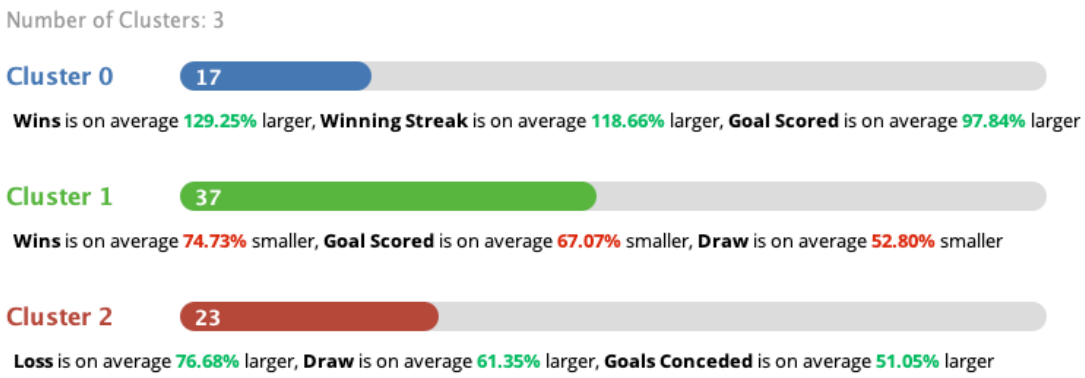
The correlation matrix shows the correlations between the predictor variables (League Position, Games Incharge, Winning Streak, Losing Streak, Wins, Loss, Draw, Goals Conceded, Goal Scored) and the response variable (Sacked). The values in the matrix range from -1 to 1, where -1 indicates a perfect negative correlation, 0 indicates no correlation, and 1 indicates a perfect positive correlation.

	League Position	*Games Incharge*	*Winning Streak*	*Losing Streak*	*Wins*	*Loss*	*Draw*	*Goals Conceded*	*Goal Scored*	*Sacked*
League Position	1									
Games Incharge	-0.59996072	1								
Winning Streak	-0.642543169	0.499398721	1							
Losing Streak	0.61600727	-0.148976679	-0.384783021	1						
Wins	-0.858645111	0.799960289	0.754228924	-0.472013345	1					
Loss	0.103066152	0.50106203	-0.083897604	0.391068171	0.176682864	1				
Draw	-0.336632107	0.686785283	0.192468096	-0.049271147	0.476924843	0.487441856	1			
Goals Conceded	-0.332464366	0.727061173	0.18597035	0.026889789	0.595944299	0.77832392	0.649942607	1		
Goal Scored	-0.762137129	0.788287176	0.687299898	-0.36306017	0.903330675	0.297985111	0.534547101	0.554143186	1	
Sacked	0.667208335	-0.767718269	-0.56224207	0.321003682	-0.81645458	-0.255484526	-0.587372329	-0.557469146	-0.755992661	1

Based on the matrix, it can be seen that League Position has a positive correlation of 0.67 with Sacked, indicating that higher league positions are associated with a higher likelihood of being sacked. In contrast, Games Incharge, Winning Streak, Losing Streak, Wins, Loss, Draw, Goals Conceded, and Goal Scored all

have negative correlations with Sacked, indicating that higher values in these predictor variables are associated with a lower likelihood of being sacked. Specifically, Wins, Games Incharge and Goal Scored have the strongest negative correlations with Sacked, at -0.82, -0.77 and -0.76, respectively.

Based on the clustering analysis, it was found that K=3 was the optimal number of clusters for the data. The Naive Bayes model was then applied using this clustering and achieved an accuracy of 98.75%. The data has been clustered into three clusters using K-means clustering with K=3. Each cluster represents a group of football managers who have similar characteristics based on the features used in the clustering analysis.



Cluster 0 represents managers who have a higher number of wins, longer winning streaks, and more goals scored compared to the other two clusters. This cluster contains 17 managers.

Cluster 1 represents managers who have a lower number of wins, fewer goals scored, and fewer draws compared to the other two clusters. This cluster contains 37 managers.

Cluster 2 represents managers who have a higher number of losses, more goals conceded, and more draws compared to the other two clusters. This cluster contains 23 managers.

Based on the whole analysis, iIt seems that both the K=3 Naive Bayes clustering and Naive Bayes classification approaches have their strengths and weaknesses.

The K=3 Naive Bayes clustering approach can help identify different groups of managers based on their performance indicators, which can provide valuable insights into why some managers are more likely to be sacked than others. On the other hand, Naive Bayes classification approach can predict whether a manager will be sacked or not based on their performance indicators, which is the main objective of our research.

What's missing from the model?

Apart from the data set that we've included, sentiments can also influence in various ways. Positive sentiments towards a manager can increase job security, while negative sentiments can lead to increased scrutiny and pressure from the club's owners or fans. Sentiments can also influence the team's performance, and if the team is underperforming, negative sentiments towards the manager may increase, resulting in a higher likelihood of being sacked. Moreover, sentiments can affect the team's morale and motivation, which in turn can impact their performance and ultimately the manager's job security.

Sentiments expressed in news headlines or statements made by club officials could potentially influence the decision to sack a manager. For example, the headline "Antonio Conte SACKED by Tottenham" suggests a negative sentiment towards the manager, while the statement "Jurgen Klopp was not sacked because he had given successful results as a manager and the board trusted him" suggests a positive sentiment towards the manager. In addition to factors such as league position and team performance, it is important to consider the potential influence of sentiments expressed in news and media coverage when analyzing the likelihood of a manager being sacked.

7 | References

WHEN SHOULD YOU SACK THE MANAGER? by Chris Hope.