



UTILIZING MACHINE
LEARNING TECHNIQUES
IN FOOTBALL
PREDICTION

Ryan Duarte – A00247351



AUGUST 20, 2022


DECLARATION

I hereby certify that the material, which is submitted in this thesis towards the award of MSc. in Data Analytics, is entirely my own work and has not been submitted for any academic assessment other than the part fulfilment of the above-named award.

Future student may use the material contained in this thesis, provided that the source is acknowledged in full.

Student ID Number: A00247351

Name of Candidate: Ryan Duarte

Signature of Candidate: 

Date: 20th August 2022

ACKNOWLEDGEMENTS

Firstly, I would like to thank my supervisor, Kevin Gaylard, for all his advice and guidance throughout this research project. A thank you must also be extended to all my lecturers throughout my time in TUS who have helped and encouraged me to get to this stage in my academic career.

On a more personal note, a thank you to my partner, Vanessa, for being so tolerant and understanding during this time and encouraged me from the beginning. Thank you to my mother, without whom I simply would not have been able to accomplish all that I have.

Finally, to Eoin, Chris, Michael and Rachael, thank you. Your friendship and support helped make this possible.

In memory of Martin McKeon, a beloved grandfather. You are missed by all of us everyday

Ad Meliora.

Match

An event where two teams compete to win in a game of football.

Home Win

A result of a match whereby the team playing in their home stadium scored more goals than the opponent

and won.

Away Win

A result of a match whereby the team that has travelled to the opposition's stadium, scored more goals than

the opponent and won.

Draw

A result of a match whereby the two teams scored the same of goals, and the result is a stalemate.

ML/Machine Learning

A form of Artificial Intelligence employed to learn from a data set to make predictions.

Models

Refers to the methods, or types, of machine learning employed throughout the study.

kNN

k-Nearest Neighbour, a supervised machine learning model.

DT

Decision Tree, a supervised machine learning model.

RF

Random Forest, a supervised machine learning model.

SVM

Support Vector Machine, a supervised machine learning model.

NN

Neural Network, a supervised machine learning model.

NB

Naïve Bayes, a supervised machine learning model.

xGB

xGBoost, a supervised machine learning model.

MLR

Multinomial Logistic Regression, a supervised machine learning model.

Target Variable

Refers to the FTR column in all the data sets.

FA

Football Association; English football's governing body.

EFL

English Football League; Comprised of all the professional football divisions in England.

UEFA

Union of European Football Associations; The organisation that governs all the European football leagues.

FIFA

Federation Internationale de Football Association; The organisation above UEFA that is the governing body of all association football, beach football and futsal. However, in this study FIFA typically refers to the football simulator of the same name which provided the ratings for all football players.

Barclays Premier League/Premier League

The top division of the English Football League, the focus of the study.

La Liga

The top division of the Spanish Football League.

Season

Refers to the thirty-eight-game domestic calendar where teams compete for points to finish first and become champions.

GK/Goalkeeper

The one player on a team allowed to use their hands. Their main aim is to prevent the opposition from scoring.

Def/Defenders

The defensive minded players that focus on preventing the opposition from scoring.

Mid/Midfielders

The players in the middle of the pitch that must contribute to both scoring goals and preventing goals from being scored against them.

Att/Attackers

The attack minded players that focus on scoring as many goals as possible.

Subs/Substitutions

Players that are currently not playing but may be introduced should those currently playing get fatigued, injured or the manager wishes to change tactics.

Ratings

A numeric value attributed to every player that is based upon six core facets of football, which themselves are calculated from twenty-nine individual data points.

2019/20

Refers to a specific season; the season started in the Autumn before the partition (2019) and finished in the summer after the partition (2020). The same applies for 2020/21 and 2021/22.

TITLE

Utilizing Machine Learning Techniques in Football Prediction – A00247351

TABLE OF CONTENTS

Declaration.....	1
Acknowledgements.....	2
Glossary of Terms, Abbreviations and Acronyms.....	3
Title.....	7
Table of Figures	13
Table of Tables	15
Abstract.....	16
Chapter 1 – Introduction	18
Background	18
Rationale of the Study, Research Aim and Objectives	21
Rationale of Study	21
Research Question	22
Research Objectives	22
Focus and Limitation.....	23
Chapter 2 – Literature Review	25
Introduction	25
Data Analytics.....	25
Models.....	31
Multinomial Logistic Regression	32

k-Nearest Neighbours – kNN	32
Neural Networks.....	33
Naïve Bayes.....	34
Decision Trees	35
Support Vector Machines	37
Random Forests	39
xGBoost.....	40
Summary	41
Chapter 3 – Methodology	42
Introduction.....	42
Research Questions and Objections Revisited.....	43
Research Philosophy	44
Positivist Research	45
Data Model.....	46
Business Understanding	47
Data Understanding.....	48
Base Football Statistics.....	48
Updated Statistics	50
Data Preparation.....	55
Base Football Statistics – Data Cleansing	55
Updated Statistics – Data Cleansing.....	60

Modelling	61
Model 1.1: kNN – Base Football Statistics	62
Model 1.2: kNN – Updated Football Statistics.....	63
Model 2.1: Decision Tree – Base Football Statistics.....	63
Model 2.2: Decision Tree – Updated Football Statistics.....	63
Model 3.1: Random Forest – Base Football Statistics.....	64
Model 3.2: Random Forest – Updated Football Statistics.....	64
Model 4.1: Support Vector Machine – Base Football Statistics.....	65
Model 4.2: Support Vector Machine – Updated Football Statistics.....	65
Model 5.1: Neural Network – Base Football Statistics	65
Model 5.2: Neural Network – Updated Football Statistics	66
Model 6.1: Naïve Bayes – Base Football Statistics.....	66
Model 6.2: Naïve Bayes – Updated Football Statistics	66
Model 7.1: xGBoost – Base Football Statistics.....	66
Model 7.2: xGBoost – Updated Football Statistics	67
Model 8.1: Multinomial Logistic Regression – Base Football Statistics	67
Model 8.2: Multinomial Logistic Regression – Updated Football Statistics	67
Evaluation.....	68
Deployment.....	69
Summary	69
Chapter 4 – Analysis of Findings.....	72

Introduction	72
Data Understanding.....	72
Base Football Statistics Structure	72
Updated Football Statistics Structure	73
Null Values	74
Outliers	75
Set-Pieces.....	76
Defence and Stability Over Offence.....	76
Ratio of Results	77
Accuracy Over Quantity	78
Average Player Quality in the Premier League	79
Data Preparation.....	80
Training and Testing.....	80
Normalisation	81
Conversions	82
Modelling	82
Model 1.1: kNN – Base Football Statistics	83
Model 1.2: kNN – Updated Football Statistics.....	84
Model 2.1: Decision tree – Base Football Statistics	84
Model 2.2: Decision Tree – Updated Football Statistics.....	86
Model 3.1: Random forest – Base Football Statistics	86

Model 3.2: Random forest – Updated Football Statistics.....	87
Model 4.1: Support Vector machine – Base Football Statistics.....	88
Model 4.2: Support Vector Machine – Updated Football Statistics.....	89
Model 5.1: Neural Network – Base Football Statistics	89
Model 5.2: Neural Network – Updated Football Statistics	89
Model 6.1: naïve Bayes – Base Football Statistics.....	89
Model 6.2: naïve Bayes – Updated Football Statistics.....	90
Model 7.1: xGBoost – Base Football Statistics.....	91
Model 7.2: xGBoost – Updated Football Statistics	91
Model 8.1: Multinomial Logistic Regression – Base Football Statistics	91
Model 8.2: Multinomial Logistic Regression – Updated Football Statistics	92
Final ACCURACY Table.....	92
Evaluation.....	93
Summary	97
Chapter 5 – Discussion and Findings.....	99
Introduction	99
Resarch Questions and Objectives Revisited II	99
Data Understanding.....	100
Data Preparation.....	103
Test Statistics – Cleansing.....	103
Modelling	105

Evaluation.....	106
Summary	114
Chapter 6 – Conclusion and Reflection	117
Research Questions and Objectives Revisited – Final	117
Contribution to the Field	120
Limitations	121
Reflection and Future Work.....	124
References.....	126
Appendices.....	132
Model 1.1: kNN – Base Football Statistics	132
Model 1.2: kNN – Updated Football Statistics.....	133
Model 2.1: Decision tree – Base Football Statistics	134
Model 2.1: Decision Tree – updated football statistics	135
Model 3.1: Random forest – Base Football Statistics	136
Model 3.2: Random forest – Updated football statistics	137
Model 4.1: Support Vector machine – Base Football Statistics	138
Model 4.2: Support Vector Machine – Updated Football Statistics.....	139
Model 5.1: Neural Network – Base Football Statistics	140
Model 5.2: Neural Network – Updated Football Statistics	141
Model 6.1: naïve Bayes – Base Football Statistics.....	142
Model 6.2: naïve Bayes – Updated Football Statistics.....	143

Model 7.1: xGBoost – Base Football Statistics	144
Model 7.2: xGBoost – Updated Football Statistics	145
Model 8.1: Multinomial Logistic Regression – Base Football Statistics	146
Model 8.2: Multinomial Logistic Regression – Updated Football Statistics	147
Final Prediction Model: kNN – Test Football Statistics	148

TABLE OF FIGURES

Figure 1: Published findings of Reep and Benjamin [21].....	26
Figure 2: Visual Representation of Neural Network [38].....	34
Figure 3: An example of a DT created for this dissertation.....	36
Figure 4: An example of how SVM choose the best bisection line [39]	37
Figure 5: Further example of how SVM choose the best bisection line [39]	38
Figure 6: Example of how RF come to their prediction [41].....	39
Figure 7: An example of how RF splits features and makes decisions, known as bagging [41]	40
Figure 8: Visual representation of the CRISP-DM model [47]	47
Figure 9: An example of the data source used to build the updated data set [50]	54
Figure 10: Outliers found in data sets using box plots.....	56
Figure 11: Full list of features in base statistics data set.....	59
Figure 12: Full list of features in updated statistics data set	60
Figure 13: Depiction of values in a data set would be normalized [53]	62
Figure 14: Image of a confusion matrix utilizing the Iris data set [54]	68

Figure 15: TP = True Positives, TN = True Negatives, FP = False Positives and FN = False Negatives.....	69
Figure 16: Structure of Base Statistics data set.....	73
Figure 17: Structure of Updated Statistics data set.....	74
Figure 18: Code to check for NA/Null/Missing values	74
Figure 19: Box plot of outliers for Full-Time Home Goals.....	75
Figure 20: Highlighting importance of set-pieces	76
Figure 21: Highlighting how important the defence and midfield is compared to attack	77
Figure 22: Break down of the results of the Premier League over two seasons.....	78
Figure 23: Correlation between shots and shots on target and winning	78
Figure 24: Breakdown of average quality of player in defence, midfield and attack.....	79
Figure 25: Data before normalization	81
Figure 26: Data after normalization.....	81
Figure 27: An example of the different forms a confusion matrix can take	82
Figure 28: Example of a DT using the “fancyRpartPlot” function.....	84
Figure 29: The logic and splitting operations of the more successful DT model for the base statistics data set.....	85
Figure 30: The logic and splitting operations of the more successful DT model for the updated statistics data set.....	86
Figure 31: Top ten most important variables to determine the result, as found by the RF model for the base statistics data set.....	87

Figure 32: Top ten most important variables to determine the result, as found by the RF model for the updated statistics data set	88
Figure 33: Most important features when deciding each individual result as per the NB model utilizing the base statistics data set	90
Figure 34: Most important features when deciding each individual result as per the NB model utilizing the base statistics data set	90
Figure 35: Most important features when deciding each individual result as per the NB model utilizing the updated statistics data set	91
Figure 36: Full list of features of test statistics data set	103
Figure 37: The results of each game according to the trained kNN model	106
Figure 38: Four Manchester City games including unexpected Crystal Palace result	119

TABLE OF TABLES

Table 1: Breakdown of size of data sets before and after splitting	80
Table 2: Accuracy ratings of all models for both the base statistics data set and updated statistics data set.....	92
Table 3: Predicted points in alphabetical order.....	107
Table 4: Final table predicted by kNN sorted by points	108
Table 5: Predicted league table with position differential from real table.....	110
Table 6: Final predicted points tally and points difference from the real total.....	113
Table 7: Objectives success/failure rate.....	118

ABSTRACT

From debates between football analysts, to opposing team's fans, predicting the winner between two football teams has always been intrinsically tied to the sport. In more recent times, predicting the winner between two teams, and by extension why this team won, has taken on added importance. From football teams looking for ways to gain a competitive advantage, to attain championships and increased revenue, to fans looking to back their team and see a return in the form of betting, accurately predicting the winner between two teams takes on increased importance.

With the expanded availability of machine learning techniques, it is now possible to build multiple models that can learn and interpret a data set to provide a prediction as to who will win between any two football teams. The models in this paper will be provided with base football statistics, those statistics that are gathered after every match such as the number of shots a team has or how many fouls were committed, and additional psychological and non-psychological factors. This is due to a football match being determined by more than just base statistics, with a team's mentality and ability to deal with external factors a key part of the modern game.

This paper aims to not only provide an accurate prediction for which team would win in any given match, but also provide some answer as to why they won. Showing what variables and features most determine why one team is selected to win over another, providing some explanation and logic behind the predictions. Utilizing seminal works in the field, such as Razali et al. (2017) and Gangal et al. (2015), to be more informed on which models have been used previously and how they performed, this paper seeks to build upon all that came before.

CRISP-DM was the methodology used to keep the research structured and focused, while a positivist research approach was utilized to ensure that only unbiased quantitative data

was used, data that is entirely built upon facts and figures. This quantitative data set was compiled and curated by the author utilizing two seasons of Premier League football data and supplied to the eight models that were selected to allow them to learn. This learning was then applied to one final data set to assess the predictive power it has gained in learning.

The author was successfully able to predict the results of 72.37% matches across the three-hundred and eighty game Premier League season, comparable to the literature. The model that performed best returned 85% accuracy when trained on nothing but base statistics, and 75% accuracy when trained on the additional factors that were included. This resulted in a predicted final league table that closely resembles the real final league table, with most discrepancies between the two able to be understood and explained.

CHAPTER 1 – INTRODUCTION

This chapter will enlighten the reader on what the author will accomplish by undertaking this study, and by what means this end goal will be attained. This chapter will outline the underlying problems in relation to the field of study, which is performing accurate predictions of football matches to assess winners, losers the possibility of draws and final league position.

BACKGROUND

Football, as it is almost universally known, or soccer in some regions such as Northern America or Ireland, can trace the modern lineage of the sport to England in 1863 with the birth of the FA or Football Association, English footballs governing body, [1]. In 1888 the English Football League (EFL) was created. Four years later a second division was added as more teams joined the EFL, creating the basis of the modern football pyramid in England. The pinnacle of this pyramid is the Barclays Premier League.

Today, football is the most popular sport in the world, with an estimated 3.5 billion fans globally as per Sourav Das, [2]. They arrived at this figure based on a wide variety of criteria such as global fan base and audience, viewership on television, television rights deals, popularity on social media and number of professional leagues in the world. Taking so many things into account, football is the most popular sport in the world by some margin. The popularity of football across the globe at this point in time even transcends gender in the sport. With the women's European final at Wembley arena, in which England beat Germany to become champions, registering a record attendance for a UEFA tournament, with a total attendance of 87,192 people, [3]. This is coupled with the fact that the three highest attended matches this year were all women's football matches. With the European final attendance figure only being surpassed by the attendance figures at the Barcelona Feminí versus Real Madrid Femenino match, 91,553 in attendance, and the Barcelona Feminí versus VfL Wolfsburg Women match, 91,648 in attendance.

Within this popularity and viewership, “the Premier League is the most watched football league in the world,” [4], and “draws the highest global television audience of any football league and has the most live coverage of all European leagues.” It is easy to see that the Premier League is the most watched league in the world as any given game day could have upwards of 14 million viewers, [5], which, over the thirty-eight game weeks in a season, could contribute to 532 million viewers. On average, the 2018 World Cup commanded 517 million views, [6].

As the popularity of the sport rises, the field of sports analytics has also seen comparable growth. Professional teams use analytics to gain an advantage over an opponent, pundits use sports analytics to form their predictions for a match and betting companies utilize analytic methods to stack the odds in their favour. There are many methods used in the field of sports analytics to estimate the most likely winner between two teams. These range from the first forays into prediction using Poisson distribution, to more modern techniques utilizing machine learning algorithms such as Bayesian Networks.

The current online gambling industry is valued at approximately \$59 billion, with a predicted growth of 57% in the next year, [7]. Gambling and football are so intrinsically linked that almost every facet of the game of football can be bet upon. As per the UK’s gambling commission, “sports betting remains the most popular gambling activity,” [8] and sports betting makes up “almost 35% of the money spent in online betting as well as in physical establishments.” Football makes up 47% of all sports betting in the United Kingdom, the next closest figure is horse racing which accounts for 27.3%, [8]. Horse racing used to be the market leader in sports betting but has since been overtaken by football due to the abundance of in-play betting and early cash-out options, [9]. Coupled with the plethora of betting options such as number of yellow or red cards shown, number of corners, who will be winning at half-time and full-time, handicap betting and building accumulators, football gambling has never been

more attractive a proposition. With such an increase in value and personal loss, more reliable methods to accurately predict the victor between any two teams is of increased importance. This is only heightened when gambling companies such as Paddy Power saw an increase in their share price of greater than 8% after they declared “pre-tax profits had trebled to £72 million sterling in the first six months of 2021,” [10]. Ladbrokes saw an increase of 31% in net gaming revenue in 2022, [11], while Bet365 saw profit before tax rise from £137 million in 2020 to £470 million in 2021, [12].

With such an increase in profits seen by bookmakers, and with increased marketing around easier ways to gamble, especially in relation to football, the question arises; is it possible to build a model that is statistically accurate enough to be able to bet relatively safely? This is the area the author has identified and will address over the course of this study. Can a model be produced with application potential that can aid in gambling decisions to help mitigate losses? Additionally, if the model does turn out to be comparable, or better, than the models used in other studies, stepwise regression or features contained within some of the R Studio packages could be used to better understand which metrics are the most important when it comes to winning. If a model such as this can be produced, certain ethical boundaries must also be acknowledged as the aim of this dissertation is not to enable those who gamble or those who have gambling addictions, but simply to aid in providing a more informed decision.

The impact this dissertation will have, is that it may help provide a basis for the accurate prediction of football match outcomes. Additionally, the work may help teams to understand the variables they should aim to have higher values in to provide the best chance of winning. If the model is sufficiently accurate, emulating or surpassing previous attempts to predict results, there is application potential as an aid for making more educated bets. This can reduce personal losses and hopefully ease the strain gambling can have on certain individuals and their families. Issues such as these are more pertinent today than ever as gambling becomes more

readily accessible due to the implementation of easily downloaded applications and a higher presence of gambling advertisements on television and on sports teams' jerseys. Today in the United Kingdom there are over 400,000 gambling addicts with a further two million in danger of developing an addiction, [13]. The work done throughout this dissertation will also be able to identify what metrics are the most important in the sport today. Areas will be identified that can provide a team a platform from which they can build a tactical game plan to beat another, informing how teams should set up and play today. This model, if successful, could be tweaked and altered to be adapted for any other sport, encompassing each sports unique set of features and variables.

In summary, the aim of this dissertation is to produce a prediction model that can successfully predict the result between two teams in a Premier League football match. Several predictive analytical methods will be used on two seasons worth of football data, to ascertain which is the most successful method for predicting the outcome of any given match. If the models are sufficiently accurate, being comparable to models created in other studies, then it could potentially in the future be utilized as an aid for gambling, for studying trends in football or any given sport, assessing what statistics are the best for a team to perform well in and as a method of opposition analysis.

RATIONALE OF THE STUDY, RESEARCH AIM AND OBJECTIVES

RATIONALE OF STUDY

As mentioned previously in the Background section, the gambling industry is seeing growth year-on-year, seeing profits rise into the hundreds of millions. The increased marketing, ease of access and plethora of gambling options, from before any match starts to in-game options, have seen those at risk of forming gambling addictions and unique gamblers rise. To help alleviate losses, stress and the strain on relationships that comes with gambling, [14], this paper

serves to provide a platform that can provide a more informed decision when it comes to predicting the result of any game of football. While the final model produced will be able to inform the user on more than the result of a match, such as the variables of a match that most strongly correlate to a win, aiding those who need help and guidance is the primary concern.

RESEARCH QUESTION

Can a machine learning model be produced that can accurately gauge the winner between two Premier League teams?

RESEARCH OBJECTIVES

1. Create a machine learning model that can predict the result of any given Premier League game utilizing standard in game statistics

There are a number of classification-based machine learning models, most of which are described in Chapter 2 the Literature Review. The first objective is to create a machine learning model that can predict the result of a match between two specified teams.

2. Create a machine learning model that can predict the result of any given Premier League game utilizing standard in game statistics and additional statistics and psychological factors

Building upon the first objective, Objective 2 is concerned with fine tuning and upgrading the models created before to assess whether they now perform better or worse given the new information. In theory, more information can aid a model in making the correct decision. However, increasing the number of features in a data set could also confuse and complicate the model to the point where it results in reduced accuracy.

3. Assess whether the model which performs best in Objective 2 can be used to predict the results of the most recent Premier League season.

Once the best model has been found, the model that is most accurate with the expanded data set, it must be tested to establish whether it can accurately predict the winners between two teams in the most recent Premier League season.

FOCUS AND LIMITATION

As has been discussed up to this point, the Premier League is the most watched, broadcasted and supported league in the world. As such, while the research will be conducted across multiple regions and territories, the data that will be used will focus in its entirety on the Premier League. Limiting the research to one competition, the Premier League, keeps all base statistics homogenous. This is due to the disparate quality levels between domestic competitions. Premier League teams, or higher ranked teams, tend to submit weakened sides in the domestic cup competitions, reserving their typical first-team players for the more important league games.

While domestic cup glory is an alluring prospect, the financial prospects of doing better in the league take precedence. For winning the FA Cup a team takes home £4.5 million, [15], for winning the Carabao Cup they can expect to be rewarded £100,000, [16], while finishing in 17th in the Premier League and just about avoiding relegation in to the Championship a team can earn around £8.8 million, [17]. As such, the lower ranked teams prioritise staying in the Premier League as opposed to winning titles and thus the quality of their play is diminished. The same can be said of the more successful Premier League teams as they aim to win the league and therefore send out youth teams to preserve fitness and reduce minutes for senior players. This is evident in the Aston Villa versus Liverpool FA Cup game in January 2021 as Aston Villa sent out their youngest ever squad with an average age of 18 years and 294 days,

[18]. Therefore, due to the varying quality of teams, players used, age and quality of play, the Premier League will remain the sole focus.

Additionally, Covid-19 brought about a unique season. Due to the global pandemic, football was changed to represent the new rules and regulations. This included a larger substitutes' bench, coupled with additional substitutes allowed during a game. This provides the bigger teams with greater player depth even more of a chance to impact the game from the bench. Additionally, and more importantly, the pandemic brought with it a stadium ban for fans. This removed the crowd element of the sport and removed the importance a home game can have, as without fans there is no atmosphere in support of the home team or in opposition of the away team.

Finally, in a similar manner to Razali et al., discussed in detail in Chapter 2, this model will incorporate 3 seasons worth of data. This is comprised of two seasons of data to learn from, the 2019/20 and 2020/21 Premier League seasons, and one season to test the predictive capabilities of the model, the most recent 2021/22 season. This encompasses one complete and uninterrupted season, 2019/20, with an additional season interrupted by Covid-19, the new substitute rules and a lack of fans, 2020/21, to learn from. With the predictions being performed on another uninterrupted regulation season, the 2021/22 season.

CHAPTER 2 – LITERATURE REVIEW

This chapter provides references to key seminal studies and documents that are related to the subject being researched. The findings from these seminal papers will provide the framework for modelling and analysis that will be carried out in this paper.

INTRODUCTION

“The best way to predict the future, is to create it.” This quote is attributed to Peter Drucker, (1909-2005), one of the most widely known and influential voices in modern business management, [19]. The author agrees with Peter Drucker’s quote, and it is this ideology that formed the basis of this research.

The literature review chapter of this dissertation aims to provide an in-depth review of a wide range of literature and workings that are related, and as current as possible, to the topics and methods discussed and utilised throughout. Firstly, the literature review will aim to focus on predictive analytics in relation to football. Secondly, it will focus on the methods being used, these being: kNN, Naïve Bayes, Multinomial Logistic Regression, Decision Trees, Neural Networks, Support Vector Machines, Random Forests and xGBoost. Finally, all studies and texts researched in this dissertation are focused on of the end goal of creating a machine learning model that can predict the result of any given Premier League game utilizing standard in game statistics and additional psychological and non-psychological features.

DATA ANALYTICS

The study of football from an analytical standpoint, dates back as far as 1956 when M. J. Moroney used Poisson distribution to determine the average amount of goals scored per team per game [20]. While predicting scores, winners and even the number of goals was not the main aim of his book, it is nonetheless one of the earliest noted entries of analytics being applied to football. In his book “Facts from Figures” [20], Moroney noted that there are other mitigating

factors in relation to predicting the number of goals any team would score such as the quality of the opposition or the weather. Far from the global dominating sport it is today, analytics in relation to football have come a long way in the 66 years since Moroney originally conducted this research. Not totally discredited in the field, Moroney’s study helped provide the basis for other studies to evolve and grow.

One of the next earliest entries in football analytics was “Skill and Chance in Association Football,” by Reep and Benjamin [21]. In their research they primarily focused on the passing aspect of football. They note that teams that began passing in one of the opposition quarters, the oppositions half of the pitch when the pitch is broken into quadrants, more often led to shots than passes that began in their own quadrants. They ascertained that “30% of regained possessions led to shots at goal,” and most notably that, “of all goals scored against them 50% come from such failures to move the ball into the defender’s half,” [21].

TABLE 4

Season	Total matches	Detail	Ratio of shooting area origin goals to all goals	Average No. of shots to score 1 goal	Ratio of shots from regained possessions to all such moves	Ratio of shots from shooting area origin attacks to all such attacks	Ratio of all shots to all attacks reaching the shooting area	Ratio of goals “against” conceded by “own half” breakdowns to all goals against	Ratio of shooting area origin attacks to all attacks reaching the shooting area	Ratio of regained possessions in shooting area to all shooting area origin attacks
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
1952-53	26	Miscellaneous	.624	10.1				.516		
1953-54	51	Miscellaneous	.528	9.4				.503		
1954	8	World Cup	.308	9.1	.314	.231	.170	.359	.245	48.0
1954-55	43	Miscellaneous	.494	9.0				.466		
1954-55	14	Sheffield W.	.510	9.8	.296	.223	.125	.449	.301	52.5
1955-56	15	Miscellaneous	.397	9.8	.319	.250	.157	.349	.296	54.3
1955-56	42	Sheffield W.	.558	9.6	.327	.244	.137	.515	.287	52.2
1956-57	18	Miscellaneous	.533	10.8	.266	.211	.133	.567	.329	52.8
1956-57	42	Sheffield W.	.535	9.5	.299	.230	.140	.500	.299	51.6
1957-58	42	Sheffield W.	.547	9.0	.282	.218	.128	.472	.313	52.3
1957-58	15	Miscellaneous	.546	8.0	.293	.228	.153	.468	.299	54.7
1958	11	World Cup	.525	9.7	.297	.225	.141	.525	.313	53.1
1958-59	36	Miscellaneous	.538	7.9	.267	.206	.133	.506	.292	51.3
1959-60	41	Miscellaneous	.544	9.4	.272	.217	.138	.513	.310	52.0
1960-61	38	Miscellaneous	.556	8.9	.291	.222	.149	.469	.305	52.8
1961-62	42	Miscellaneous	.546	10.6	.265	.208	.141	.511	.299	52.7
1962	18	World Cup	.520					.340		
1962-63	42	Miscellaneous	.551	10.4	.275	.216	.143	.497	.299	51.8
1963-64	26	Miscellaneous	.536	8.8	.275	.216	.146	.409	.299	53.6
1964-65	17	Miscellaneous	.480	11.8	.257	.213	.146	.540	.265	48.6
1965-66	50	Miscellaneous	.575	10.0						
1966	11	World Cup	.600	16.2	.246	.218	.152	.560	.311	50.9
1966-67	18	Miscellaneous	.426	10.4	.269	.221	.153	.475	.293	52.1

Note: Blank cells in this table indicate that the necessary records were not compiled for the particular set of matches.

Figure 1: Published findings of Reep and Benjamin [21]

Reep and Benjamin made some noteworthy assumptions, however, these assumptions cannot be applied to modern day football analytics. Assumptions such as an individual player's quality does not vary by that much compared to other players in the top tier of football. While this may have been the case, at least to a certain extent in the late 50s and early 60s when this research was initially undertaken, at a time when most teams in the First Division of English football were of a similar calibre, that is not the case today. In the modern game, one need only look at how the Premier League is now comprised of "The Big Six" who annually battle for championships and prestige, while the rest of the league fight to be crowned "The Best of the Rest" and finish in seventh. This is largely due to the amount of money these bigger teams possess, as with these extensive resources comes the ability to construct state of the art facilities and even more crucially, purchase a higher calibre of player.

While Reep and Benjamin may not have had access to FIFA statistics that rate players on every facet of the game, the quality of any player is a huge factor when it comes to determining the accuracy and quality of a pass, or adversely the chance of intercepting a pass or breaking up play. Papers such as "An Improved Prediction System for Football a Match Result" [22], or "Predicting match outcomes in association football using team ratings and player ratings" [23], utilise player ratings as an integral metric when it comes to prediction, to omit such factors in the research undertaken by Reep and Benjamin seems like an oversight. This study does intend to incorporate the quality of a player when predicting the result of a football match through the use of FIFA player ratings combined with the areas of the pitch in which they operate. It is only compounded and highlighted further when in more recent times other facets of football are utilising player quality as an indicator or in prediction, such as in xG (expected goals). This is most present in the paper "Creating a Model for Expected Goals in Football using Qualitative Player Information," by Pau Madrero Pardo [24], in which the author utilized FIFA player ratings to help build his Expected Goals models.

In more recent times, access to machine learning models has paved the way for more intricate methods of prediction, accounting for more and a wider range of parameters to be utilized. In the paper “Football Result Prediction with Bayesian Network in Spanish League-Barcelona Team” by Owrampur et al. [25], the authors attempt to predict the results of every Barcelona match throughout the 2008-2009 season utilizing a Bayesian Network. “Bayesian networks are a type of probabilistic graphical model that uses Bayesian inference for probability computations,” they depict “conditional dependence, and therefore causation,” [26].

The authors rightly noted that, “there are a large number of factors which could affect the outcome of a football match,” and introduced into their network more than base statistics, and due to the success outlined in their study, this will be done in this dissertation. Owrampur et al. (2013) rightly insert psychological factors into their network. These psychological factors include weather, whether it is a home game, and the current form over the previous five games, a common indicator for performance. In addition to these psychological factors, they include non-psychological factors, though only some of these statistics will be utilized in the model built for this dissertation. Owrampur et al. (2013), in lieu of using typical statistics such as shots on target or possession, use non-psychological statistics such as performance of all players, average number of home goals and how many main players are injured. In this study, the models being built will encompass, and develop, many of these factors. For example, in place of the number of home goals variable this research will use goal difference. This is a greater example of a team’s goal scoring abilities while removing the already accounted for home advantage which will be found due to the method used to clean the data, “one hot encoding,” which is discussed in greater detail in the Data Preparation section of this dissertation. Additionally, weather affects both teams equally and is more beneficial to research performed in Spain where the weather metric is most likely intended to document if it rained.

As rain is a far more common occurrence in the UK, weather is a less beneficial variable to derive knowledge in prediction.

The model used in the aforementioned paper by Owrampur et al. (2013) attempted to predict the result of every game of the thirty-eight game season in La Liga that Barcelona played, with the statistics and figures being updated for every game to reflect the previous game's results. With this approach, Owrampur et al. (2013) had a model that was correct 92% of the time. This is an incredibly high accuracy rating for any model; however, there is a caveat to this success. Barcelona of the 2008-2009 era were one of the most dominant teams in European football, under head coach Pep Guardiola, one of the most successful and innovative managers. In that 08/09 season Barcelona won not only the league (La Liga), but also the Champions League and the Copa Del Rey [27]. The season after, they retained the league title while also winning the FIFA Club World Cup, European Super Cup, and the Spanish Super Cup, [27]. Additionally, Pep Guardiola himself has won 35 major honours throughout his career [28]. The problem with hyper-focusing on a team as strong as this one, is that if the model always predicts that the team wins in every match it is more than likely to be correct. With the 92% success seen less likely to be replicated had they focused on Espanyol, the team that finished 10th that season and experienced a very balanced distribution of wins, draws and losses at 12, 11 and 15 respectively.

Due to the issues raised with the paper by Owrampur et al. (2013), another paper utilizing a Bayesian Network by Razali et al. (2017) in the paper entitled "Predicting Football Matches Results using Bayesian Networks for English Premier League (EPL)," [29], was studied. However, they utilized base statistics only, the metrics that are most commonly discussed when pundits analyse performance. Metrics such as team shots, shots on target, fouls committed and corners, for both home and away teams were included. All these metrics, and

more, are incorporated into the models utilized in this dissertation, as they are the most concrete statistics for each team that directly feed into the final result.

Razali et al. (2017) achieved an average of 75.09% accuracy across the 3 seasons worth of data they used. It is hoped that the inclusion of external, psychological, factors, in addition to other metrics that this accuracy can be increased. Additionally, while this is quite an accurate model, it lends credence to the issue raised previously, that by not focusing this model on a talented team the accuracy sees a drop off. It is also noteworthy that the data used in the models are different while ultimately attempting to perform the same action.

Finally, in the paper “Analysis and Prediction of Football Statistics using Data Mining Techniques,” by Gangal et al. (2015) [30], the authors utilized three machine learning techniques in an effort to predict match outcomes, and in turn, player performance. As opposed to previous uses of prediction methods in football, the ultimate aim of the authors was to predict the winner of a match so that they could assert which players were most likely to play well. This was done so that they would be able to make accurate transfers and therefore perform better in the Premier League’s Fantasy Football game and claim the rewards that came from winning this competition. Although their ultimate aim was different, as stated previously, they were still attempting to predict which team will win or lose using machine learning techniques. In the paper by Gangal et al. (2015) they utilized Bayesian Networks again, while also using Neural Network and Genetic Programming, two of which are utilized in this paper: Bayesian Networks and Neural Networks.

Neural Networks will be explained in more detail later in this segment, but Genetic Programming, according to a leading researcher in the field, Dr John Koza, is a “method for automatically creating computer programs,” [31]. Koza et al. stated that “it starts from a high-level statement of what needs to be done and uses the Darwinian principle of natural selection

to breed a population of improving programs over many generations.” In more simple terms, the model learns through each iteration it is run, how best to optimise and perform the task asked of it, in this case predicting the winner between two football teams.

The study from Gangal et al. (2015), [32], made one key choice when building their models. They built their models based upon a binary outcome that one team will win and one team will lose. This is not the case in football with the third option being that the two teams draw. Due to the fact that the result cannot be a binary choice, methods such as Logistic Regression are not applicable when predicting the result of a football match, as they require a binary decision to be made at the end, either a home win or an away win. As such, Logistic Regression, while a powerful classification tool, will not be utilized in this study, and has been replaced with Multinomial Logistic Regression. This will be discussed in depth in Chapter 2, but in short it provides a similarly powerful model as Logistic Regression but allows for more than one class to be identified which is ideal for this scenario.

MODELS

The aim of this dissertation is to attempt to find the best models for predicting the winner in any given football match. Due to this, multiple methods will be tried and tested. Every method will be tested using the base statistics, and then again with the updated data set that contains the psychological factors and extra metrics. All the following methods are classified as “Supervised Machine Learning” models, this means they all require the data sets being supplied to them for learning to be split into a training and testing data set. The typical split for this training and testing divide is an 80/20 split respectively. Author Brett Lantz utilizes an 80/20 split in his book “Machine Learning with R,” [33] (pg. 268) and in the paper entitled “An Efficient Data Partitioning to Improve Classification Performance While Keeping Parameters Interpretable” by Korjus et al. (2016), an 80/20 training and testing split is also used [34]. Due to this percentage being the typical partition, with regards to machine learning and data

splitting, this is the size of the splits that will be used going forward and will be referenced as “the split” from henceforth. The methods that will be used are as follows:

MULTINOMIAL LOGISTIC REGRESSION

Multinomial Logistic Regression (MLR) is an extension of Logistic Regression. As mentioned previously, Logistic Regression (LR) is a form of predictive analytics used when the dependent variable is dichotomous (binary). LR “is used to describe data and explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables,” [35].

MLR, similarly to LR, is used to predict a categorical dependent variable. As an extension of LR, it allows for more than two categories of dependent variable to be predicted. In the case of football prediction this is ideal as there are three ways a match can end; a home win, an away win, or a draw. Dr Jon Starkweather and Dr Amanda Kay Moske assert that, “like binary LR, MLR uses maximum likelihood estimation to evaluate the probability of a categorical membership,” [36]. They also state that it is an attractive analysis because, “it does not assume normality, linearity or homoscedasticity,” (pg. 1).

As stated previously, when reviewing the study undertaken by Gangal et al. (2015), [32], MLR will be used for this research as it is an upgrade to their solution of LR which can only produce binary results, a win or a loss. As football can result in three different results, a home win, an away win, or a draw classified as H, A and D respectively, MLR is the preferred method between the two classification models for this study.

K-NEAREST NEIGHBOURS – KNN

“In a single sentence, nearest neighbour classifiers are defined by their characteristic of classifying unlabelled examples by assigning them the class of similar labelled examples,”

[33]. In *Machine Learning with R*, the author likens the model to a human like ability to remember things that came before, to make assumptions about what is to come next. “Despite the simplicity of the idea, nearest neighbour methods are extremely powerful,” [33].

kNN methods have been used by companies like “Amazon or Netflix when recommending books to buy or movies to watch,” [37], whereby they see what types or forms of media a user consumes and suggest something that is similar that the user may also enjoy. The book, show or movie that they suggest will be located in a similar set, or contain similar features, to those that were looked at previously. If Netflix can see a user likes action films for example, they will more often suggest action films.

Due to the successful implementation of kNN at billion-dollar companies like Amazon, Netflix and Spotify, it makes sense to attempt to harness the power it possesses for this research. kNN attempts to find data points that contain similar features and groups them together. However, it performs in the original stage of testing with the base statistics, there is a chance that the additional features added to the data set help the model and improve its accuracy.

NEURAL NETWORKS

“Neural networks reflect the behaviour of the human brain, allowing computer programs to recognise patterns and solve common problems in the fields of AI, machine learning, and deep learning,” [38]. Neural Networks, also called Artificial Neural Networks or Simulated Neural Networks, (ANN) and (SNN) respectively, is a machine learning algorithm. A Neural Network (NN) is comprised of input and output layers, separated by one or more layers in between. Each input parameter is connected to the next layer, with every input from that layer connected to the next layer and so on, until the model arrives at the final output, pictured in Figure 2.

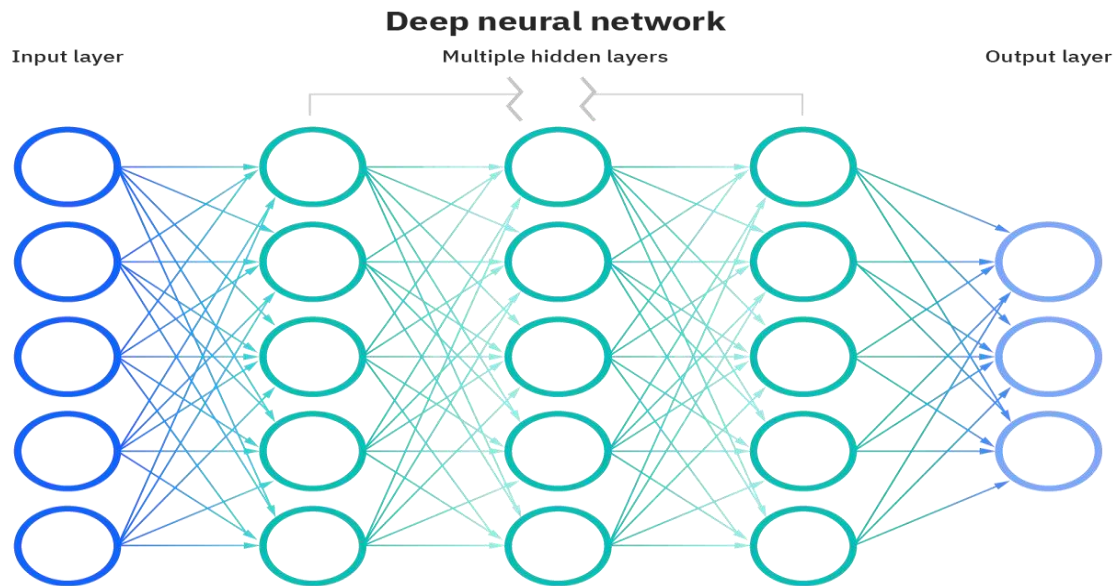


Figure 2: Visual Representation of Neural Network [38]

For this dissertation, the base statistics complete with psychological factors, are fed into the input layer. Each further layer assigns a weight or value that the input must achieve to activate the node on that layer. If this is achieved then the node permits the data to go through to the next layer, and this continues until the final output nodes ascertain whether the data being provided is most likely to end in a home win, an away win or a draw.

NNs will be used in this paper due to the success this model has seen in other studies, including Gangal et al. (2015), [32]. Additionally, in theory it is possible to get a highly accurate model by including more and more hidden layers as each layer should improve the accuracy of the last in a manner akin to Decision Trees. As long as the criteria the model uses to split the features is sufficient, each additional layer should only aid in the overall accuracy.

NAÏVE BAYES

Naïve Bayes (NB) was built upon the basis of Bayes Theorem, a theorem put forward by Thomas Bayes, posthumously in 1763. In essence, it states that the probability of an event occurring should be revised as any new information is provided. Returning a value between 0

and 1, this value represents a percentage that any event may occur. A 0 being returned depicting a scenario which categorically will not occur, and a 1 indicating a scenario which will occur with absolute certainty. “Typically, Bayesian classifiers are best applied to problems in which the information from numerous attributes should be considered simultaneously in order to estimate the overall probability of an outcome,” [33].

NB has been used repeatedly in the field of Football Prediction, as shown previously. This is most likely because it is simple to implement and due to the assumptions it makes. NB assumes that all features are equally important and independent. While this is rarely true, in the models attempting to predict the winner of a match assigning equal importance to variables is acceptable.

The use of this model was endorsed by the number of other researchers and studies who utilized it when building other football result prediction models. These include Razali et al. (2017) and Owrampur et al. (2013) who were mentioned earlier for example.

DECISION TREES

Decision Trees (DT) are a supervised machine learning technique. Supervised machine learning means “the model is trained and tested on a set of data that contains the desired categorization,” [39]. In the case of this dissertation and data set, the desired categorisation is the FTR or Full-Time Result column. It is used to make predictions based on how a previous set of questions or checks were answered. Every DT is comprised of similar elements. A root node which, as with any tree, is the base of every DT and is made up of the whole data set. From this a series of questions are asked, based on the data set being used. These resulting nodes branch off and produce decision nodes. At each node the model asks itself, what feature will allow the data to be split in such a way that the resulting groups are as different from each

other as they can be, while the data items of each group are as similar to each other as possible? This process continues until reaching the final leaf or terminal node.

An example of this, as created for this dissertation, is pictured below in Figure 3. Each decision node and terminal node is one of three colours, representing the three possible outcomes a football match can have. In the example below it starts with the most likely outcome, being that of a home win, with each subsequent decision altering the prediction of the model, akin to how human brains work. With each new piece of information a human, and the model, are more informed and better equipped to predict the result. Each node also includes a percentage value dictating how confident the model is in the decision, allowing those using the model to be more informed.

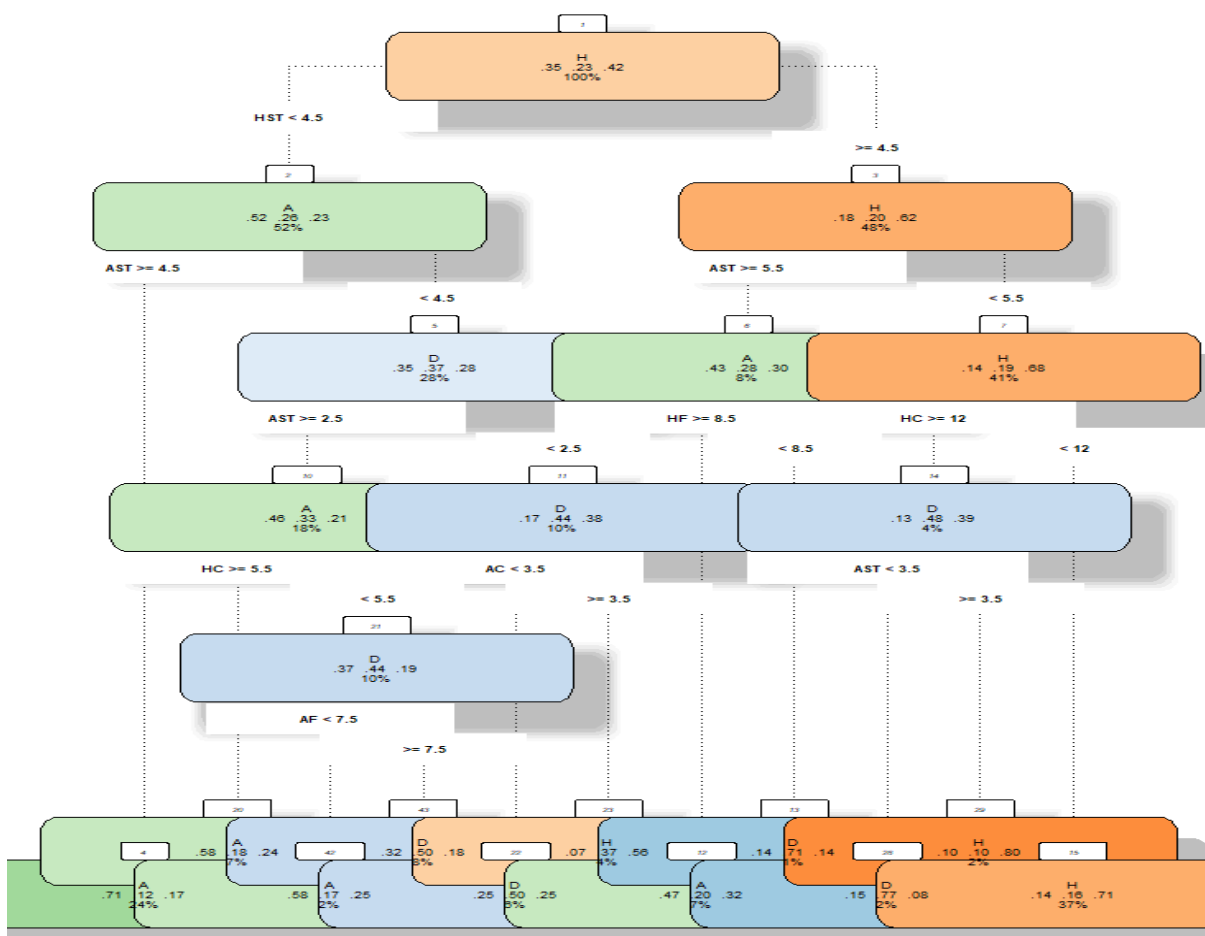


Figure 3: An example of a DT created for this dissertation

Due to the papers studied in the literature review, DT's will be used in this study. Like Neural Networks they seek to find the best method of splitting data to derive actionable and useable information. In Figure 3 this process can be seen in action. One of the uses of this study is that it has the potential capacity to assess what way to set up or play against certain teams, or what statistics are most important to have high values in. DTs are a way of visualising these metrics.

SUPPORT VECTOR MACHINES

Support Vector Machines (SVM), are a highbred machine learning model, utilizing aspects from both kNN and Linear Regression. Essentially, it lays the users' data on a plane and creates boundaries that separates the data into partitions on both sides. Each row of data becomes a data item that is plotted on this created plane. Classification is performed by finding the hyper-plane that best differentiates the classes, in the case of this paper A, D and H are the classes being differentiated. While mostly used in binary problems, yes or no scenarios, it can be used for multi-class differentiation problems.

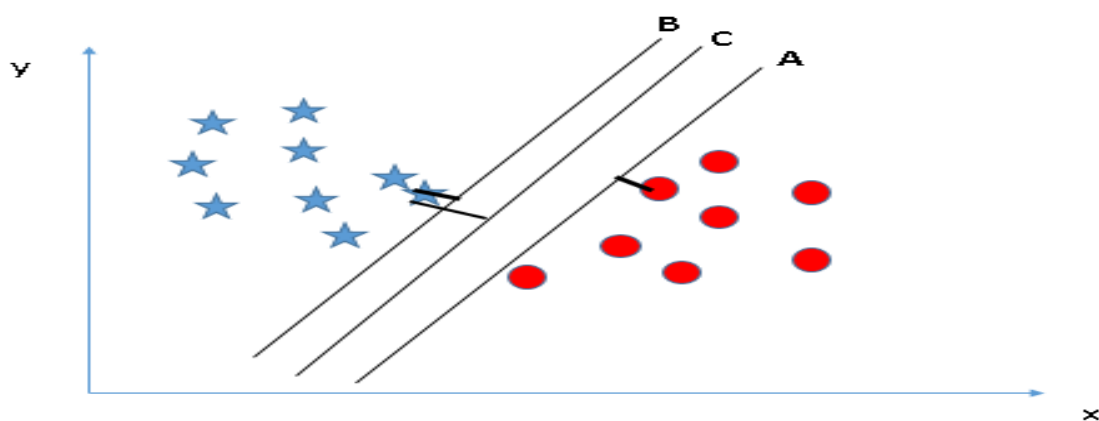


Figure 4: An example of how SVM choose the best bisection line [39]

SVM attempts to keep all those data items with similar features together, bisecting the groups containing similar features with a hyper-plane. The hyper-plane that best does this, is the one that best follows the Maximum Margin Hyperplane (MMH). In Figure 4 above, it is evident

that line C is the most accurate hyper-plane as it has the highest margin between two points in either categorisation cluster. SVM bisects data using MMH as it reduces the chance of misclassification errors. However, in the case of certain data it will forgo the MMH bisection line in an effort to increase accuracy. In the case of Figure 5 below, instead of splitting the data points along the hyper-plane B, as would typically be the case using MMH, SVM instead partitions the data using hyper-plane A as it attempts to prioritise accuracy.

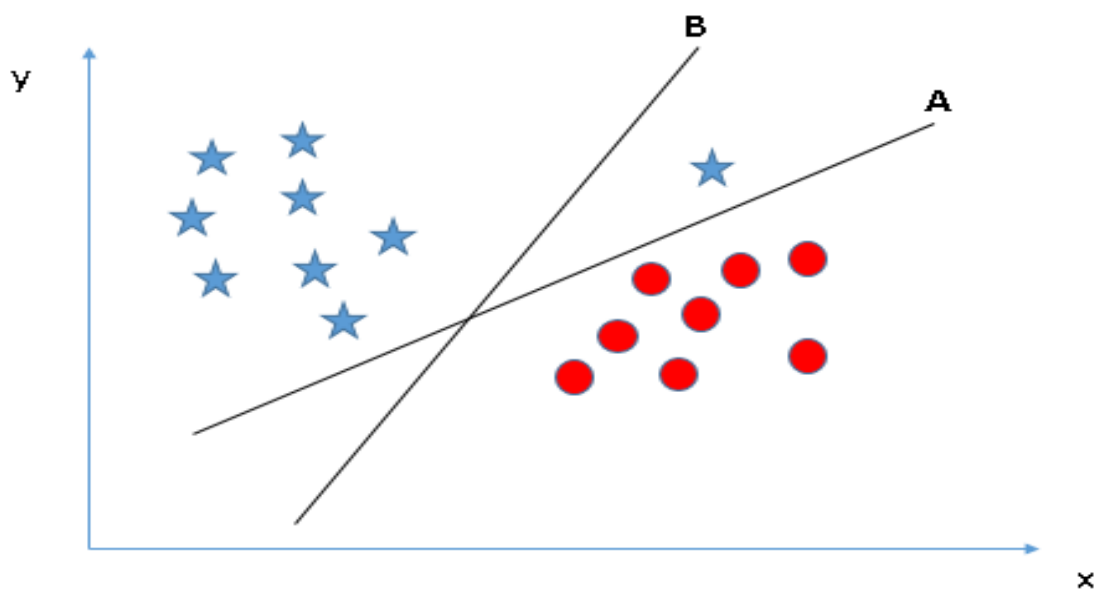


Figure 5: Further example of how SVM choose the best bisection line [39]

This method is similar in principle to kNN, splitting data points based on similar features, but it also prioritises accuracy. Should a data point contain features that closely resemble another classification, an away win's set of features resembles a home win's set of features, the model will choose accuracy over splitting data points evenly. As such this model will be used for this study.

RANDOM FORESTS

Created by Leo Breiman and Adele Cutler, Random Forests (RF) functions similarly to Decision Trees, discussed earlier. However, RF utilizes many Decision Trees (DT), to come to its prediction, using trees to build a forest, whereby the answer that most trees predict is the answer RF will return. It operates under the thought process of “a large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models,” [40]. This exact methodology can be viewed in Figure 6 below.

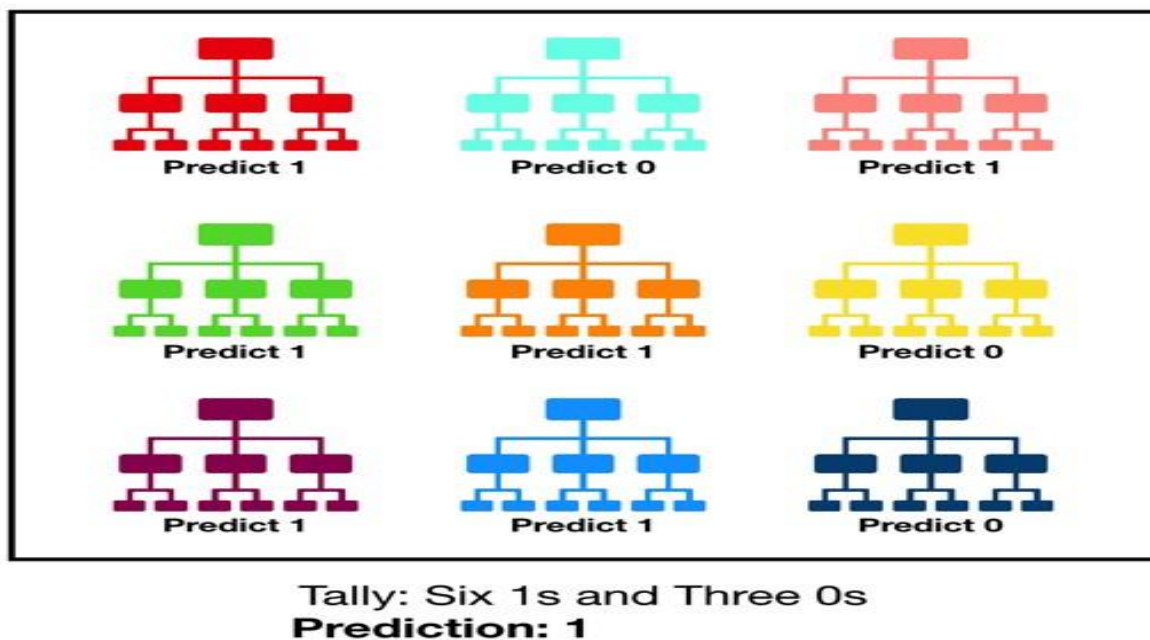


Figure 6: Example of how RF come to their prediction [41]

It is the fact that each tree is uncorrelated with another that allows this model to perform well, producing ensemble or group predictions that are more powerful and accurate than any one singular prediction may be. Each tree protects another from its individual mistakes or errors as the majority decision is the final choice that is made. While some trees may perform to a lesser standard than others, as long as these trees are the minority, the model itself will offset these shortcomings by selecting the trees that performed better and come to the correct prediction.

The main difference between typical DTs and those that are used in RF, is that feature selection has been altered. In a standard DT each and every feature is used in making the prediction. As such it is highly sensitive to changes made to the base data it is trained on, as one small change to the data source can alter the way a DT makes its predictions or the way the model asks itself the next question on how to split the data further. Within the scope of RF, each DT picks random features from the data set from which to grow its tree using replacement. This results in diverse trees of different shapes and sizes; this act is known as bagging. This process aids each DT to remain as uncorrelated from another as possible while still driving towards the same end result. A basic image of this process is shown in Figure 7 below.

RF, in theory, is more powerful than any singular DT. Utilizing the power of a cluster of DTs to select a majority decision, and as such will be used in this study.

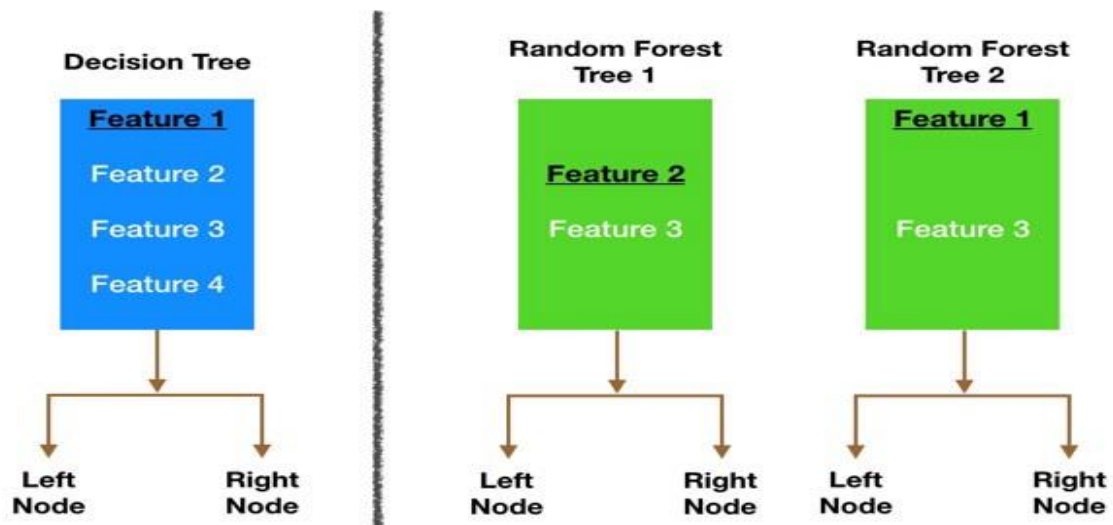


Figure 7: An example of how RF splits features and makes decisions, known as bagging [41]

XGBOOST

xGBoost is short form for the term “eXtreme Gradient Boosting.” Created by Tianqi Chen as a method to improve an original machine learning model. It does so by “combining it with a number of other weak models in order to generate a collectively strong model,” [42]. A similar

concept to the previously discussed Random Forest methodology. In his paper, “XGBoost: A Scalable Tree Boosting System,” Tianqi Chen (2016) described the thought process through which he went in order to create xGBoost, and why it is better than previous models that have been implemented. Of the boosting methods that came prior, xGBoost is faster and more optimized than its contemporaries and provides a scalable and accurate model for data scientists.

xGBoost has been used to win numerous Kaggle competitions due to the way it functions and the power it possesses. As such, this model will also be incorporated into this study.

SUMMARY

For this study to be successful, all models mentioned above will be utilized. This allows the author to ascertain which model is the best based purely on accuracy and avoid any biases that may linger. It is possible to form opinions on the models prior to their implementation, such as assuming that Naïve Bayes will be less effective than DTs because it assumes all variables are of equal importance. Or that RF will be more effective than a DT because it builds multiple trees to form a group decision. Or that SVM will be more effective than kNN because they function similarly, but SVM prioritises accuracy. As such, all models will be used and the model that performs best will be the model used to predict the most recent seasons results.

CHAPTER 3 – METHODOLOGY

This chapter details the source of the data used throughout this research and how it was cleansed in order to give a more accurate data set to the machine learning models. This chapter also details the overarching guiding strategy and philosophy used by the author to accomplish their end goals.

INTRODUCTION

The data being used in this dissertation is a complete set of statistics from every match over the last two seasons in the English Premier League. The most recent two seasons, not including the current domestic season, were selected so that the model had more than one season to learn from. This was made more important as the 2020/21 season was played behind closed doors, which is to say without fans in attendance, due to the ongoing global pandemic. Due to the nature of the season, that season's results should be viewed in a bubble as it is well known that supporters can impact a game of football, regularly being cited as the "12th man." In addition to the 2020/21 season, at least one other fully completed, fan attended, season is also used to help balance the model. This additional season is the 2019/20 season. If the models are to be used to predict the current season's results, a season more similar to a regulation season needs to be included.

The data will be used in all eight machine learning methods to establish their accuracy, both before and after the extra variables have been included, such as possession, five game form and goal difference. The aim is that the strongest model is then used to predict the results of the current season and assess how accurate it has become. Features such as formation have been omitted as firstly, they will be relatively covered and represented by including the quality of the defence, midfield and attackers. With a team using five defenders more likely to have a higher defence rating than those who use four defenders for example. Secondly, the formation is more indicative of the current footballing landscape, with 4-4-2 once being the default

formation, it has now given way to the 4-2-3-1, 4-3-3 or 3-5-2. Swapping formation for the quality of the players in the areas is a more apt metric.

Additional non-psychological features will be added, as will psychological variables as the result of a match can be swayed by psychological factors also. Psychological factors such as suspensions to both major and minor players, form over the last 5 matches and the result of the last fixture between two teams will also be included in the data set. While non-psychological features will be included such as attempting to account for the quality of a team with defence ratings, midfield ratings and attacker ratings also added for each game along with a Goal Difference statistic to chart the team's capacity for scoring goals.

RESEARCH QUESTIONS AND OBJECTIONS REVISITED

The author set out to answer one research question:

- Can a machine learning model be produced that can accurately gauge the winner between two Premier League teams?

The answer to the research question will be attained by achieving three core research objectives, those being:

1. Create a machine learning model that can predict the result of any given Premier League game utilizing standard in game statistics

Produce the eight different machine learning models, using all of the different methods discussed above, that can predict the result of any game of football in the Premier League.

2. Create a machine learning model that can predict the result of any given Premier League game utilizing standard in game statistics and also additional statistics and psychological factors

Secondly, once those models have been created, they will be repurposed to perform the same task on the newly created and updated table of data that now encompasses additional features of a psychological and non-psychological variety.

3. Assess whether the model which performs best in Objective 2 can be used to predict the results of the most recent Premier League season.

Finally, the model that performed the best under research Objective 2 will be utilized to predict the results of all matches in the Premier League for the latest season, and as yet unseen by any of the models, the 2021/22 season.

RESEARCH PHILOSOPHY

“Research philosophy is associated with assumption, knowledge and nature of the study. It deals with the specific way of developing knowledge,” [43]. Essentially, it determines the way the researcher approaches, collects, analyses and uses data. There are 4 core research philosophies, those being interpretivism, positivism, pragmatism and realism. The four core philosophy branches can be explained as follows:

1. Interpretivism: focuses on qualitative data and are “prepared to sacrifice reliability and representativeness for greater validity,” [44]. Taking on a subjective view and focusing on meaning, interpretivism is more humanistic than any of the other models.
2. Positivism: focuses on quantitative data that can lead to statistical analysis, such that one action causes a particular outcome given the environment. “Positivism relates to the viewpoint that the researcher needs to concentrate on facts,” [45]. Positivist studies utilize a structured methodology so that replication of the study can be performed at a future date.
3. Pragmatism: while positivism and interpretivism are mutually exclusive, those who follow a pragmatic philosophy believe there are numerous ways to conduct the research.

“Pragmatism involves research designs that incorporate operational decisions based on ‘what will work best’ in finding answers to the questions under investigation,” [46].

4. Realism: containing two branches, direct realism and critical realism, it is founded on scientific approach in the pursuit of answers, in a manner similar to positivism.

It must be noted that all studies contain traces, and work within, the scope of other philosophy branches, for example axiology. “Axiology is a branch of philosophy which is concerned about judgements, aesthetics and ethics,” [47]. Derived from the Greek word *axi*, meaning value or worth, axiology is primarily concerned with humans flourishing as a result of the study undertaken. This paper, while not its sole intention, also seeks to limit the losses of those who gamble, to alleviate debt and stress on themselves and their loved ones.

Richard Purtill, author of the “The Purpose of Science” asserted that the main aim of science is to explain and predict, [48]. Explaining and predicting is the underlying intent of the author for this study, forming facts and predictions based on what is known and quantifiable. To this end a positivist research approach will be undertaken by the author. The aim of this research is to turn what is known, the statistics and metrics, derive knowledge from them and perform accurate predictions.

POSITIVIST RESEARCH

The philosophy adopted for this paper is that of positivism as it focuses on statistical analysis. Positivism follows the notion that an action in one environment produces a specific outcome, and this can be determined through astute measuring and observation. To assume a positivist approach to this study, the author must remain independent of the study, and all studies and findings must be purely objective, with no room for bias. As the author has no impact on any of the data used and has instead collected and built upon existing statistics and measurements a positivistic approach can be maintained.

Two of the core positives of adopting a positivistic approach are firstly, the focus on quantitative data that is more reliable than qualitative data. Adopting a more scientific approach and using more scientifically sound and objective data. Due to this, all findings and analysis are more trustworthy than assumptions and findings founded on qualitative data. The second core positive is, as mentioned before, positivism follows a structured approach whereby the study can be replicated again with a reduced margin of error as it is built upon specific rules and scientific method.

Since the ultimate goal is to develop a model that can be used to predict the winner between any two teams in the Premier League using quantitative data, turning base facts into actionable usable knowledge, positivism is the correct research philosophy to follow. Coupled with the fact the author has no impact on statistics or measurements, other than building the second data set from sourced data, positivism as the correct research philosophy is only heightened further.

DATA MODEL

Before beginning to implement the research, a methodology needs to be selected to determine how the study is approached. To keep the study structured and organised, the CRISP-DM methodology was selected. CRISP-DM, an acronym for Cross-Industry Process for Data Mining, is a cyclical model used to keep structure and order to most analytics or data mining projects.

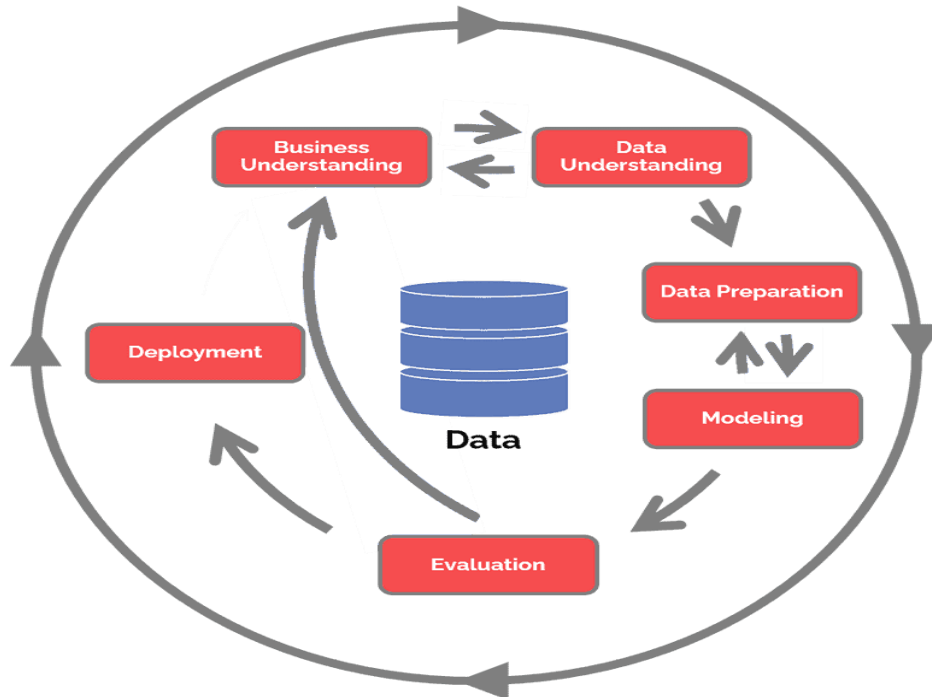


Figure 8: Visual representation of the CRISP-DM model [47]

The model is comprised of six sequential stages:

1. Business understanding – what is it the research is trying to do?
2. Data understanding – what data is needed and is it clean?
3. Data preparation – how is the data organized for modelling?
4. Modelling – what modelling techniques are to be used?
5. Evaluation – which model best achieves the research objectives?
6. Deployment – how to apply the best model to attain research objectives?

BUSINESS UNDERSTANDING

Before beginning with any research or study, a fundamental understanding of what the aim of the research is must be understood. In this context, business understanding assumes the form of the research question and objectives. The research question remains:

- Can a machine learning model be produced that can accurately gauge the winner between two Premier League teams?

For the author to answer this question the three research objectives that were previously stated must be met:

1. Create a machine learning model that can predict the result of any given Premier League game utilizing standard in game statistics
2. Create a machine learning model that can predict the result of any given Premier League game utilizing standard in game statistics and also additional statistics and psychological factors
3. Assess whether the model which performs best in objective 2 can be used to predict the results of the most recent Premier League season.

Establishing exactly what the end goals of the research are helps to alleviate any deviation from these goals. This primes all work and study undertaken to be in the service of achieving the objectives established here.

DATA UNDERSTANDING

BASE FOOTBALL STATISTICS

The data being used was sourced from "football-data.co.uk" and will require cleansing and the removal of certain columns. Certain variables within the data set are unnecessary, such as the referee, bookies odds for a specific event or the date and time. Additionally, the half-time home and away goals will need to be removed as there would be no way of knowing this prior to kick-off. Before these were removed the data set contained seven-hundred and sixty individual observations with one-hundred and six different variables forming a data set that contains 80,560 individual data points. The final base statistics data set contained seven-

hundred and sixty unique observations with fifty-nine features combining to form 44,840 individual data points. The data set, as described by the notes provided from the source, contains the following features for every Premier League match, with additional betting odds that have been omitted, described in detail below.

Div = League Division

Date = Match Date (dd/mm/yy)

Time = Time of match kick-off

HomeTeam = Home Team

AwayTeam = Away Team

FTHG and HG = Full Time Home Team Goals

FTAG and AG = Full Time Away Team Goals

FTR and Res = Full Time Result (H=Home Win, D=Draw, A=Away Win)

HTHG = Half Time Home Team Goals

HTAG = Half Time Away Team Goals

HTR = Half Time Result (H=Home Win, D=Draw, A=Away Win)

Attendance = Crowd Attendance

Referee = Match Referee

HS = Home Team Shots

AS = Away Team Shots

HST = Home Team Shots on Target

AST = Away Team Shots on Target

HHW = Home Team Hit Woodwork

AHW = Away Team Hit Woodwork

HC = Home Team Corners

AC = Away Team Corners

HF = Home Team Fouls Committed

AF = Away Team Fouls Committed

HFKC = Home Team Free Kicks Conceded

AFKC = Away Team Free Kicks Conceded

HO = Home Team Offsides

AO = Away Team Offsides

HY = Home Team Yellow Cards

AY = Away Team Yellow Cards

HR = Home Team Red Cards

AR = Away Team Red Cards

UPDATED STATISTICS

To create a data set that encompasses more aspects of the game, inclusive of certain psychological factors and additional non-psychological factors, the cleaned data set will be updated to include new data features. Once again, this data set contained seven-hundred and sixty individual observations with one-hundred and six unique variables forming a data set that contains 80,560 individual data points. Once these were removed, as they were for the base statistics data set, the updated statistics data set contained seven-hundred and sixty unique observations with eighty-three features to form a data set containing 63,080 individual data points. To combine the data from both sources the author created additional variables and

calculated the values of each, this accounted for 21,280 of the data points. The new features that are added to the base statistics data set are as follows:

H.GK = Home Goal Keeper

H.Def = Home Defence

H.Avg_Def = Home Average Defence

H.Mid = Home Midfield

H.Avg_Mid = Home Average Midfield

H.Att = Home Attack

H.Avg_Att = Home Average Attack

H.Subs = Home Substitutes

H.Avg_Sub = Home Average Substitutes

A.GK = Away Goal Keeper

A.Def = Away Defence

A.Avg_Def = Away Average Defence

A.Mid = Away Midfield

A.Avg_Mid = Away Average Midfield

A.Att = Away Attack

A.Avg_Att = Away Average Attack

A.Subs = Away Substitutes

A.Avg_Sub = Away Average Substitutes

Five_Game_Form_H = 5 Game Form for Home Team

Rev_Fixture_H = Reverse Fixture Home Team

Maj_Suspension_H = Major Suspensions Home Team

Min_Suspension_H = Minor Suspensions Home Team

GD_H = Goal Difference Home

Five_Game_Form_A = 5 Game Form for Away Team

Rev_Fixture_A = Reverse Fixture Away Team

Maj_Suspension_A = Major Suspensions Away Team

Min_Suspension_A = Minor Suspensions Away Team

GD_A = Goal Difference Away

HP = Home Possession

AP = Away Possession

To assess the ratings of each player, a compendium of all players in the Premier League and their associated ratings must be created. These ratings were sourced from “fifaindex.com” a website charting every player to have been registered in a FIFA competition and therefore featured in a game of FIFA from 2005 up to and including the most recent season of football in 2022. As rating a player is subjective, no one person will have the exact same opinion on any player and their qualities as another. The best way to provide a numeric rating for each player is utilizing the most played and discussed football simulator in the world. FIFA is the international governing body of association football. It is therefore fitting as the title of the foremost football simulation game in which all professionally registered players are rated, assessed and ranked based on their skills in the game.

“The ratings are produced by The Ratings Collective, which serves as a talent scouting network for the FIFA games, assessing more than 17,000 players worldwide and assigned numerical values to more than 30 of their attributes to define their in-game skill

levels,” [49]. Ratings are based upon six core facets of being a football player, (Pace, Shooting, Passing, Dribbling, Defending and Physical), with each of these categories being rated based upon twenty-nine subcategories. These sub-categories range from acceleration and sprint speed in the Pace category to shot power and finishing in the Shooting category. As the ratings are unbiased and produced through collaboration, they serve as the most fair and accurate representation of every player in the league, with all values in the H.GK, H.Def, H.Mid, H.Att, H.Subs, A.GK, A.Def, A.Mid, A.Att and A.Subs deriving their value courtesy of these ratings. These values represent non-psychological statistics that could help the models derive which team is statistically better and therefore more likely to win in any game. For example, Manchester City who contain world class players for all positions on the pitch and coming off the bench, are more likely to have higher ratings across these categories than a newly promoted side. There is also a small psychological impact of these statistics as it is always in a player’s head that they are currently playing against a player that is considered world class, such as Mohamed Salah, Kevin De Bruyne or Alisson Becker. The quality of player may also dictate things such as tactics, with attack minded defenders such as Kieran Tierney being instructed to ignore their attacking impetus to focus on defending higher quality forwards such as the aforementioned Mohamed Salah. The ratings from the aforementioned “fifaindex.com” are coupled with the data from “m.football-lineups.com” to see each and every line up over the previous two seasons and therefore attribute a value to each section of the pitch and bench.

The football line-ups website was also the location of all the remaining features of the data set. The site contained a very useful colour scheme whereby if the user selects the team they are searching for, Aston Villa for example, the result is colour coordinated. The colour system used is red for a loss, green for a win and yellow for a draw. An example of such is pictured below in Figure 9.

2021-08			
28	1-1	Brentford	4141
21	2-0	Newcastle	433
14	2-3	Watford	4231

Figure 9: An example of the data source used to build the updated data set [50]

This colour coding allowed for a quick transfer of data from source to the data set at a glance. The form was calculated in the same manner to points distribution in football; 3 points for a win, 1 for a draw and 0 for a loss. Each season every team would start off on 0, as they are not in any current form having come back from holiday, and the first game result would be reflected in the following match. In the example provided above in Figure 9, Aston Villa are bringing no form into the Newcastle match, represented by a 0 in the data set, having lost the first game of the season to Watford. However, they are bringing better form into the Brentford match, represented by a 3 in the data set, after beating Newcastle in their second game. The form over five games variable will impact all games in the same manner. Staying with Aston Villa from Figure 9, they would be bringing a rating of 4 into their match after Brentford, having picked up a draw and an additional point to their five game form variable.

This website was also used to determine the Reverse Fixture H and Reverse Fixture A variables as this source contains all previous games and their scores. The Reverse Fixture features are psychological variables as it is in a team's mind that they beat, or were beaten by, the team in question. It is also easy using this source to calculate the GD H and GD A variables. Using Figure 9 again, it can be quickly seen that after game one against Watford the Goal Difference would be at -1, after game two it would be at +1 and after game three it would remain at +1. These figures would be placed in the correct variable of GD H or GD A depending on whether they are currently playing home or away. GD H and GD A represent both statistical and psychological aspects of the game.

The same source also contains match breakdowns, including whether a player has been sent off, and whether it was a straight red card or two yellow cards. This is significant as a straight red card represents a two-game ban, while two yellow cards result in a one game ban. The Maj/Min Suspension H and Maj/Min Suspension A are psychological factors. It is expected that if a team is missing players, they must call upon inferior players to fill the team resulting in a less able squad. This will be reflected in the ratings attributed to all the places on the pitch or bench, but also if a team's leading goal scorer, team captain or central defensive pillar is suspended it is sure to have a psychological impact on the squad.

Finally, the inclusion of possession statistics for both home and away teams was also included as this is a very common metric used in football analysis. Incorporating this as a feature of the data set is beneficial as in real world football it dictates which team is the most dominant and controlled the ball for more of a match. The logic is that a team that a team that has a lesser possession percentage is less likely to score as they see less of the ball in a game.

DATA PREPARATION

BASE FOOTBALL STATISTICS – DATA CLEANSING

Before the first models can be built, the data set needs to be cleansed. Firstly the 2020/21 and 2019/20 seasons data sets are combined to form the data that will be manipulated and built upon. This in itself brought to light anomalies in the data. Although both data sets were gathered from the same source, between seasons the company that collected the data swapped naming conventions for the two Manchester clubs, Manchester City and Manchester United. In the 2019/20 season they are referred to as Man City and Man United, whereas in the 2020/21 season the full names of Manchester City and Manchester United were used. To remedy this issue the “gsub” function in R Studio was utilized. By calling the column, and the specific text the user wishes to alter, the author was able to change the names throughout the entire data set

into the shorthand versions of their respective names. This same process was carried out to fix the name of Wolverhampton Wanderers as in one data set they are labelled as Wolverhampton, whilst in the other they are referred to as Wolves.

The data set also contained betting odds from a number of bookmakers, forming the largest features of this data set ranging from column 25 to column 106 inclusive. These columns can be immediately dropped as they will provide no actionable information for the purposes that these models are intended. Once the data set had been condensed into this more precise form, box plots were used to assess whether any of the features contained any outliers. The resultant box plots flagged two outliers: there was one match in which one team beat the other 9 goals to 0, and another which was won 8 goals to 0. Upon further investigation, these results were real match results that occurred but due to being so much larger than the standard result they showed up in the tails of the box plots. Box plots viewable below in Figure 10.

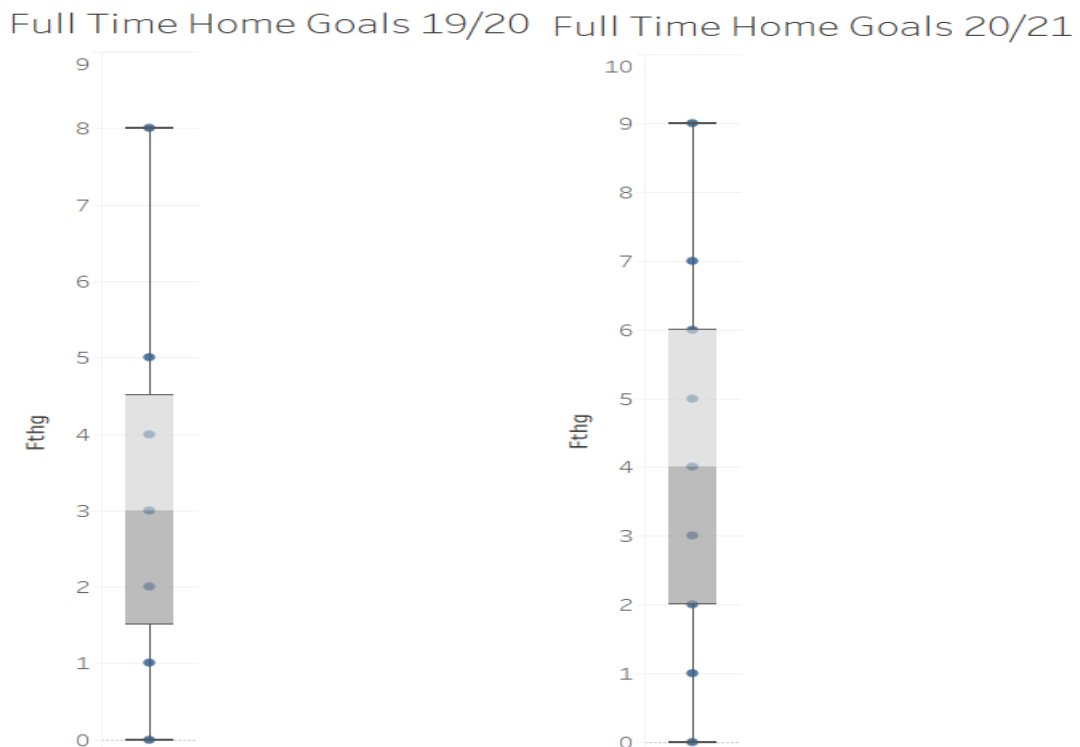


Figure 10: Outliers found in data sets using box plots

The next columns to drop are also columns that provide no meaningful insight into predicting the result of a football match. These columns are the Referee, Division, Date and Time. The referee can potentially have an impact on any game, breaking up play or allowing the game to flow however they see fit. However, being an unbiased entity and presuming all referees are of a similar standard, their impact should be minimal as the two teams compete. The Division variable can be removed with little thought as everyone is in the same division, the Premier League. Date can be removed also as in the original data set this is just used to dictate the day the game is played. Over the course of two seasons worth of data this variable is another that provides very little input. If the models were being created over a number of seasons worth of data, then perhaps trends could be formed, such as noting a team that perform consistently poorly over the busy Christmas period. However, due to the fact this a relatively small sample of only two years, these trends would be hard to identify and are therefore removed. The same rationale as was displayed for the removal of the Date variable is applied to the Time variable.

The next set of cleaning is aimed at producing an unbiased model. The data set contains two variables Full Time Home Goals (FTHG) and Full Time Away Goals (FTAG). As the models are trying to predict the result of a match between two teams, the results being either A, H or D representing an away win, a home win, and a draw respectively, including the final goal tally for home and away teams will provide too much information. The models will be able to establish that if FTHG is greater than FTAG, then the Full Time Result (FTR) is always a H. This renders the other features completely superfluous as the models will always look to these features to inform their decision. As such, it must be blinded from these features, so they have been removed.

A similar logic is applied to Half Time Home Goals (HTHG), Half Time Away Goals (HTAG) and Half Time Result (HTR). These features may be incorporated for a second model used mid-game to predict the winner. However, as the intent of this research is to predict the

winner prior to kick-off they have been removed. It could be argued that these features could help the prediction model, but at the risk of reducing the effectiveness of the other more commonly used statistics they have been dropped.

It is also worth noting that the data sets that were used for this research did not contain HFKC or AFKC, nor did it contain HO, AO or HHW and AHW. Only certain seasons of data, with the source containing every Premier League, Championship, Division 1, Division 2, Division 3, League 1, League 2 and Conference League from the 1993/94 season until the most current 2021/22 season, contain all the base statistics. (Between the 2002/03 season and 2005/06 season the English Football Leagues underwent a revamp and the names of the divisions were changed, with the previous Division 1 becoming the Championship and Division 2 became the new Division 1 and so on.) While the inclusion of these statistics would be ideal in allowing the model more learning opportunities from commonly used football metrics, finding a comprehensive data source that contains everything this data set does while including additional features proved difficult. This resulted in many matches containing null or N/A values. These values contribute nothing and would need to be removed from any data set before any machine learning model would function and were removed.

The final stage of cleaning involves the team names themselves. Some machine learning models find it difficult dealing with categorical variables, variables that have “two or more categories, but there is no intrinsic ordering to the categories,” [51], such as the team names. As such, a more binary or numeric method must be introduced to the data set in order for the model to be able to comprehend the data set and its unique features. While it is possible to convert each team to a binary representation, it is much easier and beneficial to one hot encode them. “One hot encoding is a process of converting categorical data variables so they can be provided to machine learning algorithms to improve predictions,” [52]. Essentially this method takes all 23 different teams and creates 46 new columns, as each team plays home and

away. For example, instead of now having a HomeTeam and AwayTeam column, the data set will now contain “Liverpool H” and “Liverpool A” as two separate columns. These columns will contain a 1 if the team is featured in the match and a 0 if they are not. This informs the model that of the 23 different teams, only the two teams with a 1 are in play, hypothetically Liverpool H and Man City A. With a now complete and clean data set, all 8 machine learning models can be created. The complete data set is viewable below in Figure 11.

\$ FTR	: num	0 1 1 0 0 0 0 0 1 1 ...	\$ Watford.H	: num	0 0 0 0 0 0 0 0 0 0 ...
\$ HS	: num	0.133 0.133 0.7 0.467 0.2 ...	\$ West.Brom.H	: num	0 0 0 0 1 0 0 0 0 0 ...
\$ AS	: num	0.444 0.296 0.185 0.519 0.444 ...	\$ West.Ham.H	: num	0 0 0 1 0 0 0 0 0 0 ...
\$ HST	: num	0.1429 0.2143 0.4286 0.2143 0.0714 ...	\$ Wolves.H	: num	0 0 0 0 0 0 0 0 0 0 ...
\$ AST	: num	0.4 0.333 0.2 0.133 0.467 ...	\$ Arsenal.A	: num	1 0 0 0 0 0 0 0 0 0 ...
\$ HF	: num	0.522 0.609 0.391 0.565 0.522 ...	\$ Aston.Villa.A	: num	0 0 0 0 0 0 0 0 0 0 ...
\$ AF	: num	0.478 0.435 0.217 0.261 0.348 ...	\$ Bournemouth.A	: num	0 0 0 0 0 0 0 0 0 0 ...
\$ HC	: num	0.118 0.412 0.529 0.471 0.118 ...	\$ Brighton.A	: num	0 0 0 0 0 0 0 0 0 0 ...
\$ AC	: num	0.188 0.188 0 0.438 0.312 ...	\$ Burnley.A	: num	0 0 0 0 0 0 0 0 0 0 ...
\$ HY	: num	0.333 0.333 0.167 0.333 0.167 ...	\$ Chelsea.A	: num	0 0 0 0 0 0 1 0 0 0 ...
\$ AY	: num	0.333 0.167 0 0.333 0.167 ...	\$ Crystal.Palace.A	: num	0 0 0 0 0 0 0 0 0 0 ...
\$ HR	: num	0 0 0 0 0 0 0 0 0 0 ...	\$ Everton.A	: num	0 0 0 0 0 1 0 0 0 0 ...
\$ AR	: num	0 0 0 0 0 0 0 0 0.5 0 ...	\$ Fulham.A	: num	0 0 0 0 0 0 0 0 0 1 ...
\$ Arsenal.H	: num	0 0 0 0 0 0 0 0 0 0 ...	\$ Leeds.A	: num	0 0 1 0 0 0 0 0 0 0 ...
\$ Aston.Villa.H	: num	0 0 0 0 0 0 0 0 0 0 ...	\$ Leicester.A	: num	0 0 0 0 1 0 0 0 0 0 ...
\$ Bournemouth.H	: num	0 0 0 0 0 0 0 0 0 0 ...	\$ Liverpool.A	: num	0 0 0 0 0 0 0 0 0 0 ...
\$ Brighton.H	: num	0 0 0 0 0 0 1 0 0 0 ...	\$ Man.City.A	: num	0 0 0 0 0 0 0 0 0 0 ...
\$ Burnley.H	: num	0 0 0 0 0 0 0 0 0 0 ...	\$ Man.United.A	: num	0 0 0 0 0 0 0 0 0 0 ...
\$ Chelsea.H	: num	0 0 0 0 0 0 0 0 0 0 ...	\$ Newcastle.A	: num	0 0 0 1 0 0 0 0 0 0 ...
\$ Crystal.Palace.H	: num	0 1 0 0 0 0 0 0 0 0 ...	\$ Norwich.A	: num	0 0 0 0 0 0 0 0 0 0 ...
\$ Everton.H	: num	0 0 0 0 0 0 0 0 1 0 ...	\$ Sheffield.United.A	: num	0 0 0 0 0 0 0 0 0 0 ...
\$ Fulham.H	: num	1 0 0 0 0 0 0 0 0 0 ...	\$ Southampton.A	: num	0 1 0 0 0 0 0 0 0 0 ...
\$ Leeds.H	: num	0 0 0 0 0 0 0 0 0 1 ...	\$ Tottenham.A	: num	0 0 0 0 0 0 0 0 0 0 ...
\$ Leicester.H	: num	0 0 0 0 0 0 0 0 0 0 ...	\$ Watford.A	: num	0 0 0 0 0 0 0 0 0 0 ...
\$ Liverpool.H	: num	0 0 1 0 0 0 0 0 0 0 ...	\$ West.Brom.A	: num	0 0 0 0 0 0 0 0 1 0 ...
\$ Man.City.H	: num	0 0 0 0 0 0 0 0 0 0 ...	\$ West.Ham.A	: num	0 0 0 0 0 0 0 0 0 0 ...
\$ Man.United.H	: num	0 0 0 0 0 0 0 0 0 0 ...	\$ Wolves.A	: num	0 0 0 0 0 0 0 1 0 0 ...
\$ Newcastle.H	: num	0 0 0 0 0 0 0 0 0 0 ...			
\$ Norwich.H	: num	0 0 0 0 0 0 0 0 0 0 ...			
\$ Sheffield.United.H	: num	0 0 0 0 0 0 0 1 0 0 ...			
\$ Southampton.H	: num	0 0 0 0 0 0 0 0 0 0 ...			
\$ Tottenham.H	: num	0 0 0 0 0 1 0 0 0 0 ...			

Figure 11: Full list of features in base statistics data set

UPDATED STATISTICS – DATA CLEANSING

This data set was cleansed in the same manner the base statistics data set was, due to it being an extension of this data set. Removing N/A columns and features that provide no actionable information After including the author collated and variables and performing an analysis on the data, it was found that the updated statistics data set does not contain null or missing values and represents two complete seasons worth of data. The complete data set is viewable below in Figure 12.

\$ FTR	: num	0	1	1	0	0	0	0	0	1	1	...	\$ EvertonA	: num	0	0	0	0	0	1	0	0	0	0	...													
\$ HS	: num	0.133	0.133	0.7	0.467	0.2	...	\$ FulhamA	: num	0	0	0	0	0	0	0	0	1	...	\$ LeedsA	: num	0	0	1	0	0	0	0	0	0	0	...						
\$ HST	: num	0.1429	0.2143	0.4286	0.2143	0.0714	...	\$ LeicesterA	: num	0	0	0	1	0	0	0	0	0	0	...	\$ LiverpoolA	: num	0	0	0	0	0	0	0	0	0	0	...					
\$ HF	: num	0.522	0.609	0.391	0.565	0.522	...	\$ ManCityA	: num	0	0	0	0	0	0	0	0	0	0	...	\$ ManUnitedA	: num	0	0	0	0	0	0	0	0	0	0	...					
\$ HC	: num	0.118	0.412	0.529	0.471	0.118	...	\$ NewcastleA	: num	0	0	1	0	0	0	0	0	0	0	...	\$ NorwichA	: num	0	0	0	0	0	0	0	0	0	0	...					
\$ HY	: num	0.333	0.333	0.167	0.333	0.167	...	\$ SheffieldUnitedA	: num	0	0	0	0	0	0	0	0	0	0	...	\$ SouthamptonA	: num	1	0	0	0	0	0	0	0	0	0	...					
\$ HR	: num	0	0	0	0	0	0	0	0	0	0	...	\$ TottenhamA	: num	0	0	0	0	0	0	0	0	0	0	...	\$ WatfordA	: num	0	0	0	0	0	0	0	0	0	0	...
\$ HP	: num	0.435	0.194	0.484	0.645	0.306	...	\$ WestBromA	: num	0	0	0	0	0	0	0	1	0	...	\$ WestHamA	: num	0	0	0	0	0	0	0	0	0	0	...						
\$ AS	: num	0.444	0.296	0.185	0.519	0.444	...	\$ WolvesA	: num	0	0	0	0	0	1	0	0	...	\$ H_GK	: num	0.36	0.6	0.96	0.64	0.48	0.88	0.48	0.36	0.6	0.4	...							
\$ AST	: num	0.4	0.333	0.2	0.133	0.467	...	\$ H_Def	: num	0.3682	0.4229	0.6766	0.4975	0.0299	...	\$ H_Avg_Def	: num	0.382	0.492	1	0.641	0.422	...															
\$ AF	: num	0.478	0.435	0.217	0.261	0.348	...	\$ H_Mid	: num	0.591	0.423	0.28	0.686	0.812	...	\$ H_Avg_Mid	: num	0.213	0.315	0.584	0.44	0.251	...															
\$ AC	: num	0.188	0.188	0	0.438	0.312	...	\$ H_Avg_Mid	: num	0.01	0.445	0.975	0.055	0.025	0.105	0.045	0.38	0.88	0.05	...	\$ H_Avg_Att	: num	0.0909	0.4773	0.8941	0.5	0.2273	...										
\$ AY	: num	0.333	0.167	0	0.333	0.167	...	\$ H_Att	: num	0.26	0.269	0.373	0.33	0.199	...	\$ H_Sub	: num	0.26	0.269	0.373	0.33	0.199	...															
\$ AR	: num	0	0	0	0	0	0	0	0.5	0	...	\$ H_Avg_Sub	: num	0.0624	0.0646	0.0894	0.0792	0.0478	...	\$ A_GK	: num	0.76	0.4	0.4	0.44	0.8	0.6	0.72	0.72	0.48	0.68	...						
\$ AP	: num	0.565	0.806	0.516	0.355	0.694	...	\$ A_Def	: num	0.102	0.444	0.439	0.423	0.531	...	\$ A_Avg_Def	: num	0.195	0.135	0.13	0.114	0.223	...															
\$ ArsenalH	: num	0	0	0	0	0	0	0	0	0	0	...	\$ A_Mid	: num	0.439	0.417	0.63	0.425	0.68	...	\$ A_Avg_Mid	: num	0.618	0.545	0.557	0.572	0.689	...										
\$ AstonVillaH	: num	0	0	0	0	0	0	0	0	0	0	...	\$ A_Att	: num	0.6464	0.3286	0.0464	0.3071	0.0786	...	\$ A_Avg_Att	: num	0.843	0.778	0.769	0.722	0.935	...										
\$ BournemouthH	: num	0	0	0	0	0	0	0	0	0	0	...	\$ A_Sub	: num	0.324	0.231	0.24	0.303	0.335	...	\$ A_Avg_Sub	: num	0.0314	0.0217	0.0226	0.0293	0.0326	...										
\$ BrightonH	: num	0	0	0	0	0	1	0	0	0	0	...	\$ Five_Game_Form_H	: num	0	0	0	0	0	0	0	2	0	...	\$ Five_Game_Form_A	: num	0	0	0	0	0	0	0	0	0	...		
\$ BurnleyH	: num	0	0	0	0	0	0	0	0	0	0	...	\$ Rev_Fixture_H	: num	0	0	0	0	0	0	0	0	0	...	\$ Rev_Fixture_A	: num	0	0	0	0	0	0	0	0	0	...		
\$ ChelseaH	: num	0	0	0	0	0	0	0	0	0	0	...	\$ GD_H	: num	0.407	0.407	0.407	0.407	0.407	...	\$ GD_A	: num	0.423	0.423	0.423	0.423	0.423	...										
\$ CrystalPalaceH	: num	0	1	0	0	0	0	0	0	0	0	...	\$ Maj_Suspension_H	: num	0	0	0	0	0	0	0	0	0	...	\$ Min_Suspension_H	: num	0	0	0	0	0	0	0	0	0	...		
\$ EvertonH	: num	0	0	0	0	0	0	0	1	0	0	...	\$ Maj_Suspension_A	: num	0	0	0	0	0	0	0	0	0	...	\$ Min_Suspension_A	: num	0	0	0	0	0	0	0	0	0	...		
\$ FulhamH	: num	1	0	0	0	0	0	0	0	0	0	...																										
\$ LeedsH	: num	0	0	0	0	0	0	0	0	1	...																											
\$ LeicesterH	: num	0	0	0	0	0	0	0	0	0	0	...																										
\$ LiverpoolH	: num	0	0	1	0	0	0	0	0	0	0	...																										
\$ ManCityH	: num	0	0	0	0	0	0	0	0	0	0	...																										
\$ ManUnitedH	: num	0	0	0	0	0	0	0	0	0	0	...																										
\$ NewcastleH	: num	0	0	0	0	0	0	0	0	0	0	...																										
\$ NorwichH	: num	0	0	0	0	0	0	0	0	0	0	...																										
\$ SheffieldUnitedH	: num	0	0	0	0	0	0	1	0	0	0	...																										
\$ SouthamptonH	: num	0	0	0	0	0	0	0	0	0	0	...																										
\$ TottenhamH	: num	0	0	0	0	1	0	0	0	0	0	...																										
\$ WatfordH	: num	0	0	0	0	0	0	0	0	0	0	...																										
\$ WestBromH	: num	0	0	0	1	0	0	0	0	0	0	...																										
\$ WestHamH	: num	0	0	1	0	0	0	0	0	0	0	...																										
\$ WolvesH	: num	0	0	0	0	0	0	0	0	0	0	...																										
\$ ArsenalA	: num	1	0	0	0	0	0	0	0	0	0	...																										
\$ AstonVillaA	: num	0	0	0	0	0	0	0	0	0	0	...																										
\$ BournemouthA	: num	0	0	0	0	0	0	0	0	0	0	...																										
\$ BrightonA	: num	0	0	0	0	0	0	0	0	0	0	...																										
\$ BurnleyA	: num	0	0	0	0	0	0	0	0	0	0	...																										
\$ ChelseaA	: num	0	0	0	0	0	1	0	0	0	0	...																										
\$ CrystalPalaceA	: num	0	0	0	0	0	0	0	0	0	0	...																										

Figure 12: Full list of features in updated statistics data set

MODELLING

During this stage of the CRISP-DM cycle, the models that would be used to achieve the research objectives are typically chosen, built, and assessed. However, for the purposes of this study the models have already been selected, those being:

1. kNN
2. Decision Trees
3. Random Forest
4. Support Vector Machines
5. Neural Networks
6. Naïve Bayes
7. xGBoost
8. Multinomial Logistic Regression

All of these models are supervised learning models and require being split into a training and testing set. This split is an 80%/20% split in the data, as mentioned in the Modelling section in Chapter 2, to allow the maximum opportunity for learning, and capacity to demonstrate the knowledge the model gained. Additionally, some models require the data to be normalized. This is done to ensure that all the data looks and reads the same across all variables in the data set and allows the models to perform to a higher standard with easier to read and process information. It also helps alleviate data redundancy and duplication. An image showing how values are normalized is viewable in Figure 13.

Data Values	Normalized
12	0
19	12.5
21	16.07
23	19.64
25	23.21
35	41.07
47	62.5
48	64.29
59	83.93
65	94.64
66	96.43
67	98.21
68	100

Figure 13: Depiction of values in a data set would be normalized [53]

Note that the smallest value when normalized changes from 12 to 0, while the largest value changes from 68 to 100. Instead of having values which could mean anything depending on context, the model now better understands what each value in between the largest and smallest value represents, and to what extent they contribute.

MODEL 1.1: KNN – BASE FOOTBALL STATISTICS

Before any analysis or model building can begin, the appropriate packages must be downloaded and installed in R Studio. Packages are built for specific purposes and contain data, documentation and tests for all programming or machine learning models. For the kNN model, the only packages required are the “class” and “gmodels” packages.

The kNN model required the creation of two functions, a normalize and accuracy function. The normalize function was created to normalize the data set as kNN performs to a better standard after this action has been performed. Additionally, a confusion matrix is the best way to visualise the results of the kNN model, as such an accuracy function was created which calculates the accuracy of this matrix. Finally, to get the kNN model to function

correctly, the target variable must be converted into a numeric representation. An away win, a draw and a home win being represented by a 0, 0.5 and 1 respectively.

MODEL 1.2: KNN – UPDATED FOOTBALL STATISTICS

Using the updated statistics data set required minimal tweaking to the model. Mostly, it consisted of tweaking the ranges the normalize and data frame functions needed to be applied to.

MODEL 2.1: DECISION TREE – BASE FOOTBALL STATISTICS

Just as with kNN, Decision Trees require the download and installation of packages before they can be built or pruned. The packages required for this model were “DAAG, party, rpart, rpart.plot, mlbench, caret, pROC, tree and dplyr.”

Decision Trees, DT, proved to be a unique model as there are many ways of implementing a DT, all of which are correct, but result in different accuracy levels. The first model required transforming the target variable into a factor before the DT would function correctly. As with kNN previously, the model was split into training and testing and also utilizes a confusion matrix to view the accuracy of the model. It was also possible to turn the base DT into a more aesthetically pleasing model using the “fancyRpartPlot” function contained within the “rpart” package, this can be viewed below in Figure 28 in Chapter 4.

Upon completion of the first model, the second model, which instead altered the data in its entirety into a data frame, was also able to provide a picture of the decision process, Figure 29, in addition to a text version of the model’s decision process.

MODEL 2.2: DECISION TREE – UPDATED FOOTBALL STATISTICS

Just as with the two originally produced DT, the two DT variants again required separate tuning to function. The first model required no additional tweaks, aside from the conversion of the

target variable, while the second model again required altering the data into a data frame to function correctly.

MODEL 3.1: RANDOM FOREST – BASE FOOTBALL STATISTICS

The packages required for the successful creation of a Random Forest model were “data.table, mlr, tidyverse, xgboost, caret, randomForest, vcd and ROCR.”

Random Forest, RF, as with kNN before it required fine tuning the data set so the model would operate correctly, converting the target variable to a numeric feature. As with Naïve Bayes, the RF packages contain functions to assess the most important factors in its decisions, visible in Figure 31 below in Chapter 4. Additional steps to get the model to run smoother, such as rounding the results of the prediction to more accurately fall in line with the assigned values of 1, 2 and 3 for an away win, draw and home win respectively, had to be taken. This was due to the results lying somewhere between these values and therefore producing a class of its own, meaning the model could not assert whether a prediction was one result or another, rounding these values helped the model function correctly.

MODEL 3.2: RANDOM FOREST – UPDATED FOOTBALL STATISTICS

RF required the same rounding as was needed when using the model on the base statistics for the it to run properly. Without this, separate classes or predictions are created rendering the model unusable. It is possible to view the variables considered the most important with regards to predicting the outcome of any match below in Figure 31 and Figure 32 in Chapter 4.

MODEL 4.1: SUPPORT VECTOR MACHINE – BASE FOOTBALL

STATISTICS

For the successful creation of the Support Vector Machine model, the packages “e1071, rpart, kernlab and caret,” were needed.

The Support Vector Machine, SVM, model required no alteration of the data set before or after the splitting of the data. A training control was needed to be created, using the “trainControl” function, to inform the model by what method it was to use in training and how many times it was to perform this method.

MODEL 4.2: SUPPORT VECTOR MACHINE – UPDATED FOOTBALL

STATISTICS

Built using the same techniques as were used previously; using the same percentage split for training and testing, creating the same training control and utilizing a confusion matrix to view the accuracy of both training and testing, this SVM model required very minimal modifications to function correctly utilizing the updated statistics data set.

MODEL 5.1: NEURAL NETWORK – BASE FOOTBALL STATISTICS

Requiring the fewest packages, Neural Networks required only the installation of the “neuralnet” package.

As with kNN, the Neural Network model, NN, requires a normalize function in order to be able to operate at full capacity. After converting the data into a data frame, just as in kNN the model requires returning the result as a number, as opposed to a letter of A, D or H, instead replacing these values with 0, 0.5 and 1 respectively.

MODEL 5.2: NEURAL NETWORK – UPDATED FOOTBALL STATISTICS

NN was created using the same steps that were used in the base statistics model. The target variable was converted to a number, the data set was morphed into a data frame and then normalized.

MODEL 6.1: NAÏVE BAYES – BASE FOOTBALL STATISTICS

Naïve Bayes required a plethora of packages during creation for the model to function correctly. These packages were “tidyverse, ggplot2, caret, caretEnsemble, psych, Amelia, mice, Ggally, rpart, randomForest, e1071, klaR, naivebayes and dplyr.”

Naïve Bayes, NB, required no fine tuning of the base statistics data set. Once the data had been split into the training and testing sets, a new training control was formed, and the model could be created. The NB packages in R Studio do contain a useful feature that allows the user to assess the most important variables for each result or classification, visible in Figures 33 and 34 below.

MODEL 6.2: NAÏVE BAYES – UPDATED FOOTBALL STATISTICS

The NB model was similarly created to the previous model. Just as performed with the base statistics model it is possible to view the factors the model considered the most important. Due to their being significantly more variables this time, however, only the text form of these variables will be shown. Figure 35 contains the updated list of variables the new NB model utilized in order to derive its predictions.

MODEL 7.1: XGBOOST – BASE FOOTBALL STATISTICS

xGBoost required the “data.table, mlr, tidyverse, xgboost and caret” packages for a model to be constructed.

xGBoost, as some of the models prior, required converting the target variable to a numeric representation of itself. Once again, this results in 0, 0.5 and 1 as a representation for an away win, draw and home win respectively. Once split into the training and testing set, additional steps were needed to be performed. Each training and testing set needed to be converted into a matrix and after this converted into a specific xGBoost matrix so the model can run.

MODEL 7.2: XGBOOST – UPDATED FOOTBALL STATISTICS

xGBoost was created in the same manner as it was for the base statistics data set. The data was split into a training and testing data set, with these subsets then converted into a matrix and then morphed into a specific xGBoost matrix.

MODEL 8.1: MULTINOMIAL LOGISTIC REGRESSION – BASE FOOTBALL STATISTICS

Multinomial Logistic Regression (MLR) models require only the “readr, caret and dplyr” packages to allow a functional model to be created.

MLR required altering the target variable into a factor. Upon splitting the data, the data sets were able to be run through a prediction function to assess the accuracy of the model.

MODEL 8.2: MULTINOMIAL LOGISTIC REGRESSION – UPDATED FOOTBALL STATISTICS

The MLR model was created by changing the target variable to a factor and the partitioning in the same way as was performed in the base model.

EVALUATION

This evaluation stage is aimed finding the model that best serves the objectives of the study. Prior to this phase, in the previous modelling stage, each model was created and contrasted, with an accuracy rating attributed to each. In most cases, the best way to view this accuracy was with a confusion matrix.

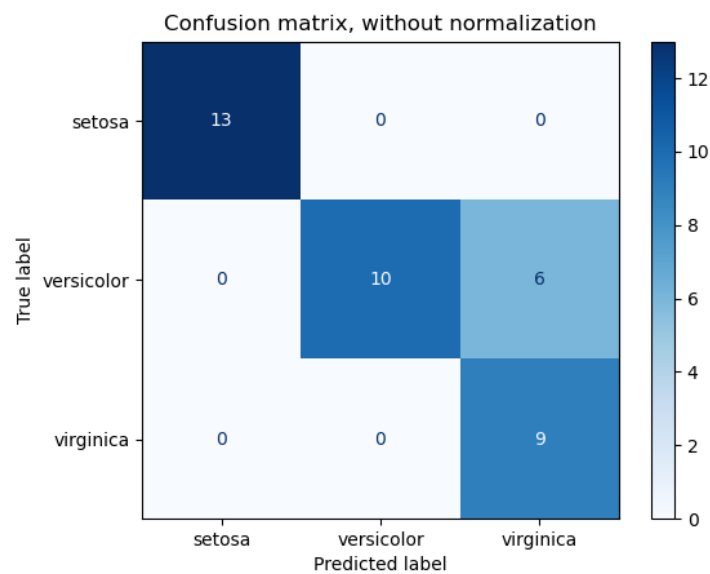


Figure 14: Image of a confusion matrix utilizing the Iris data set [54]

Figure 14 is a confusion matrix for the Iris data set and was selected as it incorporates the same number of classifications as this study. The numbers along the diagonal of the matrix, 13, 10 and 9, represent the number of data points for which the predicted value and the real value were the same. This is otherwise known as correct prediction based on the data the model had to work with. Any number that is not along the diagonal, represents a misclassified data point and is, a wrong prediction. The higher the values along the diagonal, compared to those that are not along the diagonal, would indicate more correct decisions have been made, therefore the overall model is accurate and makes correct predictions.

The overall accuracy of a confusion matrix is calculated as the number of correct predictions, values on the diagonal, divided by the total number of predictions, all values in the matrix. This can also be interpreted as:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$
$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}$$

Figure 15: TP = True Positives, TN = True Negatives, FP = False Positives and FN = False Negatives

DEPLOYMENT

Deployment can vary from study to study, in some cases it can take the form of a report whilst in some it may take on the form of a repeatable process. In the context of this study, deployment is centred around implementing repeatable predictions utilizing the model selected, kNN. As such, the deployment section is focused on the ability of the author to utilize the created model, with the learning it has accrued, and apply it to the final test data set.

The author will take the model that has been created and will find the best method to incorporate the final test statistics data set. This data set is the set that has been altered to reflect the knowledge that would be known prior to a football match starting, so as to assess the true prediction power of the model selected. This implementation process will need to be repeatable so that if any new information comes to light, or more seasons of data were to be added, the prediction process can be run again.

SUMMARY

At the end of this chapter, the author has established three key research objectives, those objectives being:

1. Create a machine learning model that can predict the result of any given Premier League game utilizing standard in game statistics

2. Create a machine learning model that can predict the result of any given Premier League game utilizing standard in game statistics and also additional statistics and psychological factors
3. Assess whether the model which performs best in Objective 2 can be used to predict the results of the most recent Premier League season

By this point, two objectives have been achieved with the third and final objective to be accomplished in the coming chapters. Both Objective 1 and Objective 2 were fulfilled upon the successful build of each model for the base statistics and then the updated statistics.

Additionally, the positivist research approach was also decided upon. This was important as it helps guide the study and ensure the author remains impartial on all research, allowing all findings to be objective with no bias. It ensured the use of purely quantitative data, facts and figures, and removed the option of incorporating qualitative, or interpretation based, data. This allows the study to be replicable, a core concept of the paper, ensuring the models produced can be used season after season while only requiring updating the data source.

After the approach to the research was asserted, the data sets were cleaned and made ready for purpose. For the base statistics, this included ensuring all naming traditions were similar between data sets and the removal of variables that either provide too much information, such as the number of goals scored at full time, or provide no actionable information, such as the division or referee. For the updated statistics this process included the same steps as performed on the base data set, but it was also required to include all additional variables the data set required. This included adding in player ratings for each position on the pitch, the form over the previous five games and goal difference amongst other features. This was a meticulous process but allows the model to incorporate more than base statistics and learn from these additional statistics.

The models were then created and run. All eight models were created twice, totalling sixteen models: eight for the base statistics data set and eight for the updated statistics data set. Lastly, the basis upon which each model will be assessed was described, whereby the accuracy of each model is the focus. The greater the accuracy, the more likely the model will be in successfully predicting the final results of the test statistics data set which none of the models have seen.

CHAPTER 4 – ANALYSIS OF FINDINGS

This chapter presents the results and analysis of the methods chosen in the previous chapter, and will allow the author to select which model will be used for the ultimate aim of this study, prediction. Utilizing the CRISP-DM methodology, as discussed and used previously, where it applies, Chapter 4 is concerned with the prediction capabilities and aspects of this study.

INTRODUCTION

The main objective of this study is to perform predictions of football matches utilizing both psychological and non-psychological factors. Before this can be done, the model that is best suited for this task needs to be selected. As discussed in Chapter 3, the overall accuracy of each model is what determines its capabilities for successful predictions. As such, utilizing the confusion matrix, as described previously, or other methods, where a confusion matrix cannot be implemented, the overall accuracy of each model will be determined and ranked.

DATA UNDERSTANDING

At this stage in the study, the analysis provided by the models has returned some meaningful insights. Before moving to the modelling, understanding what the data has returned is key.

BASE FOOTBALL STATISTICS STRUCTURE

Once again, the Base Football Statistics data set is comprised of seven-hundred and sixty unique observations with fifty-nine features combining to form 44,840 individual data points, once the superfluous columns were removed. The data set's variables are all numbers, although some of these are morphed into factors in some models. Other models require moulding the data set into other forms like matrices or data frames. The structure is visible below in Figure 16.

\$ FTR	: num	0 1 1 0 0 0 0 0 1 1 ...	\$ Watford.H	: num	0 0 0 0 0 0 0 0 0 0 ...
\$ HS	: num	0.133 0.133 0.7 0.467 0.2 ...	\$ West.Brom.H	: num	0 0 0 0 1 0 0 0 0 0 ...
\$ AS	: num	0.444 0.296 0.185 0.519 0.444 ...	\$ West.Ham.H	: num	0 0 0 1 0 0 0 0 0 0 ...
\$ HST	: num	0.1429 0.2143 0.4286 0.2143 0.0714 ...	\$ Wolves.H	: num	0 0 0 0 0 0 0 0 0 0 ...
\$ AST	: num	0.4 0.333 0.2 0.133 0.467 ...	\$ Arsenal.A	: num	1 0 0 0 0 0 0 0 0 0 ...
\$ HF	: num	0.522 0.609 0.391 0.565 0.522 ...	\$ Aston.Villa.A	: num	0 0 0 0 0 0 0 0 0 0 ...
\$ AF	: num	0.478 0.435 0.217 0.261 0.348 ...	\$ Bournemouth.A	: num	0 0 0 0 0 0 0 0 0 0 ...
\$ HC	: num	0.118 0.412 0.529 0.471 0.118 ...	\$ Brighton.A	: num	0 0 0 0 0 0 0 0 0 0 ...
\$ AC	: num	0.188 0.188 0 0.438 0.312 ...	\$ Burnley.A	: num	0 0 0 0 0 0 0 0 0 0 ...
\$ HY	: num	0.333 0.333 0.167 0.333 0.167 ...	\$ Chelsea.A	: num	0 0 0 0 0 0 1 0 0 0 ...
\$ AY	: num	0.333 0.167 0 0.333 0.167 ...	\$ Crystal.Palace.A	: num	0 0 0 0 0 0 0 0 0 0 ...
\$ HR	: num	0 0 0 0 0 0 0 0 0 0 ...	\$ Everton.A	: num	0 0 0 0 0 1 0 0 0 0 ...
\$ AR	: num	0 0 0 0 0 0 0 0 0.5 0 ...	\$ Fulham.A	: num	0 0 0 0 0 0 0 0 0 1 ...
\$ Arsenal.H	: num	0 0 0 0 0 0 0 0 0 0 ...	\$ Leeds.A	: num	0 0 1 0 0 0 0 0 0 0 ...
\$ Aston.Villa.H	: num	0 0 0 0 0 0 0 0 0 0 ...	\$ Leicester.A	: num	0 0 0 0 1 0 0 0 0 0 ...
\$ Bournemouth.H	: num	0 0 0 0 0 0 0 0 0 0 ...	\$ Liverpool.A	: num	0 0 0 0 0 0 0 0 0 0 ...
\$ Brighton.H	: num	0 0 0 0 0 0 1 0 0 0 ...	\$ Man.City.A	: num	0 0 0 0 0 0 0 0 0 0 ...
\$ Burnley.H	: num	0 0 0 0 0 0 0 0 0 0 ...	\$ Man.United.A	: num	0 0 0 0 0 0 0 0 0 0 ...
\$ Chelsea.H	: num	0 0 0 0 0 0 0 0 0 0 ...	\$ Newcastle.A	: num	0 0 0 1 0 0 0 0 0 0 ...
\$ Crystal.Palace.H	: num	0 1 0 0 0 0 0 0 0 0 ...	\$ Norwich.A	: num	0 0 0 0 0 0 0 0 0 0 ...
\$ Everton.H	: num	0 0 0 0 0 0 0 0 1 0 ...	\$ Sheffield.United.A	: num	0 0 0 0 0 0 0 0 0 0 ...
\$ Fulham.H	: num	1 0 0 0 0 0 0 0 0 0 ...	\$ Southampton.A	: num	0 1 0 0 0 0 0 0 0 0 ...
\$ Leeds.H	: num	0 0 0 0 0 0 0 0 0 1 ...	\$ Tottenham.A	: num	0 0 0 0 0 0 0 0 0 0 ...
\$ Leicester.H	: num	0 0 0 0 0 0 0 0 0 0 ...	\$ Watford.A	: num	0 0 0 0 0 0 0 0 0 0 ...
\$ Liverpool.H	: num	0 0 1 0 0 0 0 0 0 0 ...	\$ West.Brom.A	: num	0 0 0 0 0 0 0 0 1 0 ...
\$ Man.City.H	: num	0 0 0 0 0 0 0 0 0 0 ...	\$ West.Ham.A	: num	0 0 0 0 0 0 0 0 0 0 ...
\$ Man.United.H	: num	0 0 0 0 0 0 0 0 0 0 ...	\$ Wolves.A	: num	0 0 0 0 0 0 0 1 0 0 ...
\$ Newcastle.H	: num	0 0 0 0 0 0 0 0 0 0 ...			
\$ Norwich.H	: num	0 0 0 0 0 0 0 0 0 0 ...			
\$ Sheffield.United.H	: num	0 0 0 0 0 0 0 1 0 0 ...			
\$ Southampton.H	: num	0 0 0 0 0 0 0 0 0 0 ...			
\$ Tottenham.H	: num	0 0 0 0 0 1 0 0 0 0 ...			

Figure 16: Structure of base statistics data set

UPDATED FOOTBALL STATISTICS STRUCTURE

After the columns that provide no value were removed, the Updated Football Statistics data set is comprised of seven-hundred and sixty unique observations with eighty-three features to form a data set containing 63,080 individual data points. Just as with the Base Football Statistics, the data set's variables are all numbers, although some of these are morphed into factors in some models. Other models require moulding the data set into other forms like matrices or data frames. This structure is visible below in Figure 17.

```

$ FTR      : num 0 1 1 0 0 0 0 0 1 1 ...
$ HS       : num 0.133 0.133 0.7 0.467 0.2 ...
$ HST      : num 0.1429 0.2143 0.4286 0.2143 0.0714 ...
$ HF       : num 0.522 0.609 0.391 0.565 0.522 ...
$ HC       : num 0.118 0.412 0.529 0.471 0.118 ...
$ HY       : num 0.333 0.333 0.167 0.333 0.167 ...
$ HR       : num 0 0 0 0 0 0 0 0 0 ...
$ HP       : num 0.435 0.194 0.484 0.645 0.306 ...
$ AS       : num 0.444 0.296 0.185 0.519 0.444 ...
$ AST      : num 0.4 0.333 0.2 0.133 0.467 ...
$ AF       : num 0.478 0.435 0.217 0.261 0.348 ...
$ AC       : num 0.188 0.188 0 0.438 0.312 ...
$ AY       : num 0.333 0.167 0 0.333 0.167 ...
$ AR       : num 0 0 0 0 0 0 0 0.5 0 ...
$ AP       : num 0.565 0.806 0.516 0.355 0.694 ...
$ ArsenalH : num 0 0 0 0 0 0 0 0 0 ...
$ AstonVillaH : num 0 0 0 0 0 0 0 0 0 ...
$ BournemouthH : num 0 0 0 0 0 0 0 0 0 ...
$ BrightonH : num 0 0 0 0 0 1 0 0 0 ...
$ BurnleyH : num 0 0 0 0 0 0 0 0 0 ...
$ ChelseaH : num 0 0 0 0 0 0 0 0 0 ...
$ CrystalPalaceH : num 0 1 0 0 0 0 0 0 0 ...
$ EvertonH : num 0 0 0 0 0 0 0 0 1 ...
$ FulhamH : num 1 0 0 0 0 0 0 0 0 ...
$ LeedsH : num 0 0 0 0 0 0 0 0 1 ...
$ LeicesterH : num 0 0 0 0 0 0 0 0 0 ...
$ LiverpoolH : num 0 0 1 0 0 0 0 0 0 ...
$ ManCityH : num 0 0 0 0 0 0 0 0 0 ...
$ ManUnitedH : num 0 0 0 0 0 0 0 0 0 ...
$ NewcastleH : num 0 0 0 0 0 0 0 0 0 ...
$ NorwichH : num 0 0 0 0 0 0 0 0 0 ...
$ SheffieldUnitedH : num 0 0 0 0 0 0 1 0 0 ...
$ SouthamptonH : num 0 0 0 0 0 0 0 0 0 ...
$ TottenhamH : num 0 0 0 0 0 1 0 0 0 ...
$ WatfordH : num 0 0 0 0 0 0 0 0 0 ...
$ WestBromH : num 0 0 0 0 1 0 0 0 0 ...
$ WestHamH : num 0 0 0 1 0 0 0 0 0 ...
$ WolvesH : num 0 0 0 0 0 0 0 0 0 ...
$ ArsenalA : num 1 0 0 0 0 0 0 0 0 ...
$ AstonVillaA : num 0 0 0 0 0 0 0 0 0 ...
$ BournemouthA : num 0 0 0 0 0 0 0 0 0 ...
$ BrightonA : num 0 0 0 0 0 0 0 0 0 ...
$ BurnleyA : num 0 0 0 0 0 0 0 0 0 ...
$ ChelseaA : num 0 0 0 0 0 1 0 0 0 ...
$ CrystalPalaceA : num 0 0 0 0 0 0 0 0 0 ...
$ EvertonA : num 0 0 0 0 0 1 0 0 0 ...
$ FulhamA : num 0 0 0 0 0 0 0 0 1 ...
$ LeedsA : num 0 0 1 0 0 0 0 0 0 ...
$ LeicesterA : num 0 0 0 0 1 0 0 0 0 ...
$ LiverpoolA : num 0 0 0 0 0 0 0 0 0 ...
$ ManCityA : num 0 0 0 0 0 0 0 0 0 ...
$ ManUnitedA : num 0 0 0 0 0 0 0 0 0 ...
$ NewcastleA : num 0 0 0 1 0 0 0 0 0 ...
$ NorwichA : num 0 0 0 0 0 0 0 0 0 ...
$ SheffieldUnitedA : num 0 0 0 0 0 0 0 0 0 ...
$ SouthamptonA : num 0 1 0 0 0 0 0 0 0 ...
$ TottenhamA : num 0 0 0 0 0 0 0 0 0 ...
$ WatfordA : num 0 0 0 0 0 0 0 0 0 ...
$ WestBromA : num 0 0 0 0 0 0 0 0 1 ...
$ WestHamA : num 0 0 0 0 0 0 0 0 0 ...
$ WolvesA : num 0 0 0 0 0 0 1 0 0 ...
$ H_GK : num 0.36 0.6 0.96 0.64 0.48 0.88 0.48 0.36 0.6 0.4 ...
$ H_Def : num 0.3682 0.4229 0.6766 0.4975 0.0299 ...
$ H_Avg_Def : num 0.382 0.492 1 0.641 0.422 ...
$ H_Mid : num 0.591 0.423 0.28 0.686 0.812 ...
$ H_Avg_Mid : num 0.213 0.315 0.584 0.44 0.251 ...
$ H_Avg_Att : num 0.01 0.445 0.975 0.055 0.025 0.105 0.045 0.38 0.88 0.05 ...
$ H_Subst : num 0.0909 0.4773 0.8941 0.5 0.2273 ...
$ H_Avg_Subst : num 0.26 0.269 0.373 0.33 0.199 ...
$ A_GK : num 0.0624 0.0646 0.0894 0.0792 0.0478 ...
$ A_Def : num 0.76 0.4 0.4 0.44 0.8 0.6 0.72 0.72 0.48 0.68 ...
$ A_Avg_Def : num 0.102 0.444 0.439 0.423 0.531 ...
$ A_Mid : num 0.195 0.135 0.13 0.114 0.223 ...
$ A_Avg_Mid : num 0.439 0.417 0.63 0.425 0.68 ...
$ A_Att : num 0.618 0.545 0.557 0.572 0.689 ...
$ A_Avg_Att : num 0.6464 0.3286 0.0464 0.3071 0.0786 ...
$ A_Subst : num 0.843 0.778 0.769 0.722 0.935 ...
$ A_Avg_Subst : num 0.324 0.231 0.24 0.303 0.335 ...
$ Five_Game_Form_H : num 0.0314 0.0217 0.0226 0.0293 0.0326 ...
$ Five_Game_Form_A : num 0 0 0 0 0 0 0 0 2 0 ...
$ Rev_Fixture_H : num 0 0 0 0 0 0 0 0 0 ...
$ Rev_Fixture_A : num 0 0 0 0 0 0 0 0 0 ...
$ GD_H : num 0.407 0.407 0.407 0.407 0.407 ...
$ GD_A : num 0.423 0.423 0.423 0.423 0.423 ...
$ Maj_Suspension_H : num 0 0 0 0 0 0 0 0 0 ...
$ Min_Suspension_H : num 0 0 0 0 0 0 0 0 0 ...
$ Maj_Suspension_A : num 0 0 0 0 0 0 0 0 0 ...
$ Min_Suspension_A : num 0 0 0 0 0 0 0 0 0 ...

```

Figure 17: Structure of updated statistics data set

NULL VALUES

The base statistics data set was tested to see if any of the columns that were kept contained any null or missing values using a simple function in R Studio, viewable below in Figure 18. The updated statistics data set was created by the author, but was tested with the same code just to be certain, and also contained no null or missing values.

```

> sum(is.na(data))
[1] 0
> sum(is.na(data2))
[1] 0

```

Figure 18: Code to check for NA/Null/Missing values

OUTLIERS

As mentioned previously, there were two matches that raised curiosity as they appeared incorrect producing scores that were represented as outliers. Shown below in Figure 19 are matches where there were eight home goals scored in 2019/20 and nine home goals scored in 2020/21. These figures are so far removed from the average of approximately three and four respectively, that they needed to be double checked to assess whether an input error had occurred. Upon further research, both of these scores were in fact legitimate with Manchester City having scored eight without reply against Watford in the 2019/20 season, while Manchester United scored nine against Southampton in the 2020/21 season.

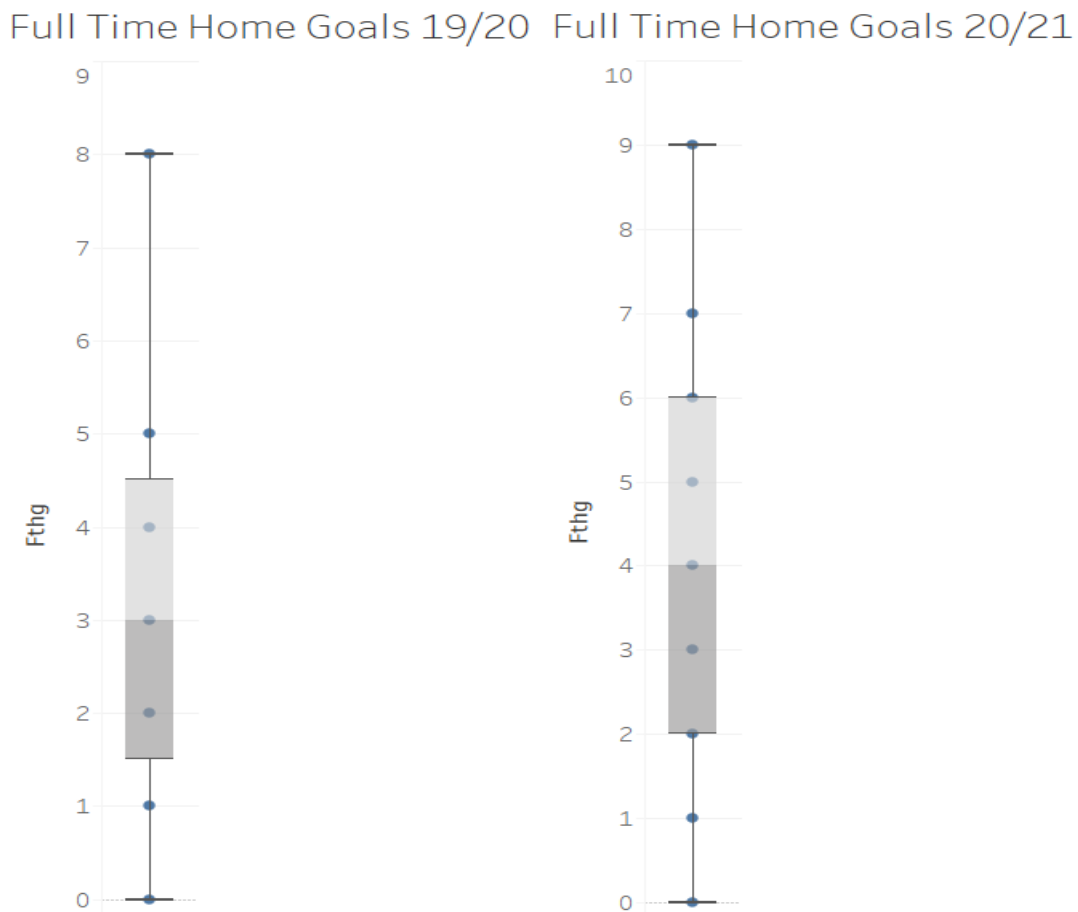


Figure 19: Box plot of outliers for Full-Time Home Goals

SET-PIECES

The image below in Figure 20 is taken from the base statistics data set. Of the variables it shows, aside from Liverpool or Manchester City participating in the match, set-pieces are more integral to determining the winner than who is actually playing. This is valuable information as when it comes to the prediction phase of the study, the model will use statistics such as these instead of relying on the inherent strength of any given squad.

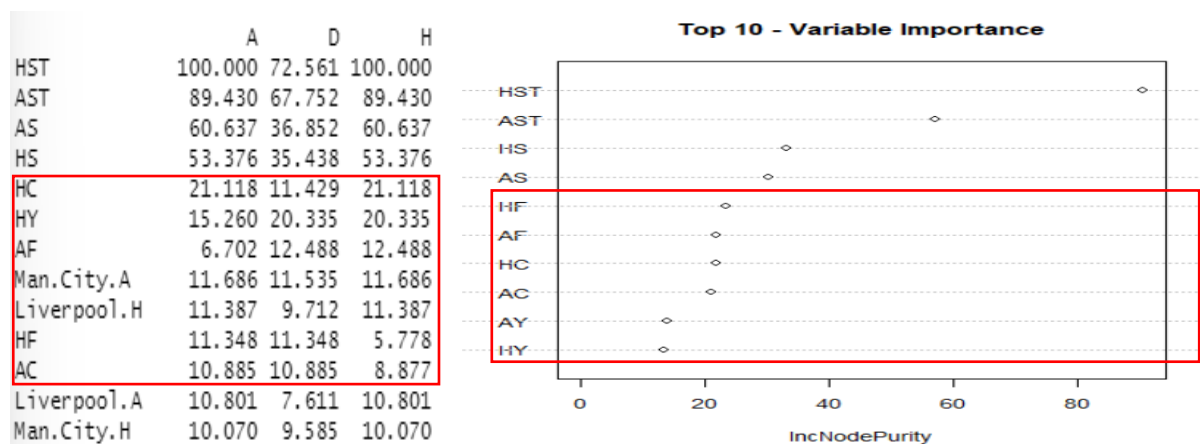


Figure 20: Highlighting importance of set-pieces

Set-pieces are taking on an increased importance in the modern game, regardless of position on the pitch. With high values in corners and free kicks being used to determine the most likely winner between two teams. Information such as this could dictate transfer policies and tactics as clubs seek to benefit from set piece routines. This could also sway the way people would bet on a team to win, adjusting a bet if one team has a high amount of set piece goals or another has conceded from a lot of set-pieces for example. (Data that was unavailable and is not included in the data set.)

DEFENCE AND STABILITY OVER OFFENCE

A stronger defensive presence on the pitch was more associated with winning than having a stronger attack. This suggests providing a strong platform from which the team can build upon

without conceding goals, is better than committing to attack and attempting to outscore opponents. Similarly, the calibre of midfield player was more influential in a win than the calibre of attacking player. Meaning gaining control of the momentum of a game and having better players who can contribute in both ends of the pitch is more important than having the best striker to finish chances. This is demonstrated below in Figure 21 where in both charts each team's defence and midfield is cited as more important than either team's respective attackers.

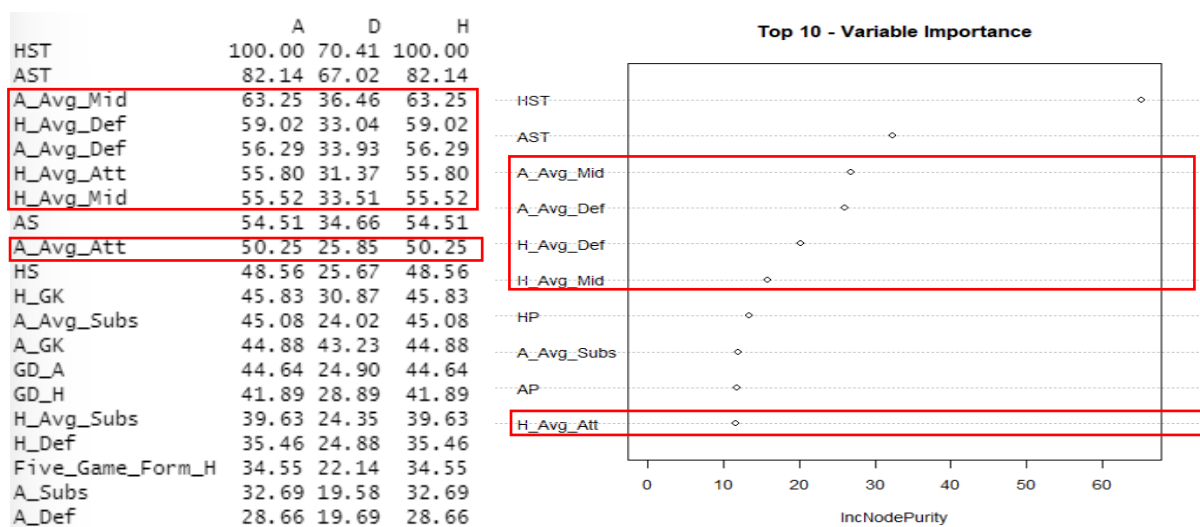


Figure 21: Highlighting how important the defence and midfield is compared to attack

Further to this, Liverpool and Manchester City are the most successful teams over the last five years. It is telling that both teams contain the strongest midfield and defence in the league. With players like Virgil van Dijk and Ruben Dias in their backlines nullifying even the strongest of forwards such as Heung Min Son or Cristiano Ronaldo, widely considered the best player in the world.

RATIO OF RESULTS

Both data sets can provide a breakdown of the full-time results. Figure 22 shows that a home win is approximately 14.9% more likely than an away win, showing that having home advantage is a real phenomenon in football. This reinforces the authors decision to try to

account for this via one hot encoding and would insinuate that the Covid-19 pandemic could have altered a lot of results with no home support in the stadium. A draw makes up almost a quarter of all results, but is approximately 34.9% less likely than an away win and 44.6% less likely than a home win.

```
> prop.table(table(data$FTR)) * 100
      A      D      H
35.39474 23.02632 41.57895
```

Figure 22: Break down of the results of the Premier League over two seasons

ACCURACY OVER QUANTITY

No matter what data set is being analysed, having shots on target, for either the home or away side, is the most important metric when attempting to predict who the winner of any match will be.

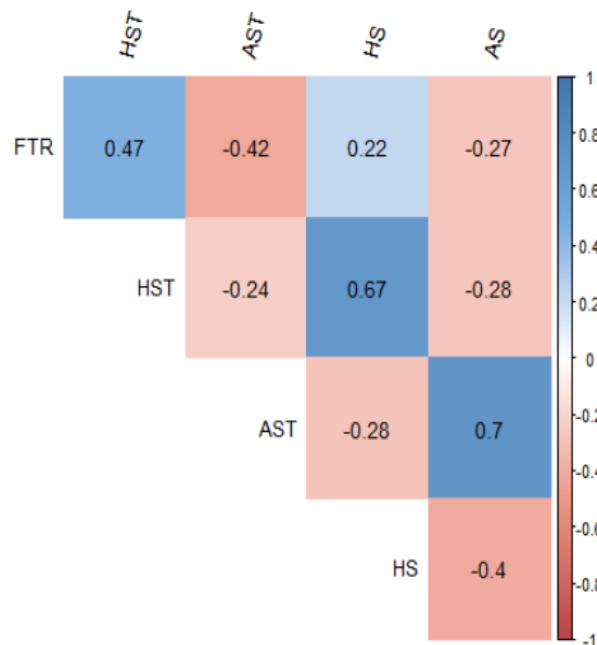


Figure 23: Correlation between shots and shots on target and winning

Figure 23 above shows that having shots that actually test the goalkeeper, or shots on target, have almost twice the correlation with a win than simply shooting. This would suggest that a more patient, or methodical approach to any match is statistically better for a team looking to

win. Holding off from shooting from distance and crafting clear cut opportunities is the clearest path to winning. While it could be argued that a better calibre of attacker would be more accurate with their shooting, this observation more so backs up the importance of the defence and midfield. The defence are tasked with stopping the clear-cut chances, while the midfield assists in this while also attempting to create these higher expected goal (xG) chances.

AVERAGE PLAYER QUALITY IN THE PREMIER LEAGUE

While the data seems to point to defenders being more important to winning than attackers, it would seem that teams still prioritise signing the star forwards. Prioritising attacking players that score goals, sell shirts and entice the crowds has led to the average attacker having an overall of 79.2 compared to the defence being 77.5. This shows that teams are unaware of the advantage they could gain from purchasing a higher quality defender over a higher quality forward, believing that a greater goal scoring potential is more important than stopping than those players.

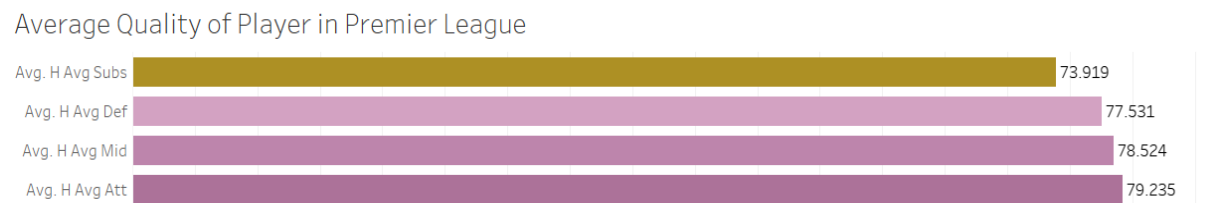


Figure 24: Breakdown of average quality of player in defence, midfield and attack

DATA PREPARATION

The main preparations to be performed at this stage now the data has been cleansed are as follows:

TRAINING AND TESTING

As mentioned in the Literature Review, training and testing was always an 80% to 20% split in the data. With seven-hundred and sixty observations, two full seasons worth of matches, the data takes on the sizes depicted in Table 1 below.

	Number of Observations	Number of Features	Unique Data Points	Total Unique Data Points
Training Base Football Statistics	608	59	35,872	
Testing Base Football Statistics	152	59	8,968	44,840
Training Updated Football Statistics	608	83	50,464	
Testing Updated Football Statistics	152	83	12,616	63,080
Training Before Removal of Unnecessary Columns	608	106	64,448	
Testing Before Removal of Unnecessary Columns	152	106	16,112	80,560

Table 1: Breakdown of size of data sets before and after splitting

This table shows how large the potential for learning is when compared to the whole data set, and how much smaller the testing set is. This could explain the drop off in accuracy experienced by some models.

NORMALISATION

For some models to work, the data needs to be normalized. This transforms the data from the figures reported in the match, Figure 25, to the normalised version in Figure 26. The best example of how this normalisation works is represented in the AF column in rows ten and thirteen. Eighteen away fouls is quite a high amount and this is indicated by the value of .739. The closer the value is to 1, the higher the base value. This is reflected in AF row fifteen as the value of six is quite low and therefore returns a value of .217.

	FTR	HS	HST	HF	HC	HY	HR	HP	AS	AST	AF	AC	AY	AR	AP
1	A	5	2	12	2	2	0	45	13	6	12	3	2	0	55
2	H	5	3	14	7	2	0	30	9	5	11	3	1	0	70
3	H	22	6	9	9	1	0	48	6	3	6	0	0	0	52
4	A	15	3	13	8	2	0	58	15	2	7	7	2	0	42
5	A	7	1	12	2	1	0	37	13	7	9	5	1	0	63
6	A	9	5	15	5	1	0	51	15	4	7	3	0	0	49
7	A	13	3	8	4	1	0	52	10	5	13	3	0	0	48
8	A	9	2	13	12	2	0	55	11	4	7	5	1	0	45
9	H	17	7	9	11	1	0	72	6	4	11	1	0	1	28
10	H	10	7	13	5	1	0	50	14	6	18	3	2	0	50
11	A	13	4	13	9	2	0	76	14	5	10	3	1	0	24
12	H	7	3	11	7	0	0	63	14	3	13	5	1	0	37
13	A	14	7	16	3	4	0	53	9	6	18	2	3	0	47
14	A	6	0	16	7	3	0	47	13	6	15	1	0	1	53
15	A	5	3	10	1	0	1	39	18	6	6	11	0	0	61

Figure 25: Data before normalization

	FTR	HS	HST	HF	HC	HY	HR	HP	AS	AST	AF	AC
1	0.0	0.13333333	0.14285714	0.5217391	0.11764706	0.33333333	0.0	0.43548387	0.44444444	0.40000000	0.47826087	0.1875
2	1.0	0.13333333	0.21428571	0.6086957	0.41176471	0.33333333	0.0	0.19354839	0.29629630	0.33333333	0.43478261	0.1875
3	1.0	0.70000000	0.42857143	0.3913043	0.52941176	0.16666667	0.0	0.48387097	0.18518519	0.20000000	0.21739130	0.0000
4	0.0	0.46666667	0.21428571	0.5652174	0.47058824	0.33333333	0.0	0.64516129	0.51851852	0.13333333	0.26086957	0.4375
5	0.0	0.20000000	0.07142857	0.5217391	0.11764706	0.16666667	0.0	0.30645161	0.44444444	0.46666667	0.34782609	0.3125
6	0.0	0.26666667	0.35714286	0.6521739	0.29411765	0.16666667	0.0	0.53225806	0.51851852	0.26666667	0.26086957	0.1875
7	0.0	0.40000000	0.21428571	0.3478261	0.23529412	0.16666667	0.0	0.54838710	0.33333333	0.33333333	0.52173913	0.1875
8	0.0	0.26666667	0.14285714	0.5652174	0.70588235	0.33333333	0.0	0.59677419	0.37037037	0.26666667	0.26086957	0.3125
9	1.0	0.53333333	0.50000000	0.3913043	0.64705882	0.16666667	0.0	0.87096774	0.18518519	0.26666667	0.43478261	0.0625
10	1.0	0.30000000	0.50000000	0.5652174	0.29411765	0.16666667	0.0	0.51612903	0.48148148	0.40000000	0.73913043	0.1875
11	0.0	0.40000000	0.28571429	0.5652174	0.52941176	0.33333333	0.0	0.93548387	0.48148148	0.33333333	0.39130435	0.1875
12	1.0	0.20000000	0.21428571	0.4782609	0.41176471	0.00000000	0.0	0.72580645	0.48148148	0.20000000	0.52173913	0.3125
13	0.0	0.43333333	0.50000000	0.6956522	0.17647059	0.66666667	0.0	0.56451613	0.29629630	0.40000000	0.73913043	0.1250
14	0.0	0.16666667	0.00000000	0.6956522	0.41176471	0.50000000	0.0	0.46774194	0.44444444	0.40000000	0.60869565	0.0625
15	0.0	0.13333333	0.21428571	0.4347826	0.05882353	0.00000000	0.5	0.33870968	0.62962963	0.40000000	0.21739130	0.6875

Figure 26: Data after normalization

CONVERSIONS

Certain models require the data set to be moulded so that the model can function to the best of its ability. These conversions take the forms of:

- **Data Frames:** Converts data into tabular data which allows the machine learning model to learn from it and manipulate it.
- **Numeric:** Converts the target variable into a numeric representation. This is done due to some models only being able to function when predicting a numeric value.
- **Factor:** Factorising the target variable converts the data to a vector. Similarly to the numeric function, this is done to convert string categorical variables into numeric values that allow the models to function.
- **xGB Matrix:** An optimized matrix specifically for xGBoost that increases efficiency and training speed.

MODELLING

Modelling the data now centres on collating and interpreting all of the accuracy metrics that were returned, as opposed to creating and editing the models. It must be noted that where a confusion matrix could be used, it was used. Additionally, multiple methods of incorporating a confusion matrix were used depending on the model, correlation for example.

```
Total Observations in Table: 153
```

data_test_labels	data_test_pred			Row Total
	0	0.5	1	
0	42 0.955 0.808 0.275	2 0.045 0.095 0.013	0 0.000 0.000 0.000	44 0.288
0.5	8 0.229 0.154 0.052	18 0.514 0.857 0.118	9 0.257 0.112 0.059	35 0.229
1	2 0.027 0.038 0.013	1 0.014 0.048 0.007	71 0.959 0.887 0.464	74 0.484
Column Total	52 0.340	21 0.137	80 0.523	153

Confusion Matrix and Statistics

```
Reference
Prediction A D H
A 20 4 4
D 0 0 0
H 33 31 59
```

Overall Statistics

Accuracy : 0.5232

[1] 85.62092

Figure 27: An example of the different forms a confusion matrix can take

Some models return a tidy accuracy metric, seen on the right-hand side of Figure 27, while some return the accuracy as a percentage, as seen on the left-hand side of Figure 27. Other models, such as those that use correlation to determine accuracy, produce the accuracy metric as a decimal which requires the value to be multiplied 100 to determine the overall accuracy.

MODEL 1.1: KNN – BASE FOOTBALL STATISTICS

As mentioned before, the best way to view the results of the kNN model is with a confusion matrix. The X and Y axis of this confusion matrix are made up of the training and testing partitions in the data. This matrix shows the user what was predicted, what the true result was and the accuracy of the prediction for each result in the form of a decimal. A k value of 19 was also selected. This is due to the fact that the best k values are typically an odd number close to the square root of the data set being used, with the value being odd so as to prevent a “classification tie.” A classification tie being where the model has produced an even number of nearest neighbours for one classification, and an even number of nearest neighbours for another classification. This means the model is unsure which classification to place the data point and this can decrease accuracy. To avoid this an odd number is used forcing the model to choose which classification best suits the data point. An accuracy function was created which calculates the sum of the diagonal of the matrix, divides this value by the sum of the rows and multiplies it by 100 to return a percentage value for accuracy. kNN also requires the target variable to be converted into a numeric representation for the model to work correctly, as such, in this model an away win, a draw and a home win are represented by a 0, 0.5 and 1 respectively. The model returns an approximate 85% accuracy.

MODEL 1.2: KNN – UPDATED FOOTBALL STATISTICS

The model was tweaked as to incorporate the larger number of variables. The same k value of 19 was applied to this model also as the length of the data set, or number of observations or data points, remained the same. After running the new model an accuracy of approximately 75% was found.

MODEL 2.1: DECISION TREE – BASE FOOTBALL STATISTICS

The first model produced an accuracy of approximately 55%, this model is viewable below in Figure 28. The “fancyRpartPlot” function has been applied to the tree as was mentioned in the previous chapter.

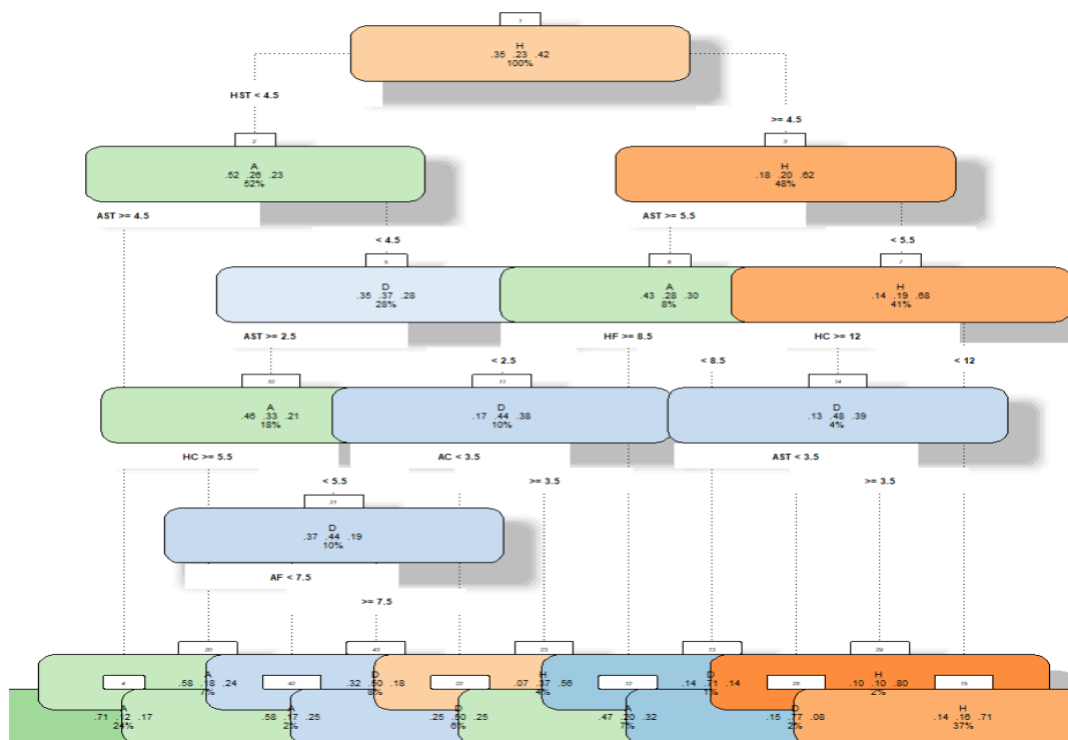


Figure 28: Example of a DT using the “fancyRpartPlot” function

The second model, which instead altered the data in its entirety into a data frame, was also able to provide a text version of the model’s decision process. Coupled with the standard depiction of a DT, it is easy to understand the logic the DT is using to get to the results that is predicting

for each match. These can both be viewed in Figure 29 below. This DT model produced an accuracy of just over 63%.

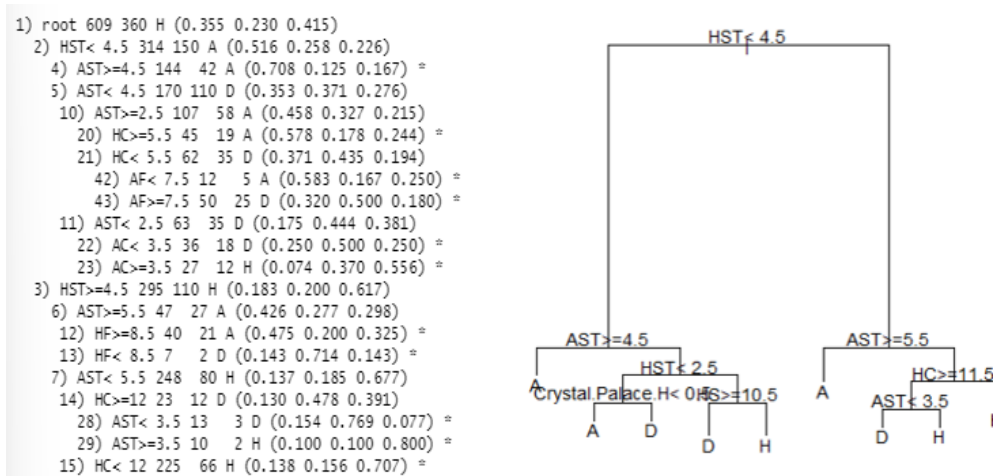


Figure 29: The logic and splitting operations of the more successful DT model for the base statistics data set

The left side of Figure 29 is the text form of the DT. Starting with the statistically most likely outcome, a home win as shown by the “H”, coupled with the probability of it being this result at 41.5%. Each new line denotes the decision process, and what factor the model used for splitting the data. Line 2, for example, uses the splitting criteria of having the home team have an average of less than 4.5 shots on target. This turns the predicted value at this point from a home win to an away win, as denoted by the A. The probabilities also shift to reflect this new decision with the 41.5% home win reduced to 22.6%, while the away win probability has risen from 35.5% to 51.6%. The “*” at the end of each line identifies the final result predicted based upon the splits performed, or terminal nodes.

The right side of the image shows the same logic, without showing what way the model is voting at each stage. Instead, it shows each decision on its way to achieving the purest classification nodes possible. Following the logic of tree on the right, it is possible to see that should the home team have greater than 4.5 shots on target, while the away team have greater than 5.5 shots on target in a game where the home team accrues more than 11.5 corners, is

statistically more likely to result in a home win. This logic follows, attaining this number of corners would insinuate that the home team has been dominant and has seen many blocked shots leading to corners.

MODEL 2.2: DECISION TREE – UPDATED FOOTBALL STATISTICS

Just as with the two originally produced DT, when using the updated statistics data set the second model was more powerful than the first. The first model required no additional tweaks to the created model and returned an accuracy of 53%. The second model returned an accuracy of 60%. Both the text and traditional view of the updated DT can be viewed in Figure 30, just as they were previously, in Figure 29.

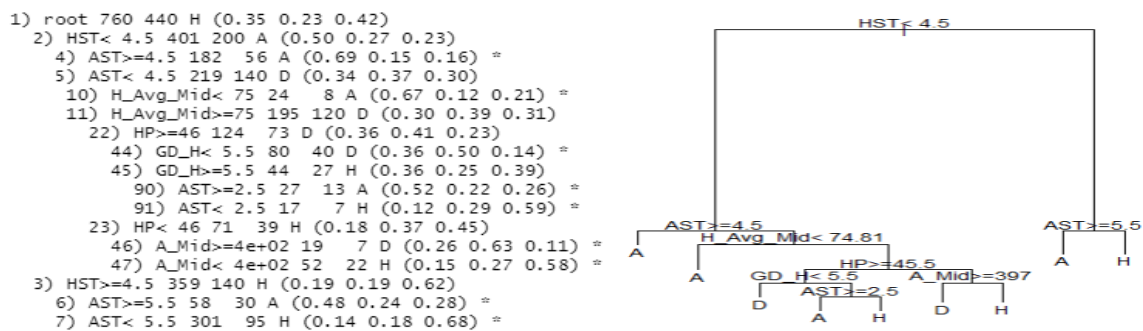


Figure 30: The logic and splitting operations of the more successful DT model for the updated statistics data set

MODEL 3.1: RANDOM FOREST – BASE FOOTBALL STATISTICS

This model in training operated as predicted resulting in an 86% accuracy rating, performing better than the DT method as it is supposed to. However, in testing there was a noteworthy drop off resulting in a 46% accuracy metric. This is a considerable drop for a method expected to perform better than the DT model.

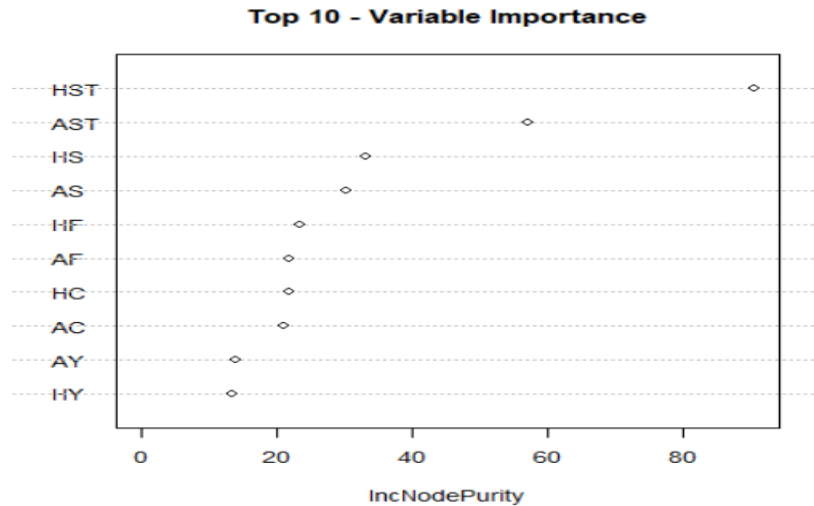


Figure 31: Top ten most important variables to determine the result, as found by the RF model for the base statistics data set

It is possible to view the variables the models believe are the most pertinent to any result being predicted. The title attributed to the x-axis, “IncNodePurity,” denotes how much the model would increase in error if this variable was randomly changed. Indicating that Home Shots on Target is especially important to the final prediction of this model, as discussed previously.

MODEL 3.2: RANDOM FOREST – UPDATED FOOTBALL STATISTICS

The updated statistics data set resulted in a 91% accuracy in training. However, the drop off in testing was substantial returning an accuracy of approximately 45%. Just as was done with the base statistics data set, it is possible to view the most important variables that the RF model used when making its predictions in Figure 32.

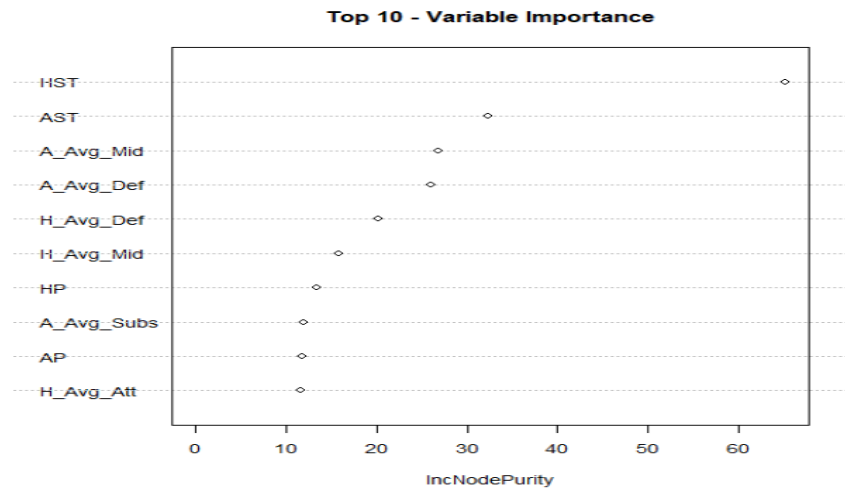


Figure 32: Top ten most important variables to determine the result, as found by the RF model for the updated statistics data set

Just as was shown in Figure 32, it is now possible to view what the most important metrics are in this new RF model. The HST IncNodePurity variable has decreased from approximately 90% to around 65%, indicating the new variables have a greater impact on the prediction than they had previously.

MODEL 4.1: SUPPORT VECTOR MACHINE – BASE FOOTBALL STATISTICS

SVM proved to be a more computationally heavy method than the other models, taking significantly longer to render a prediction, frequently requiring a soft reboot of the R Studio environment to get the model to work and produce its predictions. Once again utilizing a confusion matrix, it was found that SVM proved to be a consistent model producing approximately 61% accuracy in training and a little over 62% in testing.

MODEL 4.2: SUPPORT VECTOR MACHINE – UPDATED FOOTBALL STATISTICS

Utilizing the updated statistics data set, the model remained consistent in its accuracy by producing approximately 57% in training and 58% in testing. While less than the base statistics model, the two results remain within 1% of each other.

MODEL 5.1: NEURAL NETWORK – BASE FOOTBALL STATISTICS

NN returned an accuracy of just over 61%. It must be noted that the final model used was not the first model created. Due to computational limitations, the author was unable to insert additional nodes and layers between the input and output layers which may have been able to produce a more accurate model.

MODEL 5.2: NEURAL NETWORK – UPDATED FOOTBALL STATISTICS

After all the tweaks had been done to accommodate the updated statistics data set, the model returned an accuracy of approximately 56%.

MODEL 6.1: NAÏVE BAYES – BASE FOOTBALL STATISTICS

Of all the models NB returned one of the lower accuracy reports, returning a 52% accuracy having not once predicted a draw. As mentioned in the previous chapter, NB contains functions for viewing the variables most important when performing its predictions, viewable below in Figure 33.

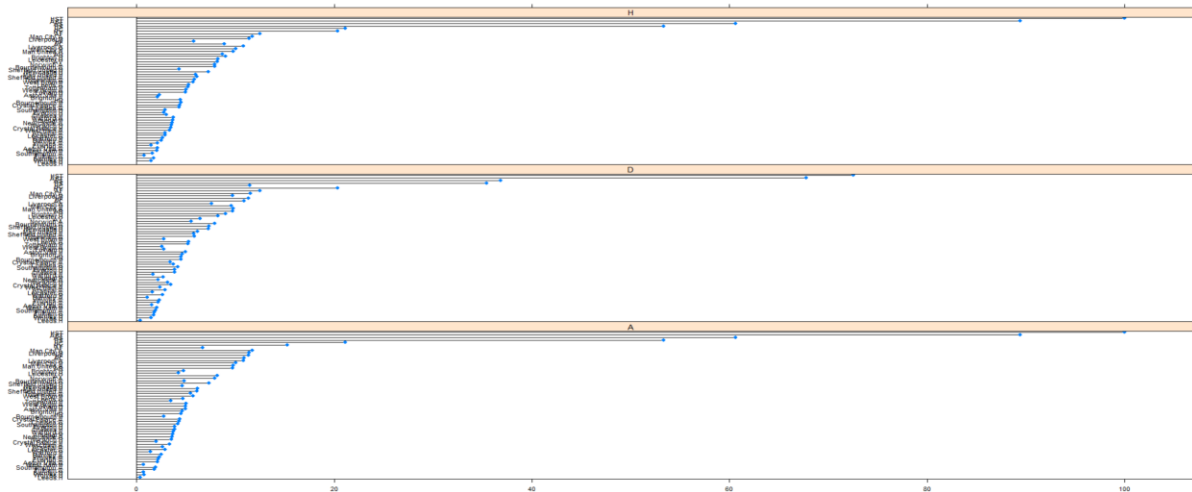


Figure 33: Most important features when deciding each individual result as per the NB model utilizing the base statistics data set

While this image is difficult to interpret, there is another function that performs similarly, relaying the following information.

	A	D	H
HST	100.000	72.561	100.000
AST	89.430	67.752	89.430
AS	60.637	36.852	60.637
HS	53.376	35.438	53.376
HC	21.118	11.429	21.118
HY	15.260	20.335	20.335
AF	6.702	12.488	12.488
Man.City.A	11.686	11.535	11.686
Liverpool.H	11.387	9.712	11.387
HF	11.348	11.348	5.778
AC	10.885	10.885	8.877
Liverpool.A	10.801	7.611	10.801
Man.City.H	10.070	9.585	10.070
Man.United.A	9.804	9.782	9.804
AR	9.750	9.750	8.685
Brighton.H	4.742	8.996	8.996
Leicester.H	4.239	8.267	8.267
AY	8.194	6.460	8.194
Bournemouth.A	4.800	7.886	7.886
Norwich.A	7.886	5.505	7.886

Figure 34: Most important features when deciding each individual result as per the NB model utilizing the base statistics data set

Just as with RF in Figures 31 and 32, NB can display in this list format the most important variables that it utilized when deciding the outcomes of matches.

MODEL 6.2: NAÏVE BAYES – UPDATED FOOTBALL STATISTICS

This NB model utilizing the updated statistics data set returned an accuracy of approximately 58%. Representing an increase in accuracy.

	A	D	H
HST	100.00	70.41	100.00
AST	82.14	67.02	82.14
A_Avg_Mid	63.25	36.46	63.25
H_Avg_Def	59.02	33.04	59.02
A_Avg_Def	56.29	33.93	56.29
H_Avg_Att	55.80	31.37	55.80
H_Avg_Mid	55.52	33.51	55.52
AS	54.51	34.66	54.51
A_Avg_Att	50.25	25.85	50.25
HS	48.56	25.67	48.56
H_GK	45.83	30.87	45.83
A_Avg_Subst	45.08	24.02	45.08
A_GK	44.88	43.23	44.88
GD_A	44.64	24.90	44.64
GD_H	41.89	28.89	41.89
H_Avg_Subst	39.63	24.35	39.63
H_Def	35.46	24.88	35.46
Five_Game_Form_H	34.55	22.14	34.55
A_Subst	32.69	19.58	32.69
A_Def	28.66	19.69	28.66

Figure 35: Most important features when deciding each individual result as per NB model utilizing the updated statistics data set

Figure 35 contains the updated list of variables the new NB model utilized in order to derive its predictions.

MODEL 7.1: XGBOOST – BASE FOOTBALL STATISTICS

The first xGBoost model returned an accuracy of approximately 57% when using the base statistics data set.

MODEL 7.2: XGBOOST – UPDATED FOOTBALL STATISTICS

The second xGBoost model returned an accuracy of approximately 59% when using the updated statistics data set.

MODEL 8.1: MULTINOMIAL LOGISTIC REGRESSION – BASE FOOTBALL STATISTICS

The training model resulted in an accuracy of approximately 68% while testing resulted in an approximate accuracy of 59%. A significant but not unexpected decrease in accuracy which can be explained by the data set being comprised of 20% of the overall data set compared to the 80% used in training.

MODEL 8.2: MULTINOMIAL LOGISTIC REGRESSION – UPDATED

FOOTBALL STATISTICS

This model, incorporating the updated statistics, returned an accuracy of approximately 72% in training and 66% in testing.

FINAL ACCURACY TABLE

Upon collecting all of the accuracy metrics, the author compiled this data into the following table, Table 2, ready for analysis:

Model	Base Statistics Accuracy (%)	Updated Statistics Accuracy (%)	Increase / Decrease (%)
kNN	85	75	-10
Decision Tree	63	60	-3
Random Forest	46	45	-1
Support Vector Machines	62	58	-4
Neural Networks	61	56	-5
Naïve Bayes	52	58	+6
xGBoost	57	59	+2
Multinomial Logistic Regression	59	66	+7

Table 2: Accuracy ratings of all models for both the base statistics data set and updated statistics data set

Preparing the data like this provides an interpretable and easy to read framework from which the basic analysis can begin.

EVALUATION

The first eight models were built utilizing the base statistics data set. This stage returned kNN as the most powerful model with an accuracy of 85%. This is, however, a drop of 10% from the previous model. The accuracy of this model, even after seeing a decrease of 10%, is on par with the model created by Razali et al. (2017) discussed previously and was a paper that influenced the study performed here. This model was 22% more accurate than the next most accurate model, the DT model.

This stage of modelling brought to light some unexpected accuracy ratings, such as the RF method performing so accurately in training, 86%, but being less accurate than guessing in testing, 46%. This is brought into starker contrast when it was also expected that RF would operate to a higher level than the DT model due to the way it works. Making multiple decision trees and choosing the result based on a general consensus or a group decision, as opposed to just one DT coming to its conclusion. Additionally, the NB model which is utilized in many other studies also performed rather poorly, being marginally better than guessing at random with an accuracy of 52%. This was another accuracy that was unexpected as the model tended to be more powerful in other studies.

It was interesting to note at this stage the variables that the models deemed the most important. The base statistics implied that shots on target for both home and away teams, followed by shots in general were the most important metrics when determining who would win any given game. If the data at this stage was to be used to inform teams on a transfer strategy, or what ways to set up for a match, it would imply that prioritising attack minded players or focusing on increasing the volume of shots would be the best way to win, statistically speaking. Additionally, the number of set-pieces a team has is also a more valuable indicator of who will win than who the team is. (It was also interesting to note that RF found free kicks, or fouls, a more important metric than NB, which found corners more important. This implies

that set-pieces are important in the modern game, regardless of the position they are being taken from.) This information could inform teams on training ground tactics and training regimes, especially if it becomes apparent that some teams lose the majority of games where the opposition has a high number of set-pieces. These insights are the types of information that this study was looking to discover. So that teams, sports analysts or those utilizing the model for gambling purposes have a statistical insight that others do not.

At this point in the study kNN is the model that will be used as the prediction model for the most recent season, the test statistics data set.

At this stage, the updated models were created. These models were adapted versions of the models that were created previously that allow the inclusion of the updated statistics data set. kNN was once again the most powerful model with an accuracy of 75%. The extra statistics served to confuse the model, but not to the point where they were completely detrimental to the learning process with kNN remaining 9% more accurate than the next closest model of MLR.

MLR was the model which benefitted the most from the additional variables as these brought the model to be the second most accurate for both the updated statistics, and the second strongest model compared to the strength of the base statistics models also. 75% accuracy is comparable to the papers studied in the literature review, such as Razali et al. (2017) [29] and Igiri (2014) [22]. The additional statistics did also see decreases in accuracy in the DT, RF, SVM and NN models. However, the remaining NB, xGB and MLR models all saw increases in accuracy. The increases in accuracy can be explained by some models needing the extra variables to learn adequately, whilst the models that saw decreases, which brought their accuracy ratings closer to 50%, are unsuited to this type of prediction. As noted, however, kNN is the exception. This model saw a decrease with the added variables but remained the most

powerful model by a good margin, proving the strength of kNN and why it is used in large corporations.

Once again, at this stage it is interesting to note the variables that the models now consider the most important to prediction. It is interesting to note that of the ten most important variables, in Figure 32 above, eight of them are now from the updated statistics. As explained earlier, the newly created possession metrics back up the logic that having more of the ball is important. It is also interesting that the average quality of the midfield and defence is more important than the quality of the attackers. This seems to suggest that having a stronger foundation for your team, or a better calibre of midfielder coupled with a solid defensive base to build from, is more important than having higher quality goal scorers. This sort of information can drive transfer strategies for teams in the future.

In the new data sets, while shots on target for both home and away teams were the top two variables when deciding the result of a game, they are followed by completely different features which frames these statistics differently. Previously, utilizing nothing but base figures, the data implied that having forwards who could get multiple shots in a game was the statistically best way to win any given match. However, the inclusion of the updated variables asserts that shots on target are important, provided the team has the defensive stability to build off of, in addition to a higher calibre of midfield player to help dictate a match. The important variables imply that it is the quality of the midfield that helps a team to raise the number of shots on target, with the knowledge that a good defensive foundation behind them will be available should the attack fail.

As stated before, information such as this can help drive transfer decisions and would also imply that the current landscape of transfers and their prices is misguided. At the time of writing, of the top fifty transfers in football the first defender to feature is Harry Maguire as the

17th most expensive transfer of all time, [55]. In fact, of the fifty most expensive transfers only seven of these players are considered defenders, with only ten being midfielders. Information found in this study could have the impact of altering the transfer landscape, changing the most expensive players in the world from those who score goals, to those who create them or, adversely, stop them.

Once again, it is interesting to note in Figure 35 that of the 20 most important variable only four are from the base statistics. This further drives home the importance of these additional psychological and non-psychological factors. Just as was found in the most important variables in the RF model, this list implies a strong defensive foundation coupled with being able to control the midfield with better quality players in this area is more important than higher quality attackers. This information further could further inform a team's transfer strategy, with this being backed by the fact that goal difference is also a key metric.

Upon the conclusion of this stage of the study, some interesting results have been brought to light. RF being consistently inaccurate is noteworthy. It was expected that this model would function to a higher standard than the DT model due to the way in which these two models' function. The fact that the model performed so similarly across the two data sets would indicate, however, that no matter what additional variables were added the model would continue to function at this poor level. No other literature selected RF as a model to use for football prediction, and this study would seem to indicate why. The NB model also performed underwhelmingly. This model was utilized in other studies and produced results that were more accurate than the kNN model created in this study. However, the most accurate NB model in the literature was created by Owrapipur et al. (2013), and the issues with that paper and the accuracy of their model have already been raised. It remains to be seen whether if the author had access to more variables, would the NB model produced be comparable to other NB models utilized for the same objectives.

After the results of these models, kNN is the model that will be selected going forward to perform the prediction section of this paper. Although it saw a decrease in accuracy, it was still by far the most accurate. The update statistics and increased variables may have served to confuse the model; however, a 75% accuracy rating is the most accurate measurement and just as accurate as some of the literature studied for this dissertation.

SUMMARY

In this chapter, all the results and accuracy ratings that had been created in Chapter 3 were collated and analysed. This analysis was performed so as to be able to accomplish Objective 3. In this analysis it was also found that the additional variables added to the data set became the leading contributors to most of the model's internal logic for discerning the result of a football match. This proves that football is so much more than the base statistics that are available, and these additional psychological and non-psychological factors play a large part in any final result.

It was also found that the current transfer market may place value on the wrong areas of the pitch, paying higher values for attack minded than defensive minded players. The models found that a higher calibre of defender was more beneficial than having a higher rated attacker, therefore, better defenders should be worth more than better attackers. It was also found that set-pieces are a large part of the modern game, no matter their location on the field. While teams should prioritise improving the positions from where they shoot, to have a higher chance of hitting the target as opposed to taking numerous shots that fail to make the goalkeeper perform a save.

Finally, kNN was the model that was selected as the best model to perform the final objective of this study. This model returned an accuracy metric comparable to the literature that was studied and was the strongest performer for both sets of statistics. It was 22% stronger

than the next strongest model when dealing with the base statistics, while being 9% stronger than the next strongest model when learning from the updated statistics data set. With a final accuracy of approximately 75%, kNN will apply the learning it has gathered to the final test statistics data set and its prediction powers will be assessed.

CHAPTER 5 – DISCUSSION AND FINDINGS

Chapter 5 is focused upon the final phase of this dissertation, utilizing the model selected in the previous chapter to perform prediction, assess the results and discuss the key findings.

INTRODUCTION

In the previous chapter the model that was chosen to be brought forward to fulfil this study was kNN. With an accuracy rating of approximately 75%, comparable to the literature that informed this study, kNN was consistently the most accurate and powerful model. In this chapter, the final test statistics data set will be cleansed and prepared, to be made ready to utilize the learning the kNN model has accrued in alignment with the research questions and objectives.

The data being used in this section of the study is the most recent completed season of the Premier League, the 2021/22 season. This data has been moulded into a data set that best suits the author's needs, as outlined in the "Research Questions and Objectives." The process of creating this data set will be discussed in greater detail in the "Data Preparation" section of this chapter but has already been touched upon in the previous chapter. All methodologies that were utilized previously, and the rationale behind them, was applied to this final data set.

RESEARCH QUESTIONS AND OBJECTIVES REVISITED II

The author set out to answer one overarching research question:

- Can a machine learning model be produced that can accurately gauge the winner between two Premier League teams?

The answer to the research question will be attained by achieving three core research objectives, those being:

1. Create a machine learning model that can predict the result of any given Premier League game utilizing standard in game statistics

2. Create a machine learning model that can predict the result of any given Premier League game utilizing standard in game statistics and also additional statistics and psychological factors
3. Assess whether the model which performs best in Objective 2 can be used to predict the results of the most recent Premier League season.

At this stage of the study, both Objective 1 and Objective 2 have been achieved, leaving only the final Objective to be attained. Machine learning models were built that could analyse and learn from both the base statistics data set and the updated statistics data set. With the best performing of these models, across both data sets, being kNN. Achieving an 85% accuracy in the base statistics data set, and a 75% accuracy in the updated statistics data set, kNN was the strongest performing model.

Due to the strength of its performance across the data sets, kNN was selected to move forward to perform the ultimate aim of this study, predicting the winner of any two teams in the Premier League, and thus, achieve Objective 3.

DATA UNDERSTANDING

The data utilized in this section of the study is the test statistics data set. This data set is comprised of all the statistics that made up the updated test statistics data set. The final test statistics data set was comprised of three-hundred and eighty unique observations with eighty-three variables combining to make 31,540 data points. (This is exclusive of the additional six dummy columns). The final data set was comprised of the following variables:

FTR = Full Time Result (H=Home Win, D=Draw, A=Away Win)

HS = Home Team Shots

AS = Away Team Shots

HST = Home Team Shots on Target

AST = Away Team Shots on Target

HC = Home Team Corners

AC = Away Team Corners

HF = Home Team Fouls Committed

AF = Away Team Fouls Committed

HY = Home Team Yellow Cards

AY = Away Team Yellow Cards

HR = Home Team Red Cards

AR = Away Team Red Cards

H.GK = Home Goal Keeper

H.Def = Home Defence

H.Avg_Def = Home Average Defence

H.Mid = Home Midfield

H.Avg_Mid = Home Average Midfield

H.Att = Home Attack

H.Avg_Att = Home Average Attack

H.Subs = Home Substitutes

H.Avg_Subs = Home Average Substitutes

A.GK = Away Goal Keeper

A.Def = Away Defence

A.Avg_Def = Away Average Defence

A.Mid = Away Midfield

A.Avg_Mid = Away Average Midfield

A.Att = Away Attack

A.Avg_Att = Away Average Attack

A.Subs = Away Substitutes

A.Avg_Sub = Away Average Substitutes

Five_Game_Form_H = 5 Game Form for Home Team

Rev_Fixture_H = Reverse Fixture Home Team

Maj_Suspension_H = Major Suspensions Home Team

Min_Suspension_H = Minor Suspensions Home Team

GD_H = Goal Difference Home

Five_Game_Form_A = 5 Game Form for Away Team

Rev_Fixture_A = Reverse Fixture Away Team

Maj_Suspension_A = Major Suspensions Away Team

Min_Suspension_A = Minor Suspensions Away Team

GD_A = Goal Difference Away

HP = Home Possession

AP = Away Possession

In addition to these features, there are forty team variables. This is due to the one hot encoding process, whereby each team was given a separate variable to demonstrate whether they were playing at home or away. The original twenty teams become, twenty teams that play at home,

LiverpoolH for example, and twenty teams that play away, LiverpoolA, which creates forty unique teams. The full data set is viewable below in Figure 36.

\$ FTR	: chr	"H" "H" "A" "H" ...	\$ EvertonA	: int	0 0 0 0 0 0 0 0 0 0 ...
\$ HS	: num	8 16 14 13 14 9 13 14 17 13 ...	\$ LeedsA	: int	0 1 0 0 0 0 0 0 0 0 ...
\$ HST	: num	3 8 3 6 6 5 7 3 3 3 ...	\$ LeicesterA	: int	0 0 0 0 0 0 0 0 0 0 ...
\$ HF	: num	12 11 10 15 13 6 18 4 4 11 ...	\$ LiverpoolA	: int	0 0 0 0 0 0 0 1 0 0 ...
\$ HC	: num	2 5 7 5 6 5 2 3 7 3 ...	\$ ManCityA	: int	0 0 0 0 0 0 0 0 0 1 ...
\$ HY	: num	0 1 2 0 2 1 3 1 1 2 ...	\$ ManUnitedA	: int	0 0 0 0 0 0 0 0 0 0 ...
\$ HR	: num	0 0 0 0 0 0 0 0 0 0 ...	\$ NewcastleA	: int	0 0 0 0 0 0 0 0 0 0 ...
\$ HP	: num	35 49 36 62 48 56 38 50 47 34 ...	\$ NorwichA	: int	0 0 0 0 0 0 0 0 0 0 ...
\$ AS	: num	22 10 14 4 6 17 11 19 8 18 ...	\$ SouthamptonA	: int	0 0 0 0 1 0 0 0 0 0 ...
\$ AST	: num	4 3 8 1 3 3 2 8 9 4 ...	\$ TottenhamA	: int	0 0 0 0 0 0 0 0 0 0 ...
\$ AF	: num	8 9 7 11 15 10 13 14 3 8 ...	\$ WatfordA	: int	0 0 0 0 0 0 0 0 0 0 ...
\$ AC	: num	5 4 6 2 8 4 4 11 6 11 ...	\$ WestHamA	: int	0 0 0 0 0 0 0 0 1 0 ...
\$ AY	: num	0 2 1 0 0 2 1 1 0 1 ...	\$ WolvesA	: int	0 0 0 0 0 1 0 0 0 0 ...
\$ AR	: num	0 0 0 0 0 0 0 0 0 0 ...	\$ H_GK	: int	75 87 81 86 81 83 72 76 74 87 ...
\$ AP	: num	65 51 64 38 52 44 62 50 53 66 ...	\$ H_Def	: int	372 323 309 244 316 313 291 294 367 313 ...
\$ ArsenalH	: int	0 0 0 0 0 0 0 0 0 0 ...	\$ H_Avg_Def	: num	74.4 80.8 77.2 81.3 79 ...
\$ AstonVillaH	: int	0 0 0 0 0 0 0 0 0 0 ...	\$ H_Mid	: int	223 411 301 497 395 406 368 220 229 238 ...
\$ BrentfordH	: int	1 0 0 0 0 0 0 0 0 0 ...	\$ H_Avg_Mid	: num	74.3 82.2 75.2 82.8 79 ...
\$ BrightonH	: int	0 0 0 0 0 0 0 0 0 0 ...	\$ H_Att	: int	152 79 151 83 81 86 76 226 160 250 ...
\$ BurnleyH	: int	0 0 1 0 0 0 0 0 0 0 ...	\$ H_Avg_Att	: num	76 79 75.5 83 81 ...
\$ ChelseaH	: int	0 0 0 1 0 0 0 0 0 0 ...	\$ H Subs	: int	616 702 633 732 652 682 653 616 659 682 ...
\$ CrystalPalaceH	: int	0 0 0 0 0 0 0 0 0 0 ...	\$ H_Avg Subs	: num	68.4 78 70.3 81.3 72.4 ...
\$ EvertonH	: int	0 0 0 0 1 0 0 0 0 0 ...	\$ A_GK	: int	81 77 77 79 75 81 84 89 82 89 ...
\$ LeedsH	: int	0 0 0 0 0 0 0 0 0 0 ...	\$ A_Def	: int	310 302 304 303 302 234 313 338 317 333 ...
\$ LeicesterH	: int	0 0 0 0 0 1 0 0 0 0 ...	\$ A_Avg_Def	: num	77.5 75.5 76 75.8 75.5 ...
\$ LiverpoolH	: int	0 0 0 0 0 0 0 0 0 0 ...	\$ A_Mid	: int	392 388 383 301 306 306 383 236 401 250 ...
\$ ManCityH	: int	0 0 0 0 0 0 0 0 0 0 ...	\$ A_Avg_Mid	: num	78.4 77.6 76.6 75.2 76.5 ...
\$ ManUnitedH	: int	0 1 0 0 0 0 0 0 0 0 ...	\$ A_Att	: int	67 78 77 157 151 237 80 265 80 256 ...
\$ NewcastleH	: int	0 0 0 0 0 0 0 0 1 0 ...	\$ A_Avg_Att	: num	67 78 77 78.5 75.5 ...
\$ NorwichH	: int	0 0 0 0 0 0 0 1 0 0 ...	\$ A Subs	: int	670 644 624 612 668 628 652 697 656 759 ...
\$ SouthamptonH	: int	0 0 0 0 0 0 0 0 0 0 ...	\$ A_Avg Subs	: num	74.4 71.6 69.3 68 74.2 ...
\$ TottenhamH	: int	0 0 0 0 0 0 0 0 0 1 ...	\$ Five_Game_Form_H	: int	0 0 0 0 0 0 0 0 0 0 ...
\$ WatfordH	: int	0 0 0 0 0 0 1 0 0 0 ...	\$ Five_Game_Form_A	: int	0 0 0 0 0 0 0 0 0 0 ...
\$ West.Ham.H	: int	0 0 0 0 0 0 0 0 0 0 ...	\$ Rev_Fixture_H	: num	0 0 0 0 0 0 0 0 0 0 ...
\$ WolvesH	: int	0 0 0 0 0 0 0 0 0 0 ...	\$ Rev_Fixture_A	: num	0 0 0 0 0 0 0 0 0 0 ...
\$ ArsenalA	: int	1 0 0 0 0 0 0 0 0 0 ...	\$ GD_H	: int	0 0 0 0 0 0 0 0 0 0 ...
\$ AstonVillaA	: int	0 0 0 0 0 0 1 0 0 0 ...	\$ GD_A	: int	0 0 0 0 0 0 0 0 0 0 ...
\$ BrentfordA	: int	0 0 0 0 0 0 0 0 0 0 ...	\$ Maj_Suspension_H	: int	0 0 0 0 0 0 0 0 0 0 ...
\$ BrightonA	: int	0 0 1 0 0 0 0 0 0 0 ...	\$ Min_Suspension_H	: int	0 0 0 0 0 0 0 0 0 0 ...
\$ BurnleyA	: int	0 0 0 0 0 0 0 0 0 0 ...	\$ Maj_Suspension_A	: int	0 0 0 0 0 0 0 0 0 0 ...
\$ ChelseaA	: int	0 0 0 0 0 0 0 0 0 0 ...	\$ Min_Suspension_A	: int	0 0 0 0 0 0 0 0 0 0 ...
\$ CrystalPalaceA	: int	0 0 0 1 0 0 0 0 0 0 ...	\$ blank_1	: int	0 1 1 1 1 1 1 1 1 1 ...
			\$ blank_2	: int	0 1 1 1 1 1 1 1 1 1 ...
			\$ blank_3	: int	0 1 1 1 1 1 1 1 1 1 ...
			\$ blank_4	: int	0 1 1 1 1 1 1 1 1 1 ...
			\$ blank_5	: int	0 1 1 1 1 1 1 1 1 1 ...
			\$ blank_6	: int	0 1 1 1 1 1 1 1 1 1 ...

Figure 36: Full list of features of test statistics data set

DATA PREPARATION

TEST STATISTICS – CLEANSING

The third, and final, set of data that needs to be altered is the most recent Premier League season. This season of statistics is the set that will be used for the prediction portion of this dissertation to assess the power of the final model.

The logic applied to the base statistics data set and the cleaning performed there, and also in the updated statistics data set, is also applied to this data set. The same process of collating and creating the updated statistics data set was also reproduced for this data set, adding

in all additional features to the base statistics to create this test statistics data set that is similar in structure to the updated statistics data set.

However, additional steps need to be performed on this data set. Up to this stage, the data sets have been utilizing historical data to provide the models with the most accurate measurements to learn from. This means that the data sets show all the final statistics of a match, such as how many shots each team had for example. While all the additional, author created, statistics are all readily available up to one hour before any given Premier League game would start, statistics such as each teams' line-ups or who is suspended, statistics such as how many shots a team will have would not be available. This dissertation is focused on the prediction power of the models that are created, as such, a method to fill in the statistics that are unknown before kick-off needs to be implemented.

The author utilizes the form over five games as a metric in the data set to help determine the most likely winner between the two teams. In the same manner, the base statistics have been averaged out over the previous five games so that their impact is felt in the sixth game. (An allowance of the first two matches using the base statistics was permitted for each team so that the model had figures to start getting average values from game three onwards). This was the best method to alter the base statistics portion of the data set as, just like the five game form feature, this shows the average way a team is playing. Doing this ensures the model is actually performing a prediction as it has been blinded from the true values, which wouldn't be available as the game hasn't been played yet, and is instead basing its decision on the average values of each feature.

Finally, the kNN model will not work on a data set that does not contain the same dimensions as the data set it was built upon. The updated statistics data set contained information across two seasons of Premier League football, due to this, instead of containing

the statistics of twenty teams, it contains the data of twenty-three teams. As mentioned previously, in the Data Understanding section, the way the data is manipulated in the one hot encoding process creates two variables for each team representing them as home and away. As such, this means the updated statistics data set will contain six additional columns. This is because three teams got relegated and thus removed from the next seasons statistics and are replaced by three new teams producing the six additional columns. To account for this, and for the model to work, six dummy columns are added to the data set that contain no meaningful data. This allows the kNN model to function, without altering the methodology it was following or altering its logic processes. These six columns are visible in Figure 36.

MODELLING

Just as was performed before, the appropriate packages must be loaded into the R environment. These packages are “class” and “gmodels.”

This model is built as it was before for the updated statistics data set. It is created in exactly the same way as the model before as the model needs to be trained in the same manner as it was previously on the updated statistics data set. As such, the data was normalized with the same k value of 19 being passed into the model. The accuracy function was utilized again to assess whether model has performed to the same standard as previously. Returning the same accuracy of 75%, the test statistics data set can now be altered.

As was mentioned, the kNN functions better once the data is using has been normalized. Due to this, the normalize function is applied to the test statistics data set before it can be passed through the model. This normalize function is also applied to the dummy columns that are appended to the data set, with the values in these being so minute they are negligible. With the model trained on the updated statistics data set, the learning it has accrued can be applied to the test statistics data set. This produces our prediction results for the season.

EVALUATION

The objective of this section is to analyse how well the kNN model performed its prediction task, utilizing the new test statistics data set. Once the data set was run through the trained kNN model, it returns a list of all the final results in the form kNN has converted them to, 0 for an away win, 0.5 for a draw and 1 for a win. These results can be viewed below in Figure 37.

```
> new_test_pred
[1] 1 1 0 1 1 1 1 0 0 1 1 1 0.5 1 1 1 0 0 0 1 1 0.5 0.5 0.5 0 1 1 1 1 0 0.5 1 0 0 1 0 0 1 0 1 0 0 0 1 0 0 0.5
[48] 1 0 0 0 0 1 0 0 0 0 0 1 0.5 0.5 1 1 1 1 0.5 0.5 1 0 0 0 0 1 1 1 1 0 0 0 0.5 1 1 0 0 1 0 0.5 0 1 0 0 1 1 0
[95] 0 0 0 0 0 0.5 1 0 0 1 1 0 1 1 0 1 0 1 0 0.5 1 1 0.5 1 1 1 1 0 1 1 0.5 1 1 1 0.5 0.5 1 0.5 0 0.5 0 0 0 1 1 1 1
[142] 0 0 0 0 1 1 1 1 1 1 1 1 1 1 0 1 1 1 0 1 0.5 0.5 1 0.5 1 0 0 0.5 0 1 0 1 0 0 0.5 1 1 0 0 1 0.5 0 1 0 0 0 1 0
[189] 1 0.5 0 1 1 0.5 1 0 1 1 0.5 1 0 0.5 0 0 0 0 0 0 0 1 0.5 1 0 0.5 1 0 1 1 0 1 0.5 0 0 1 0 1 0.5 1 0 0 0 1 0 1 1 0.5
[236] 1 0 0.5 0 1 1 0 0 1 1 0 1 1 1 0 0 1 1 0 0 1 0 1 1 0 1 0 1 0 0 1 0 1 1 0 0 1 0 0 1 1 1 1 0.5 1 0 1 1 0.5 0
[283] 0 1 0 0 1 1 1 0.5 0 0 0 1 0.5 1 1 0.5 1 1 1 0.5 0 0 0 1 1 1 1 0 1 1 0 1 1 1 1 0 0.5 1 1 1 1 1 1 0 0 0.5 1 1
[330] 1 0 1 0 1 0 0 0 0 1 1 0 1 1 0 1 1 0.5 0 1 0 0 1 0 0 1 0 0 1 1 0.5 1 0 0.5 0 0 1 0 1 0 1 0 1 1 0 1 1 0 1 1
[377] 1 1 1 0
Levels: 0 0.5 1
```

Figure 37: The results of each game according to the trained kNN model

At this stage, the results need to be turned in to meaningful football related data. The easiest method to convert these values into points for teams was to line them side-by-side with the test statistics data set. The data set dictates which two teams played, and the results inform which team won, or if they drew. These results are displayed in the order the matches were fed into the model, or the same order they are in the data set. As such, it is a simple task of attributing the wins and draws to the correct teams. Doing this created the table viewable in Table 3.

Team	W	L	D	GP	PTS
Arsenal	27	11	0	38	81
Aston Villa	14	22	2	38	44
Brentford	9	26	3	38	30
Brighton	11	11	16	38	49
Burnley	4	20	14	38	26
Chelsea	32	5	1	38	97
Crystal Palace	11	14	13	38	46
Everton	10	27	1	38	31
Leeds	7	23	8	38	29
Leicester	16	17	5	38	53
Liverpool	37	1	0	38	111
Man City	35	3	0	38	105
Man Utd	26	9	3	38	81
Newcastle	14	17	7	38	49
Norwich	4	32	2	38	14
Southampton	8	19	11	38	35
Tottenham	25	11	2	38	77
Watford	5	32	1	38	16
West Ham	22	13	3	38	69
Wolves	16	20	2	38	50

Table 3: Predicted points in alphabetical order

Once rearranged into the final table order based on points across the thirty-eight-game season, the final table as predicted by the kNN model was produced in Table 4.

Teams	W	L	D	GP	PTS
Liverpool	37	1	0	38	111
Man City	35	3	0	38	105
Chelsea	32	5	1	38	97
Arsenal	27	11	0	38	81
Man Utd	26	9	3	38	81
Tottenham	25	11	2	38	77
West Ham	22	13	3	38	69
Leicester	16	17	5	38	53
Wolves	16	20	2	38	50
Brighton	11	11	16	38	49
Newcastle	14	17	7	38	49
Crystal Palace	11	14	13	38	46
Aston Villa	14	22	2	38	44
Southampton	8	19	11	38	35
Everton	10	27	1	38	31
Brentford	9	26	3	38	30
Leeds	7	23	8	38	29
Burnley	4	20	14	38	26
Watford	5	32	1	38	16
Norwich	4	32	2	38	14

Table 4: Final table predicted by kNN sorted by points

The table viewable above in Table 4, depicts the final league positions of each team in the Premier League. At this stage, it is interesting to note that two of the teams used as an important

decider of whether teams win or lose both finished in the top two, Figure 34, those being Liverpool and Manchester City. This is most likely due to both teams being the two most dominant teams in the league across the last three years, with one league title going to Liverpool and two going to Manchester City in that time.

Additionally, both teams boast some of the highest rated players in the world in the defensive and midfield positions. Virgil van Dijk and Ruben Dias are the top-two rated center backs in the world and play for Liverpool and Manchester City respectively, while the likes of Trent Alexander-Arnold, Andrew Robertson and Joao Cancelo are amongst the best full-backs in the world. Both teams also boast some of the best midfield options in the world with Kevin de Bruyne, Fabinho and Bernardo Silva playing for the two teams. While they both have access to a high calibre of player in the forward positions, in the top-ten forward players in the world between the two clubs only two players appear, Mohamed Salah and Sadio Mane both of Liverpool. This lends further weight to the notion that the best teams are built upon a stronger defence and midfield, rather than containing the best attacking players.

Adversely, the teams who are predicted to be relegated from the league, Norwich, Watford and Burnley, contain some of the weakest calibre of players in the defensive and midfield positions. Norwich's average player rating for their first team quality players is approximately 71.27 in defence and 70.53 in midfield. Watford and Burnley have ratings of 71.75 and 73.81 in defence respectively and 73.5 and 74.89 in midfield, amongst the lowest in the league. This is made more apparent when noting that the average quality in the Premier League for players in these positions is 77.53 and 78.52 in defence and midfield respectively, as shown in Figure 24, showing just how far below the standard these teams are.

Before contrasting the real league table and the predicted league table positions, it is interesting to note that all the teams that comprise the group of teams known as "The Big Six"

all finished in the top six positions. While teams that have either only recently been promoted to the league or have just avoided relegation make up the bottom of the table. In between these groups are the teams that consistently push for European places, the top seven positions, such as West Ham and Leicester, and also those that do just enough each season to avoid relegation battles, Southampton for example.

Pos	Team	GP	W	L	D	Pos Diff
1	Liverpool	38	37	1	0	+1
2	Man City	38	35	3	0	-1
3	Chelsea	38	32	5	1	0
4	Arsenal	38	27	11	0	+1
5	Man Utd	38	26	9	3	+1
6	Tottenham	38	25	11	2	-2
7	West Ham	38	22	13	3	0
8	Leicester	38	16	17	5	0
9	Wolves	38	16	20	2	+1
10	Brighton	38	11	11	16	-1
11	Newcastle	38	14	17	7	0
12	C Palace	38	11	14	13	0
13	Aston Villa	38	14	22	2	+1
14	Southampton	38	8	19	11	+1
15	Everton	38	10	27	1	+1
16	Brentford	38	9	26	3	-3
17	Leeds	38	7	23	8	0
18	Burnley	38	4	20	14	0
19	Watford	38	5	32	1	0
20	Norwich	38	4	32	2	0

Table 5: Predicted league table with position differential from real table

Table 5 shows the predicted league table and how close the model was to predicting the real final league table. The model predicted that Liverpool would finish as champions ahead of Manchester City in a close race. With both teams breaking points records in recent seasons, such as the largest points gap over second place or the most points won over thirty-eight games [56], being consistently the top two teams in the country, this is an accurate start. In the real season Manchester City narrowly beat Liverpool to the title, winning by just one point, ninety-four points to ninety-three. Due to this, the model predicting a Liverpool win shows that both teams are so close in quality that they are almost inseparable.

Chelsea finished third in real life, just as they were predicted to by the model. Chelsea were expected to challenge for the title during the season, having just won the Champions League, possessing a quality squad and had just added the veteran Thiago Silva to their defence, increasing the quality of players at their disposal in this position. Arsenal and Manchester United meanwhile both jumped one position, from fifth and sixth to fourth and fifth respectively, in the predicted table compared to the real table. This could be in relation to their transfer dealings. Arsenal signed two starting eleven quality players in defence and one in midfield, and Manchester United signed Raphael Varane, a much higher calibre player than any other they had in defence. The diminishing importance of the quality of attacker is further highlighted here as Manchester United signed Ronaldo and Jadon Sancho, one of the highest rated forwards in the world and another attacker that is one of highest rated attackers in the league, while Arsenal sold their highest rated forward in Pierre-Emerick Aubameyang. These position rises saw a drop in Tottenham's final position by these two places. This again could coincide with losing their best rated defenders in Juan Foyth and Toby Alderweireld.

West Ham and Leicester finished as "the best of the rest," finishing in seventh and eighth in the real table, just as they were predicted to. After this, Wolves were predicted to narrowly finish ahead of Brighton by just one point, fifty points to forty-nine respectively. In

real life, Brighton finished above Wolves, but purely on goal difference with both teams finishing on fifty-one points. This another case of the model understanding just how close these teams are in terms of quality, like Liverpool and Manchester City, and how inseparable they are. Newcastle finished eleventh in real life, just as the model predicted, with a total of forty-nine points in both the predicted and real tables.

Between twelfth and seventeenth, sees only Brentford disrupt the overall accuracy of the model. Brentford were predicted to finish sixteenth, however, in the real table they finished thirteenth. This could be explained by the fact the model had no prior learning about Brentford or how they play, just as in real life the teams had never encountered Brentford in the Premier League. All other teams in the model had at least one season of learning before the predictions were made. Even teams such as Norwich and Watford, who have been promoted, relegated and promoted again season on season, had previous games to learn from. Brentford however, had less information to give the model and this could have contributed to the final predicted league position being the furthest away from the real league position. If Brentford had been predicted to finish in the position they did in real life, every league position from eleventh to twentieth would have been accurately predicted by the model.

Finally, the model successfully predicted the three relegated teams, and the teams that were locked in the relegation battle with them. As mentioned previously, all teams with access to a lower calibre of defender and midfielder were relegated from the league, those being Norwich, Watford and Burnley. Leeds, a team only promoted to the Premier League the season before, found themselves struggling to survive this season, finishing three points clear of relegation in real life, just as the model predicted. While Everton, who typically finish somewhere in the middle of the table, finished seventeenth in real life four points above relegation. The model predicted both the position and points differential.

Team	PTS Pred	Points Diff
Liverpool	111	+18
Man City	105	+13
Chelsea	97	+23
Arsenal	81	+10
Man Utd	81	+12
Tottenham	77	+19
West Ham	69	+13
Leicester	53	+1
Wolves	50	-1
Brighton	49	-2
Newcastle	49	0
C Palace	46	-2
Aston Villa	44	-2
Southampton	35	-10
Everton	31	-9
Brentford	30	-9
Leeds	29	-9
Burnley	26	-9
Watford	16	-7
Norwich	14	-8

Table 6: Final predicted points tally and points difference from the real total

At the end of the season, the model successfully predicted two-hundred and seventy-five games of the full three-hundred and eighty game season. This brings the overall accuracy of the prediction on the test statistics data set to 72.37%. A small percentage drop from the approximate 75% accuracy of the model prior to the prediction, which may be attributed to the model not having any prior learning in regard to Brentford as mentioned previously.

Judging from Table 6, the quality of the team tends to be the leading contributor to the winner between any two teams. With the teams who utilize a higher calibre of player performing better over the course of the season than those without players who are rated as highly. This is demonstrated by Liverpool who were predicted to drop points in only one match in the whole season. So, as the good teams win a few extra games, it would follow that the lesser teams lose a few extra games. However, while the model may lean on the quality of player, especially in defence and midfield, shown in Figures 32 and 35, as a large variable to determine outcome, it is still able to utilize other variables to influence the results. Producing the final predicted table, which is so close to the real table, balancing out the final results over the course of the season.

SUMMARY

In summary, kNN was the most powerful machine learning model across both the base statistics data set and the updated statistics data set, returning the highest accuracy across both sets of testing. Due to this power and high accuracy, comparable to the literature studied for this dissertation, it was selected to move forward and perform the final phase of this study, prediction. It is evident at this stage as to why the biggest companies such as Netflix or Amazon, identified in Chapter 2, utilize a kNN model in their companies. The power and accuracy it possesses and the fact it is so adaptable is hugely beneficial.

At this stage in the study the data was analysed and cleansed. To perform prediction, the historic values and statistics cannot be used as they would not be known prior to the game being completed. The metrics and variables that could not be known prior to the game were made in to averages over the previous five games played. This was done to firstly show symmetry with one of the additional variables added to the data set, but also to show the current form the team is in. A team performing well is more likely to come into a game with added confidence, while a team performing poorly is likely to underperform. These confidence issues

are likely to manifest themselves in the statistics at the end of each game. It is also the closest metric that can be used to assess how teams play and provide values for these variables, with the previous five games a common length of time to assess a team. After this, the rest of the data set was collated and included in the same manner was performed on the updated statistics data set. Finally, dummy columns were included containing no real value so the model can perform correctly.

The model functioned very well. Of the approximate 75% accuracy it was operating at before the test statistics data set was used, it returned an accuracy of 72.37%, correctly predicting the right result two-hundred and seventy-five times out of the full three-hundred and eighty game season. Potential reasons for the slight decrease in accuracy include the fact the model had never been exposed to Brentford before. Due to this it had no prior knowledge of them, their players or how they play, leading to the model having to return a prediction based on no prior learning. Another reason for this decrease in accuracy may have been the fact the base statistics were averaged out across the last five games, producing different results to the true values. There is no way to account for this however and is the best way the author could substitute these values.

The model successfully predicted the top six clubs, noting the close race for the title that Liverpool and Manchester City endured. The model predicted Liverpool to narrowly win the league, in real life however, Manchester City won the league by one point on the final day of the season, proving just how close these two teams are and how accurate the model is as the league title could have gone either way. A reason for selecting Liverpool as the winner could be attributed to the data sets used, as in the 2019/20 season Liverpool brushed competitors aside in the league and won it with the largest points tally in Premier League history, [56]. One main update to the top six was Tottenham dropping two positions while Arsenal and Manchester United rose those two positions. This could be attributed to Tottenham being the

only team of the three to lose some of its main defensive players, while both Arsenal and Manchester United added higher quality players in that area of the pitch.

The model was also able to discern who would be relegated, and in exactly which positions, accurately asserting that Norwich, Watford and Burnley would be relegated. kNN was also able to assess which teams would be in the relegation battle, such as Leeds and Southampton. It also was able to predict the surprising Everton relegation battle. In the data sets the model learned from, Everton finished tenth and twelfth, comfortably in the middle of the table leading to very little information as to the upcoming relegation scare. Proving the model was able to assess the variables and features accurately to predict this unexpected drop in quality. Additionally, Brentford were predicted to finish in sixteenth, but in reality, they finished thirteenth. This again could be attributed to the model having no prior learning with regards to them, with Brentford having been in the Championship in the years preceding the data sets that were used. However, if the model had successfully predicted their true final position, the model would have accurately placed all teams from eleventh to twentieth. Additionally, Brighton and Wolves were predicted to finish within a point of each other, with Wolves finishing higher. In reality, Brighton finished above Wolves, but this was purely on goal difference. Once again, this shows the understanding of the league and the various statistics the model has accrued.

Over the course of a season, it was predicted that the teams with a higher calibre of player all earned a few extra points than they did in real life, which directly relates to the teams without those players dropping some extra points. However, the fact that final predicted table is so close to the final real table would indicate that the model was able to look past the quality of player as a variable to determine the winner and was able to balance the results over the course of a season. This is demonstrated as the middle of the table team's, eighth to thirteenth, final point total were all within two points of the real final points total each team accrued.

CHAPTER 6 – CONCLUSION AND REFLECTION

This closing chapter addresses the overall success of the study in service of the Objectives laid out by the author. This chapter will also attempt to address key observations pertaining to the study while also acknowledging some of the difficulties and limitations the author encountered.

RESEARCH QUESTIONS AND OBJECTIVES REVISITED – FINAL

The author set out to answer one research question during the course of this study:

- Can a machine learning model be produced that can accurately gauge the winner between two Premier League teams?

In order to answer this question, three research objectives were identified. They were designed in a way that aide's progression in a step-by-step manner, guiding the study along in a natural way. To answer the research question, the following three objectives were decided upon:

1. Create a machine learning model that can predict the result of any given Premier League game utilizing standard in game statistics
2. Create a machine learning model that can predict the result of any given Premier League game utilizing standard in game statistics and also additional statistics and psychological factors
3. Assess whether the model which performs best in Objective 2 can be used to predict the results of the most recent Premier League season.

To address the overall success of this study, addressing the success of each Objective must be done first.

Objective	Success/Failure
Objective 1	Success
Objective 2	Success
Objective 3	Success

Table 7: Objectives success/failure rate

1. Objective 1: Can be viewed as highly successful. Overall, eight models were created with one model, kNN, boasting an accuracy metric of approximately 85%, viewable in Table 2. The model showed clear understanding of the base statistics data set, providing a solid foundation upon which to build.
2. Objective 2: Can also be viewed as very successful. The eight models were remodelled to account for the new updated statistics data set. Some models, such as the Multinomial Logistic Regression and Naïve Bayes models, saw an increase in accuracy. Some other models, such as the Decision Trees and Neural Network models, saw decreases. However, kNN returned an accuracy metric of approximately 75%, comparable to the literature studied, and was the best performing model again.
3. Objective 3: Is also a success. Bringing forward the strongest model, kNN, the author was able to achieve 72.37% accuracy across the whole season of the test statistics data set. The final model accurately predicted the close fight at the top of the league as Liverpool battled Manchester City for the title. It was also able to predict who would be relegated and in what order they would be relegated in. If not for the unseen as of this stage Brentford, the model would have accurately placed over half of the league, successfully deducing who would finish where and what other close battles would be fought. Wolves and Brighton is one such battle, or Everton being dragged into a fight to avoid relegation that was not seen by anyone prior to the season beginning.

As each independent research objective can be clearly viewed as a success, the overall study can be deemed a success. All objectives were met, and the fundamental research question of “can a machine learning model be produced that can accurately gauge the winner between two Premier League teams?” can be answered: yes.

Additionally, although the accuracy of the final model was 75%, with it producing a 72.37% accuracy on the final test statistics data set, it must be noted that football is still a game of chance. Successfully predicting three games out of every four is still incredibly high when in football any team can win on any given day, or a smaller club can produce a big result. On 14th March 2022 Manchester City played away at Crystal Palace, a game that would be included in the test statistics data set. Manchester City finished the game with 74% possession, nineteen shots, four shots on target and twelve attacking set-pieces, although this further enforces the analysis in Chapter 4 that the accuracy of shots is more important than the volume. In any other game, statistics such as these result in a win just as the model predicted, while in actuality Crystal Palace managed to get a draw from this match.

02	2-0	Burnley
2022-03		
14	0-0	C Palace
06	4-1	Man.Utd
2022-02		
26	1-0	Everton

Figure 38: Four Manchester City games including unexpected Crystal Palace result

The Manchester City and Crystal Palace game is viewable above in Figure 38. The model successfully predicted the results against Everton, Manchester United and Burnley, those being wins for Manchester City as evidenced by the green highlight. However, it incorrectly predicted a win at Crystal Place with the result in reality being a draw resulting in a 75% accuracy for

this subset of matches, which matches the overall accuracy metric of the model that was calculated.

CONTRIBUTION TO THE FIELD

The study brought some interesting observations to light. During the duration of the study, it was found that any set-piece that can be considered attacking are an important variable when deducing who will win a game. As such, teams that have better set-piece routines are more likely to do well. This is backed up by the likes of Liverpool who hired Thomas Grønnemark in 2018 as a throw-in coach, in the search of taking advantage of even these set-pieces. An aspect of the game that has gone relatively unaddressed since Rory Delap utilized his signature long throw-ins to cause issues that teams had no experience defending from. Delap directly assisted five goals from these throw-ins, [57], an advantage teams seem to be acknowledging today, with Delap having since retired in 2013. Additionally, Liverpool have employed the help of Neuro11, a data-based sports science company, which aims to extract the best quality from their set-piece takers.

It was also found that defensive players and midfielders are more important than attackers, when it comes to deciding who will win a football match. This was found in the variable importance charts but was also backed up by the transfer data behind the final predicted table. With teams who had strengthened the quality of their defence performing better than those who recently let their better defenders leave but had strengthened their attacking ranks. This information could inform transfer strategies, especially to teams struggling financially, due to Covid-19 or for other reasons, who may only be able to afford one marquee signing a season. Prioritising a defender would prove astute, over the signing of a striker. It also seems that this is most fiscally responsible as attackers are much more expensive than defenders. It would seem that attackers, while contributing less to a win, sell more tickets or merchandise and are therefore more expensive than defenders, who are far more valuable to

success. As a final point, to show the accuracy of the model and its important variables with regards to defenders being more valuable than attackers, looking at the expected goals, xG, from the 2018/19 season highlights this further. Of the five worst performing attacking teams, teams who did not score as many goals as they should have based on the currently implemented xG models, three of them were not relegated. Implying that while they were unable to score goals, they were able to grind out results by being resilient in defence.

It was stated that this study could be used to help teams with transfer policies or tactical analysis, it could aid sports analysts in their dissection of the game or help those with gambling issues. At the conclusion of this study, this paper can provide the basis or fundamentals of an application that functions as a betting guide or indicator. As more seasons of data get added to the learning model, there will be fewer teams it hasn't encountered and can only grow in accuracy. One hour before kick-off for any given match, it could be updated to provide the most likely result of any game. Additionally, it is believed that this model could be applied to any other sport in order to achieve the same function, it need only substitute the different metrics and variables that are used in the new sport to function.

LIMITATIONS

The research did, however, bring to light some issues during this study. Methods such as Random Forest performed so poorly, with an accuracy metric of less than pure chance. The reason for this was unclear, yet it may be due to the data being used not being suitable for this model. Additionally, Naïve Bayes is a model that was used in the literature to achieve high accuracy ratings, Owrampur et al. (2013), [25]. While the issues relating to that study have been mentioned, hyper focusing on one very successful team, the model was still so much more powerful than the model that was created in this study.

It must also be noted that both the Support Vector Machine model and the Neural Network model proved computationally stressful on the machine used to build the models. The SVM model frequently slowed the machine down or froze the R Studio environment, requiring a soft reboot. Additionally, the NN model which boasts incredible power in other tasks was made less powerful in this study due to the hardware limitations. NN functions best when additional node layers can be placed into the model, as seen in Figure 2. However, due to the power the machine possessed, it was impossible to insert any additional node layers into the model. As such, a work around model was produced. It was as accurate as some of the other models in the end, but it remains to be seen whether more powerful hardware which could insert these additional nodes would have produced a more accurate model, comparable to kNN.

Aside from hardware or computational issues, there are other noteworthy limitations in this study. Firstly, all the player ratings used are based on historical data. This means that a players true rating is represented in the following season. Ruben Dias, for example, was rated 81 in FIFA 21, but, after performing well all season, earning his multi-million-pound move to Manchester City, his rating in FIFA 22 rose to 89. So, while rated at 81, he was performing to the standard of an 89 rated player. In the data set though, no matter how well he is playing his rating of 81 will be used, which directly alters the values of the variables and thus the accuracy of the model. Instances such as these cause discrepancies between the data provided and the learning that can be gained.

This rating can also be decreased through no fault of the player. Mohamed Salah, for example, saw his overall decrease from 90 in FIFA 21 to 89 FIFA 22 after Liverpool had a disappointing season finishing fourth. However, while the team may have performed below the standards expected of them, Salah himself was the top-scorer in the league that season. His rating decreased, even though he had a good goal-scoring season. Additionally, these ratings do not consider how long it may take a player to adapt to their new clubs or managers in the

event of a transfer or manager change. This means a player may have a high rating but perform poorly due to getting used to new tactics or their team-mates which again directly effects the variable ratings. Finally, on player rating limitations, the ratings attributed to players do not account for players being played out of position. For instance, Virgil van Dijk missed most of the 2019/20 season through injury, with players like Fabinho and Jordan Henderson, midfield players, filling the injured defender's position. While each player is rated highly, center back is neither players natural, nor is it their preferred, position to play. As such, it would follow that each player would not be able to perform to their full capabilities as they perform in a position that is uncommon to them. This decrease in full ability is unaccounted for across the whole season.

There are also noteworthy matches such as Norwich and Manchester City, where Manchester City named nine players to their substitute bench, most of which were youth players. This number of youth players brings down the average quality of the bench, but none got any minutes in the match as the senior established players were turned to.

Part of this study was incorporating psychological variables. These psychological variables all took on factors that could still be quantified, however, there are other psychological factors that cannot be input into a data set that could play a significant part on a player's mental state. "Manager Bounce" the phenomena where teams play better when a new manager is brought into the club mid-season, is unaccounted for as it is unknown whether teams will adapt to the new manager or not. Of the four teams that replaced managers during the 2020/21 season, those teams being Tottenham, Chelsea, Wolves and West Brom, each team won more points per game than they had prior to the change, [58]. Whereas when Crystal Palace hired Frank de Boer, results were poor, and he was fired within seventy-seven days. Additional outside variables that most certainly affect players psychologically also have not been accounted for as they are difficult to quantify. Events such as the proposed Super League

discussions which led to many players issuing statements against their clubs, deaths of family members, losing key players, such as Liverpool losing Sadio Mane who was African Player of the Year in 2022, or extreme cases such as Mason Greenwood's legal battle. All psychologically impactful, but difficult to quantify and account for in this study based on prediction.

The study also does not take into account player fatigue, or whether a team's priorities are on the Premier League or one of the domestic or European cup competitions. Additional variables were also sought out to be incorporated into the study, chiefly injuries. However, quotes for thousands of euros ended the chance of getting these variables for the study. This is unfortunate as injuries would have been a significant feature, as injuries to first team players can have a huge impact on a season. Liverpool saw this first hand as in the 2020/21 season they saw first team players miss a combined two-hundred and twenty-seven games, decimating any chance they had at fighting for the title.

Finally, Covid-19 also altered the 2019/20 season of data as more substitutions were allowed and, more importantly, there was a suspension of all games played between the 13th March 2020 and 17th June 2020, [59]. Impacting player wellbeing, psychologically and physically, and providing a season's experience that no player had to deal with prior.

REFLECTION AND FUTURE WORK

Upon the completion of this study, it can be viewed as a success. Certain issues arose in the first modelling stage, but R Studio is a comprehensive enough language to allow these issues to be worked around. The CRISP-DM methodology helped provide structure and worked in tandem with the positivist research philosophy. The most time-consuming aspect of this study was the compilation of the updated statistics data set. This required gathering facts and figures

from many different sources, building player compendiums and accurately placing everything where it needed to be for all 1140 matches across the three seasons of selected data.

The MSc in Data Analytics course set this study up for the success it achieved. Developing code writing skills from zero knowledge in both R Studio and Python, to a level where writing a dissertation requiring these skills is compulsory, is a manageable and achievable task. The guidance provided prior to any undertaking was also pivotal, informing the choice of studying an issue that there is an interest in and has a need to be addressed. The author will continue to update this model, game after game, to assess its power in the coming season now being able to utilize three seasons worth of learning. Taking note of whether Nottingham Forest will cause the same effect as Brentford, having also not been viewed by the model previously.

Going forward, this paper can provide the basis for future research in this field. Adapting this model to include additional features such as injuries, a way of accounting for more psychological factors and events, adjusting player ratings depending on the position they are playing and whether it is their natural position or not and in general create a more robust data set, which could also be applied to other sports such as basketball or American football. Altering the model to be able to be used during the course of a game, as statistics get added or players get substituted out would also be a worthwhile progression of this research. Additionally, a development of an application that could be used as aide for making more informed gambling decisions could be created, however, certain ethical boundaries must be addressed in this instance.

REFERENCES

- [1] PFSA, “<https://thepfsa.co.uk/>,” [Online]. Available: <https://thepfsa.co.uk/football-history/>. [Accessed 15 June 2022].
- [2] S. Das, “sportsbrowser.net,” 23 January 2022. [Online]. Available: <https://sportsbrowser.net/most-popular-sports/>. [Accessed 7 April 2022].
- [3] C. Wright, “<https://www.espn.com/>,” 1 August 2022. [Online]. Available: <https://www.espn.com/soccer/blog-the-toe-poke/story/4710453/womens-attendances-have-dominated-european-football-in-2022>. [Accessed 4 August 2022].
- [4] “<https://www.premierleague.com/>,” [Online]. Available: <https://www.premierleague.com/this-is-pl/the-fans/686489?articleId=686489#:~:text=The%20Premier%20League%20is%20the,90%20minutes%20of%20unpredictable%20action.&text=Neville%2C%20Sky%20Sports-,The%20Premier%20League%20draws%20the%20highest%20global%20telev>. [Accessed 15 June 2022].
- [5] sportingindex, “<https://www.sportingindex.com/>,” 21 October 2021. [Online]. Available: <https://www.sportingindex.com/spread-betting-blog/premier-league-viewing-figures>. [Accessed June 2022].
- [6] K. Fansler, “<https://worldsoccertalk.com/>,” 4 May 2022. [Online]. Available: <https://worldsoccertalk.com/2022/05/04/how-many-people-watch-the-world-cup/>. [Accessed 15 June 2022].
- [7] S. Lock, “<https://www.statista.com/>,” 10 March 2022. [Online]. Available: <https://www.statista.com/statistics/270728/market-volume-of-online-gaming-worldwide/#:~:text=Market%20size%20of%20online%20gambling%20worldwide%202019%2D2023&text=The%20global%20online%20gambling%20market,double%20in%20the%20upcoming%20years..> [Accessed 19 March 2022].
- [8] businessfirstonline, “<https://www.businessfirstonline.co.uk/>,” 23 July 2020. [Online]. Available: <https://www.businessfirstonline.co.uk/articles/what-are-the-most-popular-sports-to-bet-on-in-the-uk/>. [Accessed 15 June 2022].
- [9] pledgesports, “<https://www.pledgesports.org/>,” [Online]. Available: <https://www.pledgesports.org/2020/05/most-popular-sports-to-bet->

- [30] A. T. A. D. A. V. Z. & K. A. Gangal, "Analysis and Prediction of Football Statistics using Data Mining Techniques," *International Journal of Computer Applications*, vol. 132, pp. 8-11, 2015.
- [31] J. K. M. S. M. M. W. Y. J. & L. G. Koza, "Genetic Programming IV: Routine Human Competitive Machine Intelligence," 2003.
- [32] A. T. A. D. A. Z. V. & K. A. Gangal, "Analysis and Prediction of Football Statistics using Data Mining Techniques," *International Journal of Computer Applications*, vol. 132, pp. 8-11, 2017.
- [33] B. Lantz, *Machine Learning with R*, Birmingham: Packt Publishing, 2019.
- [34] H. & V. K. M. a. R. Korjus, "An Efficient Data Partitioning to Improve Classification Performance While Keeping Parameters Interpretable," *PLoS One*, vol. 1, no. 8, p. e0161788, 2016.
- [35] Statisticsolutions, "www.statisticssolutions.com," [Online]. Available: <https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/what-is-logistic-regression/>. [Accessed 07 April 2022].
- [36] J. & K. M. A. Starkweather, "https://it.unt.edu/sites/default/files/mlr_jds_aug2011.pdf," [Online]. Available: https://it.unt.edu/sites/default/files/mlr_jds_aug2011.pdf. [Accessed 7 April 2022].
- [37] M. Chatterjee, "www.mygreatlearning.com," Great Learning, 3 February 2020. [Online]. Available: <https://www.mygreatlearning.com/blog/knn-algorithm-introduction/>. [Accessed 7 April 2022].
- [38] IBM Cloud Education, "www.ibm.com," IBM, 17 August 2020. [Online]. Available: <https://www.ibm.com/cloud/learn/neural-networks>. [Accessed 7 April 2022].
- [39] mastersindatascience, "https://www.mastersindatascience.org," 2022. [Online]. Available: <https://www.mastersindatascience.org/learning/introduction-to-machine-learning-algorithms/decision-tree/>. [Accessed 09 June 2022].
- [40] T. Yiu, "https://towardsdatascience.com," 12 June 2019. [Online]. Available: <https://towardsdatascience.com/understanding-random-forest-58381e0602d2>. [Accessed 10 June 2022].

- [41] S. Ray, “<https://www.analyticsvidhya.com/>,” 13 September 2017. [Online]. Available: <https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/>. [Accessed 15 June 2022].
- [42] nvidia, “<https://www.nvidia.com/>,” 2022. [Online]. Available: <https://www.nvidia.com/en-us/glossary/data-science/xgboost/>. [Accessed 10 June 2022].
- [43] J. Dudovskiy, “<https://research-methodology.net/>,” 2022. [Online]. Available: <https://research-methodology.net/research-philosophy/>. [Accessed 16 June 2022].
- [44] K. Thompson, “<https://revisesociology.com/>,” 18 May 2015. [Online]. Available: <https://revisesociology.com/2015/05/18/positivism-interpretivism-sociology/>. [Accessed 16 June 2022].
- [45] J. Dudovskiy, “<https://research-methodology.net/>,” [Online]. Available: <https://research-methodology.net/research-philosophy/positivism/>. [Accessed 16 June 2022].
- [46] “<https://www.nottingham.ac.uk/>,” [Online]. Available: [https://www.nottingham.ac.uk/helmpen/rlos/research-evidence-based-practice/designing-research/types-of-study/understanding-pragmatic-research/section03.html#:~:text=Pragmatism%20involves%20research%20designs%20that,find%20solutions%20to%20research%20prob](https://www.nottingham.ac.uk/helmpen/rlos/research-evidence-based-practice/designing-research/types-of-study/understanding-pragmatic-research/section03.html#:~:text=Pragmatism%20involves%20research%20designs%20that,find%20solutions%20to%20research%20prob.). [Accessed 16 June 2022].
- [47] S. & C. P. Guha Thakurta, “<https://www.projectguru.in/>,” 25 June 2015. [Online]. Available: <https://www.projectguru.in/research-philosophy/>. [Accessed 16 June 2022].
- [48] R. Purtil, “The Purpose of Science,” *Philosophy of Science*, vol. 37, no. 2, pp. 301-306, 1970.
- [49] C. Wright, “<https://www.espn.com/>,” 14 September 2021. [Online]. Available: <https://www.espn.com/soccer/blog-the-toe-poke/story/4474417/fifa-22-player-ratings-ronaldo-bumped-out-of-top-two-messi-still-no-1>. [Accessed 12 June 2022].
- [50] football-lineups, “<https://m.football-lineups.com/team/Aston-Villa/FA-Premier-League-2021--2022/fixture/>,” [Online]. Available: <https://m.football-lineups.com/team/Aston-Villa/FA-Premier-League-2021--2022/fixture>. [Accessed 16 June 2022].

- [51] U. A. R. C. S. M. a. D. Analytics, “<https://stats.oarc.ucla.edu/>,” 2021. [Online]. Available: <https://stats.oarc.ucla.edu/other/mult-pkg/whatstat/what-is-the-difference-between-categorical-ordinal-and-interval-variables/>. [Accessed 12 June 2022].
- [52] A. Fawcett, “<https://www.educative.io/>,” 11 February 2021. [Online]. Available: <https://www.educative.io/blog/one-hot-encoding>. [Accessed 12 June 2022].
- [53] Zach, “<https://www.statology.org/normalize-data-between-0-and-100/>,” 30 November 2020. [Online]. Available: <https://www.statology.org/normalize-data-between-0-and-100/>. [Accessed 4 August 2022].
- [54] Scikit, “<https://scikit-learn.org/>,” 2022. [Online]. Available: https://scikit-learn.org/stable/auto_examples/model_selection/plot_confusion_matrix.html. [Accessed 25 July 2022].
- [55] F. Desk, “<https://www.footballtransfers.com/>,” 2022 July 14. [Online]. Available: <https://www.footballtransfers.com/en/transfer-news/it-serie-a/2021/04/football-50-biggest-transfers-of-all-time>. [Accessed 15 July 2022].
- [56] PremierLeague, “<https://www.premierleague.com/>,” 26 July 2020. [Online]. Available: <https://www.premierleague.com/news/1561869#:~:text=Most%20points%20won%20over%2038,in%202018%20and%202005%20respectively..> [Accessed 26 July 2022].
- [57] “<https://www.espn.com/>,” 3 May 2020. [Online]. Available: <https://www.espn.com/soccer/english-premier-league/story/4090887/how-rory-delaps-long-throw-ins-ruffled-arsene-wengers-parka-and-showed-us-all-their-importance>. [Accessed 11 August 2022].
- [58] B. McAleer, “<https://www.theguardian.com/>,” 25 November 2021. [Online]. Available: <https://www.theguardian.com/football/who-scored-blog/2021/nov/25/investigating-new-manager-bounce-premier-league>. [Accessed 27 July 2022].
- [59] J. Prince-Wright, “<https://soccer.nbcsports.com/>,” 20 March 2021. [Online]. Available: <https://soccer.nbcsports.com/2021/03/20/premier-league-covid/>. [Accessed 27 July 2022].
- [60] R. & Benjamin, “Skill and Chance in Association Football,” *Journal of the Royal Statistical Society. Series A*, vol. 131, no. 4, pp. 581-585, 1968.
- [61] N. Hotz, “<https://www.datascience-pm.com/>,” 16 April 2022. [Online]. Available: <https://www.datascience-pm.com/crisp-dm-2/>. [Accessed 16 June 2022].

MODEL 1.1: KNN – BASE FOOTBALL STATISTICS

```
data = read.csv("UpdatedCombinedSeasonsData.csv")
#knn

library(class)
library(gmodels)

#remove extra columns
data = data[, -c(1,2,3)]

#create normalize and accuracy functions
normalize <- function(x){
  return((x - min(x))/ (max(x)-min(x)))
}

accuracy = function(x){
  sum(diag(x)/(sum(rowSums(x))))* 100
}

#convert target variable to numeric and morph to data frame
data$FTR = as.numeric(as.factor(data$FTR))

data = as.data.frame(lapply(data[1:59], normalize))

#data train and test split
data_train = data[1:608,]
data_test = data[608:760,]

data_train_labels = data[1:608,1]
data_test_labels = data[608:760,1]

#train knn
data_test_pred = knn(train = data_train, test = data_test,
                     cl = data_train_labels, k = 29)

#confusion matrix
CrossTable(x = data_test_labels, y = data_test_pred,
           prop.chisq = F)

#assess accuracy
target = as.factor(data[608:760,1])
tb = table(data_test_pred, target)
accuracy(tb)
```

MODEL 1.2: KNN – UPDATED FOOTBALL STATISTICS

```
data = read.csv("UpdatedCombinedSeasonsDataAdditional.csv")
#knn

#import packages
library(class)
library(gmodels)

#remove extra columns
data = data[, -c(1,2,3)]
data <- data[-c(761,762,763,764,765), ]

#create normalize and accuracy functions
normalize <- function(x){
  return((x - min(x))/ (max(x)-min(x)))
}

accuracy = function(x){
  sum(diag(x)/(sum(rowSums(x))))* 100
}

#convert target variable to numeric and morph to data frame
data$FTR = as.numeric(as.factor(data$FTR))

data = as.data.frame(lapply(data[1:89], normalize))

#data train and test split
data_train = data[1:608,]
data_test = data[608:760,]

data_train_labels = data[1:608,1]
data_test_labels = data[608:760,1]

#train knn
data_test_pred = knn(train = data_train, test = data_test,
                     cl = data_train_labels, k = 19)

data_test_pred

#confusion matrix
CrossTable(x = data_test_labels, y = data_test_pred,
           prop.chisq = F)

#assess accuracy
target = as.factor(data[608:760,1])
tb = table(data_test_pred, target)
accuracy(tb)
```

MODEL 2.1: DECISION TREE – BASE FOOTBALL STATISTICS

```
data = read.csv("UpdatedCombinedSeasonsData.csv")
#DT

#import packages
library(tidyverse)
library(caret)
library(rpart)
library(rattle)
library(dplyr)

#drop columns
data = data[, -c(1,2,3)]

data$FTR = as.factor(data$FTR)

#training and testing split
set.seed(123)
trainingdata = data$FTR %>%
  createDataPartition(p = .8, list = F)
train.data = data[trainingdata, ]
test.data = data[-trainingdata, ]

#create model
model1 = rpart(FTR ~., data = train.data, method = "class")
plot(model1)
text(model1, digits = 3)

#make more aesthetically pleasing
fancyRpartPlot(model1, box.palette = "GnRd", type = 4, cex = .6)

#create confusion matrix
predicted.classes = model1 %>%
  predict(test.data, type = "class")
head(predicted.classes)

mean(predicted.classes == test.data$FTR)

printcp(model1)

#assess accuracy
confusionMatrix(predicted.classes, test.data$FTR)

#import packages
library(DAAG)
library(party)
library(rpart)
library(rpart.plot)
library(mlbench)
library(caret)
library(pROC)
library(tree)
library(dplyr)

#drop columns
data = data[, -c(1,2,3)]

#data train test split
data_train = data[1:608,]
data_test = data[608:760,]

#create model
model2 = rpart(FTR ~., data = train.data, method = "class")

data = as.data.frame(data)

data_train_labels = data[1:608, 1]
data_test_labels = data[608:760, 1]

data$FTR = as.factor(data$FTR)

#new decision tree variant
dtree = rpart(FTR ~., data)
dtree = rpart.plot(dtree, yesno = T, tweak = 1.2, type = 5, fallen.leaves = T)

#create confusion matrix
predicted.classes = model2 %>%
  predict(data_test, type = "class")
head(predicted.classes)

#assess accuracy
mean(predicted.classes == data_test$FTR)
confusionMatrix(predicted.classes, as.factor(data_test$FTR))

model <- rpart(FTR ~., data, method = "class")
par(xpd = NA)
plot(model)
text(model, digits = 3)
print(model, digits = 2)
```

MODEL 2.1: DECISION TREE – UPDATED FOOTBALL STATISTICS

```
data = read.csv("UpdatedCombinedSeasonsDataAdditional.csv")
#DT

#import packages
library(tidyverse)
library(caret)
library(rpart)
library(rattle)
library(dplyr)

#drop columns
data = data[, -c(1,2,3)]
data <- data[-c(761,762,763,764,765), ]

data$FTR = as.factor(data$FTR)

#training and testing split
set.seed(123)
trainingdata = data$FTR %>%
  createDataPartition(p = .8, list = F)
train.data = data[trainingdata, ]
test.data = data[-trainingdata, ]

#create model
model1 = rpart(FTR ~., data = train.data, method = "class")
plot(model1)
text(model1, digits = 3)

#make model more aesthetically pleasing
fancyRpartPlot(model1, box.palette = "GnRd", type = 4, cex = .6)

#create confusion matrix
predicted.classes = model1 %>%
  predict(test.data, type = "class")
head(predicted.classes)

mean(predicted.classes == test.data$FTR)

printcp(model1)

#assess accuracy
confusionMatrix(predicted.classes, test.data$FTR)

data = read.csv("UpdatedCombinedSeasonsDataAdditional.csv")

#drop columns
data = data[, -c(1,2,3)]
data <- data[-c(761,762,763,764,765), ]

#import accuracy
library(DAAG)
library(party)
library(rpart)
library(rpart.plot)
library(mlbench)
library(caret)
library(pROC)
library(tree)
library(dplyr)

#train test split
data_train = data[1:608,]
data_test = data[608:760,]

#create model
model2 = rpart(FTR ~., data = data_train, method = "class")

data = as.data.frame(data)

data_train_labels = data[1:608, 1]
data_test_labels = data[608:760, 1]

data$FTR = as.factor(data$FTR)

#new decision tree variant
dtree = rpart(FTR ~., data)
dtree = rpart.plot(dtree, yesno = T, tweak = 1.2, type = 5, fallen.leaves = T)

#create confusion matrix
predicted.classes = model2 %>%
  predict(data_test, type = "class")
head(predicted.classes)

#assess accuracy
mean(predicted.classes == data_test$FTR)
confusionMatrix(predicted.classes, as.factor(data_test$FTR))

model <- rpart(FTR ~., data, method = "class")
par(xpd = NA) # otherwise on some devices the text is clipped
plot(model)
text(model, digits = 3)
print(model, digits = 2)
```

MODEL 3.1: RANDOM FOREST – BASE FOOTBALL STATISTICS

```
data = read.csv("UpdatedCombinedSeasonsData.csv")
#randomForest

#drop columns
data = data[, -c(1,2,3)]

#import packages
library(data.table)
library(mlr)
library(tidyverse)
library(xgboost)
require(xgboost)
library(caret)
library(randomForest)
library(vcd)
library(ROCR)

#convert target variable to numeric
set.seed(1000)
data$FTR = as.numeric(as.factor(data$FTR))

#train test split
index = createDataPartition(data$FTR, p = .8, list = FALSE)
train = data[index,]
test = data[-index,]

#rf model
rf <- randomForest(FTR~., data=train, proximity=TRUE)

print(rf)

#rounding so model works correctly, keep to 3 classes
p1 <- predict(rf, train)
p1 <- round(p1)
p1 <- as.factor(p1)

#training confusion matrix
confusionMatrix(as.factor(p1), as.factor(train$FTR))

#rounding so model works correctly, keep to 3 classes
p2 = predict(rf, test)
p2 = round(p2)
p2 = as.factor(p2)

#testing confusion matrix
confusionMatrix(as.factor(p2), as.factor(test$FTR))

#variable importance plot|
varImpPlot(rf,
            sort = T,
            n.var = 10,
            main = "Top 10 - Variable Importance")
```

MODEL 3.2: RANDOM FOREST – UPDATED FOOTBALL STATISTICS

```
data = read.csv("UpdatedCombinedSeasonsDataAdditional.csv")
#randomForest

#drop columns
data = data[, -c(1,2,3)]
data <- data[-c(761,762,763,764,765), ]

#import packages
library(data.table)
library(mlr)
library(tidyverse)
library(xgboost)
require(xgboost)
library(caret)
library(randomForest)
library(vcd)
library(ROCR)

#convert target variable to numeric
set.seed(1000)
data$FTR = as.numeric(as.factor(data$FTR))

#train test split
index = createDataPartition(data$FTR, p = .8, list = FALSE)
train = data[index,]
test = data[-index,]

#rf model
rf <- randomForest(FTR~., data=train, proximity=TRUE)

print(rf)

# rounding for both training and testing confusion matrices
p1 <- predict(rf, train)
p1 <- round(p1)
p1 <- as.factor(p1)

p2 = predict(rf, test)
p2 = round(p2)
p2 = as.factor(p2)

#assessing accuracy of each matrix
confusionMatrix(as.factor(p1), as.factor(train$FTR))
confusionMatrix(as.factor(p2), as.factor(test$FTR))

#variable importance plot|
varImpPlot(rf,
           sort = T,
           n.var = 10,
           main = "Top 10 - Variable Importance")
```

MODEL 4.1: SUPPORT VECTOR MACHINE – BASE FOOTBALL

STATISTICS

```
data = read.csv("UpdatedCombinedSeasonsData.csv")
#SVM

#drop columns
data = data[, -c(1,2,3)]

#import packages
library(e1071)
library(rpart)
library(kernlab)
library(caret)

#train test split
data_train = data[1:608,]
data_test = data[608:760,]

#training control
trainctrl <- trainControl(method = "repeatedcv", number = 10, repeats = 3)

#train svm model
svm_Linear <- caret::train(FTR ~., data = data_train, method = "svmLinear",
                           trControl=trainctrl,
                           preProcess = c("center", "scale"),
                           tuneLength = 10)

svm_Linear

#test model
test_pred = predict(svm_Linear, newdata = data_test)
test_pred

#assess accuracy
confusionMatrix(table(test_pred, data_test$FTR))

#refine model
grid <- expand.grid(C = c(0,0.01, 0.05, 0.1, 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2,5))
svm_Linear_Grid <- caret::train(FTR ~., data = data_train, method = "svmLinear",
                                trControl=trainctrl,
                                preProcess = c("center", "scale"),
                                tuneGrid = grid,
                                tuneLength = 10)

svm_Linear_Grid
plot(svm_Linear_Grid)
test_pred_grid <- predict(svm_Linear_Grid, newdata = data_test)
test_pred_grid

#assess accuracy
confusionMatrix(table(test_pred_grid, data_test$FTR))
```

MODEL 4.2: SUPPORT VECTOR MACHINE – UPDATED FOOTBALL

STATISTICS

```
data = read.csv("UpdatedCombinedSeasonsDataAdditional.csv")
#SVM

#drop columns
data = data[, -c(1,2,3)]
data <- data[-c(761,762,763,764,765), ]

#import packages
library(e1071)
library(rpart)
library(kernlab)
library(caret)

#train test split
data_train = data[1:608,]
data_test = data[608:760,]

#training control
trainctrl <- trainControl(method = "repeatedcv", number = 10, repeats = 3)

#train svm model
svm_Linear <- caret::train(FTR ~., data = data_train, method = "svmLinear",
                           trControl=trainctrl,
                           preProcess = c("center", "scale"),
                           tuneLength = 10)

svm_Linear

#test model
test_pred = predict(svm_Linear, newdata = data_test)
test_pred

#assess accuracy
confusionMatrix(table(test_pred, data_test$FTR))

#refine model
grid <- expand.grid(C = c(0,0.01, 0.05, 0.1, 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2,5))
svm_Linear_Grid <- caret::train(FTR ~., data = data_train, method = "svmLinear",
                                trControl=trainctrl,
                                preProcess = c("center", "scale"),
                                tuneGrid = grid,
                                tuneLength = 10)

svm_Linear_Grid
plot(svm_Linear_Grid)
test_pred_grid <- predict(svm_Linear_Grid, newdata = data_test)
test_pred_grid

#assess accuracy
confusionMatrix(table(test_pred_grid, data_test$FTR))
```

MODEL 5.1: NEURAL NETWORK – BASE FOOTBALL STATISTICS

```
data = read.csv("UpdatedCombinedSeasonsData.csv")
#Neural Networks

#import packages
library(neuralnet)

#drop columns
data = data[, -c(1,2,3)]

#create normalize function
normalize <- function(x){
  return((x - min(x))/ (max(x)-min(x)))
}

#convert target variable to a factor then numeric
data$FTR = as.numeric(as.factor(data$FTR))

#morph to data frame, normalize
data = as.data.frame(lapply(data, normalize))

#train test split
data_train = data[1:608,]
data_test = data[608:760,]

#create model
data_model = neuralnet(FTR ~., data = data_train)

#plot(data_model)

#assess accuracy
model_results = neuralnet::compute(data_model, data_test[1:59])

predicted_FTR = model_results$net.result

cor(predicted_FTR, data_test$FTR)
```

MODEL 5.2: NEURAL NETWORK – UPDATED FOOTBALL STATISTICS

```
data = read.csv("UpdatedCombinedSeasonsDataAdditional.csv")
#Neural Networks

#import packages
library(neuralnet)

#drop columns
data = data[, -c(1,2,3)]
data <- data[-c(761,762,763,764,765), ]

#create normalize functions
normalize <- function(x){
  return((x - min(x))/ (max(x)-min(x)))
}

#convert target variable to a factor then numeric
data$FTR = as.numeric(as.factor(data$FTR))

#morph to data frame, normalize
data = as.data.frame(lapply(data, normalize))

#train test split
data_train = data[1:608,]
data_test = data[608:760,]

#create model
data_model = neuralnet(FTR ~., data = data_train)

#plot(data_model)

#assess accuracy
model_results = neuralnet::compute(data_model, data_test[1:89])

predicted_FTR = model_results$net.result
cor(predicted_FTR, data_test$FTR)
```

MODEL 6.1: NAÏVE BAYES – BASE FOOTBALL STATISTICS

```
data = read.csv("UpdatedCombinedSeasonsData.csv")
#naive bayes

#drop columns
data = data[, -c(1,2,3)]

#import packages
library(tidyverse)
library(ggplot2)
library(caret)
library(caretEnsemble)
library(psych)
library(Amelia)
library(mice)
library(GGally)
library(rpart)
library(randomForest)
library(e1071)
library(klaR)
library(naivebayes)
library(dplyr)
library(ggplot2)
library(psych)
library(randomForest)

#make target variable factor
data$FTR = factor(data$FTR)

#train test split
indxTrain <- createDataPartition(y = data$FTR,p = 0.8,list = FALSE)
training <- data[indxTrain,]
testing <- data[-indxTrain,]

#view breakdown of results
prop.table(table(data$FTR)) * 100

#train model
x = training[,-1]
y = training$FTR
model = caret::train(x,y,'nb',trControl=trainControl(method='cv',number=10))
model

#assess accuracy
Predict <- predict(model,newdata = testing )
confusionMatrix(Predict, testing$FTR)

#view important variables
X = varImp(model)
X
plot(X)
```

MODEL 6.2: NAÏVE BAYES – UPDATED FOOTBALL STATISTICS

```
data = read.csv("UpdatedCombinedSeasonsDataAdditional.csv")
#naive bayes

#drop columns
data = data[, -c(1,2,3)]
data <- data[-c(761,762,763,764,765), ]

#import packages
library(tidyverse)
library(ggplot2)
library(caret)
library(caretEnsemble)
library(psych)
library(Amelia)
library(mice)
library(GGally)
library(rpart)
library(randomForest)
library(e1071)
library(klaR)
library(naivebayes)
library(dplyr)
library(ggplot2)
library(psych)
library(randomForest)

#make target variable a factor
data$FTR = factor(data$FTR)

#view data set
str(data)
head(data)
describe(data)

#train test split
indxTrain <- createDataPartition(y = data$FTR,p = 0.8,list = FALSE)
training <- data[indxTrain,]
testing <- data[-indxTrain,]

#view breakdown of results
prop.table(table(data$FTR)) = 100

#train model
x = training[,-1]
y = training$FTR
model = caret::train(x,y,'nb',trControl=trainControl(method='cv',number=10))
model

#assess accuracy
Predict <- predict(model,newdata = testing )
confusionMatrix(Predict, testing$FTR)

#view important variables|
X = varImp(model)
X
plot(X)
```

MODEL 7.1: XGBOOST – BASE FOOTBALL STATISTICS

```
data = read.csv("UpdatedCombinedSeasonsData.csv")
#xgBoost

#drop columns
data = data[, -c(1,2,3)]

#import packages
library(data.table)
library(mlr)
library(tidyverse)
library(xgboost)
require(xgboost)
library(caret)

#convert target variable to a factor then numeric
data$FTR = as.numeric(as.factor(data$FTR))

#train test split
index = createDataPartition(data$FTR, p = .8, list = FALSE)
train = data[index,]
test = data[-index,]

#convert train test sets to matrix
train_x = as.matrix(train[,-1])
train_y = train[,1]

test_x = data.matrix(test[,-1])
test_y = test[,1]

#create train test xgb matrix from created matrix
xgb_train = xgb.DMatrix(data = train_x, label = train_y)
xgb_test = xgb.DMatrix(data = test_x, label = test_y)

#build model
watchlist = list(train=xgb_train, test=xgb_test)
model = xgb.train(data = xgb_train, max.depth = 3, watchlist=watchlist, nrounds = 70)
final = xgboost(data = xgb_train, max.depth = 3, nrounds = 50, verbose = 0)
final

#assess accuracy
pred_y = predict(final, xgb_test)
mean((test_y - pred_y)^2)
caret::MAE(test_y, pred_y)
caret::RMSE(test_y, pred_y)
```

MODEL 7.2: XGBOOST – UPDATED FOOTBALL STATISTICS

```
data = read.csv("UpdatedCombinedSeasonsDataAdditional.csv")
#xgBoost

#drop columns
data = data[, -c(1,2,3)]
data <- data[-c(761,762,763,764,765), ]

#import packages
library(data.table)
library(mlr)
library(tidyverse)
library(xgboost)
require(xgboost)
library(caret)

#convert target variable to a factor then numeric
data$FTR = as.numeric(as.factor(data$FTR))

#train test split
index = createDataPartition(data$FTR, p = .8, list = FALSE)
train = data[index,]
test = data[-index,]

#convert train test sets to matrix
train_x = as.matrix(train[,-1])
train_y = train[,1]

test_x = data.matrix(test[,-1])
test_y = test[,1]

#create train test xgb matrix
xgb_train = xgb.DMatrix(data = train_x, label = train_y)
xgb_test = xgb.DMatrix(data = test_x, label = test_y)

#build model
watchlist = list(train=xgb_train, test=xgb_test)
model = xgb.train(data = xgb_train, max.depth = 3, watchlist=watchlist, nrounds = 70)
final = xgboost(data = xgb_train, max.depth = 3, nrounds = 50, verbose = 0)
final

#assess accuracy
pred_y = predict(final, xgb_test)
mean((test_y - pred_y)^2)
caret::MAE(test_y, pred_y)
caret::RMSE(test_y, pred_y)
```

MODEL 8.1: MULTINOMIAL LOGISTIC REGRESSION – BASE FOOTBALL STATISTICS

```
data = read.csv("UpdatedCombinedSeasonsData.csv")
#Multinomial Regression

#drop columns
data = data[, -c(1,2,3)]

#import packages
library(readr)
library(caret)
library(dplyr)

#convert target variable to factor
data$FTR = as.factor(data$FTR)

#train test split
index = createDataPartition(data$FTR, p = .8, list = FALSE)
train = data[index,]
test = data[-index,]

require(nnet)

#create and view model
multinom.fit = multinom(FTR ~., data = train)
summary(multinom.fit)
exp(coef(multinom.fit))
head(probability.table = fitted(multinom.fit))

#train and view accuracy
train$predicted <- predict(multinom.fit, newdata = train, "class")
ctable <- table(train$FTR, train$predicted)
round((sum(diag(ctable))/sum(ctable))*100,2)

#testing accuracy
test$predicted <- predict(multinom.fit, newdata = test, "class")
ctable <- table(test$FTR, test$predicted)
round((sum(diag(ctable))/sum(ctable))*100,2)
```

MODEL 8.2: MULTINOMIAL LOGISTIC REGRESSION – UPDATED

FOOTBALL STATISTICS

```
data = read.csv("UpdatedCombinedSeasonsDataAdditional.csv")
#Multinomial Regression

#drop columns
data = data[, -c(1,2,3)]
data <- data[-c(761,762,763,764,765), ]

#import packages
library(readr)
library(caret)
library(dplyr)

#convert target to factor
data$FTR = as.factor(data$FTR)

#train test split
index = createDataPartition(data$FTR, p = .8, list = FALSE)
train = data[index,]
test = data[-index,]

require(nnet)

#create and view model
multinom.fit = multinom(FTR ~., data = train)
summary(multinom.fit)
exp(coef(multinom.fit))
head(probability.table = fitted(multinom.fit))

#accuracy in training
train$predicted <- predict(multinom.fit, newdata = train, "class")
ctable <- table(train$FTR, train$predicted)
round((sum(diag(ctable))/sum(ctable))*100,2)

#accuracy in testing
test$predicted <- predict(multinom.fit, newdata = test, "class")
ctable <- table(test$FTR, test$predicted)
round((sum(diag(ctable))/sum(ctable))*100,2)
```

FINAL PREDICTION MODEL: KNN – TEST FOOTBALL STATISTICS

```
data = read.csv("UpdatedCombinedSeasonsDataAdditional.csv")
data2 = read.csv("20212022OneHot.csv")
#knn

#import packages
library(class)
library(gmodels)

#drop columns new data set
data2 = data2[, -c(1,2,3)]
#data2 = data2[, -c(84,85,86,87,88,89)]
data2 = data2[-c(381,382,383),]

#create normalize and accuracy function
normalize <- function(x) {
  return ((x - min(x, na.rm = TRUE)) / (max(x, na.rm = TRUE) - min(x, na.rm = TRUE)))}

accuracy = function(x){
  sum(diag(x)/(sum(rowSums(x))))* 100
}

#convert target variable to factor then numeric for new data set
data2$FTR = as.numeric(as.factor(data2$FTR))

#convert new data set to data frame, normalize
data2 = as.data.frame(lapply(data2[1:89], normalize))

#drop columns original data set
data = data[, -c(1,2,3)]

#convert original data set to factor then numeric
data$FTR = as.numeric(as.factor(data$FTR))

#convert to data frame and normalize
data = as.data.frame(lapply(data[1:89], normalize))

#train test split for original knn model
data_train = data[1:608,]
data_test = data[608:760,]

data_train_labels = data[1:608,1]
data_train_labels

data_test_labels = data[608:760,1]
data_test_labels

#original model built and learning aquired
old_season_pred = knn(train = data_train, test = data_test,
  cl = data_train_labels, k = 19)

#learning applied to new data set for prediction purposes
new_test_pred = knn(train = data_train, test = data2,
  cl = data_train_labels, k = 19)

#view predictions - specifically new_test_pred as this is prediction for season
old_season_pred
new_test_pred
```