

National College of Ireland

Data Science

BSHDS4

2022/2023

Christopher Weir

X19317131

X19317131@student.ncirl.ie

Beyond The Beautiful Game:

Leveraging Clustering & Generative Adversarial
Network Techniques for Europe's Top Football
Leagues

Contents

Executive Summary	2
1.0 Introduction	3
1.1. Background	3
1.2. Aims	3
1.3. Use Cases	5
1.4. Technology	7
1.5. Structure	8
2.0 State of the Art	14
2.1. Performance Disparities in Football	14
2.2. Clustering Ideology	17
2.3. Generative Adversarial Networks Ideology	19
3.0 Data	22
4.0 Methodology & Implementation	27
4.1. Methodology	27
4.2. Implementation	27
5.0 Analysis	32
6.0 Results	45
7.0 Conclusions	53
8.0 Further Development or Research	55
9.0 References	58
10.0 Appendices	60
10.1. Project Proposal	60
10.2. Ethics Approval Application (only if required)	71
10.3. Reflective Journals	86

Executive Summary

This project is based on Europe's highest level of football. Throughout this report, multiple techniques and tools will be used to classify teams at the highest level, into a certain level of playing style. Based on multiple statistics scraped from a footballing data website named FBref¹, the aim is to extract data regarding the top 5 European football leagues to build a classification model based on the performance and the strength of each team. Once the teams are successfully classified, a Generative Adversarial Network (GAN) model is built to generate synthetic data given a team's input to push them to compete with Europe's elites. Along with this, for ease of comprehensibility, the terminology involved in the data will be discussed and how data, along with these terms, have become more prevalent in football now than ever before.

By the end of this project, the goal is to build an accurate classification and GAN model that can compare team's data and groups teams together by the level of football instilled into the club by the manager, owners, and culture of the football club itself, and then generate new data to help improve the team and gain them a step in the right direction. This project should also allow for any person, whether a die-hard football fan or a mere casual fan of the game, to understand football playing styles and the impact they have on the biggest sport in the world. With this GAN model, there is a desire to create an interactive dashboard with filters for different data metrics to highlight different data points and where a team's data sits on any given metric in comparison to the data points they should be striving to reach.

Overall, the main aim is to analyse teams' playing styles, taking data from the past five years in their respective leagues, and to provide a complete report, along with a dashboard, a classification model, and a GAN model on what is internationally known as "The Beautiful Game" to allow for any reader to gain a little more knowledge and interest on the wonderful sport, and to acknowledge the impact statistics, and data has on the biggest game in the world.

¹ <https://fbref.com/en/>

1.0 Introduction

1.1. Background

I chose to undertake this project for a couple of reasons.

Since I was a young child, from 8-9 years of age, I have been an avid supporter of Liverpool Football Club who play in the English Premier League, the number one ranked league in Europe by the Union of European Football Association (UEFA) Coefficients. UEFA is the sporting body for the European Professional Leagues and their coefficients are determined by how well teams from each country perform in European competitions, such as the Champions League.

I have grown up watching Liverpool and football in general and have always taken a liking to attempt to interpret how teams play and how football in different countries is played. Of course, when it comes to aspects of data, it is impossible to quantify luck, skill, and unpredictability with numbers. However, there is a lot of data that is now recorded in the modern game which can be quantified and using all this data, it is quite possible to model this data to uncover aspects of football not previously looked at before.

However, due to aspects of football that cannot be run through a model, you always must give space in your analysis for what the common viewer calls “the eye test”. This is purely watching the game to see how players' brains tick, how their feet move, how they dribble and create space and how quick their decision-making is. Nonetheless, data in football has become a huge factor in key operating decisions on and off the pitch, and data can tell us a lot about the way a team plays their football, and I would like to be a part of that story at some point in my career.

Through the years of growing up watching football and as I understood more of the game and the tactical roles of teams. I had a growing interest in the data analysis and the statistical side of the game. It was from this growing interest where data science took my attention, and it is where I want to lead my future career into the world of analysis and data. There are Premier League teams, such as Brentford and Brighton, who are both excelling in the league table, who's recruitment decisions have been based on the background data and the underlying statistics of players who they have brought in to help the team perform better.

It is in these kinds of statistics and this kind of data that I believe has a huge impact on the pitch and on the sport of football as a whole and it is one, that I believe, has a big future in sporting regimes.

1.2. Aims

Data science has become one of the fastest growing phenomena in the sporting industry. Elite football clubs every day are hiring data scientists, analysts, engineers etc. to upgrade their data profile and to gain an edge against their competitors. Clubs at the highest level are also hiring data analysts for the sole purpose of analysing their opposition of who they face in their next game, and understanding where they can improve [1]. It is believed with the model to be created; it will help to automate the process of identifying where their teams' football

needs to improve. From this, the analysis can become quicker, more comprehensive, and more detailed.

The data contains a host of metrics, from passing types and patterns, plays in possession and out of possession, attacking, defending, shot taking among many other metrics. This data will all be centralised using methods of Machine Learning (ML). The types of algorithms to be implemented will be that of unsupervised methods. These methods will classify the data based on the centroids of clusters along a 2D X and Y axis. Depending on where a team sits on this axis, they will be identified into a cluster along with teams boasting similar statistics.

Several methods will be introduced throughout this report, including the likes of K-Means Clustering. The K-Means algorithm is generally the most known and used clustering method. Although it is an unsupervised learning to clustering in pattern recognition and machine learning, the k-means algorithm and its extensions are always influenced by initializations with a necessary number of clusters a priori [2]. To understand the number of clusters necessary for initialization, an elbow-method model will be built in which to gain insight to the best number of clusters for the data presented.

In addition to the K-Means model, a GAN model will be built. This is a deep Neural-Network model. This algorithm involves the use of two models. One is a generator, and one is a discriminator. The generator model takes the dataset as an input and creates new data based off the dataset. This generated data is passed to the discriminator to decipher if this data is from the actual dataset or from the generated 'fake' dataset. This data is passed back and forward until an equilibrium is met where the discriminator model is 50/50 on knowing if the data is from the actual dataset or from the generated dataset. This type of network is usually assigned for the use of image datasets to render AI-generated images.

As this dataset is not very large, GANs can be used to generate additional samples for a tabular dataset, which can be useful for increasing the size of a dataset and reducing overfitting. By generating new samples that are like the existing data, GANs can help to diversify the dataset and provide more representative samples for training machine learning models [3]. Additionally, by generating additional data that captures the underlying patterns and distribution of the original data, GANs can help to improve the performance of machine learning models that are trained on the data. This is very useful as the original dataset is small, GANs help to balance the classes and improve the accuracy of the model.

They can also be used to learn the underlying distribution of the data, which can then help to extract meaningful features from the dataset. These features can then be used as inputs to other machine learning models, improving their performance. In this way, identifying prevalent features from the dataset can be extremely useful in understanding what metrics pose the biggest impact on each cluster. This can make it easier to decipher where a team may need to improve to play a certain type of football or additionally, can decipher where an opposing team's biggest strengths and weaknesses are which plans and tactics can then be set in place to protect and exploit these attributes.

The achievement for this project is to classify teams, based on the data provided by FBref, into certain categories of playing styles. Once the teams are correctly classified and labelled, a visualisation can then be built to portray how strongly teams are linked to each label as they

can transition from time to time within the one game. Additionally, from this, once the teams are correctly classified and labelled, each label will be discussed on the strengths and weaknesses and the type of playing style and transitional play that these teams from each cluster are likely to play in.

1.3. Use Cases

As stated previously, the highest level of clubs is starting to adopt data science practices into their operations. This is not to say that this is a new approach for these clubs as the use of data has always been integrated into the research behind many ongoing within a club's grounds. However, what is new, is the angle they take with their data science approach and furthermore, the effectiveness and impact that data science has on their bigger operations.

As one example, the two most decorated clubs in England are Liverpool and Manchester United. Since 2021, Liverpool have taken on a client by the name of Neuro-11. Neuro-11 is a neuro-transmitter agency set up by Dr. Niklas Hausler and his business partner Patrick Hantschke [4]. Liverpool have utilised the tools from Neuro-11 by wearing brain sensor kits whilst they train in certain situations, such as, taking free kicks, penalties, essentially any form of set-pieces. Whilst these actions take place, the brain waves from the players are observed and the data is passed into a system where analysis can be performed to understand more of a player's mental state in these situations.

This type of analysis hits data science in football at a new angle to gain a competitive edge over their opponents. As stated by Dr. Hausler himself: "The physical side in football has almost been maxed out. These guys are super fit. The next step really is to directly train the brain. That's where we come in with something that's scientifically validated."

In the case of Manchester United, in October of 2021, they hired a head Data Scientist by the name of Dominic Jordan. This was the first step in approaching data science with a much heavier priority. Additionally, in March of 2023, they have appointed a Lead Data Analyst by the name of Alex Kleyn who was formally employed at Southampton. In the wake of these employment, they are starting to build out their data science unit by interviewing for a host of roles in data science, machine learning, and data engineering.

Looking outside the Premier League, in the French Ligue 1, OGC Nice on the 21st of March of 2023 have announced a partnership with Statsbomb. Statsbomb hosts a massive database of data which FBref data is sourced from and the data in which is being used for this project. This type of partnership will allow Nice to have access to all the data provided by Statsbomb to utilise efficiently in a host of operations within the club.

Over the past 3 years, Porto from the Portuguese Primeira Liga, AS Roma from Serie A, Chicago Fire from the MLS, Frankfurt from the Bundesliga, and even international teams such as the Scottish FA, along with many more all have partnerships with Statsbomb, highlighting the growing use and necessity of data science in the world of football.

Considering a more personal approach. The likes of Manchester City's Kevin De Bruyne and Brighton's Kaoru Mitoma. They have both used data science on a more personal level to gain advantages and leverage in their careers.

In the case of Kevin De Bruyne, to discuss a new contract and a pay raise with Manchester City, he hired a data scientist to use statistics and analysis to quantify his impact and role in the team to justify a 30% pay raise¹. How this worked, was De Bruyne contacted the company Analytics FC to write up a statistical report on him. Analytics FC use a host of metrics such as that of Expected Goals (xG) and Expected Assists (xA) (these will be discussed in depth later in the report). An algorithm is built from these metrics to highlights a player's 'contribution value' to his/her team. When this value is calculated, it can then be benchmarked against players within the same age and position from across the topflight leagues. From the players De Bruyne's value matched up against, it was shown that these players were being paid significantly more than him. So, by using analytics in football, De Bruyne was able to justify a major pay raise for his contribution to the team.

For the case of Kaoru Mitoma, he started his professional footballing career in the Japanese J1 League with Kawasaki Frontale. Mitoma plays as a winger for Brighton. In this position, he would find himself out wide a lot relying on his dribbling ability to beat defenders to either cut inside into more central areas of the field or to go around the outside of the defenders to get a wide angle into the opposition's box. However, before getting into professional football at the age of 18. Mitoma decided to attend university instead at University of Tsukuba and enrolled into a Physical Education course. The most awe-inspiring part of his time at university was that he wrote his thesis on the art of dribbling. Mitoma would attach Go-Pro's to his and his friends' heads and in each dribbling situation, he would take note of where they were looking, their body language, how they shifted their opponents centre of gravity to be able to drive past them. All this research is subconscious data collection and machine learning of the mind and body.

As Brighton were far ahead of the data science game, through their analysis of players in the market, they quickly identified the attributes of lowly 23-year-old hailing from Japan and snatched him for only £3 million in 2021, which in today's market, is an absolute steal as after his first year of topflight football with Brighton, his value is likely to increase north of £50 million.

To pair with Mitoma and to show the quality of Brighton's data department was the signing of the then 16-year-old Evan Ferguson from Ireland's own Bohemians in 2021. Fast-forward to 2023 and Evan Ferguson at 18 is becoming a regular starter for Brighton. He is scoring regularly in the topflight and has gone on to make his senior debut for the Ireland national team. With his physical stature, technique, and composure on the ball at such a young age, he has all the tools to become one of the best strikers of his generation.

From all these cases of data techniques thriving in the world of football, it comes as no surprise that as the years go on, data is moving to the forefront of scouting, marketing, tactics, playing styles, and player roles. It is due to these many factors that it is felt that this project to be of great interest and importance to the footballing community.

¹ <https://trainingground.guru/articles/how-analytics-fc-helped-de-bruyne-negotiate-new-man-city-deal>

1.4. Technology

For this project, Python in Jupyter Notebook will be used. Jupyter Notebook is a server application where Python code can be written and edited and is enabled to run via the web browser environment GUI. The purpose for picking Jupyter Notebook to run is due to its many advantages over other Python environments. These advantages are the likes of interactive documentation where Jupyter Notebook allows for the mixing of code, text, and images in the same document, making it much easier to write interactive documentation and tutorials. In this, it makes it easier to understand each part of the code, its use, and the overall purpose of each block of code in reaching the end goal of the project.

Jupyter Notebook also supports easy visualization. Jupyter Notebook has built-in support for data visualization using libraries such as Matplotlib and Seaborn, which makes it easy to create charts and graphs directly in the notebook. These charts are easily accessible as each small block of code can be triggered to produce the charts and graphs.

One of the biggest advantages of Jupyter Notebook is its modularity. Jupyter Notebook allows for the breaking down of complex problems into smaller, more manageable pieces by creating separate cells for each step in the analysis. This for me, is arguably the biggest advantage in this project. As there are many steps to run from data collection, retrieval, cleaning, merging into DataFrames, appending columns, defining countless functions etc. all before any real analysis and models can be built and coded for the project. Breaking each step of code down into their own cells makes it much more manageable and can be comprehended much easier.

Through Jupyter Notebook, data will be scraped from FBref. FBref is a website that provides football statistics and data from major leagues and competitions around the world. The site is a subsidiary of Sports Reference LLC, and their advanced statistics is provided from Opta Analytics (formally partnered with Statsbomb), which is known for providing sports data to many of the top football teams in the world as stated previously.

FBref offers comprehensive data on football teams, with incredibly detailed data spreading back to 5 years ago. The site provides various statistics, such as over 250 team performance metrics, team rankings, and historical records of teams.

Python has become an integral skill for data scientists and analysts due to its flexibility, simplicity, and rich ecosystem of libraries and tools. Python's clean syntax and ease of use make it extremely accessible. Additionally, Python's extensive libraries, such as NumPy, Pandas, and Scikit-learn, provide powerful data manipulation, analysis, and machine learning capabilities, all integral for this following project. Python is quickly becoming the go-to language for data science and analysis, and it is increasingly used in industries such as finance, healthcare, and marketing. It is for these reasons why it is felt that the use of Python for this project is of utmost importance in showing the comprehension and handling on the coding language.

Additionally, Streamlit can be used to dashboard the end results. Streamlit is an open-source python library to easily create visualisations and these visualisations can be posted to a web app, with the backend running from a machine learning model. With Streamlit, the data

analysis carried out can be transformed along with machine learning models to create user-friendly dashboards. This is an extremely useful tool for the assignment at hand, as it allows for effective communications of findings and insights to users, such as managers or coaches.

Streamlit is very intuitive and allows for iterations on the visualizations created. It provides a wide range of pre-built components and widgets that allow for interactive elements such as sliders, drop-down menus, checkboxes, or buttons. These interactive elements can then be linked back to the machine learning model's parameters, allowing users to tweak inputs and observe the immediate impact on the outputs.

1.5. Structure

The introduction to this report contains all the information on the background of the report. The inspiration behind this project, the impact that data has had on sport and how this project ties into this ground-breaking industry in the sport. Why it was chosen to write up the report and the motivations behind this data project. The introduction details the aims for this project, in terms of the data source, the background behind the source, the data collection etc. This introduction also highlights the use of machine learning, the outline for how these models will be built and implemented on to the dataset, along with the desired outcomes.

The beginning of the report involves an executive summary stating the key points of the project, the models to be built for what purpose and how this model will be implemented. It also highlights the type of data to be used within the project and gives an overall sense for the feel of the entire project. It plays as an impactful summarisation of the project.

The next step highlights the introduction. The introduction itself is split up into different subsections, starting with the background. This background highlights the motivations and aspirations for the project. It details more of a personal touch to the project allowing for readers to gain more of a sense of who I am and why this project is one that is felt so passionate about as to have it as the final year thesis. It also gives a sense of the footballing structure and gives a slight insight into how analysis and statistics ties into football.

Following on from the background are the aims. These aims set out to highlight the aspirations for the project, the expected result, the models to be built and drives into detail on these models to ensure the reader gains a full understanding of the reasoning as to why these models were chosen for this project. Through these aims, the reader can gain full comprehension of the goals of this project and the models to be used to ensure these goals are met.

Additionally, in the introduction, a key area to be highlighted is that of the Use Cases. These Use Cases are imperative in highlighting the impact that data science and analysis has on the modern game of football, and how data impacts all key operations of a football club. Through these Use Cases, it highlights the need for such a model and how this project can be of increasing importance in the world of sport and to the teams looking to gain that competitive edge on their opponents.

Along with these Use Cases, the necessary tools and technologies needed are highlighted along with the advantages of using these technologies over their compatriots, and how they

will be used throughout the project. This section entails the importance of Python and the motivation behind its use, along with helping readers to understand why it is an integral skill for data scientists to have.

Essentially, the introduction summarises, from start to finish, the work done on this final year project and the motivation and ideas behind the work.

The State-of-the-Art section contains a host of literature reviews. These literature reviews are in depth evaluations of the relevance or significance of the work carried out and how this carries across to the work done within this project.

These literature reviews have been picked due to either: their relevance and motivation to an aspect of a piece of code being used, their relevance in terms of sports performance using data analysis, or in terms of building classification models for clustering and labelling data. The idea here is to take a plethora of literature reviews which hit various aspects of key points in this project to ensure each angle worked on can be backed up by a specific use case or past project.

This type of research will also allow for the making of informative decisions on how the work should be carried out, the type of approaches to take and the methods involved in building specific models. From this, it is believed that the work can be an accumulative summary of various projects and ideas from the authors of these papers, journals, articles etc.

This section also contains a piece on the project and how it differentiates to those that have been carried out prior to this report. It gives information on the different approach taken to create something new and innovating and how this can lead into a new aspect of data analysis not previously thought of in the footballing industry.

The Data section details the dataset that has been used in this report. It details how the data was gathered, where it was sourced, a small bit of background on the source, the permissions required to access the data and the steps in getting the data into a clean csv file available for use.

This section also details the tables collected and a host of metrics and their explanations from the tables, as some data metrics are calculated using data algorithms. It highlights the phases of football the data shows and why this is useful for the application to the model that is built.

In this section, aside from just data collection, it also explains why this data is of valuable use for the goals set and why this data is valid and should be held in high regards aside from the basis of data collection as there are multiple dimensions to consider about the data rather than just collecting the data into a tidy csv file. All these dimensions equate to a high-quality dataset ready for use.

The Methodology section is where the organisation of this report will be concatenated in. The Methodology contains the necessary steps following a Cross Industry Standard Process for Data Mining (CRISP-DM) method. This section of the report will allow for the steps taken to understand the decision-making process behind the project, how this model can be applied to real-world programs and processes.

The methodology section of the report is crucial for understanding what data was attained and for what purposes, the steps taken to configure this data correctly and to outline such biases or weaknesses within the dataset that will need to be addressed appropriately.

This section describes the steps of the project from start to finish. As it is the methodology, these steps will be adhered to keep organised throughout the project work. This involves the understanding of the project from a real-world context, understanding the data used and how this data is organised and sourced.

Following on from this point, this section details how this data is modelled and how machine learning practises are applied to the data to understand to come to a concrete solution on what this data can do. It is important, within this section, to evaluate the results correctly, to understand and reflect what can be done better and how this model can be used within the spaces of sports teams to better these teams.

The following steps for data preparation are cleaning, transforming, standardising, analysing, and visualising. Once the data is in cleaned and cleared of anomalies, duplicates, nulls etc. it can then be transformed into a desired format. This transformation involves the dropping of meaningless columns and then combining the DataFrames from different leagues into a singular dataset.

Arguably one of the most important steps for this project in the data preparation stage is the standardising of the data. As the German Bundesliga only play 34 matches each compared to the other leagues who all play 38, it is extremely important to 'per 90' the data. This means to divide each metric by the number of games they play a season to get an average per match. In doing this, it removes any bias against German teams within the data and ensures the data is appropriate for modelling.

The next step is to build the model on the data. This model will be a GAN model, which is a type of neural network. This type of network can be challenging to train properly as two models work against each other. So as accuracy and training improve on one model, it weakens on the other. The best course of action here is to find the equilibrium between the two models.

Once the model is trained appropriately, then the model can be evaluated on testing data. This testing data will be on 'synthetic data' which will be explained later in the report. Once the model is working accurately, it is ready to be deployed.

The Analysis section is where the core of the coding work lies. After the data is pre-processed and cleaned, the analysis picks apart the data to uncover trends, and promising insights. Following on from this, the analysis allows for work to be done to identify teams that are statistically similar, highlighting the variance of the distribution of the dataset, and showcases the variables which show the most variance and how these variables impact on the clustering models built.

Whilst the methodology section outlines the necessary steps, and the work to be done in each step to reach the end goal and to keep the project organised and within scope, the analysis section details more on the actual work done on the data before the final modelling is complete. The analysis entails use of multiple techniques and unsupervised machine learning

models to organise the data. This organisation of the data is crucial to understand where the data points for each team lies on an X-Y axis, so a deep neural networks model can be deployed to define how teams can reach a new point on this X-Y axis, closer to their desired location.

It is important to note, at this stage, the reasoning behind the choice for these approaches. Outlined in the State-of-the-Art section of this report, the empirical evidence and the research reviewed backs the use of the models and analysis techniques used in this report.

Whilst picking the analysis techniques based on study is important, it is also important to back reasoning behind the parameters chosen for these analysis techniques. As stated before, in the State-of-the-Art section, it details from research, the motivation behind the parameters set and how these parameters can be adhered to, to create a comprehensive, powerful analysis.

The Results section is important to show, in organised fashion, the output of the model, along with correct interpretations of the model's output. With added validation data to be run through the saved model, to highlight how the model would work when data is passed through it. On top of this, added justification of the model is important by detailing how a team currently plays and why certain metric improvements would bring them into a new cluster.

This section details the motivation behind the project, the analysis techniques used, even though this is reiteration, it is important to detail alone in the results section to ensure the results is fully understood with all necessary context behind it.

The uses of graphs/charts are imperative for any concise and clean results presentation. These will be put into effect to highlight metrics of the improved data in comparison to the team's data as it is currently. It is important to present the data in a visually appealing and easily interpretable manner.

Overall, detailing the reports section is crucial for this assignment. It is important for understanding the final model deployed, the results from the model, how to interpret these results and to produce visualisations so readers can easily comprehend the results in a compelling manner.

The Conclusion section details the advantages and disadvantages faced throughout the project. It details the limitations to the project and how these limitations can be minimised in future work.

Once the results have been presented, it is important at this stage to draw meaningful conclusions. As it is important to understand the results, it is extremely important to understand the limitations set and how these limitations can be improved upon in future work. The conclusions drawn are supported by the results presented.

It is important from the conclusive work to detail not only the limitations of the projects, but the strengths to the project. In the conclusion section, compiled interpretation of the results is summarised, and what worked best for the project is detailed. Certain analysis techniques

were very well received from the project whilst other work on modelling and the data itself, has some drawbacks.

It is important to detail, in a concise manner, all these drawbacks and strengths, along with a compilation of the results, for a comprehensive conclusion that readers can extract powerful summaries and information from.

The Further Research section details what could be done given more time, how the project could be expanded upon and the possibilities for delving into deeper analysis given the time and resources to do so.

It is important to reiterate the limitations of the project, and within further research, how to minimise the effects of these limitations. Detailed in this section, it is imperative to discuss how can work progress even further with the baseline set from this project, how can this work effect other industries and especially how can this work impact the industry the original data is modelled for.

Given more time, it is important to consider what additional analysis could be carried out, or what additional variables could be considered to create a more accurate reflection of real-world data as stated before, data is not football and football is not just data. There is a real element of the eye-test which readers need to understand. However, this does not mean that data is useless in football. On the contrary, it is one of the most powerful tools in understanding a player's impact on a game, the control of the game, the impact team's have on a pitch. Data in football removes all bias from the sport, which only adds to the richness of the context of football.

Finally, a short summary is presented of future work to be done, avenues this assignment could explore in future, and what could be improved upon on this assignment with the knowledge gained from working on it.

The References section is an important detail of the report. The references draw back to the literature reviews which host a plethora of motivations and reasoning behind the models built and the parameters inserted into these models, with justifications attached. This section ensures readers are aware of the prior research done into the work done in this assignment.

Having references allows for added depth and context into theories behind code along with certain ideas to be explored and justified. The literature reviews referenced highlight the motivation behind the project and allow for insight into the use of this model in real-world applications to change footballing business and cultures within their philosophies and recruitment. It is crucial for this section to be included to back up the work carried out with accurate, reliable field studies done within the space of this assignment.

The Appendices section provides additional information and context that goes beyond that of the main report. In this section, various materials that support and supplement the project report are included to give deeper insight into the process behind the work done, such as the proposal, ethics forms, and reflective journals. The Appendices section is especially useful for readers who want to gain a deeper understanding of the research process and the numerous steps involved in carrying out the project.

The proposal, ethics forms, and reflective journals are particularly important additional pieces of documentation as they provide insight into the research process and certain reflections and thought processes throughout the project. The proposal provides a high-level overview of the project and its objectives, outlining the work to be done, the methodology, and part of the modelling process. The ethics forms showcase the ethical approval received for the data, ensuring that the data has been collected in an ethical and responsible manner. The reflective journals offer a unique perspective on the research process, documenting the thoughts, challenges, and achievements over the course of the project.

2.0 State of the Art

2.1. Performance Disparities in Football

Throughout this assignment, the modelling and clustering will make use of distribution of in-game event metrics to determine clusters to be used for the GAN. To give comprehensive background and reasoning for this use case, it is very important to understand the disparity of performances from the top teams to the smaller teams and how these disparity's come about, and how these disparities become dissolved within the world of football.

In today's climate and sporting industry, the financial power of football clubs around the world can lead to having a direct impact on the success that these clubs face. A study conducted between the 2009/10 to 2013/14 footballing seasons highlights the methods and factors behind the financial growth of football clubs. These types of studies have grown increasingly popular over the past two decades, due to the commercialization of the sport. It is well known the importance and impact that winning football matches have from a league and trophy perspective, but winning football matches determines the financial safety of clubs and how clubs can evolve and grow.

Several studies have investigated the financial performance of football clubs using methods, such as financial ratio analysis, statistical analysis, and econometric modelling. For example, Nünnerich-Asmus and Bühn analysed the financial performance of German football clubs using financial ratio analysis, while Doherty and Richards used statistical analysis to examine the factors affecting the financial performance of English football clubs.

There is an element of diversity surrounding the attributes which affect financial stability of clubs. The major revenue streams for clubs include the likes of transfer fees for players, stadium capacity which allows for more matchday revenue and ticket sales, league competitiveness which provides prize money to the clubs competing within the league, broadcasting rights which clubs get a share of due to providing consumers on television, and sponsorship deals, which clubs wear on their jerseys or training kits. Revenue streams are a critical factor in the financial situations which clubs find themselves [5].

Within the top 5 leagues, teams with more success from before commercialization, along with teams with extremely wealthy owners, will attract more fans, allowing for higher revenue streaming across matchday tickets and broadcasting rights. With higher revenue, these teams can afford more premium players along with premium managers which helps teams to play their best football and propels them into Europe's elite clubs.

It is important to consider the on-field aspects that differentiate the best teams from the rest of the competition. Regarding this, a study was done on the 2016/17 English Premier League season to highlight the key performance indicators (KPIs) that differentiate the best teams from those lesser. From this, the key metrics to success could be defined.

It is becoming increasingly important in sport to perform performance analysis, to identify the value of data-driven decision-making and to understand the use of KPIs to measure and evaluate team performance. The investigation discussed uses a range of statistical methods

to identify the most important KPIs to achieve success in the Premier League. From the results received, possession, passes completed, shots on target, and tackles are the metrics that correlate most to the success of a football team.

Data-driven decisions can help guide team strategy. By highlighting the increasing importance of developing a data-driven approach into football management, it backs the motivation behind this assignment and how this investigation will be carried onwards[6].

It is important from a data point of view, and from a football point of view to understand how to evaluate different team's value and playing style, and from this, the disparities begin to show clearer. A study was conducted on player performance data from the Premier League. This study provides insights into the factors which drive player valuation in the transfer market. From understanding how performance metrics value a player worth, this can then relate into team worth, and more importantly, understanding how teams size up in comparison to each other.

There is no doubt that over the past decade, the transfer market has exploded in terms of valuations and inflation. The results highlighted from the research show that the number of goals scored, assists made, and successful passes completed are the metrics that correlate most with player valuation.

These metrics can be related into team success and to identify the differences between teams and sort these teams into clusters [7].

However, it is important to note that for the work done in this assignment, it is not how successful the team is at executing their methods, but more so what their methods originally are. Based on the results from the paper above, taking the number of shots taken, the xA, and the number of passes attempted would be an ideal place to start to understand where the teams belong within the X and Y axis.

It is an important note in the case of identifying disparities between clubs to use correct and proactive methods to achieve accurate results. A study conducted before commercialization became a big factor in disparity aims to identify the most efficient Premier League clubs and highlight areas for improvement for less efficient clubs. This is in similar fashion to the work being carried out in this assignment. This investigation was carried out on the seasons between 1998/99 and 2002/03 using Data Envelopment Analysis (DEA).

It is important to note that DEA is a very popular method for evaluating team performance in sports. It is a non-parametric technique used to evaluate the relative efficiency of decision-making units (DMUs) that convert multiple inputs into multiple outputs. DMUs in this research are sports teams. DEA evaluates the efficiency of DMUs by comparing their input-output relationships to those of the most efficient DMUs. It calculates the efficiency score of a DMU as the ratio of its weighted output to its weighted input, where the weights are determined by the most efficient DMUs. The efficiency score ranges from 0 to 1, where a score of 1 indicates that the DMU is fully efficient, while a score of less than 1 indicates less efficiency.

The input and output factors taken into consideration were those such as financial resources, squad size and quality, and performance statistics. The study provides a valuable contribution

to the literature on sports performance evaluation. It makes a good example of highlighting the value of DEA as a tool for evaluating team performance in football.

For this assignment, DEA was originally thought upon but was very quickly disparaged as there are much up-to-date and industry-based models that could be and are used. Regardless of this, the paper helped to understand the efficiency behind clubs and how they play their football and gave interesting insight into the disparity between clubs and their competitors [8].

As previously stated, the use of xG has protruded to the forefront of data analysis in football with many top teams using the method behind xG to scout potential players to bring to their teams. To give context as to why this is the case and to understand why xG has been included in the dataset used for the model built for this assignment is important to comprehend the full potential of an xG model.

Studies have been conducted based on large shot/goal datasets from the English Premier League and the German Bundesliga. This dataset provides insights into what factors drive goal scoring and shot efficiency.

It is necessary to understand the importance of goal scoring and shot efficiency in soccer. By highlighting the value of this data and the insights to be gained from it, teams and personnel can begin to extract the knowledge from the model and can make data-driven decisions based off the model. The use of performance indicators to measure and evaluate team performance has become extremely prevalent over the past 3-4 years and xG has become the forefront spokesman of data analysis in football.

xG models are designed as a metric to estimate the probability of a shot resulting in a goal based on a multitude of factors, such as shot distance, shot angle, and shot type. Shot efficiency metrics measure the ratio of goals scored to shots taken. A logistic regression model was used which gives each shot a probability (a number falling between 0 and 1) of being a goal.

The study conducted uses xG to evaluate the performance of Premier League and Bundesliga teams and players across the 2012-13 season. The results indicate that the traditional 'Big 6' i.e., Manchester United, Manchester City, Arsenal, Tottenham, Chelsea, and Liverpool all exceeded their expected goals. Manchester United exceeded their xG the most by scoring 86 goals from an estimated xG value of 60. This means that Manchester United exceeded their xG by 26, or they scored 26 more goals than expected based on the type of shots they took. It is important to remember xG is based on the average shot resulting in a goal, so exceeding their total goal count by 26 would be expected from Manchester United with above average shot takers in their team. United's main goal scorer in 2012-13 was Robin Van Persie, who scored 25 goals from 16.1 xG.

The paper states that the results can enable team strategy and decision-making, whilst also highlighting the importance of teams to develop a data-driven approach to decision-making within a footballing regime [9].

2.2. Clustering Ideology

Clustering is an important part of the process of this assignment. Clustering was used to cluster teams with similar data points together to highlight the disparities of the teams. The use of an unsupervised machine learning method in K-Means Clustering makes sense as there is no target variable required to run the model. The model builds centroids based on the data fed into the model which is plotted on to an X-Y axis and based on where teams fall on that X-Y axis, the Euclidean distance is calculated from each team's data point and whatever centroid they are closest to, they are grouped into that label. However, as found from research, there are certain limitations which must be worked around and improved when using the likes of K-Means Clustering by itself.

The research put forward shows the importance of clustering in machine learning and data mining models. The importance of these models is shown by highlighting the value of unsupervised learning techniques for identifying patterns in complex datasets with many variables to consider. For context, within the dataset used in this assignment, there are 77 variables to be considered.

The limitations noted from the research indicates that whilst the K-Means algorithm is a simple and efficient technique for clustering large datasets, it is also quite sensitive to the initial selection of centroids and that this model may converge to sub-optimal solutions to fit the number of centroids selected originally.

From the research provided on the K-Means algorithm, there are discussions of a new method for selecting the initial centroids, based on the concept of distance distribution. The research shows a significant improvement on the K-Means algorithm on complex datasets when this new method is put into place. This method is a U K-Means algorithm. This method works by selecting a random data point from the dataset and then selecting the next centroid based on the distance distribution of the centroids selected previously. The U K-Means algorithm ensures that the initial centroids are well-spaced and are representative of the dataset, which, in turn, improves the accuracy and efficiency of the clustering process [2].

Whilst this method was originally considered for use in this assignment, from further research and reading on alternative methods, it is decided to take a different route to nullify the limitations of a K-Means algorithm.

One method that was considered and implemented on to this assignment was the use of the elbow method in synchronisation with K-Means Clustering. As discussed previously, one of the main challenges when it comes to K-Means Clustering is determining the optimal number of clusters. This is where the elbow method takes shape. This method works by plotting the sum of squared distances between data points and their assigned centroids against the number of clusters represented as K. Within the plot provided, there is an "elbow point", which is a crux in the line chart. This crux represents the point at which the addition of another cluster would not significantly decrease the sum of squared distances, and therefore additional clusters would be rendered unnecessary for the data represented.

Based on research done, other heuristic models were considered. However, research which was provided shows a comparison of the elbow method with other methods such as

silhouette analysis, gap statistic, and average silhouette width. Based on research across these methods, it has been taken on a consensus that the elbow method is an effective approach for determining the optimal number of clusters in K-Means [10].

Due to this research done, this assignment makes use of the elbow method to determine the optimal number of clusters for the K-Means model. The analysis on this method is discussed in detail within the analysis section of this report.

Another heuristic method used within this assignment was that of hierarchical clustering which is a tangent based off K-Means Clustering. Studies done on this topic have taken data from Opta – a football database where FBref pull their data from – for the Turkish Super League, English Premier League, and German Bundesliga in three seasons between 2015 and 2018. With this data, the goal was to profile players in a similar fashion as this assignment focuses on profiling teams.

It is important to note that traditional player evaluation methods, for example, scouting and eye-based assessments, are subjective and blind bias could affect the accuracy of a report and of a player's/team's true abilities. To combat this, research done proposes the use of data-driven methods, such as clustering, to accurately classify players based on their performance.

The studies make use of the hierarchical clustering for each position to discover the ideal number of clusters for each position. The model found that 3 clusters, 4 clusters, and 5 clusters, for defensive, midfield, and offensive players respectively was most accurate to profile players accurately. These clusters were found by using a dendrogram which acts as a bottom-up approach of sorts [11]. For this hierarchical clustering, the Ward method was used.

Instead of measuring the distance directly, Ward method analyses the variance of clusters. Ward's is said to be the most suitable method for quantitative variables. Ward's method says that the distance between two clusters, A and B, is how much the sum of squares will increase when we merge them:

$$\Delta(A, B) = \sum_{i \in A \cup B} \|\vec{x}_i - \vec{m}_{A \cup B}\|^2 - \sum_{i \in A} \|\vec{x}_i - \vec{m}_A\|^2 - \sum_{i \in B} \|\vec{x}_i - \vec{m}_B\|^2 = \frac{n_A n_B}{n_A + n_B} \|\vec{m}_A - \vec{m}_B\|^2$$

where m_j is the centre of cluster j , and n_j is the number of points in it. Δ is called the merging cost of combining the clusters A and B. With hierarchical clustering, the sum of squares starts out at zero (because every point is in its own cluster) and then grows as we merge clusters. Ward's method keeps this growth as small as possible.

The Euclidean Distance is the "ordinary" straight-line distance between two points in Euclidean space¹.

$$d(p, q) = d(q, p) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \dots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^n (q_i - p_i)^2}$$

In other studies, on agglomerative hierarchical clustering within football, researchers use this approach to group similar sequences of team interactions together. These sequences are the

¹ <https://jbhender.github.io/Stats506/F18/GP/Group10.html>

likes of passing sequences or player movements. The clusters created were then analysed to identify common patterns of interaction between teams and to research and observe the effects of different strategic decisions, including changes in formation and player positioning.

One of the advantages outlined within the papers for agglomerative hierarchical clustering is that it allows for the identification of clusters at many different levels of granularity. This can be useful to analyse complex datasets such as the kind used in this assignment with over 70 variables. It is useful for complex datasets as the interactions between teams during a football match can be observed on many granular levels. By analysing the clusters at different levels of granularity, the researchers could identify general patterns of interaction and patterns on strategies or formations [12].

The use of hierarchical agglomerative clustering for this assignment could provide great insight into how team's play and which team's play in similar fashion depending on the transitions on the pitch and what attributes the teams have great strength in and which attributes they will greatly need to improve upon.

2.3. Generative Adversarial Networks Ideology

The practical use of GAN's was very intriguing when put forward to use on this dataset. With limited data rows as there are only 490 teams. 20 teams from the Premier League, Ligue 1, La Liga, and Serie A, and 18 teams from Bundesliga, for each season. This idea of using a GAN was very interesting but initially there were doubts surrounding the compactness and completeness of such a model on limited data, along with limitations surrounding the use of the model.

GAN's are a type of Neural Network deep learning model that consists of two networks. One model being a generator, and the other being a discriminator. The generator ideally takes random noise as input and generates synthetic data that is like the training data, so that this random noise passes the discriminator. The discriminator is trained to distinguish between this real and fake data. The ideal outcome is that the discriminator cannot tell if the data came from the generated data or from the actual dataset and an equilibrium of 0.5 is reached.

As the two networks compete, the generator learns to generate increasingly realistic data, while the discriminator learns to become more accurate at distinguishing between real and fake data.

The way to use the model that was put forward in this assignment is to treat teams not in the optimal cluster to act as fake data, whilst the teams within the optimal cluster would be the real data. In this way, the generator would generate new data based on this "fake data" for this data to look like teams' data in the optimal cluster and pass the discriminator. This generation technique is known as data synthesis.

Data synthesis is an extremely important task, especially in cases where real-world data is limited or difficult to obtain, such as that of the dataset used in this assignment. From research done, an approach has been put forward for generating synthetic tabular data using GANs.

Regarding rule-based approaches and statistical approaches, research done showed that GANs provide a more flexible and adaptable approach to data synthesis, which allows for the creation of complex and realistic data distributions which closely match the characteristics of real-world data. This is especially useful for the idea put forward for this project. As it is important to keep the integrity and relationships that variables have on each other.

It has been proposed that GANs for tabular data synthesis is the best option put forward. The generator is trained in conjunction with a discriminator. The two networks are trained iteratively, defined by epochs, with the generator continuously adjusting its parameters in response to how well the discriminator identifies the synthetic data, until the synthetic data produced by the generator is indistinguishable from the real data.

The researchers test their approach on several data sets, including a credit card dataset and a medical dataset. The findings show that the method was able to generate synthetic data that closely matches the makeup of the real data, while also ensuring the structure and relationships between variables were secure [13].

The use of GANs for data augmentation has grown incredibly over the past couple of years. Augmenting data using GANs helps to address the problem of limited training data, which is a prominent issue within the dataset used for this assignment.

The idea behind GAN-based data augmentation involves the use of the generator to produce synthetic data which can then be added to the original training set, increasing the amount of data available for training. This approach was very interesting and one that was considered heavily in this assignment.

One of the advantages of GAN-based data augmentation is that it can produce data that is representative of the underlying distribution of the training data. This can be very useful in certain situations where the training data is limited due to the type of data captured or due to lack of variability within the data.

Based on past research conducted, a technique for GAN-based data augmentation involves training a GAN on the original training set and using the generator to produce synthetic data that is then added to the training set. The discriminator is then retrained on the augmented training dataset to improve the number of rows trained on the data which in turn, helps to train the discriminator to establish differences between real and synthetic data.

The research is conducted on a host of image classifications, demonstrating that GAN augmentation can improve the performance of deep neural networks when the augmented data is added back into the training set and the model is retrained. The results show that GAN augmentation proves to be effective when the number of rows is limited, and that this technique can outperform other data augmentation techniques [14].

However, taking this data augmentation into account and doing further research on this technique, some drawbacks were found. The effectiveness of GAN-based data augmentation depends on the strategy for sampling used to select the synthetic data points. Researchers have offered alternative sampling strategies for selecting synthetic data points generated by GANs.

It is important to note that not all synthetic data points are very useful for augmenting a training dataset. If synthetic data points are selected without prior thought, bias or unwanted noise may creep into the training dataset, which could lead to poor performance of the trained GAN model. The research partaken aims to select sampling strategies that will improve the quality and the representative distribution of the synthetic data, while simultaneously attempting to minimize the risk of bias or noise creep. These sampling strategies include the likes of uniform random sampling, K-Means clustering, and Principal Component Analysis (PCA) based sampling.

Uniform random sampling is a sampling technique that selects synthetic data points randomly from the GAN-generated dataset. K-Means clustering partitions the GAN-generated dataset into clusters and selects a representative point from each cluster, whilst PCA-based sampling selects synthetic data points based on their distance to the principal components of the GAN-generated dataset. For the work put out for this assignment, along with past research on the advantages of K-Means Clustering, this seemed to be the most positive direction to take.

This research was evaluated on image classifications and these techniques were then compared to the likes of random sampling and stratified sampling. The experiments show that the improved sampling strategies, such as K-Means and PCA can improve the performance of the trained models, especially when the amount of GAN-generated data is limited [15].

Overall, the work carried out in this assignment differs from any work researched. From extensive searching, there does not seem to be a paper that has made use of GANs with footballing data to push teams towards certain clusters for them to play better and to improve on their overall game. Moreso, this assignment has not seen any use cases of processing tabular data from the original dataset and passing this data as generated “fake” data. For these reasons, it is believed that this idea is to be original, and certainly one to be investigated further for how this can be applied to many other industries outside of football. This assignment has the potential to be of great interest to readers who partake a liking for the sport.

3.0 Data

An email was sent to FBref to ask for their permission to scrape data from their website. Once the permission was granted, FBref was scraped for the necessary data. Even when data collection is complete, it is of the utmost importance to ensure data quality is assured, to ensure the data itself is sustainable and understood for the analysis that is being aspired to be carried out.

Firstly, the data was scraped by gathering the 25 URLs needed. One for each league from each of the last five seasons.

The URLs are assigned to variables and a function is defined to locate and grab every table needed within each URL through pandas read html functionality. These tables are then converted to DataFrames and the function working on each URL is assigned as a new variable to be called with each table name.

From this, there are some messy overhead grouping names on columns, so a function was defined and called to get rid of these unnecessary headers. When this occurs, some data columns that were grouped under these headings become jumbled and out of order, therefore for renaming columns to make it clearer what each column stands for, these columns need to be reordered and renamed. A function is defined using if conditional loop depending on which years have data jumbled up. The data is then untangled, organised accordingly, and renamed before concatenation.

When all the DataFrames have clean column headers, they are concatenated into one large DataFrame. Duplicates are then dropped, and columns are re-organised into a more desirable order for better data understanding. The DataFrame has been converted to a csv file, so the data is saved. This is now the base dataset. The analysis is started on the DataFrame, including comparing, visualising, discuss the findings, what these findings mean and how they visualise how teams play football.

For example, for the URL for the English Premier League, it provides 12 tables. Each URL provides the same 12 tables for each respective league. Each table represents different attributes, strengths, and weaknesses of each team in each league. For clarity, the titles of these 12 tables are: 'Regular Season', 'Squad Standard Stats', 'Squad Goalkeeping', 'Squad Advanced Goalkeeping', 'Squad Shooting', 'Squad Passing', 'Squad Pass Types', 'Squad Goal & Shot Creation', 'Squad Defensive Actions', 'Squad Possession', 'Squad Playing Time', 'Squad Miscellaneous Stats'.

These tables are for the teams, for example, let's take Arsenal Football Club from the English Premier League. These are stats to show Arsenal's pass types, their shooting patterns, their goalkeeping saves, their defensive work etc. This will be discussed in more detail below.

From this base dataset, irrelevant columns were dropped, and the data was standardised through a function, and this new data was saved to a new csv file. It is with this standardised data that the analysis will be performed on, and this data will be used for modelling.

Even after the dataset has been assembled and ready for analysis. There are many checks to undergo to ensure the data is of the best quality for the work to be carried out. This is because

having values and variables does not mean that the analysis is ready to begin. Data quality must be addressed first to ensure the dataset itself once modelled and analysed produces the best and most accurate results.

Previous season's data is used as the data for these entire seasons are available compared to this current season, which is yet to be concluded and where metrics can change as the year progresses which could potentially alter some forms of the findings and would need continuous backwards propagation of the analysis. This ensures completeness of the data and data consistency. This data will always be available as it is saved to a csv file and as more seasons be completed, the size of the dataset will only grow as the game of football evolves.

As for the currency of the data, the data is taken from the previous 5 years of football, modern football. Data currency refers to the degree to which the data is current and relevant in the world it models, and this data is taken from the most recent seasons of football at the level the football is currently sitting at. This data also comes from a very reliable source in Fbref as the data they use, is data licensed by affiliated data analytics companies who host partnerships with a plethora of leading topflight teams in football.

This data also boasts high validity. Data validity refers to the extent in which data measures up to its intent of its purpose. In this way, as the full dataset and metrics are explained below, there is a plethora of data for each phase of footballing play. It is crucial to ensure that data this is processed consistently. Any of the changes made will be properly explained. As stated previously, the source of the data is very reliable and reputable.

The final check to ensure the data is of the highest quality is data consistency. This refers to the data and the data sources. This dataset is free of errors, anomalies, contradictions, incomplete data, and biases. Consistency is imperative for ensuring that the data is accurate and that it can be relied upon for decision-making purposes.

The major bias that this dataset hosted was the fact that teams from the Bundesliga play 34 games whereas teams from the other leagues play 38. Therefore, as teams outside of Germany play more games, such metrics like 'Total X Completed' where X = Passes, Shots, Tackles, Defensive Actions, Saves etc. would be skewed in favour of those teams due to playing more matches. To negate this bias from the dataset, all these metrics were divided by the number of matches each team plays to create an equal 'per match' statistic for each metric. This ensures that the data is not skewed against the Bundesliga and all teams are compared against each other on an equal playing field.

The variables chosen for this dataset are based on their relevance to the goal being discussed and their ability to provide insights into the underlying relationships between the variables. Understanding the role and impact of these variables is essential in understanding the functionality of the project and data analysis. This data can help identify patterns in play, relationships between defence, midfield, and attacking transitions, and discover correlations between these attributes to inform decision-making processes.

Based on this, having the capability and the knowledge to fully understand and analyse this dataset is imperative to understanding this assignment and so, below is a brief explanation of each table and the key variables within these tables.

As previously stated, the dataset is made up of 12 tables, each highlighting key phases for each team. There are 12 tables for each of the five leagues and there are five seasons, so this dataset is compiled of data from 300 tables.

	Squad	League	Avg. Possession	Expected Goals	NP Expected Goals	Expected Assisted Goals	Shots on Target Against	Saves Made	Clean Sheets	GK Post-Shot Expected Goals	Miscontrol	Dispossessed	Passes Received
0	Arsenal 2021 - 2022	Premier League	52.8	1.771053	1.639474	1.221053	3.868421	2.631579	0.342105	1.252632	14.657895	8.842105	420.736842
1	Aston Villa 2021 - 2022	Premier League	46.5	1.260526	1.207895	0.997368	3.921053	2.526316	0.289474	1.223684	15.657895	9.052632	330.947368
2	Brentford 2021 - 2022	Premier League	44.8	1.344737	1.221053	0.963158	4.710526	3.289474	0.236842	1.300000	14.815789	7.657895	302.421053
3	Brighton 2021 - 2022	Premier League	54.4	1.394737	1.250000	0.981579	3.842105	2.684211	0.289474	1.118421	15.736842	9.184211	426.157895
4	Burnley 2021 - 2022	Premier League	40.2	1.152632	1.126316	0.897368	4.684211	3.342105	0.236842	1.250000	14.184211	7.657895	244.263158
...
485	RB Leipzig 2017 - 2018	Bundesliga	55.0	1.526471	1.411765	1.164706	4.441176	2.882353	0.176471	1.420588	14.441176	12.382353	415.764706
486	Schalke 04 2017 - 2018	Bundesliga	46.6	1.373529	1.144118	0.923529	3.794118	2.676471	0.382353	1.047059	11.588235	8.529412	351.823529
487	Stuttgart 2017 - 2018	Bundesliga	46.6	1.185294	1.117647	0.773529	4.558824	3.441176	0.352941	1.197059	9.264706	8.058824	352.617647
488	Werder Bremen 2017 - 2018	Bundesliga	47.4	1.144118	1.144118	0.929412	4.764706	3.617647	0.235294	1.235294	9.323529	10.029412	348.617647
489	Wolfsburg 2017 - 2018	Bundesliga	49.3	1.005882	0.873529	0.632353	4.500000	3.058824	0.235294	1.482353	10.764706	11.264706	373.558824

Figure 1: Brief example of finalised dataset. All metrics standardised per 90*

*Per 90 means on a per game basis to ensure teams are treated equally regardless of the number of games played.

The 'Regular Season' table shows the league table as it stands for each season. This table is not too important to the dataset. This is because this table more so highlights the number of wins, draws, losses, number of points a team accumulated, the attendance etc., which does not reveal much about how a team plays their football. It details how successful a team is for each season but not too much on the actual football the team plays. However, this table does have some key metrics, such as the number of goals scored and conceded, along with the expected goals (xG). This assignment will go into more detail on this metric below.

The 'Squad Standard Statistics' table hosts a plethora of data on the playing time and the performance of teams. Like before, a lot of this data is quite redundant or is low quality for what this assignment sets out to achieve. However, with that said, the key metrics to pull from this table is the amount of possession of the ball each team has, the number of progressive carries and progressive passes i.e., progressive carries are players moving the ball towards the opponent's goal line at least 10 yards from its furthest point in the last six passes, or any carry

into the penalty area whilst progressive passes uses the same metric but for a pass instead of a carry. This metric is incredibly useful for understanding how often a team looks to break through opposition lines and is this due to a pattern of passes or individuals carrying the ball forwards.

The 'Squad Goalkeeping' and the 'Advanced Squad Goalkeeping' table shows all the goalkeeping statistics. There are incredibly useful metrics in these tables to highlight the goalkeeper is used in each squad and allows some insight into the defence of the team. Key metrics from these tables would be the likes of number of shots faced, percentage of shots saved, the type of goals they concede i.e., corner kick goals, penalty goals, or open-play goals. Another very useful statistic here is Post-Shot Expected Goals Against minus Goals Allowed +/- (PSxG+/-). This statistic calculates the number of goals a keeper is expected to concede based on the xG model, and the number of goals they conceded.

For example, Wolverhampton Wanderer's goalkeeper Jose Sa conceded 43 goals in the 2021/2022 season. Based on all the shots taken on the goalkeeper and using the xG model to predict how likely he is to save each shot; it was expected that he would concede 49.3 goals in the 2021/2022 season. This means that Jose Sa overperformed his expected Goals Against (xGA) by +7.3 (own goals are not considered). Essentially, this model shows how well a goalkeeper has performed at shot stopping.

These tables also highlight not just goalkeeping abilities but how a goalkeeper contributes to their team on the ball by highlighting the number of launches (passes longer than 40 yards) attempted and completed by a goalkeeper along with the number of passes, throws, average length of goal kicks and passes. All these metrics show the goalkeepers role in their team and what the goalkeeper's role is in each style of play.

The 'Squad Shooting' table highlights the attacking play of teams. Key metrics from this table would be the number of shots a team takes, the xG of course, the average distance of shots, like the goalkeeping statistic, the under/over performance of goals scored minus the expected goals scored.

The 'Squad Passing' and 'Squad Pass Types' tables are arguably two of the most important tables for understanding how teams play their football. There are a multitude of key metrics highlighted in these tables. It is important to note however, that for this dataset, this assignment is not looking for how many passes of each type were completed but rather how many are attempted. Passes being attempted show the attempt to play a certain style of football rather than executing that style of football well or not.

In these tables, there are a number of key metrics, such as the progressive passing distance total, the number of passes attempted (this should be taken into account with the amount of possession each team has), how many passes attempted short, medium, and long, the number of key passes i.e., passes that directly lead to a shot, the expected Assist (xA) number i.e., the likelihood each completed pass becomes a goal assists given the pass type, phase of play,

location and distance. An assist is the pass made to the player who then scores. Passes into the opponent's final third of the pitch, number of live and dead passes made (live passes indicates open play passes), number of through balls and switches i.e., the number of passes made in behind an opponent's defence and the number of passes made that travel over 40 yards of the width of the pitch respectively. All these metrics give great insight into understanding team's passing patterns.

The 'Squad Goal & Shot Creating Actions' table highlights different metrics in the attacking phase. Key metrics include the likes of the type of passes leading to shots, take-ons, fouls, defensive actions that all lead to shots. Also, the number of shot and goal creating actions. These metrics shows how teams are creating their shots and is incredibly useful for identifying team's attacking style of play.

The 'Squad Defensive Actions' table shows the defensive phase of teams. Knowing how teams defend is crucial for understanding their type of football as defensive tactics is often an under-appreciated aspect to the game. Key metrics are tackles made in the attacking third, middle third, or defensive third of the pitch. Teams with very high attacking third tackles are more pro-active and like to take more risks in behind their defence to win the ball higher up the pitch with teams with a high defensive third tackle like to sit deeper and protect their last line of defence and let the play develop in front of them. Other metrics would be the number of challenges attempted, the number of blocks, and clearances.

The 'Squad Possession' table, like the passing table, is arguably one of the most important as this table details how teams hold possession of the ball and areas of the pitch, they hold the possession. Key metrics from this table would be the number of touches and where these touches are i.e., their own or the opposition's penalty box, the defensive, middle, or attacking third. The number of take-ons attempted, and the number of passes received. Again, these statistics need to be taken into consideration with the amount of possession teams have.

The 'Squad Playing Time' and the 'Miscellaneous Statistics' tables do not boast much quality data. However, there are some interesting metrics such as the number of substitutions used, highlighting the squad depth, how early these substitutes are made, the number of aerial duels (headers) won/lost, the number of interceptions, fouls, offsides, and crosses attempted. All these variables give some insight into parts of a team's style of play.

With all the tables explained and the key metrics used, one of the most important metrics is xG. xG is based on a statistical model used in football analytics, to predict the likelihood of a goal being scored based on various factors such as the shot distance, angle, and type. It assigns a probability score to each shot taken during a match, which is based on historical data and machine learning algorithms. xG is extensively used by coaches and analysts to evaluate team performances, and to inform decision-making around the tactical side of the game.

4.0 Methodology & Implementation

4.1. Methodology

CRISP-DM is a widely used methodology for data mining and data analysis projects. It stands for Cross-Industry Standard Process for Data Mining. The methodology consists of six major phases:

1. **Business Understanding:** This phase involves understanding the project objectives and requirements from a business perspective and identifying the data mining goals and suitable data sources.
2. **Data Understanding:** In this phase, the data is collected and explored to get a better understanding of its content, quality, and structure.
3. **Data Preparation:** In this phase, the data is cleaned and transformed to get it ready for modelling. This may involve handling missing values, outliers, and duplicates, as well as selecting and constructing appropriate variables.
4. **Modelling:** In this phase, various data mining techniques are applied and evaluated to identify the most appropriate model for the problem at hand.
5. **Evaluation:** In this phase, the selected model is evaluated to ensure it meets the business objectives and satisfies all necessary quality requirements.
6. **Deployment:** In this phase, the results of the data mining project are deployed in the organization and any necessary follow-up activities are carried out.

CRISP-DM is a flexible and adaptable methodology that can be tailored to fit the needs of a wide range of data mining and analysis projects.

4.2. Implementation

The steps surrounding CRISP-DM are as follows:

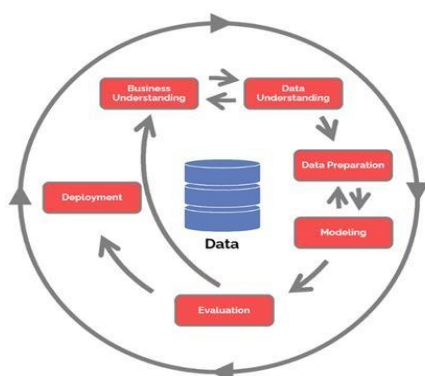


Figure 2: CRISP-DM Methodology Model¹

¹ <https://www.datascience-pm.com/crisp-dm-2/>

4.2.1. Business Understanding

This entails outlining what is to be achieved in the project. What is intended on being showcased and what direction this assignment's work will go in. Understanding the application of the project will have to the real world and to be able to apply the data and the findings into real world scenarios.

So, the goal is to analyse in-game data from Europe's top 5 football leagues over the past five seasons, and understand through analysis, where these teams lie in terms of their in-game quality statistically. From here, the end goal is to create a model to apply to real world situations so that a smaller, lesser team can learn what in-game qualities they need to improve upon to compete with the teams at the highest level.

Deeper analysis can be done after understanding where they need to improve, different radar plots will be created to highlight different parts of team's transitions, and areas of in-game qualities that are lacking. For example, if a team needs to improve on their NP expected goals i.e., creating better open play shooting opportunities more frequently. They can investigate their shot/goal creating actions data which is more in depth and understand exactly where they are struggling. From here, they can then investigate their passing data to understand where they need to improve their passing or carrying to create better opportunities. This can then be dealt with internally in training or the team could then begin to build a portfolio on the transfer market to sign players who have a positive impact on team passing and team threats.

4.2.2. Data Understanding

This involves gathering the dataset that is required to carry out this analysis and making sure the data being used is suitable for the analysis on which is being intended on carrying out. For the data selection, FBref data, a football statistics website, will be scrapped. On this website, data will be scrapped from 25 URL's. Five URLs for each league and within each URL are the same 12 tables which highlight different metrics about each team within that league.

Data understanding additionally involves exploratory analysis. Analysing the data, describing the data, and verify how clean the data is, and if there is need for change. This part of the process relies heavily on having full understanding, and description of the comprehensive dataset and all its' necessary metrics.

The first thing to be done in Jupyter Notebook is to gather all the URLs needed to scrape the data from. Variable names will be assigned to each of these URLs. A function is defined to call the URLs for scraping each specific table, so each table only needs to be defined once within the function. That function can then be called for each year of data needed.

Messy overhead columns are then cleaned and taken out of the DataFrames to ensure when a clean merge on the DataFrames. They are then merged to keep all data metrics together for each league.

Once the tables are merged by their respective years, a year column and a league column are appended to each DataFrame. Depending on how the data merges, it could come through clean, or a duplicate column can appear for 'Total Passes Completed'. If this column is duplicated it also messes up the order of the columns surrounding it. To ensure all the DataFrames are consistent for concatenating, a function is defined to change the column names in an order depending on how many columns are in each DataFrame. If the DataFrame contains an extra duplicate column, it pops one column into the other to remove all duplicates and fills in any null values. A new function is then defined to relocate all of the columns in all of the DataFrames, so they are all consistent when concatenating.

4.2.3. Data Preparation

Once the DataFrame is in the correct format, duplicate columns are then dropped. Essentially in a lot of the tables being pulled, some of the tables contained duplicate data for the purpose of each table. A line of code is run which drops any columns which have the exact same data in both. It also involves dealing with any outliers that could possibly skew the data in a positive or negative way which is not a true reflection on the dataset along with ensuring all the tables are in the correct format to be usable for data analysis. The data will be standardized before running any analysis. Renaming columns, dropping, and adding columns will all be a part of the cleaning and transformation process to ensure all the necessary data for visualising is available.

If any of the data is not in the correct format, it will be corrected and put into the correct format by passing it as string, float, numeric etc. so it can be integrated successfully with other instances into a clean and perfected dataset to produce the best accurate results so that the analysis can be as thorough and as accurate as possible. For ease of understanding which squad specifically is in the results, the Year column and the Squad column join, and the year column is dropped and then this combined DataFrame is exported to a csv file. For example, of how the new Squad column looks. For 'Paris Saint Germain', it is now 'Paris Saint Germain 2021-2022'

This saved csv file is opened, assigned to a new variable, columns are dropped which are of no use, the data is then standardised through a function which per 90's all the stats. Per 90 essentially means per game, as each game is 90 minutes long. This is necessary as in the German league, they only play 34 games compared to the other leagues which play 38. This new dataset is saved to a csv file.

A PCA is run on the data along with K-means clustering, using agglomerative clustering techniques. A heatmap of all the attributes that are most relevant to each cluster is shown and it is then assessed which teams belong to each cluster.

Based on the results, it can then be identified which teams are of possession-based playing styles for example. This will be done for all 5 leagues for all 5 years and essentially what the scatter plot will show is for each team, what is their style of play. From this, it can then be explained in footballing and real-world terms, why certain teams play a certain way and what this type of football entails. This will give context of teams' performances and their style of play.

4.2.4. Modelling

This section of the methodology involves selecting the modelling technique which is best to use the objectives set out. Generating a testing design and how is it best to split the data. It involves building the model so that it can be utilised in a project and can be used as a tool for future reference. Assessing the model is a key component to modelling data. The plan here is to build a neural network GAN model, fine tuning the model's hyperparameters, and testing the model. It will then be interpreted which model is best based on the success criteria i.e., running the generated data against the original data.

Prior research on the models used and the justification for certain approaches are outlined within the literature review, ensuring any code implemented is backed up by case studies and full comprehension of the hyper-parameters chosen is undertaken.

The test design of the GAN model will be written up, outlining the hyperparameters used, the justification and explanation of the way the model is constructed and how the model is tested. The success of the model relies on this comparison between the original and generated synthetic data and understanding where teams need to improve. This part of the modelling process will involve more explanation on the football and some added knowledge from the author to establish and explain why the results are as they are and in what context does this data highlight how a team may play.

4.2.5. Evaluation

This part of the methodology is about evaluating the results based off the model. To validate the model, additional data is called in from the Portuguese League which is currently the sixth highest league in Europe. Rows from this data will be passed into the model to generate synthetic data and these two rows of data will be compared to establish on the metrics provided, where specific teams need to improve to compete with the best teams in Europe.

Reviewing the work that has been done is extremely important, understanding what could be done better, how different issues could be tackled differently and more efficiently and highlighting drawbacks or limitations to the design or output. Acknowledging these weaknesses ensures the baseline model can be improved in future updates and helps the club to understand how trustworthy the model can be and in what areas the model can give them advantages.

In this phase, it is important to understand what was overlooked and what could possibly be improved. After this evaluation, it is then that it is decided if the model is ready for deployment. If the model has passed testing and validation, this ensures the model works correctly and that over/under fitting is not an issue. This model can then be deployed to an interactive dashboard, detailing where a team needs to improve. This dashboard will host radar plots highlighting different areas of a team's statistical in-game metrics.

4.2.6. Deployment

As stated, the model built will be deployed to a dashboard, where the outputs of the model will be compared to the team's current data. This dashboard can then be utilised by a team. From here, the management consortium can begin to understand where they are struggling through the data, understand what they need from their team to compete with the best teams in Europe and from this, they can investigate the qualities in their squad, or look externally into the transfer market to find a player that fits their profile on what they are currently lacking.

With the help of the data collected through the dashboard, the management team can investigate the qualities of their current squad. They can also look externally into the transfer market to find a player that fits their profile on what they are currently lacking. By understanding the areas where they are lacking, the team can then focus on improving their performance to achieve better results. Overall, the deployment of the model to the dashboard will provide valuable insights to the team and help them to achieve their goals.

5.0 Analysis

The analysis section of the project looks to use a plethora of data analysis techniques to mine the data and gain key insights for modelling the data. One of the main aims from this analysis is successfully clustering teams into their labels for full understanding of the modelling work to be done on the neural network. The teams in the best cluster will be used as the benchmark for the discriminator whilst the teams all the other clusters will be split for training and testing on the generator model to pass the discriminator benchmark. Through analysis, the aim is to generate new data for a team based on their current team's metrics to pass the discriminator.

During this process, many methods were considered on the clustering techniques to be implemented. The sorts of these methods considered were the use of K Nearest-Neighbour, Support Vector Machines (SVMs) and K-Means Clustering.

To establish what clustering algorithm would work best, it is important to understand the data that would be modelled on these machine learning algorithms. K Nearest-Neighbour is a non-parametric, supervised machine learning classification algorithm. For a data record t to be classified, its k nearest neighbours is retrieved, and this forms a neighbourhood of t . Majority voting among the data records in the neighbourhood is usually used to decide the classification for t with or without consideration of distance-based weighting. However, to apply kNN we need to choose an appropriate value for k , and the success of classification is very much dependent on this value. In a sense, the kNN method is biased by k . There are many ways of choosing the k value, but a simple one is to run the algorithm many times with different k values and choose the one with the best performance [16].

Due to rigorous testing, reassigning of k values, re-evaluating the results time and time again, along with the case of kNN being a supervised machine learning model, meaning that the classification points need to be identified to make kNN concise and effective, for this work being done, it is not a feasible model for this project. There are too many limitations and drawbacks with better options available.

SVMs is a machine learning algorithm that is can be used for classification analysis. The algorithm is a supervised classification machine learning algorithm. This supervised algorithm, like the kNN means it requires labelled training data to learn and make accurate predictions.

SVMs find a hyperplane, which is a classification line in the feature space based on the data points on an X-Y axis, that separates the different classes with maximum margin. The hyperplane is selected so that it the distance between the closest data points from each class is maximised, known as support vectors. This hyperplane is computed to choose the best margin for classification line so unseen data can be classified effectively.

SVMs were considered for the fact that they can handle non-linearly separable data by mapping the data into a higher-dimensional feature space using kernel functions. A kernel function is a method of using a linear classifier to solve a non-linear problem. It entails transforming linearly inseparable data to linearly separable ones The kernel function is what

is applied on each data instance to map the original non-linear observations into a higher-dimensional space in which they become separable¹.

SVMs also can handle high-dimensional data and they are effective in their handling of small datasets. However, due to the supervision required for SVM's to work accurately, and the lack of classification data in the dataset, this is not a viable option.

This then has been brought to K-Means Clustering. This machine learning algorithm seems to fit all requirements needed of an algorithm to cluster the data effectively and accurately.

K-Means Clustering is an unsupervised learning method that classifies network flow by categorizing a dataset into a definite number of clusters fixed a priori. The key idea is to select k centroids randomly, one for each cluster. Each input represented as coordinator by considering the feature values which is consisted of a group of points, each point is allocated to the closest centroid, and each group of points allocated to a centroid is a cluster the distance is the measure. The centroid of each cluster is updated later based on the points allocated to each cluster [17].

The only issue with K-Means is the is the choosing of k centroids randomly. To bypass this limitation of K-Means, simple methods can be implemented upon the data to understand what the best number of centroids would be, and this optimal figure can be deployed on the K-Means model for effective clustering.

The elbow method is a heuristic used in determining the number of clusters in a dataset. It works by fitting the model with a range of different values for the number of clusters, and then for each value, it computes an evaluation metric, such as the sum of squared errors (SSE) or the silhouette score. The elbow method is easy to implement by looking at the ideal k value graph with the position on the elbow along with the SSE (Sum of Square Error) which is less than 1. The best cluster k result will be the basis for clustering [18].

To use the elbow method, the evaluation metric is plotted for each value of the number of clusters, and then an "elbow" in the plot is identified. The elbow is the point of inflection on the curve where the improvement in the evaluation metric begins to level off. The number of clusters corresponding to this point is then chosen as the optimal number of clusters.

1

<https://towardsdatascience.com/kernel-function-6f1d2be6091#:~:text=The%20kernel%20function%20is%20what,in%20which%20they%20become%20separable.>

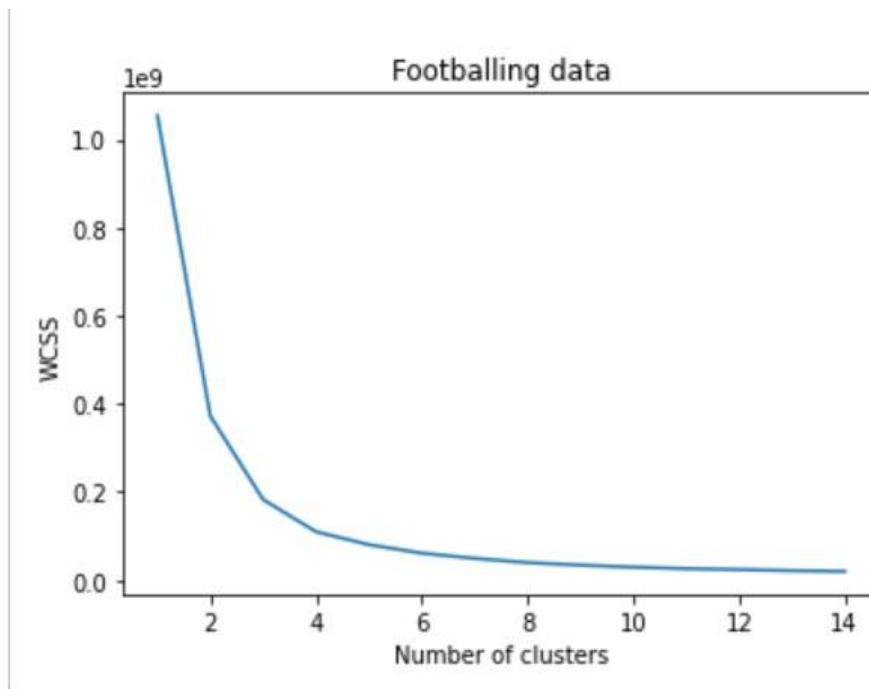


Figure 3: Elbow Method to determine the optimal value of K in K-Means model.

For Figure 3 above, the elbow begins to level out where $k = 4$. This indicates the optimal number of clusters to cover the most variance in the dataset is 4. The data is segregated into clusters to group teams with similar data together. This data highlights a team's playing style and so by choosing $k = 4$, it is identified that there are 4 clear differences in playing styles in the data across the top 5 leagues.

Once the optimal number for k has been implemented on the data and the data has been clustered into their labels, it is important to visualise the data and apply additional clustering techniques if necessary to clean and organise the clusters for deeper insights and for full understanding of the cluster points.

The features of the dataset are scaled using a standard scaler. This scaler transforms the data so that the data has a mean of zero and a variance of one. The standardization process of the input features involves subtracting the mean of each feature from each data point and then dividing by the standard deviation.

$$z = \frac{x - \mu}{\sigma}$$

Figure 4: Mathematical equation for standard scaling

In doing this, standard scaling ensures that the transformed data has a mean of zero and a variance of one. These now standardized features are often referred to as 'z-scores'.

Standard scaling is important because machine learning algorithms have an assumption that the input features are on the same scale, and that the scale of the features does not affect

their importance in the model. However, if the features are on different scales, this could result in some features being outlined as more important than other features simply because of their scale, which can then lead to biased or inaccurate models.

For example, Avg. Possession is on a scale between 40-60 as data points on average. Passes received is on a scale of anything between 150-400 as data points on average, and most other data points are on a scale between 0.5-5 on average. Standard scaling addresses this problem and reduces all these data points down to a consistent scale to remove bias from the models.

In addition to this, standard scaling can also help to improve the performance of machine learning algorithms. When features are on different scales, the algorithm may struggle to converge to a central point or may be sensitive to the scale of the features, which can lead to slower or less accurate model training.

A scatter plot is an effective graph for visualising the data points and plotting them against each other. A scatter plot uses Cartesian coordinates to display values for usually two variables for a set of data. Each point in the scatter plot represents a single data point, and the position of the point on the X-axis and Y-axis represents the values of the two variables.

A cluster in a scatter plot refers to a group of points that are closely packed together. These clusters can be visually identified by the concentration of points in a particular region along the plot. Clusters can be used to indicate the existence of any underlying patterns or relationships in the data that may not be immediately apparent from examining individual data points in tabular or array format.

To plot the data effectively, it is important to consider that there are various techniques for identifying and analysing clusters in scatter plots, such as density-based clustering algorithms or K-Means clustering. These techniques can be useful for understanding the structure of the data and for making informed decisions about how to analyse and interpret the data.

Once the features are scaled, t-Distributed Stochastic Neighbour Embedding (t-SNE) and interpretable dimensions are assigned to reduce the dimensionality of the data points to plot the X-Y scatter plot. Interpretable dimensions create lower-dimensional representations of the data that is more easily interpretable. In clustering, the goal is to group similar data points together based on similarity metrics, and these data point groupings are shown through clusters in a higher-dimensional space. However, it is often difficult to interpret and understand these higher-dimensional spaces, as they can contain many complex features.

To make clustering results more interpretable, dimensionality reduction techniques are applied to the data to reduce the number of dimensions while ensuring the relationships between the data points is kept intact. This technique aims to transform the data into a lower-dimensional space that captures the most important and informative features of the data, while discarding less relevant or redundant features.

In clustering, t-SNE can be used to visualize high-dimensional data in a lower-dimensional space, which can make it easier to interpret and analyse, like interpretable dimensions. t-SNE transforms high-dimensional data into a 2D, or 3D space. These data points are plotted on the 2D/3D space depending on the similarities/dissimilarities of their data. The t-SNE library

makes use of Principal Component Analysis (PCA) as a pre-processing step on the higher-dimensional data before deploying the t-SNE algorithm on the data to embed it to a lower-dimensional space for visualizations.

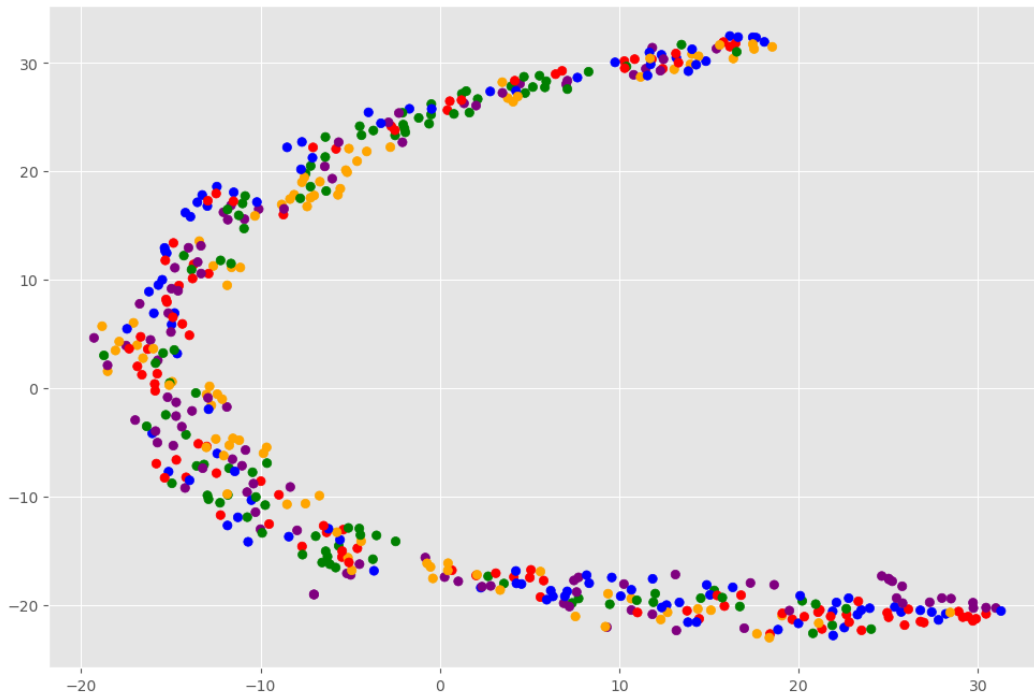


Figure 5: Scatter Plot of each team on X and Y axis

From Figure 5 above, it shows the concentration of data points across the X and Y axis. Each data point represents a team in the dataset, and the colour represents the league from which each team plays in. The colour scheme is as follows; Blue: English Premier League, Red: Spanish La Liga, Green: Italian Serie A, Purple: French Ligue 1, Orange: German Bundesliga. From this preliminary scatter plot, the data points are very intertwined, with little understanding of the clusters and the cut-off points of the clusters. To organize this cluster, a hierarchical clustering model is built.

Hierarchical clustering is an unsupervised clustering method for analysis that seeks to build a hierarchy of clusters. Hierarchical clustering algorithms construct a hierarchy of clusters by creating a multilevel tree structure or dendrogram.

There are two main types of hierarchical clustering:

1. Agglomerative clustering: This is a bottom-up approach, where each data point starts in its own cluster and pairs of clusters are merged as one moves up the hierarchy.
2. Divisive clustering: This is a top-down approach, where all the data points start in one cluster and splits are performed recursively as one moves down the hierarchy.

There are multiple heuristics that could be implemented on a hierarchical clustering model including single-linkage clustering, complete linkage clustering, average-linkage clustering, and Ward's method. Ward's method is used for classification of teams and creation of new team groups. By applying Ward method, the aim is to join elements into clusters so that the variance within clusters is minimized [19]. Ward's method is agglomerative, so this method will use the bottom-up approach.

One advantage of hierarchical clustering is that it does not require the user to specify the number of clusters in advance. However, a drawback is that it may not scale well to large datasets but fortunately, the dataset at hand is small and concise.

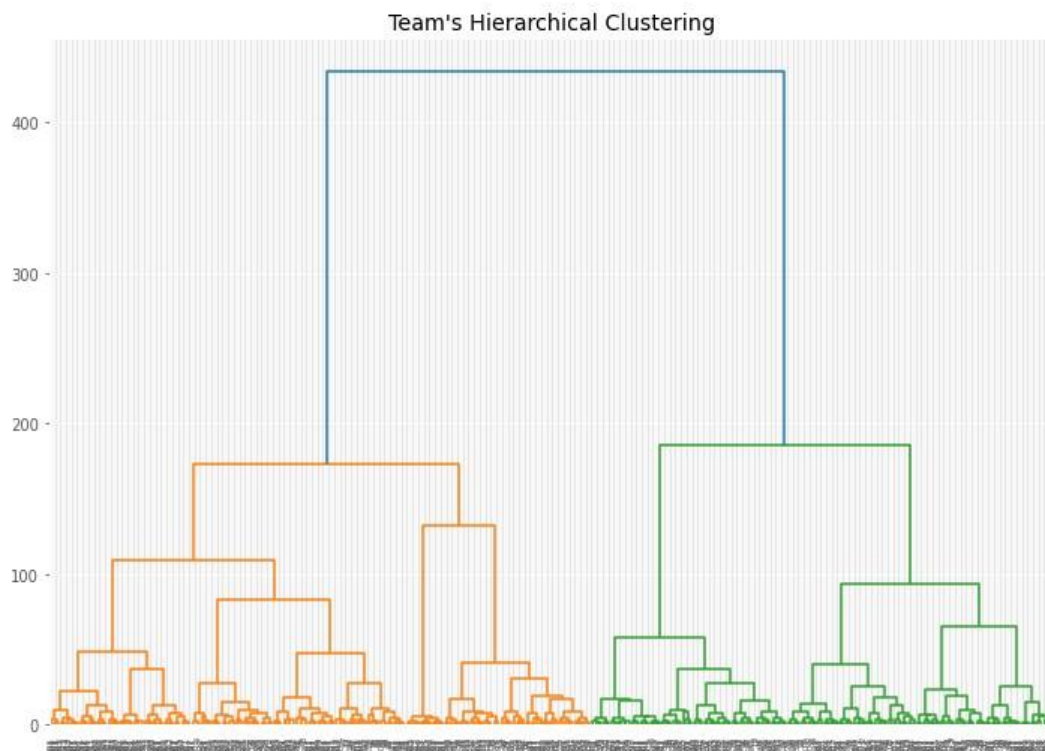


Figure 6: Dendrogram for Team's Hierarchical Clustering

For this model in Figure 6, agglomerative clustering is the approach taken. All the data point starts from the bottom of the dendrogram, and as they move up the chain, they cluster together into similarities. On the second level from the top, it shows a clustering of 4 data points, indicating the 4 clusters initialised to use in the K-Means Clustering model.

The model is created on Agglomerative Clustering setting the clusters to that initialised in the K-Means Clustering algorithm. The labels are then defined from the clusters to each team, and these labels are appended on to the DataFrame.



Figure 7: Post Hierarchical Clustering, Team's Playing Styles on X and Y axis

Based on the model in Figure 7 above, the data points are now separated and visualized based on their clusters. Teams are split quite evenly. From initial analysis and observation, the green cluster is much shorter than the red cluster below it, indicating closer data points in the green, highlighting superior data. From this analysis, it seems that green is to be the best cluster based on this small piece of analysis. In turn, this indicates blue to be the second best, yellow to be the third best and red to be last based on their footballing style.

In terms of their colours; Label 0: Blue, Label 1: Red, Label 2: Green, Label 3: Orange. As a continuation of this analysis, the teams within these clustered were identified running a small line of code to query the DataFrame. The results from each label are posted below.

```
print(df90_num.query("label == 0")['Squad'])
```

```
0          Arsenal 2021 - 2022
3          Brighton 2021 - 2022
6    Crystal Palace 2021 - 2022
9    Leicester City 2021 - 2022
12   Manchester Utd 2021 - 2022
...
485    RB Leipzig 2017 - 2018
486    Schalke 04 2017 - 2018
487      Stuttgart 2017 - 2018
488    Werder Bremen 2017 - 2018
489    Wolfsburg 2017 - 2018
```

```
print(df90_num.query("label == 1")['Squad'])
```

```
1          Aston Villa 2021 - 2022
2          Brentford 2021 - 2022
4          Burnley 2021 - 2022
7          Everton 2021 - 2022
13   Newcastle Utd 2021 - 2022
...
471          Troyes 2017 - 2018
472          Augsburg 2017 - 2018
477   Hamburger SV 2017 - 2018
478   Hannover 96 2017 - 2018
484          Mainz 05 2017 - 2018
```

```
print(df90_num.query("label == 2")['Squad'])
```

```
5          Chelsea 2021 - 2022
10         Liverpool 2021 - 2022
11    Manchester City 2021 - 2022
23         Barcelona 2021 - 2022
35         Real Madrid 2021 - 2022
...
446          Roma 2017 - 2018
465          Nice 2017 - 2018
466    Paris S-G 2017 - 2018
473    Bayern Munich 2017 - 2018
474         Dortmund 2017 - 2018
```

```
print(df90_num.query("label == 3")['Squad'])
```

```
8          Leeds United 2021 - 2022
18         West Ham 2021 - 2022
22    Atlético Madrid 2021 - 2022
27             Elche 2021 - 2022
28         Espanyol 2021 - 2022
...
451         Udinese 2017 - 2018
454         Bordeaux 2017 - 2018
457         Guingamp 2017 - 2018
467             Rennes 2017 - 2018
469         Strasbourg 2017 - 2018
```

It is clear from reading the data along with general knowledge of football that cluster 2 is the superior cluster. This is the green cluster in Figure 5 above and the teams in this cluster will be set as the benchmark for the discriminating model.

Cluster 0 which is the blue cluster in Figure 5 is the second-best cluster based on the data points for the teams contained in this cluster. These teams contain the likes of Arsenal, and Manchester United who had lacklustre seasons by their standards in 2021-22 but they finished 6th and 5th in the league respectively.

Cluster 3 which is the orange cluster in Figure 5 is the third-best cluster. These teams contain the likes of Leeds, Espanyol, Udinese, and Rennes. These teams played very promising, attacking, dynamic, vertical football. Their positions on the league table were not the strongest but their playing style and approach to football is promising based off the 2021-22 campaign for these smaller clubs.

Cluster 1 is the least cluster of the pack. This is the red cluster in Figure 5. Teams like Aston Villa, Brentford, Everton, Newcastle, who all sacked their managers at some point over the previous season end up in this cluster. Some of these teams have taken drastic steps for improvements this season under new managers whilst others remain below par for their goals.

Whilst this graph can give great insights into where clubs lie on the graph given their current data points, it is also important to consider the metrics themselves. Certain metrics have greater impacts on teams reaching new clusters than others, to visualize this, the use of a heatmap is very effective for this purpose. To make effective use of a heatmap, factor analysis is an important piece of implementation.

The goal of factor analysis is to identify underlying factors that explain the relationships between a set of observed variables. Factor analysis assumes that there are certain unobserved factors that affect the behaviour of the observed variables, and these factors can then be extracted and identified using factor analysis. The factor analysis method involves the creation of a correlation matrix of the observed variables, and then using mathematical techniques to identify the underlying factors that are responsible for the observed correlations.

Heatmaps are a graphical representation of data where the individual values contained in a matrix are represented as colours. They can be used to visualize patterns in data and help identify trends or correlations.

In the context of clustering, a heatmap can be used to visualize the clusters that have been identified in the data. Each row and column of the heatmap represents a single sample or feature, respectively, and the colour of each cell represents the strength of the relationship between the row and column elements. For example, if a heatmap is being used to visualize the results of a clustering algorithm, cells that are coloured similarly may indicate that the samples represented by those rows and columns belong to the same cluster.

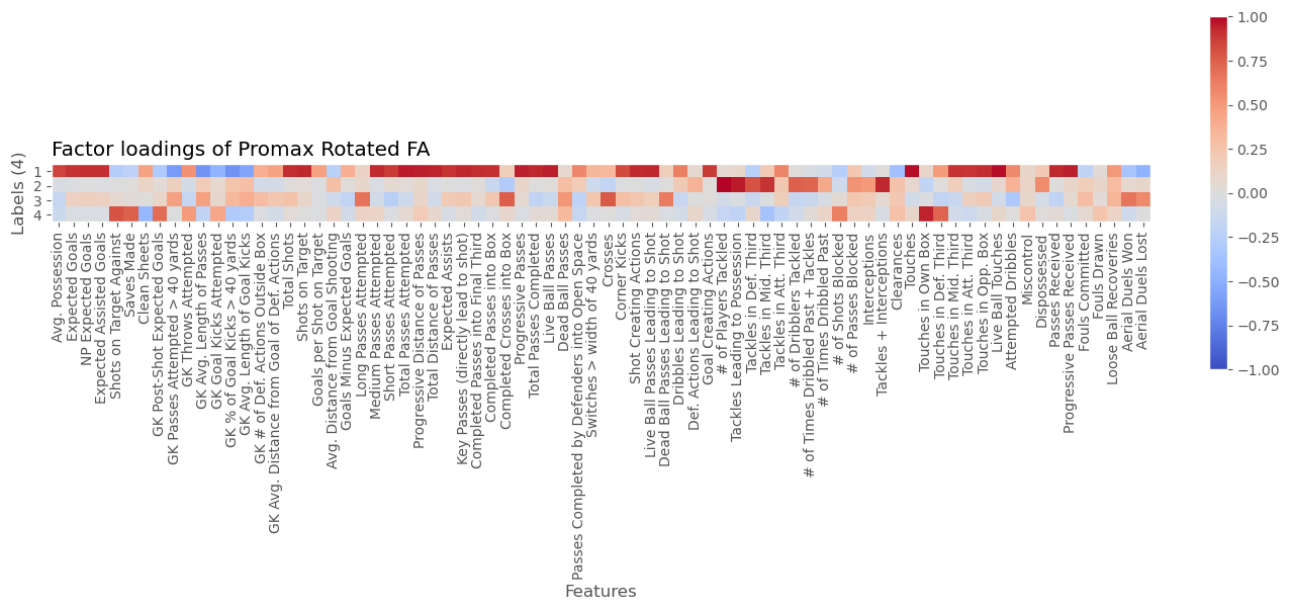


Figure 8: Heatmap for attributes contributing most to each cluster

This heatmap provides all the columns used in the analysis. For each of the four clusters represented in the graph above, the strength of the redness and blueness of the attribute identifies how strong of a relationship it correlates to the cluster. The redness identifies bigger numbers indicate a positive correlation whilst blueness identifies smaller numbers indicated a positive correlation for each cluster. This heatmap gives great insight to understand the variance in the data and it explains well the impact the data points have on each cluster to provide context when looking at the data below.

For cluster 1, from personal knowledge of football and the assessment on the scatter plot graph post hierarchical clustering, cluster 1 on the heatmap matches best to label 2. Cluster 1 has high possession, high number of expected goals, high number of passes etc. indicating a dominating, possession based, high passing team. These teams also have low number of shots against, touches in their own box, and low number of goal kicks going beyond 40 yards.

Cluster 2 matches to label 0. Cluster 2 has high percentage of tackles and regains of possession. Whilst they do not control as much possession as cluster 1, they demand a lot of possession and when not in possession which is more often than cluster 1, they work well to gain possession back with high correlation to tackles and interceptions.

Cluster 3 matches to label 3. Cluster 3 attempt a lot of long passes and crosses, indicating little play on the ball and more, long balls towards the attack to bypass the more organised teams. Their goalkeepers attempt more throws and kicks highlighting the goalkeeper being under pressure more and they win more aerial duels backing up the statement of long goalkeeper kicks and long balls into the attackers to challenge for the ball in the air.

Cluster 4 matches to label 1. Cluster 4 have a lot of shots taken on them, and have a lot of saves made, along with having little clean sheets. Most of their touches are taken either within their box or within their defensive third indicating constant pressure on their back line. They block a high number of shots, not due to good defensive actions but more so due to constant pressure on their last line of defence.

Once all initial analysis, clustering, and exploration of the data is complete, the data is saved to a csv file to start the modelling process. As explained previously, neural networks are the implementation for this assignment. Originally the idea was to use Self-Organising Maps (SOMs) which is a type of artificial neural network (ANN). This network works on competitive learning between data points. These data points compete to plot on to a 2D shape. It is a form of dimensionality reduction which when applied to the data, it displays how far certain cluster is from another. In this way, it is possible to show how far a team is from reaching a better cluster.

However, from initial analysis, consideration, and implementation on SOMs, this did not seem feasible for the result and end goal from this project. The SOM is a difficult technique to implement on high-dimensional data, the results are not very intuitive on individual data points of teams to give a detailed analysis of where they may need to improve, and from a piece of implementation, the accuracy scores were very low and would not pass for deployment.

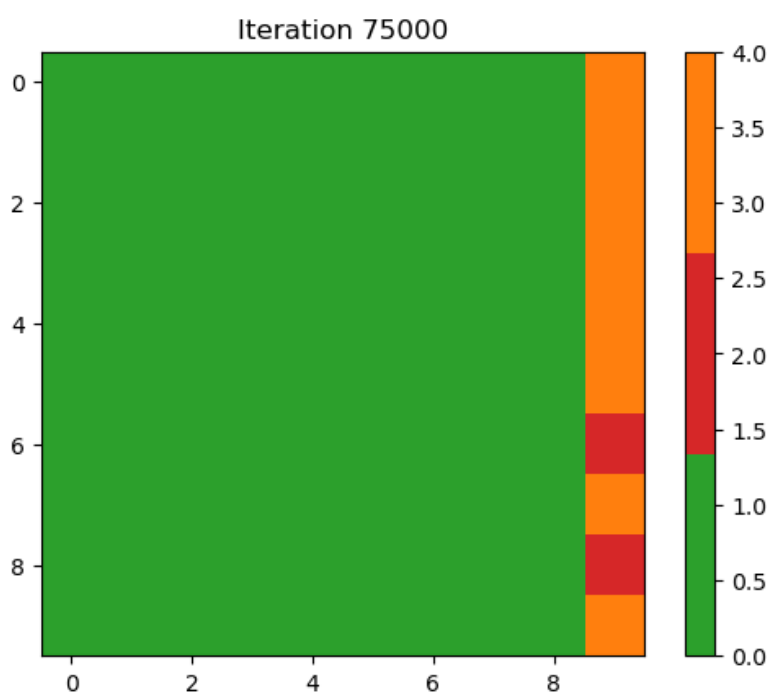


Figure 9: Label Map based on SOM with 25% accuracy est.

From further consideration on the implementation required to suit the end goal of this project, it was then decided that a GAN would be the best approach. A GAN architecture is a Neural Network algorithm for generative modelling. A GAN model consists of two independent sub-models, the Generator G and its adversary, the Discriminator D . The generative model G understands the data distribution $p(x)$ of the real data space x . Then, considering an input noise variable, the Generator G generates new adversarial examples $G(z)$ that have the same distribution of x . The Generator G is trained to maximize the probability that the Discriminator D could correctly predict generated samples as real samples, while the Discriminator D is trained to distinguish if the given sample is real or generated by the Generator G [20]. Mathematical expression for the GAN:

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{data}(x)} [\log D(x)] + \mathbb{E}_{z \sim p_z(z)} [\log (1 - D(G(z)))]$$

In the above equation, the discriminator $D(x)$ is attempting to maximize the quantity $V(G, D)$ for a given generator $G(z)$, while $G(z)$ is the generator's output when given z . $\mathbb{E}_{x \sim p_{data}(x)}$ and $\mathbb{E}_{z \sim p_z(z)}$ correspond to the expected values over all real data instances and over all generated fake instances, respectively. The real data instances in the case of this dataset are all teams within the best cluster labelled '2'. Everything outside of this cluster is labelled as the fake data. The global optimum for this task is when $p_{data} = p_g$, and this corresponds to the global minimum of the training criterion.

To avoid overfitting when training finite datasets, the Discriminator D must be optimized simultaneously with the Generator G . During the start of the learning process, the Discriminator D may reject the high confidence samples created by the Generator G , because they are different from the training data. To address that, Generator G can be trained to maximize $\log D(G(z))$ instead of minimizing $\log(1 - D(G(z)))$.

Firstly, the GAN relies on the data being completely numeric, as this data is for training and testing only, the team and the league they play in currently is quite irrelevant. Both columns are dropped for the purpose of testing and training. The data is split according to the clusters, teams in cluster 2 are set to one variable, detailed as real data, whilst all the other teams contained in clusters outside of cluster 2 are set to a separate variable, detailed as synthetic data. The clustering column is then dropped from both datasets. At this stage, the data is ready for train-test split to run on the model. The synthetic data is split 80%-20% for training and testing. The data is scaled using standard scaler and this scaled data is converted back into a DataFrame to be read into the GAN model.

The GAN model is built using the keras library. The full structure of the GAN model is as follows: the generator and discriminator are defined as functions. The GAN model is defined as a function with the arguments called within this function being the generator and discriminator functions. These functions are set to variables, the batch size and number of epochs are defined and, in a loop, based on the number of epochs, a batch of synthetic data, and a batch of real data are called. The discriminator is trained on the batch of synthetic data set with an appendix of 0 and a batch of the real data set with an appendix of 1. The generator is then trained taking a batch of synthetic data and training this data to pass the discriminator.

The generator function takes two arguments, an input dimension, and an output dimension. This function creates an instance of the Keras Sequential model. This sequential function details the linear steps from input to output, allowing for passing through a series of neural layers, one after the other. A dense layer is added with 256 units and a Leaky ReLU activation function to the model. The use of the Leaky ReLU activation function is used instead of the standard ReLU because this prevents the neurons from failing if there are negative inputs. An alpha parameter is defined set to 0.2. This defines the slope of the function for negative inputs.

Two more dense layers are defined with Leaky ReLU with 512 and 1024 units respectively. The purpose of doubling the number of units through each layer in deep learning architectures is used to increase the capacity of the model to learn complex features and patterns in the data. In doing this, the model can understand more complex relationships between the input features defined and the target variable. A final dense layer with the output dimension and a linear activation function is defined at the end of the model. The linear activation function is used because it can give the output of the generator as continuous numerical values, which can be any real number. The model is then compiled with the mean squared error (MSE) loss function and the Adam optimizer to use during training.

The MSE loss function is used across the space of data science for regression problems. The loss function works by measuring the average squared difference between the predicted output and the actual output.

The use of the Adam optimizer for this model is important. This optimizer is an adaptive learning rate optimization algorithm, and it can adapt the learning rate of each parameter during training. The algorithm computes individual adaptive learning rates for different parameters based on estimates of the first and second moments of the gradients. This then allows for the optimizer to converge faster and more reliably than traditional optimization methods like stochastic gradient descent (SGD).

The discriminator takes a single argument, being the input dimension. This function also makes use of the Sequential function. This takes four dense layers like the generator model. The first three dense layers make use of the Leaky ReLU function, but the number of units is swapped. The first layer takes 1024 units, second layer takes 512, and the third layer takes 256. This scales down the opposite way so that an equilibrium between the generator and discriminator can be found. The alpha value is set to 0.2 and makes use of the linear activation function. It makes use of the MSE loss function and the Adam optimizer as well.

The GAN function makes use of the sequential function and adds in both the generator and discriminator models. The model is compiled using the MSE loss function and the Adam optimizer with a learning rate of 0.0002.

The input and output dimensions are defined as the shape from the real dataset. The three models are then set to variables and a batch size of 100 and epoch of 1000 is defined. The batch size is the number of rows to be taken in each epoch of training. The epoch is the number of iterations done on the training data. For each iteration, 100 rows from the real dataset and 100 rows from the fake dataset.

For each 100 rows taken, the discriminator is trained with the fake data being appended with values of 0 and the real data being appended with values of 1. The model is trained on differentiating the difference between the real data and the fake data. Whilst this is model is being trained, the generator is being trained, taking in this same fake data with an appendix of 1 to generate new data based on the data being fed into the model to pass the discriminating model.

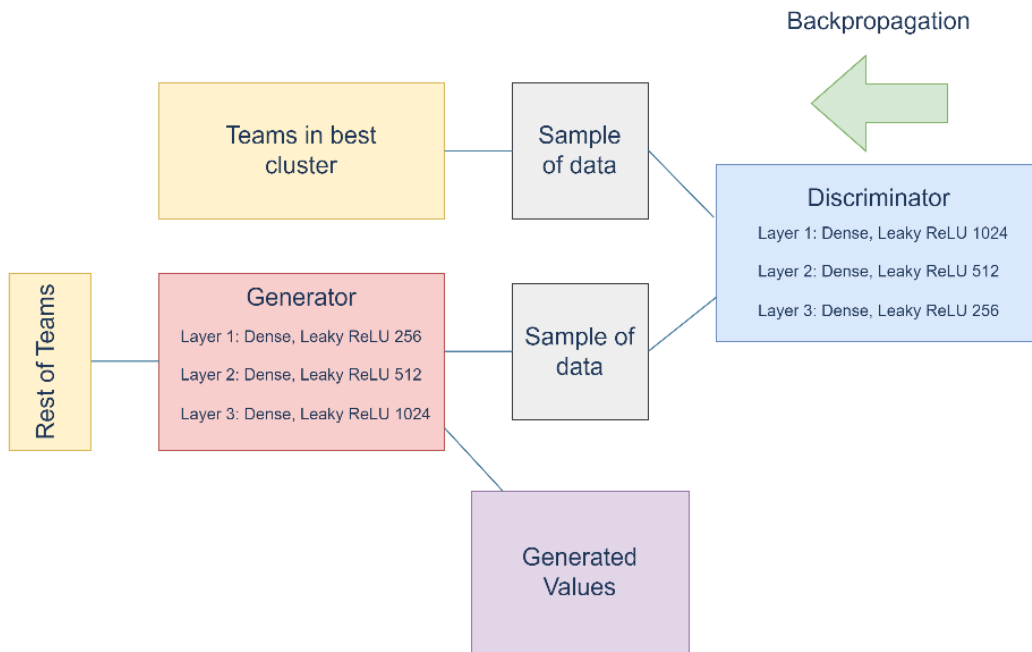


Figure 10: Visualisation of the overview architecture of the GAN model

Once the model is trained on the datasets, the generator model is tested on the test data. The data is transformed using the scaler used originally on the training data to remain consistent. This scaled data is then converted to a DataFrame, and then converted to tensors for testing using tensorflow. Generated data is then produced based on the input of the test data. This data is then inversed from the scaler back to the original format and converted to a DataFrame.

The testing data looks promising, with average possession, expected non penalty goals etc. drastically improving on the generated data. These two variables were chosen as from the factor analysis, they had powerful correlation to the label 2 cluster. Teams that were averaging 44-47% possession in their initial data, should be aiming for possession retention of between 52-56% possession depending on their data points.

To finalise this analysis and to produce results, data from the Portuguese Premeira Liga has been called from the URL, with all pre-processing steps taken from that of the original dataset. This Portuguese data will be used as part of the validation for highlighting how a single team could use this model for production within their business model.

6.0 Results

In this project, the overarching goal to be achieved was the successful implementation of a Generative Adversarial Network on to football in-game event data for each team from Europe's Top 5 Leagues over the past five seasons. For this purpose, the generative model would model and create synthetic data based on the data fed into the system as an input. Such input data would be a certain team's data to attempt to understand where they need to improve in-game to compete with the best outfits in Europe. Using K-Means Clustering and Agglomerative Hierarchical Clustering, it was found that from the data, there are four distinct clusters for grouping. The classification models built were able to identify these clusters and sort the teams into these clusters based on their in-game data.

The obtained dataset was then split dependant on the clusters and the dataset with the teams outside of the best cluster was split into training and test with a ratio of 80:20. The architecture of the generative model and the discriminating model was very similar. Both made use of Leaky ReLU with four dense layers. Each layer doubling units or halving units to learn and mutate complex relationships between variables. Adam was chosen as the best optimizer for the model.

The results from the finalised model looked very promising. There is some intuition involved on the expertise of football but with the help of radar plots to be deployed, these data points can be looked on quite effectively, giving good insights into what needs to be improved.

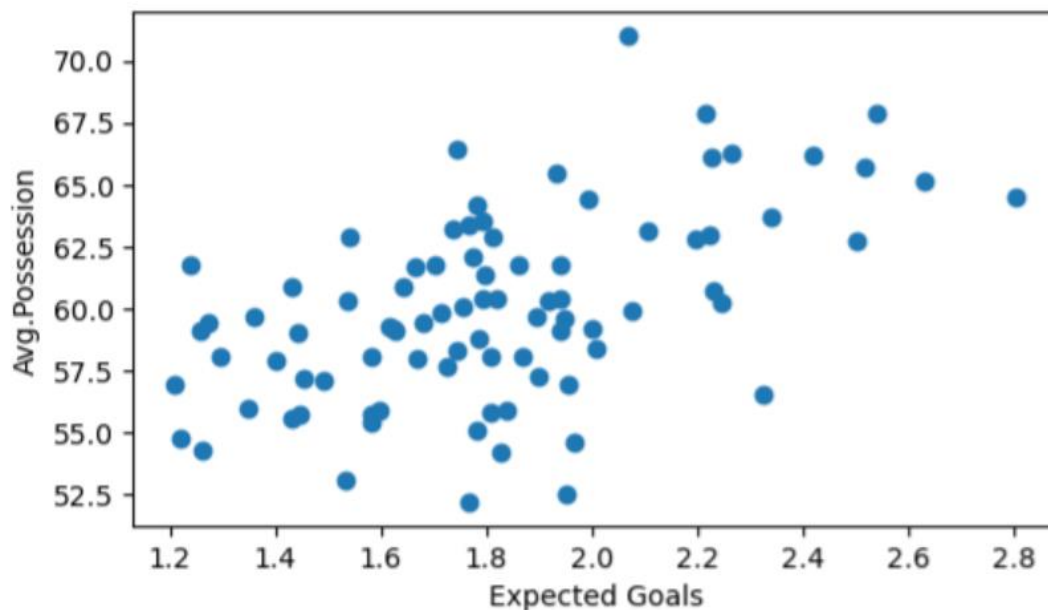


Figure 11: Scatter Plot for the 'real data' cluster

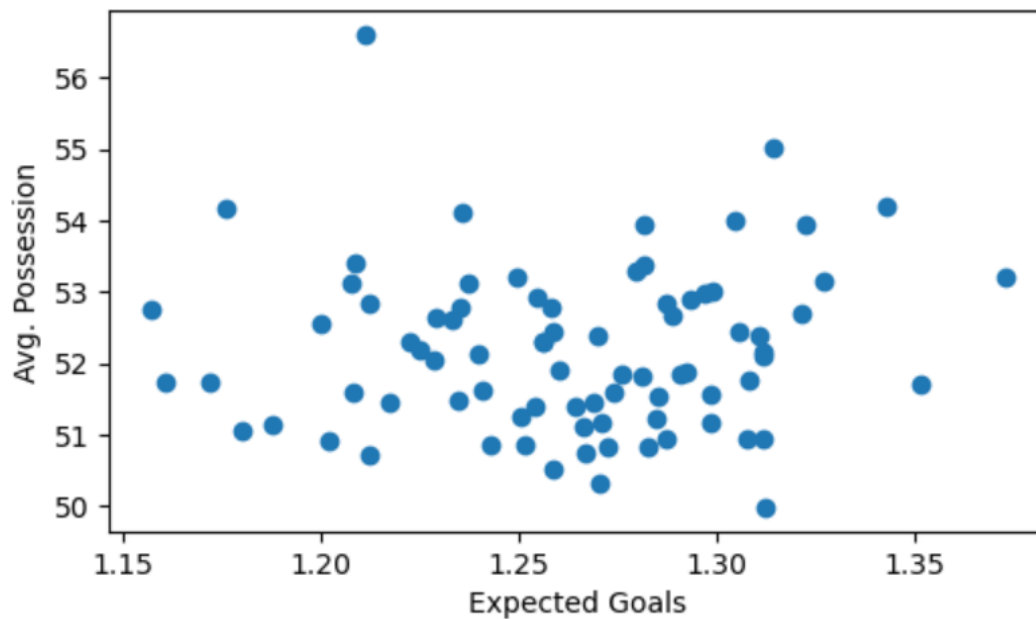


Figure 12: Scatter Plot for the testing synthetic data

From the scatter plots above, the data takes the average possession and the expected goals into account. As all metrics are influenced by the possession had or not had, it is a good weighting system to measure the rest of the variables against. From the two-scatter plots above, it is obvious that the generator models the data more towards the centre of the scatter plot whereas the original dataset was skewed to the left of the scatter plot. This implies a shift of an increase in the expected goals that is feasible for these smaller teams.

The skewness does not tend to be a major factor as it details how there are more teams within this deviation that have lower possession counts, and therefore lower expected goals, than the teams outside of this normal distribution with the likes of the team who average 70% possession. This generator gives realistic goals for these smaller teams to hit, such as the likes of having a range of 50-57% possession to dominate the game on the ball and to hit an expected goal output of more than 1 expected goal a game, the majority falling between 1.2 and 1.3 expected goals per game. As these are above average teams, and the xG model considers the average shot, this should equate to a team scoring at least 2 goals a game. This significantly improves on their previous data, where they hit less than 1 xG per game and the possession stats sitting somewhere in the 40's.

However, to understand how well the model works through plotting histograms, the skewness and distribution of the data is important to consider. The data is non-linear as showcased by the scatter plots, and this stands through for the rest of the variables in the datasets. It is important to understand that the use of standard deviations with non-linear data is not always the best idea, as well as the mean can be skewed because of outliers and the dataset could be slightly top-heavy when the model's output is designed to meet close to minimum requirements to fit into the best cluster. To understand how well the data is performing, histograms are plotted on the synthesised data and the data for the best teams.

After careful consideration and with some experimentation, it was found best to plot the 25th percentile of data from the synthesised data and the minimum value data from the best teams'

data. This is due to the model being realistic on what the lesser teams can achieve, and they will not be hitting the top-heavy pieces of the cluster without extreme overhauls of club personnel and players. So, the histograms plot the minimum requirements for the best cluster and the 25th percentile of the synthesised data to involve most of the clubs from synthesised dataset.

It is important to remember that complex relationships between data points are kept intact during the modelling training and testing process and so for the lesser teams to realistically hit these clusters, they may need to be below the minimum values of some variables to flourish in other variables to bring them into that superior cluster, which is why using the 25th percentile and not the minimum value was found to be a better gauge.

Minimum Avg. Possession for the Best Teams & 25th Percentile for the Synthesized Data

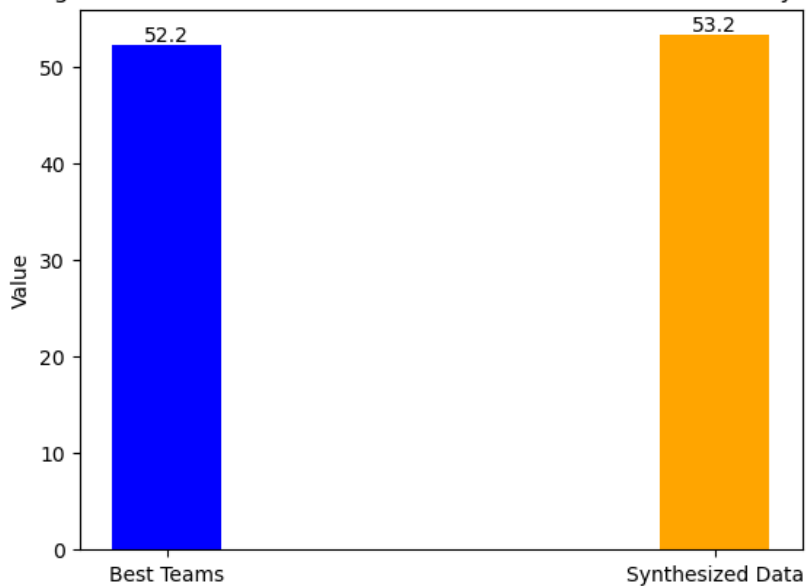


Figure 13: Histogram plotting Possession Stats from the Best Teams and Synthesised Data

Minimum Progressive Passes Received for the Best Teams & 25th Percentile for the Synthesized Data

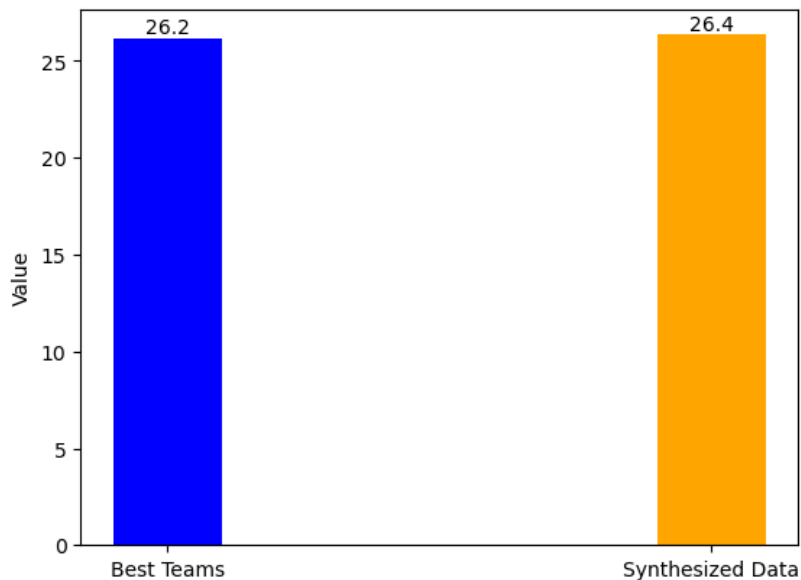


Figure 14: Histogram plotting Progressive Passing Stats from the Best Teams and Synthesised Data

Minimum Expected Goals for the Best Teams & 25th Percentile for the Synthesized Data

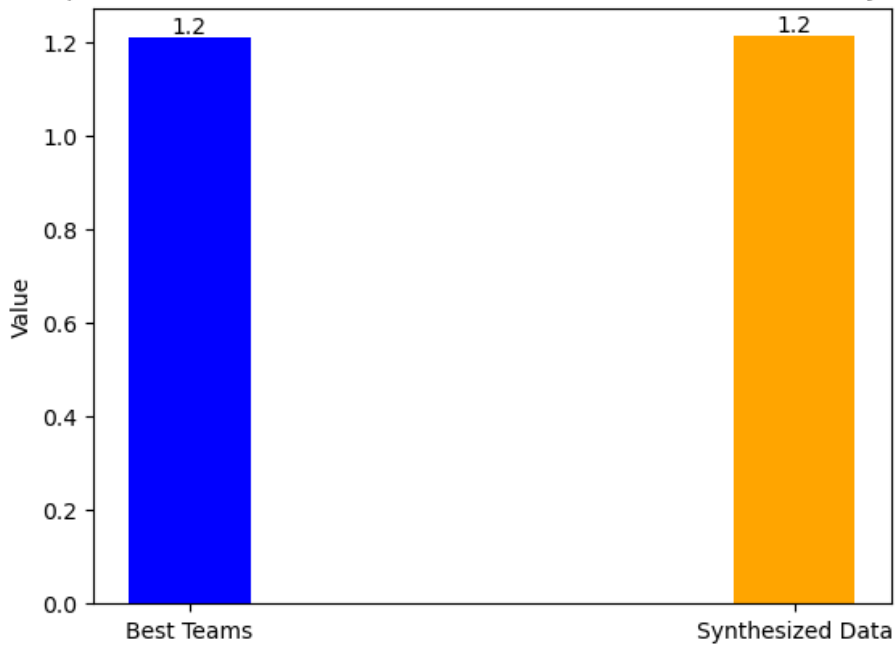


Figure 15: Histogram plotting Expected Goals Stats from the Best Teams and Synthesised Data

Minimum Key Passes for the Best Teams & 25th Percentile for the Synthesized Data

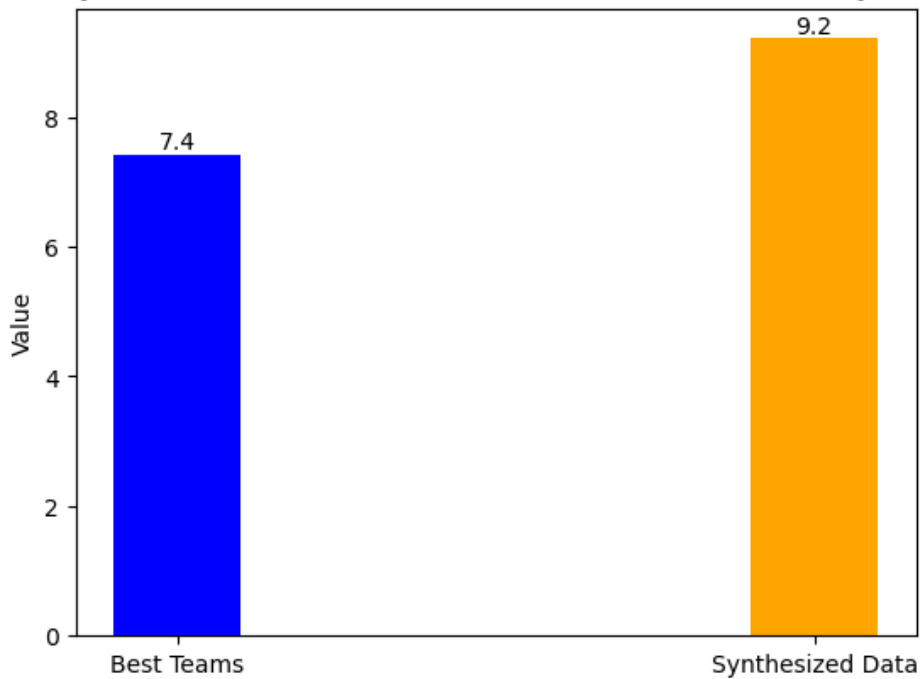


Figure 16: Histogram plotting Key Passes Stats from the Best Teams and Synthesised Data

These histograms detail that the majority of the teams' synthesised data meet or surpass the minimum requirements for some if not all metrics portrayed. These specific metrics were chosen as referred to the factor analysis heatmap. Some of the best clusters main attributes were that of possession, progressive passes, expected goals, and key passes.

For a mini case study and as an example of how a manager/club can use this model, a team was taken from the Portuguese validation set, namely Vitória de Guimaraes. Vitória de Guimaraes play in the Primeira Liga and last season, they finished 6th in the league. They finished one place below qualification for European competitions, and this season so far, they stand in this exact same predicament. They are four points off 5th place with four games to go.

Vitória like to play possession-based football with rotations between the defenders, midfield line and wingers to keep the ball moving. Last season, they played a 4-3-3 formation. This means they played four defenders, three midfielders, and three attackers. This set up is common across Europe and is synonymous with the top teams of last season. In this formation, Vitória play an attacking style with overloading their wings to open space in the centre for their central striker. If the space does not open, then the ball is played into these overloads on the wings for real damage to occur. Out of possession, in the first phase (the ball being with the opponent's back line in their own defensive third), Vitória like to press high to win the ball back further up the pitch and as previously stated, attempt to open the centre by overloading the wings. If the press is unsuccessful and the ball progresses for the opponent into the second phase of attack in the midfield area, Vitória deploy a rest midfield who block angles and passing lanes forward but rarely press the man on the ball.

At the end of the 2021-22 season, Vitória's manager was sacked due to relationship fractures with the higher ups in the club, however, their style of football remains similar but in a new formation. A 3-4-2-1 formation is deployed under their new manager, João Aroso. This means that Vitória deploy three central defenders, four across a rest midfield, two inverted wingers, and a lone striker. This new formation allows Vitória to recycle the ball in rotations between their three central defenders in the first phase. Depending on their build up, one of the two central midfielders will come short to receive the ball for progression or the left and right sided midfielders would pull short for an alternative option. Their two inverted wingers will push wide to open the space in the centre and to create attacking routes.

Out of possession, the wingers invert to create a compact box with the two central midfielders where they will press high to win the ball. If this fails, the midfield four act as a rest defence to block angles and passing lanes forward. If all else fails, the team move into a deep block, with their two wide midfielders of the four dropping back and creating a defensive line of five, the wingers drop into these vacant positions to recreate this midfield line of four, and the striker remains as the only pressing option on the opponent's defenders.

With this background context in mind, when looking at the data, the context behind the data starts to become clear and decisions based on the data can begin to be made to grow the team to get into the European spots, which will grow revenue, success, and an increase in fanbase.

	Avg. Possession	Expected Goals	NP Expected Goals	Expected Assisted Goals	Shots on Target Against	Saves Made	Clean Sheets	GK Post-Shot Expected Goals	GK Passes Attempted > 40 yards	GK Throws Attempted	...	Attempted Dribbles	Miscontrol	Dispossessed
16	48.7	1.45	1.258824	0.932353	4.235294	3.0	0.235294	1.391176	15.323529	3.5	...	16.941176	15.294118	7.617647

Above is part of Vitória's data as it was for the 2021-22 season.

0	Avg. Possession	Expected Goals	NP Expected Goals	Expected Assisted Goals	Shots on Target Against	Saves Made	Clean Sheets	GK Post-Shot Expected Goals	GK Passes Attempted > 40 yards	GK Throws Attempted	...	Attempted Dribbles	Miscontrol	Dispossessed
0	53.297264	1.155667	1.6382	0.82472	3.935609	2.479527	0.087918	1.03548	11.816778	4.731429	...	22.555769	18.485447	5.128649

Above is part of Vitória's data based on the generative model. This data is ideally as Vitória's should be to move them in the direction of playing like the best teams in Europe. The data above shows that Vitória need to ensure that when they are controlling the ball and recycling possession, that they keep the ball better. They need to ensure when they have the ball, they're picking smarter passes and when they lose the ball, that they are always pressing, breaking out of their rest midfield shape, and pressuring the ball more to win it back. Their expected goals value is good at 1.45. However, they need to improve their attacking open play, shooting accuracy and shooting decisions as their NP expected goals value is .4 off what is required. Their defensive work needs improving to ensure they face less shots per games and requires their keeper to make less saves. Their clean sheet values are good but their build up from goal kicks and from general goalkeeping play is not up to scratch. This plays into possession of kicking the ball into the opponent's half less from the goalkeeper and introduce the goalkeeper more into the possession recycling phase and the build-up phase from defence to attack. In terms of dribbles, their wingers when on the ball need to drive at the defence more instead of playing safe options and letting the team get back into their defensive shape, and dispossessions is self-explainable and ties into their control of possession.

Whilst it is easy to understand they need to improve possession and the goalkeeper needs to kick it less and they need to be dispossessed less, the highlighted issues can be drilled down into, and this is where the use of the radar plots come in.

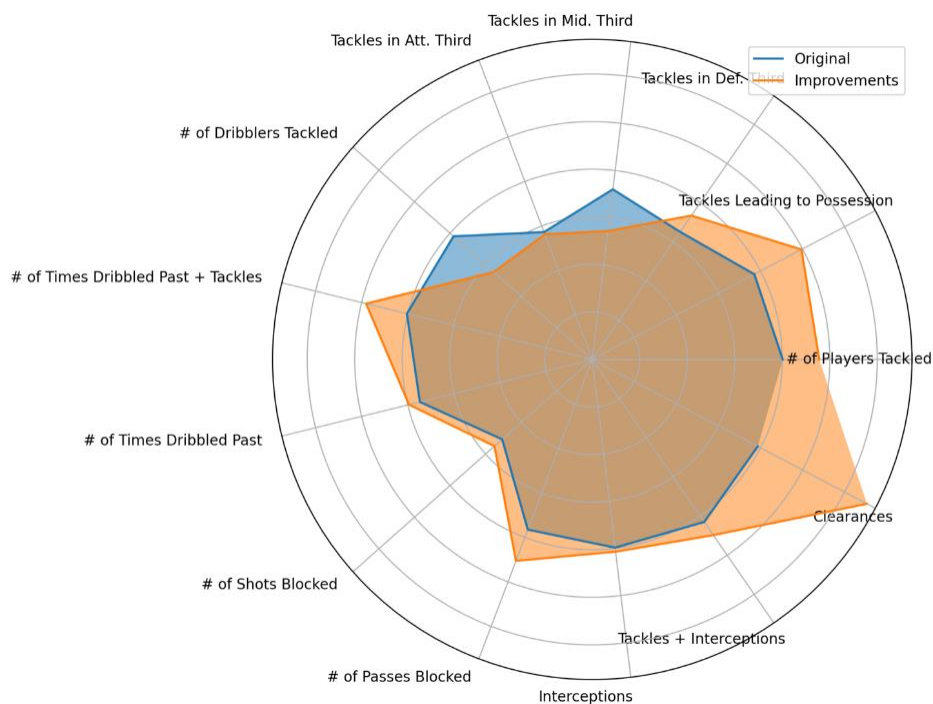


Figure 13: Radar Plot to highlight defensive actions data for Vitória FC 2021 - 2022.

From the radar plot on the team's defensive actions, the blue indicating the team's original data and the orange indicating the improvements, the team are excelling in the number of dribblers they are tackling, but where are they tackling them? From the plot, it highlights that most of the team's tackles are coming in the midfield and attacking thirds, however, not enough are coming from the defensive third. The team is too passive in attempting to tackle and press the opposition, and instead result to their mid-block once the opposition team breaks out of the early attacking press into the midfield. The midfield are doing well to recover the ball but their defence needs to shoulder the burden more in winning the ball when necessary, retaining possession and constantly keeping pressure on the opponents when building from that first phase.

Based on the results received, the interpretation behind the results, and the data points on the radar plot, the results are very promising for moving forward and with club personnel with higher understanding of football than this assignment, very promising insights can be taken from the data to deploy on to the football field to improve Vitória and push them into those European qualification spots in the league.

Dashboarding is a fundamental tool which is extremely useful for deploying on to models and to see creative outputs. These outputs are put into visualisations, in this case, radar plots, also known as spider charts or radar charts, for easy comprehensibility of the data. This dashboard can be filtered by team and by the statistics that are wished to be seen to make it extremely useful and interactive for managers to implement.

The code for the dashboard is hosted in a python file and is called in the Jupyter Notebook. This dashboard is hosted on streamlit and is very intuitive for use by personnel of a football club to bring their style of football in the right direction of the best teams in Europe.

Any teams football data can be read into the script, as the script configures the streamlit page setup, along with running the team's data through the generator model each time a team is selected on the app. By reading in the csv file with new teams' data, it is easy to use the dashboard for the team.

The file is read in, the model is opened, and the squad names contained in the file are added to a list. A sidebar select box is added to select any team from the list. Once a team is selected, their data is assigned to a variable. The team's name is assigned to a separate variable and is appended to the dashboard title. The squad and league columns are then dropped because they are categorical variables from the data.

The scaler is then fitted on the data, converted to a DataFrame and then to a tensor. The data is then run through the model to predict new values. The scaled data is then inversed and converted into a DataFrame. A second side box is created with the group of metrics to display on a radar plot. The DataFrame original data, the generated data, and the maximum values of the data are then concatenated together. The data is normalised using a variation of min-max scaling or feature scaling. This scaling technique divides the values by their corresponding maximum value from the variable and then multiplied by 100 to show their proportion of the data between 0 and 100. These values are assigned to a transposed DataFrame and re-concatenated to be set on to a radar plot to highlight the differences between each.



Figure 14: Newcastle United 2021 - 2022 Radar Plot for passing

For the radar plot above relating to Newcastle United’s passing, there are big improvements to be made on how they retain possession. Newcastle’s average possession is 40.4%, they greatly need to improve their passes and their recycling of the ball in the defensive and middle third whilst relieving pressure from their goalkeeper, always threatening to break into their second play and keeping teams pushed back. They need to be more assured on the ball and not lose possession or have mis control of the ball. Newcastle also needs to understand the importance of playing for fouls and committing fouls. These help to break up teams’ rhythm and allow Newcastle to control more tempo of the game as they see fit.

Radar plots have emerged as a valuable tool for analysing football event data in recent years. This dashboarding method offers a comprehensive overview of team performance by showcasing multiple variables simultaneously. In the context of this project, radar plots enable users to examine various aspects of footballing events, such as passing accuracy, shooting chances, defensive contributions, and much more, as it currently is for their team, and the ideal numbers they should have to compete against Europe’s best. By plotting these metrics on a radar plot, with each variable represented as a spoke, users can identify patterns, strengths, weaknesses, and tactical tendencies of teams. Radar plots provide a concise and intuitive way to summarize complex event data, making them valuable for analysing and understanding the football dynamics of teams.

7.0 Conclusions

Data science has become one of the fastest growing industries in the sporting world. Elite clubs, across a variety of sports are branching out their data divisions by hiring data scientists, analysts, engineers etc. to upgrade their data profiling, pipelines, and their insights into the sport and players. By using a Generative Adversarial Network model, it will help to these top teams in football to highlight where they need to improve to play and compete with Europe's elite. From this model, the data department's analysis can become quicker, more comprehensive, and more detailed, allowing for gains to be made on the pitch which results in more success and growth off the pitch. The results from the model look promising with the data provided, and from this small piece of analysis, with possibly huge steps being taken to succeed at the top club level, it is becoming increasingly evident that data science is going to continuously succeed at the highest level of sports.

Data science has revolutionized the sporting world, and it is now an integral part of the decision-making process for managers and team owners. With the help of data science, teams can analyse player performance, identify weaknesses in their game, and devise tactics to improve their overall performance, based on the underlying statistics from their in-game data. This model can lead the automation process of understanding where improvements need to be made as synthetic data is produced to highlight where a team needs to hit on a benchmark scale to succeed as one of Europe's elites. As data science continues to evolve, it is expected that it will become even more important in the sporting world, helping teams to achieve even greater success on and off the pitch.

However, as it is important to discuss the successes of the of the work done, it is even more important to discuss the drawbacks and limitations. The first limitation that came to fruition was the lack of data to form a dataset. The most comprehensive data from FBref only dates to the 2017-2018 season and only for Europe's Top 5 Leagues. With this means there is only a small dataset to build a model off.

Another limitation is based around the data itself. As discussed at the beginning, there is an element of the "eye test" that comes in partner with the data behind football. While the data is very intuitive for understanding how well a team performs on the ball or in defensive actions such as tackles or blocks. It is very hard through data to understand team positioning which is arguably the most important aspect of a defender's game, where to be when the ball is some place, when to press on to an opponent and when to back off, when to push up for an offside call and when to follow in behind to intercept the pass. These sort of limitations on the data is important to understand to assess accurately what the team needs and how to improve the team based on the data alone.

All data is not created equal. For anyone in data science space, it is incredibly important to understand this concept. Limiting the data to the top 5 leagues is best as the further down the leagues you go to collect the data from and implement into the one model, the data becomes less valuable. For example, Leeds United play in the Premier League and are clustered into the third cluster. If the dataset contained teams from the Irish League, such as that of Cork City or Shamrock Rovers, who won the Irish Premier Division title last year, since Shamrock Rovers competition is not relative to that of the competition Leeds face in the Premier League, including them into this model could imbalance it as their data points could

cluster them into the best cluster when they are only the best in relative to Ireland and against team's across Europe, are most likely in the third or fourth cluster.

An additional note, this concept of use for GANs is extremely flexible in how it can be applied, and it would be very small-minded to assume this model is only limited to football. It is important to understand the world of industries and applications within these industries that this model can have an impact in.

Prescriptive analytics seeks to find the best course of action for the future, has been increasingly gathering the research interest. Prescriptive analytics is often considered as the next step towards increasing data analytics maturity and leading to optimized decision making ahead of time for business performance improvements[21].

Prescriptive analytics is a dimension of data analysis that focuses on utilizing data and algorithms to provide recommendations or to prescribe actions to optimize the outcomes. This type of analytics goes beyond descriptive and predictive analytics by not only providing insights into what has happened and what is likely to happen, but prescriptive analytics also puts forwards suggestions of the best course of action to achieve a desired outcome.

The main objective of prescriptive analytics is to optimize decision-making by considering a variety of variables and potential outcomes. It considers various "what-if" scenarios and provides recommendations based on the data given to it and the business goals wanting to be achieved. Prescriptive analytics answers the questions of "What should I do?" and "Why should I do it?" [21]

Prescriptive analytics can be applied to a wide range of industries and use cases. For example, in supply chain management, it can help determine the optimal inventory levels, production schedules, and distribution strategies to minimize costs while maintaining service levels. In healthcare, prescriptive analytics can assist in optimizing treatment plans and resource allocation to improve patient outcomes and operational efficiency. This type of analysis can be applied to automate processes and grow the company's sales and profit margins.

As a simplified example, take data collected on the bestselling sweets in a shop, the data collected from all the sweets in the shop can be collected and run through the model, bestselling sweets acting as the real data and the rest acting as the generated fake data. The output of the model would highlight how the sweets with the lowest selling margins can be significantly improved. Such data could be where its stored, how its presented, is it on a shelf or does it have its own stand, the type of labels on the sweets etc. It removes the guesswork and automates the thinking process to understand how to sell the sweets.

8.0 Further Development or Research

Given further time to develop and research more on the topic of GANs, data modelling in football and the practical uses, there are a few things that could be improved upon or changed on the project outlined.

Firstly, as time moves forwards through seasons, there will only be more and more data to fit the requirements for this project based on Europe's Top 5 Leagues. More data allows for more comprehensive analysis, better model training, and a better product for output into the real world and an actual feasible end design that would work to its best when implemented into a club's institution. With additional data, the model can be retrained and refined in the future for a more accurate output and possibly for new clusters to be defined in the preliminary analysis phase of the project.

Secondly, in a similar fashion to how time will move forward increasing the range of the data in terms of years, through the research done on this project it is becoming increasingly evident that data science in football is only going to increase and become more integral and central to how clubs run. The use of data analytics in sports can be traced back to the early 2000s, but it has only been in recent years that it has gained mainstream acceptance in the world of football. This is mainly due to the datasets being analysed have become more comprehensive, and the algorithms used have become more sophisticated and robust. Since this is evidently becoming the case, greater metrics on the football field will be explored as teams understand the impact the use of such data can have in their success. Over the next couple of years, more off the ball data will be incorporated into models, this will help to understand the value of defenders through modelling, the value of dribblers and midfielders receiving the ball on the half turn, where the striker is in the box and the runs they make when so that they can meet the ball at the best possible time to have a shot on goal. As of this paper written, these metrics do not exist but as data becomes more central to operations of a club, these types of metrics will begin to be tracked which can then be implemented into the model to produce greater synthetic data for clearer analysis and to give deeper insights into the improvements that can be made on the pitch.

These first two points are based on future work that could be done if/when the data is made available. However, given more time on resources, there is very interesting work being done on clustering similar clubs/leagues to identify which teams are most like which based on how similar their leagues are also. This is work has been carried out by Ben Griffis¹. Griffis outlines that building such a model could be useful in real world application for recruitment, like the work done on this assignment where the weaknesses are highlighted and work within the squad and the transfer market may be of benefit to fix these weaknesses and turn them into strengths.

¹ <https://cafetactiques.com/2023/04/26/league-and-club-similarity-proof-of-concept-and-applications-for-recruitment/>

Griffis outlined the work done by Opta on a concept named Power Rankings¹ which rates over 13,500 domestic men's teams on a scale from 0 to 100. Griffis makes use of this power ranking system in work of his own model. In a similar fashion, this type of power ranking could work in favour of synthesising teams' data for more leagues than just the top five. In the case of Shamrock Rovers, if their data is weighted accordingly based on their power ranking, this tool could really be effective and could allow the GAN model built to have the likes of Shamrock Rovers and in fact all top division leagues from all over the world included in the dataset for training the model.

The Opta Power Rankings employ a hierarchical Elo-based rating system to gauge the performances and strength of each football team. The Elo rating system, which was originally devised for chess player ratings, has been adapted to various sports, including football, and is currently used by FIFA for its official world rankings for both men and women. To evaluate each team, the Elo algorithm works off match results from more than 2,500,000 games played since 1990, enabling it to allocate a rating that can be compared across leagues, nations, and continents.

A major challenge in comparing football teams worldwide lies in the limited opportunities for crossover between different leagues and nations. To address this issue, a hierarchical structure has been employed to facilitate the circulation of Elo points at a faster pace. This method adjusts a team's rating based on its position within its own league, as well as its league's rating, its country's rating, and its continent's rating. For instance, to determine KRC Genk's overall Elo rating in the Belgian League, four separate Elos would be added together: one for Genk, one for the Belgian League, one for Belgium, and one for Europe.

In this concept, it gives weight to clubs/leagues higher up power rankings so that clubs are only called similar if their leagues are similar and not because their data says so. In this fashion, it is interesting that Molde FC from the Norwegian League could have similar data points to Chelsea in the Premier League because the opposition they face is relative, however because Molde play in a much worse league according to the power rankings defined by Opta, these teams would never cross path in terms of their quality or similarity scores because the Premier League is so heavily weighted in comparison to the Norwegian league.

This concept served of great interest and a concept to seriously consider if this project was to be move forward given the time and resources. It will be intriguing to find out what the future holds for data in football across the world. As this industry in football is relatively new at the highest level, a lot can be learned, evolved, and made better to help teams to bridge the gap between themselves and the best in Europe. In a few years' time, it would serve of great interest to compare the results from this project done in 2023, to a new project done in the future, incorporating a greater timeframe of data, greater range of variables, and including many more teams from outside the top 5 leagues regarding the Opta Power Rankings, but it

¹ <https://theanalyst.com/eu/2023/01/power-rankings-your-club-ranked/>

is safe to say that the world of data in football is only going to grow as an industry and it is exciting to see the direction it may take.

9.0 References

- [1] U. Lichtenthaler, 'Mixing data analytics with intuition: Liverpool Football Club scores with integrated intelligence', *J. Bus. Strategy*, vol. 43, no. 1, pp. 10–16, Jan. 2020, doi: 10.1108/JBS-06-2020-0144.
- [2] K. P. Sinaga and M.-S. Yang, 'Unsupervised K-Means Clustering Algorithm', *IEEE Access*, vol. 8, pp. 80716–80727, 2020, doi: 10.1109/ACCESS.2020.2988796.
- [3] L. Tronchin, R. Sicilia, E. Cordelli, S. Ramella, and P. Soda, 'Evaluating GANs in medical imaging', in *Deep Generative Models, and Data Augmentation, Labelling, and Imperfections: First Workshop, DGM4MICCAI 2021, and First Workshop, DALI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, October 1, 2021, Proceedings 1*, Springer, 2021, pp. 112–121.
- [4] L. E. Mays and J. G. McELLIGOTT, 'Neuro-11: A computer system for relating neural spike trains to physiological and behavioral data', *Behav. Res. Methods Instrum.*, vol. 8, no. 3, pp. 325–328, 1976.
- [5] D. Alaminos, I. Esteban, and M. A. Fernández-Gámez, 'Financial performance analysis in European football clubs', *Entropy*, vol. 22, no. 9, p. 1056, 2020.
- [6] J. Oberstone, 'Differentiating the top English premier league football clubs from the rest of the pack: Identifying the keys to success', *J. Quant. Anal. Sports*, vol. 5, no. 3, 2009.
- [7] R. S. Tunaru and H. P. Viney, 'Valuations of Soccer Players from Statistical Performance Data', *J. Quant. Anal. Sports*, vol. 6, no. 2, 2010, doi: 10.2202/1559-0410.1238.
- [8] C. P. Barros and S. Leach, 'Performance evaluation of the English Premier Football League with data envelopment analysis', *Appl. Econ.*, vol. 38, no. 12, pp. 1449–1458, 2006.
- [9] A. Rathke, 'An examination of expected goals and shot efficiency in soccer', *J. Hum. Sport Exerc.*, vol. 12, no. 2, pp. 514–529, 2017.
- [10] M. Cui, 'Introduction to the k-means clustering algorithm based on the elbow method', *Account. Audit. Finance*, vol. 1, no. 1, pp. 5–8, 2020.
- [11] U. Kalenderoğlu, 'Football player profiling using opta match event data: hierarchical clustering', 2019.
- [12] L. Shaw and M. Glickman, 'Dynamic analysis of team strategy in professional football', *Barça Sports Anal. Summit*, vol. 13, 2019.
- [13] L. Xu, 'Synthesizing tabular data using conditional GAN', PhD Thesis, Massachusetts Institute of Technology, 2020.
- [14] C. Bowles *et al.*, 'Gan augmentation: Augmenting training data using generative adversarial networks', *ArXiv Prepr. ArXiv181010863*, 2018.
- [15] B. Bhattarai, S. Baek, R. Bodur, and T.-K. Kim, 'Sampling strategies for GAN synthetic data', in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 2303–2307.

- [16] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, 'KNN model-based approach in classification', in *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings*, Springer, 2003, pp. 986–996.
- [17] A. Alalousi, R. Razif, M. AbuAlhaj, M. Anbar, and S. Nizam, 'A preliminary performance evaluation of K-means, KNN and EM unsupervised machine learning methods for network flow classification', *Int. J. Electr. Comput. Eng.*, vol. 6, no. 2, p. 778, 2016.
- [18] M. A. Syakur, B. K. Khotimah, E. M. S. Rochman, and B. D. Satoto, 'Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster', *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 336, no. 1, p. 012017, Apr. 2018, doi: 10.1088/1757-899X/336/1/012017.
- [19] D. Eszergár-Kiss and B. Caesar, 'Definition of user groups applying Ward's method', *Transp. Res. Procedia*, vol. 22, pp. 25–34, Jan. 2017, doi: 10.1016/j.trpro.2017.03.004.
- [20] S. Bourou, A. El Saer, T.-H. Velivassaki, A. Voulkidis, and T. Zahariadis, 'A Review of Tabular Data Synthesis Using GANs on an IDS Dataset', *Information*, vol. 12, no. 9, Art. no. 9, Sep. 2021, doi: 10.3390/info12090375.
- [21] K. Lepenioti, A. Bousdekis, D. Apostolou, and Mentzas, 'Prescriptive analytics_ Literature review and research challenges | Elsevier Enhanced Reader'. <https://reader.elsevier.com/reader/sd/pii/S0268401218309873?token=AE13BF85D57E66C4995C69894AB065D1DE60132173860A4EECB5E1F1B061AB378F8E5514CBDD1C7C1A3012C81A84FFC&originRegion=eu-west-1&originCreation=20230509140114> (accessed May 09, 2023).

10.0 Appendices

10.1. Project Proposal



National College of Ireland

Project Proposal

Classification of football teams' playing styles
in Europe's top 5 leagues using historical data

31/10/2022

Data Science

BSHDS4

2022/2023

Christopher Weir

X19317131

X19317131@student.ncirl.ie

Contents

1.0 Objectives..... 17

2.0 Background 17

3.0 State of the Art 18

4.0 Data 19

5.0 Methodology & Analysis 20

6.0 Technical Details 22

7.0 Project Plan 23

1.0 Objectives

My project is based on the top 5 European leagues in football. Based on multiple statistics scraped from online footballing data websites, the aim is to extract data regarding the top 5 European football leagues to build a classification model based on the playing style of each team. Along with this I will discuss all the terminology involved in the data and how data, along with these terms, have become more prevalent in football now than ever before.

By the end of the report, one of my aims is for anyone, from a casual fan of the game to the most passionate supporter, all the way to the other side of the spectrum to someone who partakes no interest in the sport, to have the ability to understand football in a simplistic way, the data behind it and to gain an understanding of team's playing styles, along with creating an interactive dashboard for each league to highlight different metrics and where teams from those leagues sit on that metric.

Of course, different managers have their own tactics of the game. Of course, football is an international sport with players who share a common goal. Of course, managers want to remain unpredictable in their tactical setups, but there are always underlying trends in footballing data that can tell us how a team plays football.

Common arguments are made for what league produces the best football, but the reality is, with foreign managers coming into foreign leagues and playing a certain identity of football. Each country's league will have a variance of footballing setups. Examples of this would be the likes of Pep Guardiola, a Spaniard, and Jurgen Klopp, a German, who both manage English Premier League teams have their own identity of football brought from their home countries and implemented into their teams.

Overall, my main aim to analyse teams' playing styles from previous years data in their respective leagues, and to provide a complete, accurate statistical report, along with an interactive dashboard and a classification model, on what is internationally known as "The Beautiful Game" to allow for any reader to gain a little more knowledge and interest on the wonderful sport, and to statistically acknowledge what teams play which type of football.

2.0 Background

I chose to undertake this project for quite a few reasons.

Since I was very young, from the age of 8-9 years of age, I have been an avid supporter of Liverpool FC in the English Premier League. I have grown up watching Liverpool and football as a whole and have always taken a likeness to attempt to interpret how teams play and how football in different countries is played. Of course, when it comes to aspects of data, it will never be 100% accurate purely because you cannot quantify luck, skill, unpredictability, with numbers. Not without extensive data and all possible permutations at least. Due to this, you always must give space in your analysis for what the common viewer calls "the eye test". This is purely watching the game to see how players brains tick, how their feet move, how they dribble and create space and how quick their decision-making is. Nonetheless, data in football

has become a huge factor in key operating decisions on and off the pitch, and data can tell us a lot about the way a team likes to play football, and I would like to be a part of that story at some point in my career.

Through the years of growing up watching football and as I understood more of the game and the tactical roles of teams. I had a growing interest in the data analysis and statistical side of the game. It was from this growing interest where data science (my current course in NCI) took my attention, and it is where I want to lead my future career into the world of analysis and data. There are Premier League teams, such as Brentford and Brighton, who are both excelling in the league table, who's recruitment decisions have been based on the background data and the underlying statistics of players who they have brought in to help the team perform better.

It is in these kinds of statistics and this kind of data that I believe has a huge impact on the pitch and on the sport of football as a whole and it is one, I believe has a big future in sporting regimes.

3.0 State of the Art

Nowadays, there are papers after papers based on data. There are reports on financial performance of teams and how that relates to their success in the top 5 leagues. There are papers based on comparisons of teams/players within the same league along with studies on refereeing performances, covid impacts and comparing the top teams to the rest in a singular league.

Where my report differentiates, is comparing teams and classifying their styles of play. In doing this, it can be very powerful for managers to understand their opponents' style of play, it can also allow managers to understand their own team more as well as show how a team has transitioned from manager to manager in terms of their football. Comparing the styles of their football, all the key's components which go into a game of football and analysing their performances by diving into the data. What makes a team's identity and how did the teams perform. For any special outliers or overperformers, miniature case studies will be carried out to understand more on their playing style. Was it due to financial aid from owners? Due to new signings or due to managerial change?

In my project, I want to highlight and make the reader understand the major components and metrics which makes each team tick. This can be done by creating graphs and diving into the statistics of the teams in each league to provide insight and context into the numbers and exactly what they portray or what they indicate to portray as again, football is more than just statistics and there is that element of the "eye test" which is a key factor in understanding football.

Overall, I believe my work differs because as far as I have researched, I cannot find a paper that has done a statistical analysis of football between the playing styles of different teams in the top 5 leagues, and which has come to conclusions which have been quantified and put

into a model and published. For this reason, I believe my idea to be original and for my idea to be that of interest to readers who partake an interest in the sport.

4.0 Data

I have sent an email to FBref to ask for their permission to scrape data from their website. Once I have the permission from FBref, my plan is to scrape FBref for the necessary data.

To do so, I will gather the 25 URL's I need. One for each league from each of the last five seasons. I use previous season's data as the data for the entire season is available compared to this current season, where metrics can change as the year progresses which could potentially alter some forms of my findings and would need continuous backwards propagation of my analysis.

I will assign the URLs to a variable and define a function to locate and grab every table I need within each URL through pandas and convert these to a DataFrame and assign the function working on each URL as a new variable to be called with each table name, I require.

From this, there is some messy overhead grouping names on each column, so I will define another function to call to get rid of these unnecessary headers.

When I have all the tables from all the leagues, I will merge DataFrames together to keep all data metrics together for each league. I can then then my analysis on the tables, comparing, visualising and discuss the findings and what these findings mean and how they impact the competitiveness of the leagues.

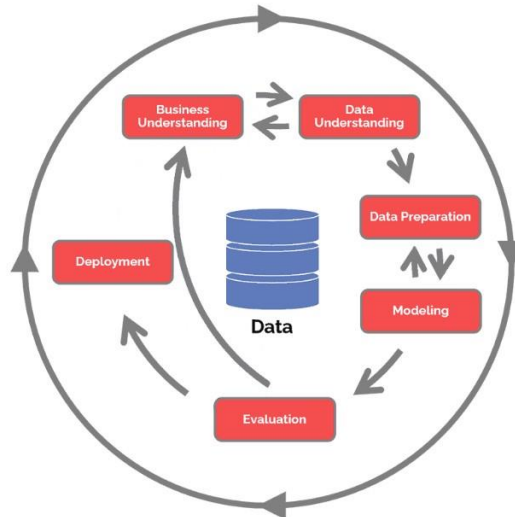
For example, for the URL I have for the English Premier League, it provides me with 12 tables. Each URL provides me with the same 12 tables for each respective league. Each table represents different attributes, strengths, and weaknesses of each team in each league. For clarity, the titles of these 12 tables are: 'Regular Season', 'Squad Standard Stats', 'Squad Goalkeeping', 'Squad Advanced Goalkeeping', 'Squad Shooting', 'Squad Passing', 'Squad Pass Types', 'Squad Goal & Shot Creation', 'Squad Defensive Actions', 'Squad Possession', 'Squad Playing Time', 'Squad Miscellaneous Stats'.

These tables are only for the teams, for example, let's take Arsenal. These are stats to show Arsenal's pass types, their shooting patterns, their goalkeeping saves, their defensive work.

I have done some testing code on pulling the necessary tables, filtering through the tables, and understanding the data given and creating some dummy visualisations and I believe with a little bit more organisation and a clear project path, with this substantial data I can write up and create a very comprehensive and interesting paper on the different footballing styles based on the analysis.

5.0 Methodology & Analysis

The methodology that best suits my project proposal would be the Cross Industry Standard Process for Data Mining (CRISP-DM). This methodology has 6 steps with all steps being carried out in order allows for me to have a comprehensive and full in-depth paper written.



The steps surrounding CRISP-DM are as follows:

Business Understanding

This entails understanding what I want to achieve from this project. What I want to showcase and what direction I want my work to go in. Understanding the application my project will have to the real world and to be able to apply my data and my findings into real world scenarios. So, my goal is to analyse Europe's top 5 football leagues, and understand through analysis, team performance and playing styles of the teams and apply this analysis to real world situations.

Data Understanding

This involves gathering the dataset that I require to carry out this analysis and making sure the data I am using is suitable for the piece of analysis on which I intend to carry out. For my data selection, I will be scraping data from FBref, a football statistics website. On this website I will be scraping from 25 URL's. A 5 URLs for each league, within each URL are the same 12 tables which highlight different metrics about each team within that league.

This is also involving exploratory analysis. Analysing the data, describing the data, and verify how clean the data is, and if there is need for change. This part of the process relies heavily on having full understanding, and description of the comprehensive dataset and all its' necessary metrics.

Data Preparation

Once I have all my tables, this involves ensuring there are no null values and if there are, how to deal with them. It also involves dealing with any outliers that could possibly skew the data in a positive or negative way which is not a true reflection on the dataset along with ensuring all my tables are in the correct format to be usable for data analysis. I will quantify if it is necessary to run a PCA on the dataset and if I will base my data on the PCA or not. Renaming columns, dropping, and adding columns will all be a part of my cleaning and transformation process to ensure I have all the necessary data for visualising.

If any of my data is not in the correct format, I will be correcting my data and putting it into the correct format by passing it as string, float, numeric etc. so it can be integrated successfully with other instances into a clean and perfected dataset to produce the best accurate results so that my analysis can be as thorough and as accurate as possible. I will be combining DataFrames to keep all my data together for ease of accessibility.

Modelling

This involves selecting the modelling technique best to use. Generating a testing design and how is it best to split the data. It involves building the model so that it can be utilised in a project and can be used as a tool for future reference. Assessing the model is a key component to modelling data. My plan is to build and run a few classification models and I will need to interpret which model is best based on my own knowledge, the success criteria, and the test design.

Evaluation

This part of the methodology is about evaluating the results. What is the best model to use when deploying. Reviewing the work, I have done, what was overlooked and what could possibly be improved will also be key in my evaluation. After this evaluation, it is then that I decide if the model is ready for deployment. I will also be building an interactive dashboard detailing the way a team plays and how they compare to other teams.

Deployment

I will develop a plan to deploy the model, hopefully to a website that then be utilised by a team where they can input their statistics and it would output their style of play, where they are least identifiable, and this can be used for certain managers to understand how their opponents play in comparison to their own team.

6.0 Technical Details

Within my project, I will be using Jupyter Notebook, and possibly SQL. Jupyter Notebook is a python environment GUI. Firstly, what I will be doing within this coding environment, is gathering all the URL's I will need to scrape my data from. I will assign variable names to each of these URL's.

I will then define a function to call the URL's for scraping each specific table, so I only need to define each table once and just call the specific URL for each league for the table I want.

I will also define a function to clean up certain tables which have messy column headers above column titles to have my tables looking aesthetically better for ease of readability from readers/viewers.

Once I have certain tables and they are assigned into a DataFrame, I will merge and append the DataFrames together into one massive dataset. I will transport this combined DataFrame into a csv file. A PCA is run on the data to check what attributes are most important for certain target variables, but this may not be used onward dependent on results, and this will be explained further as I move on with my analysis.

Once I call all the necessary tables, I will then build my classification models and determine which one is best suited to my data, and which yields the most accurate results. From this I will analyse the context of the data, the difference between teams and their playing styles and why this is the case. I will also run mathematical formulas on the data where I feel necessary to produce certain metrics to highlight different parts of the data and to give real-world references along with backing from other sources as to why this is the case and how this can impact the game of football.

To provide an example:

```
In [29]: eng_df
Out[29]:
```

	Squad	# PI	90s	Cmp	Att	Cmp%	TotDist	PrgDist	Cmp	Att	...	Att	Cmp%	Ast	xA	A-xA	KP	1/3	PPA	CrsPA	Prog
0	Arsenal	27	38.0	16110	19512	82.6	300732	87839	6886	7676	...	3307	58.6	41	36.8	4.2	425	1161	348	77	1168
1	Aston Villa	31	38.0	12521	16026	78.1	244814	86229	5111	5870	...	3329	55.5	42	33.6	8.4	348	934	305	80	1109
2	Brentford	29	38.0	11551	15894	72.7	234065	86823	4443	5378	...	3622	53.2	33	32.8	0.2	305	877	270	84	1038
3	Brighton	26	38.0	16341	20299	80.5	319622	96950	6439	7316	...	3872	60.4	30	29.0	1.0	343	1120	347	82	1211
4	Burnley	23	38.0	9508	13895	68.4	199165	82436	4020	4911	...	3864	48.3	26	27.7	-1.7	273	801	251	90	873
5	Chelsea	26	38.0	20878	24294	85.9	382347	114543	9010	9831	...	3528	66.6	52	46.9	5.1	430	1561	446	69	1614
6	Crystal Palace	24	38.0	14785	18647	79.3	292522	94227	5671	6494	...	3673	57.8	31	29.1	1.9	287	932	279	84	1037
7	Everton	33	38.0	10325	14297	72.2	206029	78513	4314	5127	...	3293	49.5	29	27.2	1.8	311	843	266	80	910
8	Leeds United	29	38.0	13830	17836	77.5	267920	90173	5535	6394	...	3578	54.6	26	30.9	-4.9	343	969	302	59	1278
9	Leicester City	28	38.0	15911	19634	81.0	301146	95451	6529	7366	...	3389	57.2	46	34.0	12.0	302	972	267	50	1144
10	Liverpool	27	38.0	20856	24964	83.5	412739	130477	7998	8930	...	4525	68.5	71	59.8	11.2	507	1824	528	93	1967
11	Manchester City	26	38.0	24032	27274	88.1	456456	119264	10056	10836	...	4223	74.2	63	59.6	3.4	525	1945	562	78	1766
12	Manchester Utd	29	38.0	16464	20092	81.9	310788	93976	7125	8053	...	3498	64.0	46	40.6	5.4	397	1253	318	68	1304
13	Newcastle Utd	29	38.0	10211	13753	74.2	208167	79329	4140	4886	...	3250	52.7	23	26.0	-3.0	297	759	207	61	743
14	Norwich City	28	38.0	11974	15658	76.5	242050	81553	4705	5440	...	3585	55.4	16	22.6	-6.6	254	705	230	47	922

Figure 8: Premier League Possession Stats for the 2021-22 season

In the table above, we can see possession statistics for all teams in the English Premier League. I will not go into detail on what the column headers mean in this proposal but for instance, I will run a PCA on this table to understand the key metrics for possession. Based on the results, I can then identify and certain amount of possession-based playing styles. I will do this for all 5 leagues and essentially what the scatter plot will show me that for each team, what is their style of play. From this I can then explain in footballing and real-world terms, why certain

teams play a certain way and what this type of football entails. This will give context of teams' performances and their style of play.

I am unsure yet on how I will go about building my dashboard but most likely the data will be inputted into either Power BI or Tableau as I have experience with both. It will have filters to choose from on different metrics you want to see or what leagues you want to see.

7.0 Project Plan

I will lay out my steps and the estimated timeline for each step based on milestones I feel important to hit to maintain readability and comprehensibility for my supervisor.

The first step will be to gather the data via web scraping.

The second step will be to clean the data: this involves taking out duplicate columns, renaming columns, changing data types if needed, and making sure the datasets are comprehensive and eligible for data analysis and combining the DataFrames.

The third step will be to run a Principal Component Analysis (PCA) and if it will be used upon my dataset moving on or not. This could take me a week or two to understand how I need to run my PCA to give me back the results I will need.

The fourth step will be to start the analysis, building the model and writing up my report, give real-world context to the analysis, and explain part of the answer from the model to the question my project is based on.

My fifth step will be to take the data, push the data into a visualisation tool such as Tensorboard, build a dashboard from this for other users to see where teams sit based on certain metrics and upload this tool with my background machine learning model running behind it.

Gathering the Data

To gather the data, I have received permission from FBref to scrape their website for data. I will do so by assigning variables to each URL and creating a function to call the URL. I will create a dictionary within the function for each table I want to scrape from the website so all I must do, is call the necessary URL along with the table variable I want to pull any table from any of the 5 URL links. This part of the process should only take me a day.

Clean the Data

Once I have all the data, I will have to rename columns, remove duplicates, remove grouping headers, merge, and join DataFrames etc. This may take me a couple of days as there is around 60 tables to clean before merging into one large DataFrame. The cleaning will all be done in Jupyter through functions and Panda's libraries. Once the data is clean and I have the

combined DataFrame, I will then push this data into a csv file and save it to my desktop. The data will also be standardized by dividing all the metrics per 90 minutes played, as each game is 90 minutes. This standardizes the data across the board as the German League, play only 34 games compared to the other leagues standards 38. This will take out bias and make the data more accurate for predictions and classifications.

Preliminary Data Analysis

I will correctly load the data and run preliminary data analysis on this such as descriptive statistics to gain more knowledge on the data, rows, columns, differential statistics for each column etc. This type of analysis should not take me too long, perhaps only a day or two to run.

Principal Component Analysis

After I have run my preliminary analysis on the data. My aim is to condense over 100 attributes. This will be done by running a Principal Component Analysis (PCA). My target variable will be the number of points attained by each team as essentially, every team's goal is to get as many points as possible across a league season. From this PCA, I will then make a judgement, backed with research from similar works, on if I will only use the attributes the PCA has given back as being statistically significant or if I will also include others in my analysis report. The PCA itself should essentially only take me a couple of days to run and get the results and to justify my use, or lack of use, of the PCA when I run my analysis on the dataset.

Main Analysis & Modelling

Once I have my revised dataset and I am happy with the decisions I have made based on the PCA, I will then take that revised data and import back into my Jupyter Notebook. From this, I will then outline my aims, what analysis to carry out and will move through different segments of my dataset, building my models and evaluating the results. Also, I will write up a piece on what the context of the data is, how important the results are and what the next steps plan to be. This will take up a good bit of time to perfect the code, perfect the model, and the analysis behind the model. I would like to give myself a good bit of time on this part of the report and I would say it could take me a couple of weeks to perfect the analysis and the modelling.

Another afterthought for modelling the data will be due to the number of rows (490) contained in the data. This may not be enough to build a Machine Learning Model for accurate predictions. This may come down to possibly collecting player data for each of the Top 5 Leagues over the past 5 seasons and matching these players to their respective teams in their respective years.

Dashboarding

The final piece of work on this project will be to build a dashboard. I will build the dashboard in Tensorboard. I will import the dataset into the dashboarding tool. My idea for this will be to have multiple graphs. From this graph, you can filter by the team and their playing style, and you can see for yourself on the dashboard where teams place in terms of their playing style. This dashboarding could be very helpful in football clubs I feel, as it can give managers and backroom staff a sense of where they stand on what metrics regarding their rivals and where they may need to improve, but it can also focus in on what they are doing well and what their identity of football is, and how their pattern of play unfolds. The dashboard itself may take me a week or two to build and perfect and get it looking the way I want it to look and feel.

10.2. Ethics Approval Application (only if required)

National College of Ireland
**Ethical Guidelines and Procedures for Research involving
Human Participants**



SEPTEMBER 2017

1. Introduction

All research involving human participants that is conducted by students or staff at the National College of Ireland should be done so in an ethical manner. The college has therefore developed an Ethics Committee, which acts as a sub-committee of the Research Committee, to ensure that ethical principles pertaining to research involving human participants are upheld and adhered to. All researchers intending to use human participants as part of their projects are thus required to reflect upon any potential ethical issues and submit their research proposals for ethical review before commencing data collection.

This document gives an overview of the core ethical principles guiding research in NCI, while also documenting the procedures required for seeking ethical approval of research involving human participants.

Am I conducting research?

Research is defined as “the attempt to derive generalisable new knowledge by addressing clearly defined questions with systematic and rigorous methods” (NHS Health Research Authority). Sometimes, we collect data in order to evaluate a service or practice we are engaged in (“service evaluation”). The main difference between research and service evaluation is in the aim: research is trying to create new generalisable knowledge, and service evaluation is trying to evaluate whether a delivered service/practice is working well. One project may have both aims included in it. It can be confusing if a service or intervention is involved, whether or not research is being conducted. If new or competing interventions are being evaluated, then it is likely to be research, whereas if an existing service is being conducted anyway, with an evaluative component, then it is likely to be a service evaluation. Research requires consideration of the below guiding principles, whereas service evaluation does not require approval from an ethics committee.

2. Guiding Principles

In line with other research institutions, there are three core guiding principles governing the ethical conductance of research involving human participants at NCI. These principles stem from the *Belmont Report* (1979) published by the National Commission for the Protection of Human Subjects of Biomedical and Behavioural Research. While it is recognised that these principles may be operationalised differently depending on the specific research discipline, it is recommended that these are consulted as a starting point for any research involving human participants.

2.1 Principle 1: Respect for Persons

This principle entails recognition that participants should be treated as autonomous individuals and hence should never be coerced or swayed into participating in a research project against their will. The participant’s right to withdraw from a research study at any time should be respected, as well as their right to dignity and protection from harm.

Respect for individuals can often be implemented in practice via the process of informed consent, whereby potential participants are made fully aware of the requirements involved in participation. While it is recognised that in certain cases deception (i.e. the withholding of certain information from participants) may take place, this should only occur when it is robustly justified for the validity of the research. In cases where deception is justified, researchers should ensure that any potential risk

resulting from this measure is minimised. Participants should also be fully debriefed on the nature of the research after it has taken place.

The principle of respect also requires researchers to protect individuals from vulnerable groups who may have diminished autonomy (see section 4.2 for more detail as to what constitutes vulnerable groups). Where full informed consent is not possible for such population groups, consent may instead be sought from their guardians. In all cases however clear assent, or willingness to participate, should be demonstrated from participants.

2.2 Principle 2: Beneficence and non-maleficence

This principle specifically focuses on the need to protect the well-being of participants. Any potential risk to participants should be minimised, whether that be risk of physical discomfort or of any psychological, emotional or social distress, while possible benefits should be maximised. Researchers adhering to this principle should thus ensure that any potential benefits derived from carrying out the study (e.g. in terms of knowledge gained) should outweigh potential risks. Even in cases where there is only a slight potential risk of harm, participants should be provided with appropriate support to alleviate this.

2.3 Principle 3: Justice

This principle emphasises the need to employ fairness in the distribution of benefits and risks to participants. The way in which participants are selected to take part in research should relate to the purpose of the study, as opposed to other factors such as availability or manipulability of participants. The exploitation of vulnerable populations should be avoided.

Where applicable, researchers are encouraged to consult guidelines stemming from their own professional bodies (e.g. The Psychological Society of Ireland) in addition to the general guiding principles above when planning their research. Researchers should also be sensitive to those issues which are specific to the population under investigation and the methodology that is employed in the project (e.g. qualitative methodologies involving the recording of data may raise issues relating to participants' right to anonymity, as well as the ethical management and use of data). Detailed consideration should be given to all these issues when planning research and when completing the Ethical Review Application form.

3. Ethics Committee

The NCI Ethics Committee was established by the Academic Council in 2012. Acting as a subcommittee to the Research Committee, its role is to oversee ethical issues arising from all research involving human participants that is conducted by students and staff of the college. The key purpose of this committee is to safeguard against any potential harm to participants, and to ensure that their rights are recognised in line with the guiding principles outlined above.

The Ethics Committee reviews all research proposals posing ethical risk to the participants involved, however the decision as to whether projects pose ethical risk is firstly made via the appropriate Filter Committee which operates at School level (see organisational structure in Figure 1 below). The Filter

Committees may review and approve research proposals which are of low ethical risk, while referring those of high ethical risk to be considered by the Ethics Committee (see categories of ethical risk in section 4.1).

While the Filter Committees are made up of staff members with subject-specific knowledge, membership of the Ethics Committee should comprise of no less than five representatives from both the School of Computing and the School of Business, including representatives from the Research Committee.

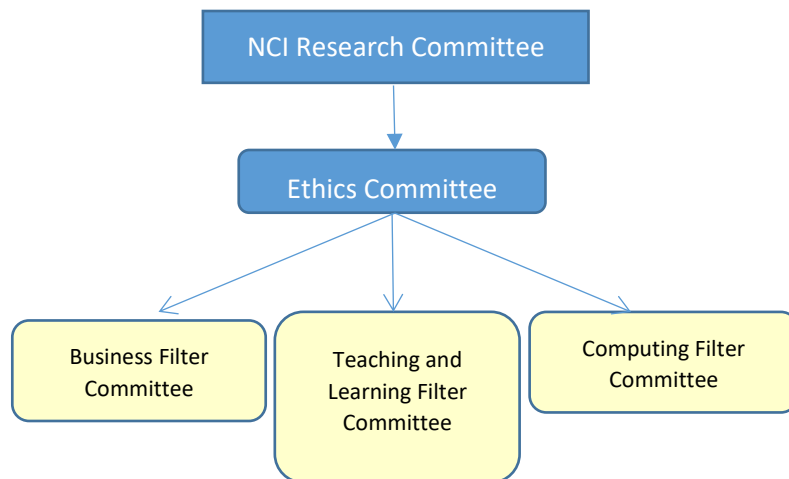


Figure 1: Committee Structures.

4. Review Process

Any staff or student of NCI wishing to conduct a study involving human participants should first submit the Ethical Review Application Form (included at the end of this document), to the relevant School Filter Committee at proposal stage. This initial review will result in a graded categorisation of ethical risk, as outlined below.

4.1 Categorisation of Ethical Risk

Research category A

Research in this category poses little ethical risk to the participants involved. Specifically, it refers to research involving human volunteers, but **excluding** studies involving:

- therapeutic interventions
- new research methodologies
- vulnerable populations (see section 4.2)
- deception of the participants
- any other significant physical, social or psychological risk to participants

Research category B

Research in this category involves human volunteers **including** studies involving:

- therapeutic interventions

- new research methodologies
- vulnerable populations (see section 4.2)
- deception of the participants
- any potentially significant risk to participants

Research Category C

This specifically refers to research involving human volunteers who are service users, patients, staff, records, etc., within the sphere of the HSE or similar setting (but not including clinical trials of investigative medicinal products).

4.2 Vulnerable groups

There are a number of participant populations that may fall under the heading of ‘vulnerable groups’. These groups require consideration of unique ethical challenges regardless of the nature of the project. Research involving such populations should therefore always be reviewed by the Ethics Committee.

Groups that may be classed as vulnerable include, but are not limited to:

- Children (under 18 years of age)
- The older old (aged 85+)
- People with an intellectual or learning disability
- Individuals or groups receiving help through the voluntary sector
- Those in a subordinate position to the researcher (e.g. employees)
- Any other groups who might not understand the research and consent process

Note: in addition to the Ethical Review process, any researchers intending to work directly with children will be required to undergo Garda Vetting in advance of the proposed research.

4.3 Exemption from Full Ethical Review

In certain limited cases, researchers can apply for an exemption from full ethical review. In such cases, the Ethical Review Exemption form should be completed, explicitly detailing why the exemption is sought.

In completing this form, researchers must declare that the research does not involve any of the following:

- Vulnerable groups
- Sensitive topics
- Risk of psychological or mental distress
- Risk of physical stress or discomfort
- Any other risk to participants
- Use of drugs or invasive procedures (e.g. blood sampling)
- Deception or withholding of information from participants
- Conflict of interest issues

- Access to data by individuals or organisations other than the researchers
- Any other ethical dilemmas

4.4 Outcomes of Review Process

Following consideration of research projects submitted for Ethical Review, each Filter Committee will submit a report to the Ethics Committee summarising the applications considered and the decisions made.

For research that is deemed to fall under Research Category A (low ethical risk), a favourable outcome at the relevant Filter Committee will be sufficient to secure ethical approval. Research falling under the other two categories must however be considered by the Ethics Committee before approval may be granted.

On the basis of this review, four key outcomes may arise:

1. Research proposal approved (no recommendations)
2. Research proposal approved pending minor revisions (to be accepted by the Chair and Research Supervisor)
3. Research proposal approved pending major revisions (to be resubmitted and approved by the Ethics Committee)
4. Research proposal rejected (resubmission necessary)

A summary of the processes involved in applying for ethical approval can be seen in Figure 2.

Appeals

Appeals against the Committee's decision may be made within ten working days. In this case, at least three members of the Ethics Committee, none of whom will have reviewed the initial application, may review this along with any additional information submitted by the applicant.

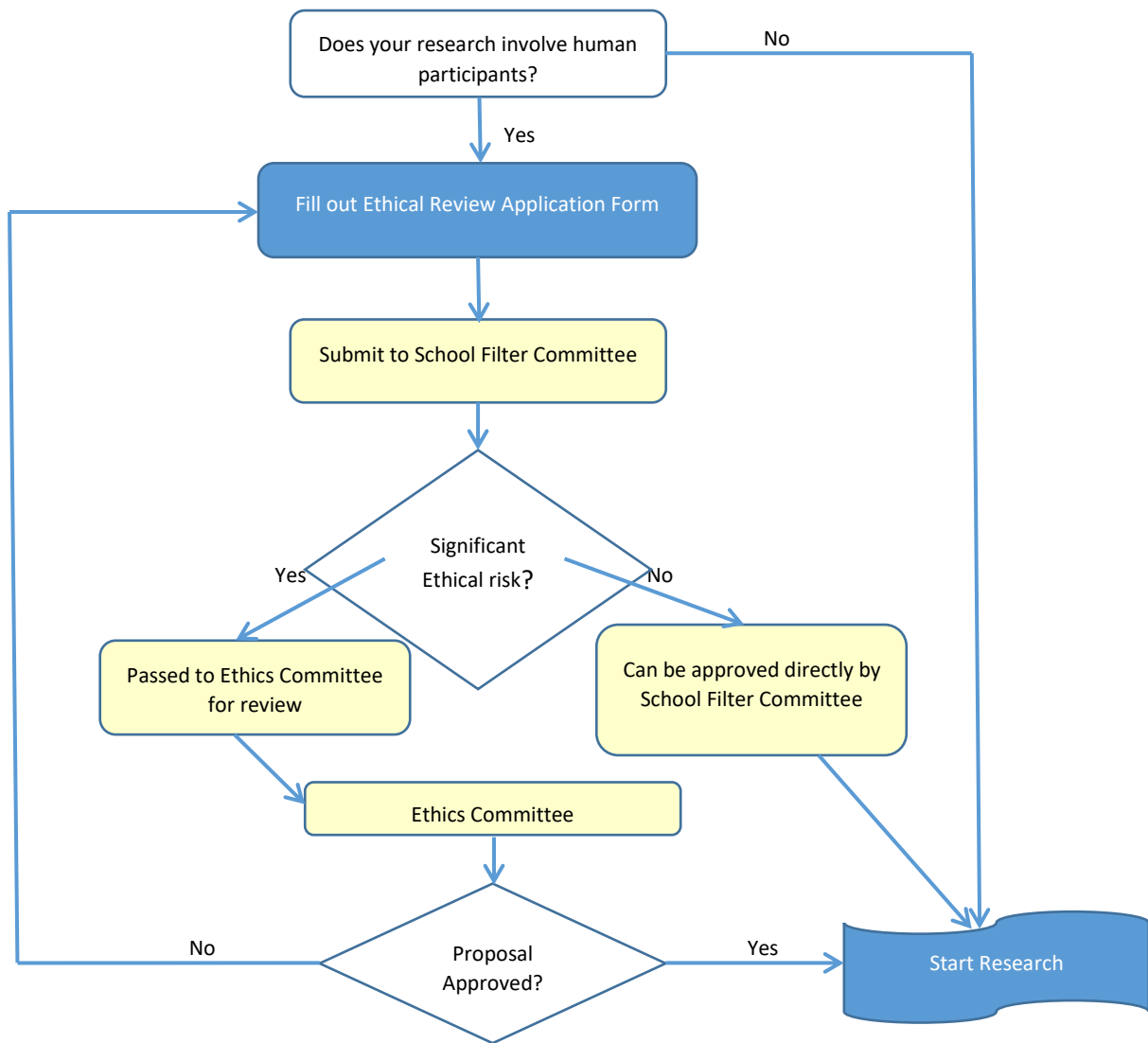


Figure 2: Process chart for seeking Ethical Approval **Ethics Application Checklist**

To be submitted alongside the ethics application.

Please complete the below checklist, ticking each item to confirm that it has been addressed.

1. I agree to obtain informed written consent from all human participants aged over 18 who are involved in this research (or if circulating digitally, I will ensure that informed consent is completed, and will have the participants indicate their informed consent by continuing with their study engagement).	<input type="checkbox"/>
	✓
	<input type="checkbox"/>
2. I agree to obtain informed written consent from the parents of anyone aged under 18 in this research (or from the schools if appropriate), and informed written assent from those under 18 in this research.	✓
3. I include a letter of agreement from a clinically responsible individual agreeing to (where appropriate) help me recruit/provide clinical support in the event that participants become distressed/host the study data collection.	<input type="checkbox"/>
4. I append a letter of agreement from an external institution or organisation agreeing to host the study.	✓
5. I agree to comply with NCI's Data Retention Policy.	<input type="checkbox"/>
6. I have appended a) information sheet, b) consent form/assent form, c) debriefing sheet.	✓
7. I have provided details of how non-anonymised data will be stored, in a safe and encrypted manner.	✓
8. I have included my contact details and those of my supervisor (where appropriate). I have only included my NCI email address and not included any personal contact information.	✓
	<input type="checkbox"/>
9. I have given sufficient details on the proposed study design, methodology, and data collection procedures, to allow a full ethical review, and I understand that my failure to give sufficient detail may result in a resubmission being required.	<input type="checkbox"/>
	<input type="checkbox"/>
10. I understand that if I make changes to my study following ethical approval, it is my responsibility to seek an ethics amendment if the change merits ethical consideration.	✓
	<input type="checkbox"/>
	✓
	<input type="checkbox"/>
	<input type="checkbox"/>

National College of Ireland

Human Participants Ethical Review Application Form

All parts of the below form must be completed. However in certain cases where sections are not relevant to the proposed study, clearly mark NA in the box provided.

Part A: Title of Project and Contact Information

Name

Christopher Weir

Student Number (if applicable)

X19317131

Email

X19317131@student.ncirl.ie

Status:

- Undergraduate
- Postgraduate
- Staff

Supervisor (if applicable)

N/A

Title of Research Project

Statistical Analysis of Europe's Top 5 Football Leagues

Category into which the proposed research falls (see guidelines)

- Research Category A
- Research Category B
- Research Category C

Have you read the NCI Ethical Guidelines for Research with Human Participants?

- Yes
- No

Please indicate any other ethical guidelines or codes of conduct you have consulted

I have read the data privacy policy and use of data policy on FBref

Has this research been submitted to any other research ethics committee?

- Yes
- No

If yes please provide details, and the outcomes of this process, if applicable:

Is this research supported by any form of research funding?

- Yes
No

If yes please provide details, and indicate whether any restrictions exist on the freedom of the researcher to publish the results:

Part B: Research Proposal

Briefly outline the following information (not more than 200 words in any section).

Proposed starting date and duration of project

31/10/2022 – 30/04/2023

The rationale for the project

Interest in football + data analytics in football/sports is becoming ever more prevalent in the business venturing and player performance/recruitment

The research aims and objectives

To identify what attributes define what footballing leagues

To add context to the meaning behind the data for the competitiveness of the leagues

To compare and analyse leagues against each other and give meaning to why the data is the way it is

The research design

Gather data from URL's via web scraping, convert these collected tables into dataframes. Run descriptive statistics on the data, clean the data and transform the data if necessary. Run analytical models on the data and provide visualisations for ease of understanding and provide in depth analysis of what the results mean

The research sample and sample size

Please indicate the sample size and your justification of this sample size. Describe the age range of participants, and whether they belong to medical groups (those currently receiving medical

treatment, those not in remission from previous medical treatment, those recruited because of a previous medical condition, healthy controls recruited for a medical study) or clinical groups (those undergoing non-medical treatment such as counselling, psychoanalysis, in treatment centres, rehabilitation centres, or similar, or those with a DSM disorder diagnosis).

98 men footballing teams with an average age of 18-38

If the study involves a MEDICAL or CLINICAL group, the following details are required:

- a) Do you have approval from a hospital/medical/specialist ethics committee? If YES, please append the letter of approval. Also required is a letter from a clinically responsible authority at the host institution, supporting the study, detailing the support mechanisms in place for individuals who may become distressed as a result of participating in the study, and the potential risk to participants.
If NO, please detail why this approval cannot or has not been sought.
- b) Does the study impact on participant's medical condition, wellbeing, or health? If YES, please append a letter of approval from a specialist ethics committee. If NO, please give a detailed explanation about why you do not expect there to be an impact on medical condition, wellbeing, or health.

The nature of any proposed pilot study. Pilot studies are usually required if a) a new intervention is being used, b) a new questionnaire, scale or item is being used, or c) established interventions or questionnaires, scales or items are being used on a new population. If no such study is planned, explain why it is not necessary.

N/A

The methods of data analysis. Give details here of the analytic process (e.g. the statistical procedures planned if quantitative, and the approach taken if qualitative. It is not sufficient to name the software to be used).

N/A

Study Procedure

Please give as detailed an account as possible of a participant's likely experience in engaging with the study, from point of first learning about the study, to study completion. State how long project participation is likely to take, and whether participants will be offered breaks. Please attach all questionnaires, interview schedules, scales, surveys, and demographic questions, etc. in the Appendix.

N/A

Part C: Ethical Risk

Please identify any ethical issues or risks of harm or distress which may arise during the proposed research, and how you will address this risk. Here you need to consider the potential for physical risk, social risk (i.e. loss of social status, privacy, or reputation), outside of that expected in everyday life, and whether the participant is likely to feel distress as a result of taking part in the study. Debriefing sheets must be included in the appendix if required. These should detail the participant's right to withdraw from the study, the statutory limits upon confidentiality, and the obligations of the researcher in relation to Freedom of Information legislation. Debriefing sheets should also include details of helplines and avenues for receiving support in the event that participants become distressed as a result of their involvement in this study.

The only risk was attaining the data ethically which I have received permission from FBref to use their data in compliance with their Use of Data policy

Do the participants belong to any of the following vulnerable groups? (Please tick all those involved).

- Children;
- The older old (85+)
- People with an intellectual or learning disability
- Individuals or groups receiving help through the voluntary sector
- Those in a subordinate position to the researchers such as employees
- Other groups who might not understand the research and consent process
- Other vulnerable groups

How will the research participants in this study be selected, approached and recruited? From where will participants be recruited? If recruiting via an institution or organisation other than NCI please attach a letter of agreement from the host institution agreeing to host the study and circulate recruitment advertisements/email etc.

Use of Data



Sports Reference <bugs_3@sports-reference.com>

21/10/2022 21:12



To: Christopher McHugh

Hi Christopher,

Thank you for your message. We are happy for you to use the data on the site in your project, provided you cite us as the source of the information (as you say you will do). It sounds like you have read the terms of use, but I will link here just in case:

https://www.sports-reference.com/data_use.html

Best of luck with your project.

Best,
Aidan

My name is Christopher Weir and I am a data science student in my final year of college. I am emailing to ask Fbref for use of their data from the top 5 european leagues for the 2021-22 season just gone. I am asking to use this data to analyse, visualise and compare the leagues statistics and to draw up conclusions on the leagues and how each league style of football differs depending on the league. This use of data will not be used to turn profit or to create an external tool. It is solely for analytical purposes for my final year research paper and I will ensure Fbref will be referenced correctly and with full regards for the data that I will use. Thank you. Kind regards, Chris

What inclusion or exclusion criteria will be used?

N/A

How will participants be informed of the nature of the study and participation?

Via email

Does the study involve deception or the withholding of information? If so, provide justification for this decision.

No

What procedures will be used to document the participants' consent to participate?

Email receipt (posted above)

Can study participants withdraw at any time without penalty? If so, how will this be communicated to participants?

N/A

If vulnerable groups are participating, what special arrangements will be made to deal with issues of informed consent/assent?

N/A

Please include copies of any information letters, debriefing sheets, and consent forms with the application.

Part D: Confidentiality and Data Protection

Please indicate the form in which the data will be collected.

- Identified Potentially Identifiable ✓ De-Identified

What arrangements are in place to ensure that the identity of participants is protected?

N/A (public data)

Will any information about illegal behaviours be collected as part of the research process? If so, detail your consideration of how this information will be treated.

No

Please indicate any recording devices being used to collect data (e.g. audio/video).

Web scraping FBref via coding with pandas

Please describe the procedures for securing specific permission for the use of these recording devices in advance.

Sent an email to FBref asking for permission and permission has been granted

Please indicate the form in which the data will be stored.

- Identified Potentially Identifiable ✓ De-Identified

Who will have responsibility for the data generated by the research?

FBref have responsibility for the data as it is public and if they remove such links/URL's then I can no longer access the data

Is there a possibility that the data will be archived for secondary data analysis? If so, has this been included in the informed consent process? Also include information on how and where the data will be stored for secondary analytic purposes.

No

If not to be stored for secondary data analysis, will the data be stored for 5 years and then destroyed, in accordance with NCI policy?

Yes ✓

No

Dissemination and Reporting

Please describe how the participants will be informed of dissemination and reporting (e.g. submission for examination, reporting, publications, presentations)?

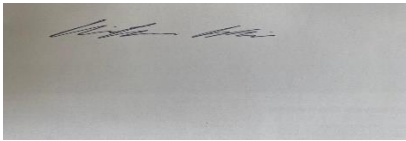
N/A

If any dissemination entails the use of audio, video and/or photographic records (including direct quotes), please describe how participants will be informed of this in advance.

N/A

Part E: Signed Declaration

I confirm that I have read the NCI Ethical Guidelines for Research with Human Participants, and agree to abide by them in conducting this research. I also confirm that the information provided on this form is correct (Electronic signature is acceptable).

Signature of Applicant _____ 

Date _____ 24/10/2022

Signature of Supervisor (where appropriate): N/A

Date _____ 24/10/2022

Any other information the committee should be aware of?

N/A

10.3. Reflective Journals

Supervision & Reflection Template	
Student Name	Christopher Weir
Student Number	X19317131
Course	BSHDS4
Supervisor	Michael Bradford

Month: October

What?

Reflect on what has happened in your project this month?

This month, I have faced a few challenges. My original pitched idea does not seem to be feasible with the available data there is right now. My idea was on housing, however, due to no permission to use company data, no permission to scrape website data, and the lack of comprehensive datasets available publicly, I have had to change my idea to one which could be done through more feasible datasets. This has since been changed to footballing datasets and I have a strong and comprehensive idea for a project that I feel, with guidance from my supervisor can result in high marks.

During this month, I have since received permission to scrape the footballing website for data, I have gathered the necessary tables I will need, I have done some exploratory analysis and some data cleaning and pre-processing. Once I have a meeting with my supervisor, I can gain more clarity on a more concise project to get moving forward.

So What?

Consider what that meant for your project progress. What were your successes? What challenges remain?

My successes were gathering a new dataset, receiving permission to use the dataset, submit all my forms and proposals on time and to outline a plan for my project.

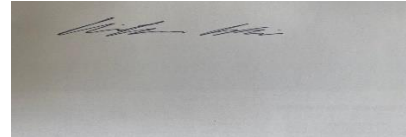
There are a lot of challenges that remain on how to merge my datasets together, how to run my PCA, what models to build and how to perfect them and how to visualise, run a model behind the visualisation etc.

Now What?

What can you do to address outstanding challenges?

I will have a meeting with my supervisor, I will do more research on similar projects to see what approaches were taken, I will address what way to construct my dataset, how to run my PCA and get all of my exploratory analysis done.

Student Signature



Supervision & Reflection Template	
Student Name	Christopher Weir
Student Number	X19317131
Course	BSHDS4
Supervisor	Michael Bradford

Month: November**What?**

Reflect on what has happened in your project this month?

This month, I have faced a few challenges. Since gathering my data from FBref, I have since moved on to fixing some functions to deal with multiple different cases when it comes to loading the data. I have created a new function which allows me to divide standardise all the statistics for a more equal and balanced dataset for more accurate models. Since building this dataframe, I have run code for K-Means Clustering, have identified the clusters, built an elbow method model to back up my use for the number of clusters and have run code to run a PCA on the data. I have built graphs to plot the clusters, built a hierarchy graph, and an agglomerative graph and the cluster returns are accurate and grouped, and have returned some results from the clusters.

So What?

Consider what that meant for your project progress. What were your successes? What challenges remain?

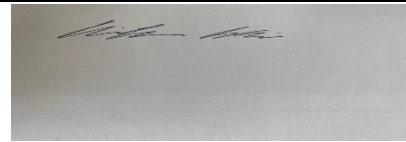
My successes were in terms of being able to define different styles of play and group the teams into a specific style of play. Returns results of the top n teams in each cluster when called. Challenges remain around the size of the data for training models. May need to investigate a solution for this as discussed with my supervisor which has been done. There are some challenges around getting the data. Challenges also remain around some test functions to be built to show the code I have built is accurately working for what I need it to do. The modelling needs to be built, and there is some background work around tidying up folders etc. to be done.

Now What?

What can you do to address outstanding challenges?

I will have a meeting with my supervisor, I will look at alternative ways to get more data if possible, I will outline and create folders for different steps of the CRISP-DM methodology and I will build some test functions to show the code I am running is working successfully to do something necessary and meaningful.

Student Signature



Supervision & Reflection Template	
Student Name	Christopher Weir
Student Number	X19317131
Course	BSHDS4
Supervisor	Michael Bradford

Month: December

What?

Reflect on what has happened in your project this month?

This month, I have faced a few challenges. I have fixed the functions that have been causing me issues this month. I have successfully standardized the data which has been saved to a csv file. Since building this DataFrame, I have run code for K-Means Clustering, have identified the clusters, built an elbow method model to back up my use for the number of clusters and have run code to run a PCA on the data. I have built graphs to plot the clusters, built a hierarchy graph, and an agglomerative graph and the cluster returns are accurate and grouped, and have returned some results from the clusters. I have also started writing up the report and have made my video and my slides for mid-point submission

So What?

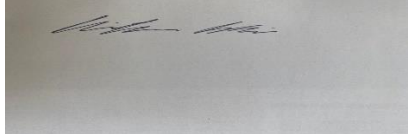
Consider what that meant for your project progress. What were your successes? What challenges remain?

My successes were in terms of being able to define different styles of play and group the teams into a specific style of play. Returns results of the top n teams in each cluster when called. I have also gotten my mid-point submission in, my videos, and my slides done. Challenges remain around the size of the data for training models. I may export the player data to sql for easier cleaning and call it back into Jupyter. Challenges also remain around some test functions to be built to show the code I have built is accurately working for what I need it to do. The modelling needs to be built, and there is some background work around tidying up folders etc. to be done.

Now What?

What can you do to address outstanding challenges?

I will try in future to clean the data in SQL as it is easier to do seemingly for player data. I will outline and create folders for different steps of the CRISP-DM methodology, and I will build some test functions to show the code I am running is working successfully to do something necessary and meaningful.

Student Signature	
--------------------------	--

Supervision & Reflection Template

Student Name	Christopher Weir
Student Number	X19317131
Course	BSHDS4
Supervisor	Michael Bradford

Month: January

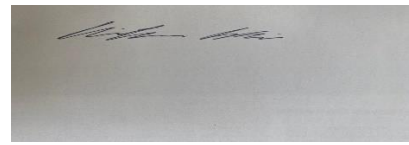
<p>What?</p> <p>Reflect on what has happened in your project this month?</p> <p>This month, I have faced a few challenges. Since gathering my data from FBref, I have fixed functions and have created new visualisations. I have learned more about the data and I have since moved on to looking for additional data to add to the dataset. From here I can then start to build models in the form of Neural Networks and more in detail of Self Organised Maps architecture.</p>
<p>So What?</p> <p>Consider what that meant for your project progress. What were your successes? What challenges remain?</p> <p>My successes were in terms of being able to fix all functions necessary so they could successfully pull in the data without any issues. Challenges that remain are that now of gathering additional data along with starting to build models based on the data along with expanding out the report.</p>

Now What?

What can you do to address outstanding challenges?

I will have a meeting with my supervisor and I will get in touch with FBref, the data provider in terms of gathering additional data. My supervisor and myself will then discuss the implementation of the models and the structure of the report.

Student Signature



Supervision & Reflection Template

Student Name	Christopher Weir
Student Number	X19317131
Course	BSHDS4
Supervisor	Michael Bradford

Month: February

What?

Reflect on what has happened in your project this month?

This month, I have faced a few challenges. I have gotten in touch with FBref and I am awaiting their response. I have a clear outline of the structure of the model to build and I

am starting to implement on build that model now. I have also started writing up and structuring more of the report.

So What?

Consider what that meant for your project progress. What were your successes? What challenges remain?

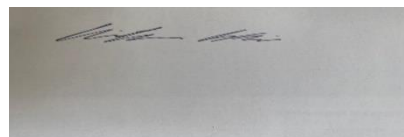
My successes were in terms of being able to identify the model I will need to build and how to implement it with a clear goal in mind. Challenges remain around the size of the data for training models. I am awaiting FBref's response on if they have the data and can provide it to me.

Now What?

What can you do to address outstanding challenges?

I will begin to implement a self-organised map and once FBref get back to me I will have a clearer idea of what I need to do and how feasible attaining this additional data will be.

Student Signature



Supervision & Reflection Template

Student Name	Christopher Weir
Student Number	X19317131
Course	BSHDS4
Supervisor	Michael Bradford

Month: March

What?

Reflect on what has happened in your project this month?

This month, I have dived further into the report and the modelling. This month I have written more into the report to bring the report up to date to where I currently am in terms of coding. I have started to build models in the form of Neural Networks and have looked deeper into GAN Neural Network structures.

So What?

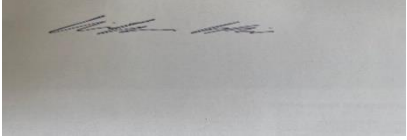
Consider what that meant for your project progress. What were your successes? What challenges remain?

My successes were in terms of bringing my report up to date. Explaining more in detail the project in multiple different sections. Challenges that remain are that now of looking more into the modelling section of the project and building on these models.

Now What?

What can you do to address outstanding challenges?

I will have a meeting with my supervisor to discuss the next steps in terms of the modelling. Get guidance and tips and hopefully will be able to build a complete and functioning model.

Student Signature	

Supervision & Reflection Template

Student Name	Christopher Weir
Student Number	X19317131
Course	BSHDS4
Supervisor	Michael Bradford

Month: April

<p>What?</p> <p>Reflect on what has happened in your project this month?</p> <p>This month, I have written up a lot of the report, I have modelled the data and have interpreted the results. I am now beginning to think up dashboarding ideas to finalise the coding repos. I have also started to clean up the code and a starting to map out the video presentation and what must be done for it. I have done the showcase profile and am starting to work on the showcase poster.</p>
<p>So What?</p> <p>Consider what that meant for your project progress. What were your successes? What challenges remain?</p> <p>My successes were in terms of bringing my report up to date. Along with building a successful model which can be correctly interpreted. This was the main piece of this project, so I am happy to get it done.</p> <p>Challenges remain on the dashboarding side of things, how this will be done. There are challenges on the showcase poster and the video presentation but these are more trivial issues that will be smoothed out easily.</p>

Now What?

What can you do to address outstanding challenges?

I will have a meeting with my supervisor to discuss the next steps in terms of the dashboarding to discuss the way to go about this. I will begin to design the poster myself and map out the way I will want the video to look and will start to address questions as well.

Student Signature

