

National College of Ireland

BSc. (Honours) Data Science

2022/2023

Thomas Cannon

x19405504

x19405504@student.ncirl.ie

Beyond the Eye Test: Improving Football
Recruitment Through The Use Of Clustering
And Support Vector Machines
Data Science Report

Contents

Executive Summary	3
1.0 Introduction	3
1.1. Background	3
1.2. Aims.....	4
1.3. Technology.....	5
1.4. Structure	6
2.0 State of the Art.....	7
3.0 Data	10
3.1 Player Performance Data	10
3.2 Player Position Data	11
4.0 Methodology.....	12
4.1 Selection.....	13
4.2 Cleaning and Pre-processing.....	14
4.3 Transformation	14
4.4 Data Mining / Information Retrieval.....	14
5.0 Analysis	18
5.1 Selection.....	18
5.2 Cleaning and Pre-processing.....	19
5.3 Transformation	20
5.4 Data Mining / Information Retrieval.....	21
6.0 Results	29
7.0 Conclusions	37
8.0 Further Development or Research	38
9.0 References	39
10.0 Appendices.....	40
10.1. Project Proposal	40
11.0 Objectives.....	42
12.0 Background	42
13.0 State of the Art.....	43
14.0 Data	43
15.0 Methodology & Analysis	44
16.0 Technical Details	46
17.0 Project Plan	46
18.0 References	48
18.1. Ethics Approval Application (only if required)	48

Appendix II	48
18.2. Reflective Journals	49
21.1. Other materials used	53
Improving Football Recruitment Through The Use Of Clustering And Support Vector Machines	53

Executive Summary

The objective for this project is to use data retrieved from a 3rd party source for analysis to try to make predictions on football players positions from the world's top football leagues, while also seeking to identify similarities between players through the use of various clustering and classification methods on past season's performance data. This report will detail a number of machine learning methods and technologies which were analysed to identify the most accurate method of predicting football players positions as well as identifying their similarities. Data from FBREF relating to the most recent season of football from the top 11 leagues around the world were combined to create a dataset which contains over 100 datapoints pertaining to each individual players performance throughout the season.

It is hoped that learnings from this project could then be used in the form of an application which could be leveraged to identify talent from other leagues or even applied to data from lower-level leagues to identify lesser-known players. This system could be used by professional football clubs and scouts to aid in the process of scouting and analysing transfer prospects.

Methods such as K Means clustering, K-Nearest Neighbour (KNN), and Support Vector Machines (SVM) are among some of the methods which were tested throughout the course of this project. Transfer learning through the use of a Support Vector Machine was applied to the raw data to create a high-level vector space in the hopes of achieving more accurate results.

The highest accuracies achieved by the various models created are as follows:

- Players Position Category through SVM: 88.3% accuracy
- Players Sub-Position through SVM: 74.84% accuracy

A more accurate model was deemed to be found to identify similar players across the leagues through the use of techniques used which was verified by a questionnaire posed to participants with an interest in football to validate the results.

Results of a T-test to identify if there was a significant difference between the choices of the questionnaire participants returned results showing that the participants asked leaned towards the output of the proposed improved model.

1.0 Introduction

1.1. Background

Statistics and data are tools which are used constantly by most, if not all large and medium sized organizations. These organizations leverage data on a daily basis to make important business decisions. Sales strategies, marketing campaigns, user experience, services and product improvements are examples of areas which companies leverage data to make decisions in these areas.

These companies have spent thousands and sometimes even millions to create infrastructures to store, handle, and gain insights from data they collect.

However, with the growth of money spent by sports teams, it may come as a surprise to many people that the use of data in sport is only a relatively new addition and still in its infancy.

It is widely regarded that Billy Beane, who was General Manager of the American Baseball team Oakland Athletics in 2002, was the first to leverage the analysis of players previous performance data in the recruitment process for a sports team. Billy established data metrics in which he would base the recruitment of players on which would be later referred to as 'Moneyball' which had a movie of the same name created. and retrieved huge backlash from baseball fans and experts because he would not sign big names for the team and instead sign players who had performed at a high level on underlying data metrics which usually went unnoticed. [1]

Data in sport has progressed from this time and football teams have started to create large data teams which are used in their recruitment and on-field strategies. Brighton & Hove Albion FC and Brentford FC are both very interesting examples of the success the use of data can bring to football clubs. Tony Bloom, the current majority share owner of Brighton, and Matthew Benham, the owner of Brentford both come from sports analytical backgrounds as they have both previously worked for and owned gambling companies. Both men took ownership of their respective clubs at a time where both clubs were in the lower levels of the English footballing leagues and implemented analytical scouting systems to their clubs recruitment systems which is deemed to be what has since got them promoted to the top level of English football and challenging with the some of the elite clubs in the world.[2]

With my passion for data and football I decided to combine the two and to work on a project to create a tool which is used to profile similar players based on previous performances.

1.2. Aims

The aim for this project is to analyse data which was retrieved data from a 3rd party source to try to make predictions on football player's positions from the top footballing leagues. This project also intended to leverage the models created to try to identify similar players based on the performance data. A number of machine learning methods and technologies were analysed throughout the course of this project with the objective of identifying the most accurate method of predicting positions as well as the similarities. Data from FBREF [3] containing performance data points pertaining to matches played during the most recent season of football from the top 11 leagues around the world to which was combined to a large dataset which contains over 100 statistics pertaining to each individual players performance. Another dataset was a collection of player positions which were listed on Transfermarkt [4], another 3rd party football data resource. This was combined with the performance dataset and used in

various classification and clustering models with the aim of making accurate predictions on players positions.



Figure 1: Representation of the Various Position Data in the Dataset

These models could then be used in order to similar players across the worlds top leagues. It is hoped that learnings from this project could then be applied in the form of an application which could be leveraged to identify talent from other leagues or even applied to data from lower-level leagues to identify lesser-known players. This system could be used by professional football clubs and scouts to aid in the process of scouting and analysing transfer prospects.

1.3. Technology

This project makes use primarily of Python code using built in, as well as external libraries for data manipulation, visualisation, and machine learning algorithm modelling, as well as a dashboarding package for creating the final interactive dashboard:

- Pandas [5]: For storage and manipulation of tabular data. The data used for this project is in tabular form and therefore needed the use of an external Python library that could store, visualise, and manipulate the data so that it could be applied to the machine learning models that would be built.
- Numpy [6]: For working with arrays and use in the computation of linear algebra. Many tasks throughout the course of the project required some of the tabular data to be converted to arrays to be manipulated and fed to the machine learning models. Numpy also has some functions that can be used for calculations that were calculated for some tasks.
- Seaborn [7], Matplotlib [8], Plotly [9]: To create visualisations of the data. Data visualisation is one of the most important components to presenting data to a reader. It was imperative to this project that libraries were used to create

informative plots and visualisations to demonstrate and present key aspects and features of the data.

- Scikit-learn [10]: For Machine Learning modelling and computations. Sklearn is a comprehensive machine learning Python library with a host of built in functions and methods to help to streamline the design and creation of machine learning models so that more focus can be placed on data cleaning and feature engineering.
- Requests [11], BeautifulSoup [12]: For web-scraping needed to obtain data. Requests and BeautifulSoup are web scraping libraries which were used hand in hand throughout the data collection process. These libraries provide the tools to read HTML code from a given web page and to analyse the code to scrape relevant data from a given page.
- Streamlit [13]: Used for dashboarding. Streamlit is a simple out of the box dashboarding library for use in Python which was used to create the final interactive dashboard to display the final similar players model created throughout the course of the project.

The project also makes use of a host of machine learning algorithms and statistical tests to come to the conclusions that can be seen in the later sections. Clustering models such as K-Nearest Neighbour (KNN),

1.4. Structure

This section will outline and detail each section of the report and provide a summary of what is discussed in each section related to the project.

- Background: This section relates to the background of the topic in context to this project. It details the importance of the project in relation to the topic and the reasoning for the choice of the project.
- Aims: This section outlines the outcomes and use cases which the project hopes to achieve. It also aligns with the rationale of the project and asserts the importance of the project to the subject.
- Technology: This section details the technologies which are used throughout the course of the project to complete the analysis that is described in the report.
- State of the Art: This section provides references to similar work carried out in the related fields and describes papers which were useful to read and understand in hopes of replicating similar work in hopes to provide accurate and researched methods for the task.
- Data: This section describes the data which was used to perform the analysis. The datasets which were used for the project are investigated and presented to the reader to provide a coherent understanding of the data which plays a massive role in the project.
- Methodology: This section provides a detailed description of the methodology that was planned and followed throughout the course of the project and describes the various steps in order of completion to arrive at the results which are presented at the end of the section.

- **Analysis:** This section details the analysis which was performed throughout the course of the project. This will provide a detailed description of the approaches which were employed throughout the work which was undertaken. The rationales will be given for each choice that was made throughout the analysis and any justifications will be stated for choices on which models were chosen and which attributes were used.
- **Results:** This section will provide a detail summary of the results gained from the work undertaken throughout the project. Tables and visualisations will be used to concisely explain the results obtained to the reader. Some necessary explanations will also be provided in this section which will be expanded on more broadly in the conclusions section.
- **Conclusions:** This section presents the key pieces of information that were found throughout the course of the analysis and provide possible reasoning for these findings. The conclusions provide a clearer understanding of the project and explain to the reader what has been achieved from the work completed throughout the project.
- **Further Development or Research:** This section provides the reader with possible steps that could be taken using the information discovered as a result of the analysis and to detail what future analysis or research could be undertaken to further the development of the subject matter.

2.0 State of the Art

Talent identification has played a role in elite sports since it's beginning. Teams seek to gain an advantage on the pitch by having the best player in every position. With the amount of money spent in the football world, it is becoming increasingly important for all elite clubs to be able to identify and recruit talented players who possess the desired skills and attributes of a top player to drive success on and off the field. At the top level of footballs elite leagues, accurate player identification and recruitment could be the slim margin between creating a successful team or creating an unsuccessful team. Player identification allows clubs to build and maintain competitive squads in order to compete at the highest levels.

With this in mind clubs are starting to an increasing amount of attention towards the creation and implementation of data analytics teams in their recruitment processes. Similar player identification in sports data has gained significant attention in recent years, as it plays a crucial role in player scouting, talent identification, and team performance analysis. The ability to accurately identify players who exhibit similar playing styles, skill sets, or positional characteristics can provide valuable insights for teams and coaches in various sports.

A number of studies have been conducted to explore different machine learning algorithms and techniques to improve the accuracy and efficiency of similar player identification. These studies have contributed to the state of the art in this field, offering innovative approaches and insights into the challenges and opportunities of player identification in sports data.

A study conducted by Yuesen Li et al titled, “Characterizing player’s playing styles based on Player Vectors for each playing position in the Chinese Football Super League” [14] which was published in the Journal of Sports Sciences in 2021 analyses problems similar to what is being undertaken during this project. The team centred their focus on trying to characterize the playing style of professional football players in the Chinese Super League. In this paper, the team describe their use of Non-Negative Matrix Factorization (NMF), which is a dimensionality reduction technique, to plot their player data to a vector space where they were able to discover 18 different aspects of playing styles which were present with the players in their dataset. The team then proposed the use of Manhattan Distance in order to calculate the most similar players in their newly created vector space. It was decided for the purpose of this project, to use Euclidean Distance instead of the Manhattan Distance method in order to calculate the most similar players in the newly created vector spaces. This was because it was found through extensive research that Euclidean Distance would be more accurate as it calculates based on a straight line between points rather than the closest unit distance in the case of the Manhattan Distance method. This was a very interesting paper to read during the research process for this project as it detailed a possible approach that could be taken towards the identification of similar players. The one problem which this analysis suffers from is that there are no mathematical techniques used in order to evaluate the performance of their similar player identification calculation. For this reason, it was decided that this project would test ways of introducing a metric upon which the various calculation methods produced by the project can be evaluated.

An article written by Mark Carey titled “Assessing midfielder similarities using machine learning” [15] details the analysis conducted by the author to assess the similarities between midfield players in the Top 5 European Leagues using machine learning techniques. The author used data provided by FBref which is similar data to what is being used for the purpose of this project. The author states their methodology to create a tool that can identify stylistic similarities between players. The author states the tool could be used by clubs in the transfer market to help them identify players who play similarly to their targets but may be more affordable. The analysis employed a cluster analysis to classify players into separate groups based on multiple variables. The author makes use of the K-Nearest Neighbour clustering algorithm and Euclidean Distance was performed on a created vector space during the analysis. The output of the analysis was a proximity matrix that shows the similarity between players. Lower numbers indicate higher similarity, while higher numbers indicate less similarity. This article highlights that a simple approach can generate a list of similar players quickly. The output of the analysis, combined with footballing expertise, can assist recruitment analysts in filtering and shortlisting potential signings. Reading this article was important for the purpose of this project as it served as a basic guideline on how a clustering algorithm could be applied to a dataset containing football players performance data which can be used to ultimately identify most similar players in a vector space. This project aimed to build on the techniques discussed in hopes of applying the analysis to every position on the football pitch while also seeking to improve on the accuracy of the methods used through the creation of an evaluation metric which was not touched upon in this article.

A paper by Garcia-Aliaga et al titled “In-game behaviour analysis of football players using machine learning techniques based on player statistics” [16] describes the approach taken by the team to tackle the problem of analysing football players in-game behaviours based on machine learning methods. The team made use of data comprising of ‘OPTA’s on-ball event records of the matches for 18 national leagues between the 2012 and 2019 seasons.’, in order to test whether a player’s position could be identified from their statistical performance. The team made use of dimensionality reduction techniques along with classification machine learning models in order to make classifications on a player’s playing position based on their in-game performance data. The team made use of U-Mapping as a dimensionality reduction technique and made use of the RIPPER machine learning method to classify their data. The team states that they retrieved accuracy scores of 97% when seeking to classify the data based on their positions. It was interesting to read this paper when researching for this project as it was clear to see the team had achieved extremely high accuracy scores from the classification model which they had created. It was decided to build upon the foundations which were laid out by this project to reduce the dimensions of the data and to use a classification model to make predictions on positions, however due to the nature of this project seeking to also group players based on their similarities, clustering methods such as K-Nearest Neighbour and other methods such as Support Vector Machines were preferred.

An article by Jiayi Fan et al titled “A Transfer Learning Architecture Based on a Support Vector Machine for Histopathology Image Classification” [17] describes the methods in which the team made use of the Support Vector Machine classification machine learning model to create a transfer learning architecture to map their dataset into a high-level vector space. The team made use of this architecture in the hopes of creating an accurate classification model for a Histopathology Image Classification problem which they were working on. The team attached their SVM classifier model to the fully-connected layer of their softmax-based transfer learning model. The team references results they achieved which were higher than the results achieved by the softmax-based approach and the Support Vector Machine models on their own which were tested in preliminary sections of the analysis. This paper was interesting to the current research as it showed ways in which researchers had already leveraged transfer learning in order to map their data to a high-level vector space in order to retrieve better results from the models they had created. Although it could not be established during research for this project that any work existed in the field of sports data analysis relating to the use of transfer learning, it could be seen from this paper and other similar work, that the use of transfer learning can help to improve the accuracies of models which are produced throughout an analysis. Therefore, it was decided to test the theory in this project in hoping to achieve a similar boost in results using the transfer learning methodology.

These papers represent a sample of the diverse research efforts dedicated to similar player identification in sports data which were identified throughout the process of research for the current problem. They highlight the utilization of similar player identification techniques, transfer learning, clustering based approaches, classification based approaches. By drawing inspiration from these papers and building upon their findings, this project

contributes to the state of the art by exploring the performance of various machine learning algorithms for similar player identification in football data. The analysis and experimentation conducted in this project hopes provide valuable insights and potential advancements in the field of player identification, aiding teams, scouts, and analysts in making data-driven decisions in the realm of football.

3.0 Data

Two datasets were necessary for the objective of this project.

1. A dataset which contains football player's performance data from the top 11 footballing leagues which were gathered during the most recently completed season.
2. A dataset containing players positions which matched the players included in the performance dataset.

3.1 Player Performance Data

The dataset which contains the Player Performance data can be found at fbref.com [3], a 3rd party football data website which hosts a collection of data for a range of various leagues around the world. This data is open source and can be scraped using web scraping tools should permission be requested and granted by the company. With the help of web scraping packages such as **Requests** and **Beautiful Soup**, Python code was written to scrape and stored data for the world's top 11 footballing leagues.

This dataset contains data relating to players from:

- Premier League (England)
- La Liga (Spain)
- Serie A (Italy)
- Bundesliga (Germany)
- Ligue 1 (France),
- Major League Soccer (USA)
- Championship (England)
- Brasileiro Série A (Brazil)
- Eredivisie (Netherlands)
- Liga MX (Mexico)
- Primeira Liga (Portugal)

This dataset contains 3308 rows and 99 columns. There are 6 identifier columns relating to each player:

- Player: First and last name.
- Nationality: Player's declared nationality.
- Team: Club which player played for during the 21/22 season.
- Position: Player's position area (GK, DF, MF, FW).
- Age: Player's age during the season.
- Comp Level: League which player played in during the season.

A list of all statistical 93 column names is:

```
'Games, Games Starts, Minutes, Minutes 90S, Goals_Per90, Assists_Per90, Goals Pens_Per90, Pens Made_Per90, Pens Att_Per90, Cards Yellow_Per90, Cards Red_Per90, Xg_Per90, Npxg_Per90, Passes Completed_Per90, Passes_Per90, Passes Pct_Per90, Passes Total Distance_Per90, Passes Progressive Distance_Per90, Passes Completed Short_Per90, Passes Short_Per90, Passes Pct Short_Per90, Passes Completed Medium_Per90, Passes Completed Long_Per90, Passes Long_Per90, Passes Pct Long_Per90, Assisted Shots_Per90, Passes Into Final Third_Per90, Passes Into Penalty Area_Per90, Crosses Into Penalty Area_Per90, Progressive Passes_Per90, Shots On Target_Per90, Shots On Target Pct_Per90, Goals Per Shot_Per90, Goals Per Shot On Target_Per90, Average Shot Distance_Per90, Shots Free Kicks_Per90, Npxg Per Shot_Per90, Xg Net_Per90, Npxg Net_Per90, Passes Live_Per90, Passes Dead_Per90, Passes Free Kicks_Per90, Through Balls_Per90, Passes Switches_Per90, Crosses_Per90, Corner Kicks Straight_Per90, Throw Ins_Per90, Passes Offsides_Per90, Passes Blocked_Per90, Sca_Per90, Sca Passes Live_Per90, Sca Passes Dead_Per90, Sca Dribbles_Per90, Sca Shots_Per90, Sca Fouled_Per90, Sca Defense_Per90, Gca_Per90, Gca Passes Live_Per90, Gca Passes Dead_Per90, Gca Dribbles_Per90, Gca Shots_Per90, Gca Fouled_Per90, Gca Defense_Per90, Tackles_Per90, Tackles Won_Per90, Tackles Def 3Rd_Per90, Tackles Mid 3Rd_Per90, Tackles Att 3Rd_Per90, Dribble Tackles_Per90, Dribbles Vs_Per90, Dribble Tackles Pct_Per90, Dribbled Past_Per90, Blocks_Per90, Blocked Shots_Per90, Blocked Passes_Per90, Interceptions_Per90, Tackles Interceptions_Per90, Clearances_Per90, Errors_Per90, Touches_Per90, Touches Def Pen Area_Per90, Touches Def 3Rd_Per90, Touches Mid 3Rd_Per90, Touches Att 3Rd_Per90, Touches Att Pen Area_Per90, Touches Live Ball_Per90, Dribbles Completed_Per90, Dribbles_Per90, Dribbles Completed Pct_Per90, Miscontrols_Per90, Dispossessed_Per90, Passes Received_Per90, Progressive Passes Received_Per90, pretty_name'
```

Figure 2: List of All Column Names Included in Player Performance Dataset

The statistical columns have been computationally standardized to be per 90 minutes using the 'Minutes 90S' column. This means that the performance statistics will represent an average value the player returned for each statistic per 90 minutes, the length of a regular adult football match. This is done to streamline the data so that a player who has played more minutes in that season will not have inflated performance data when compared to a player who has lesser minutes played.

Permission was received for the use of this data from FBREF as per the email which is linked in the ethics report. The email states that the project has permission for the use of the data from the website as long as there is sufficient citation and credit given to the website for the data and that the project is used solely for academic purposes.

3.2 Player Position Data

The dataset containing Player Position data can be found at transfermarkt.com. This data was not scraped from the Transfermarkt [4] website as the dataset had already been scraped by a twitter user by the name of Jase_Ziv83 which he kindly posted to his GitHub repository [18] and allowed me to make use of it for the purpose of this project. Permission was requested from the Transfermarkt team who kindly allowed me to use the data as long as it did not require scraping of the website. This permission was given in the form of email which can be found linked in the ethics form.

The data contains 25,625 rows relating to various different players and contains 23 columns referencing players names and positions as well as a host of Transfermarkt related attributes which are not relevant to this project.

For this reason, the dataset is limited to just 3 columns:

- Player Name: Player's standardized first and last name without accents
- Position Category: Player's position area during season (Goalkeeper, Defender, Midfielder, Forward)
- Sub Position: Players sub position during season(Goalkeeper, Left-Back, Right-Back, Centre-Back, Defensive Midfield, Central Midfield, Attacking Midfield, Right

Midfield, Left Midfield, Left Winger, Right Winger, Centre-Forward, Second Striker)

These datasets are then merged using the common column of the players name to create a complete dataset matching the players from the top 11 football leagues to their respective position categories and sub positions during the season in question.

The complete dataset contains 3189 rows and 101 columns.

Upon comparison of both original datasets and the complete dataset, it was identified that there was a small loss of instances incurred when performing the merging process. 19 players from the performance dataset were not matched successfully with the positional dataset. Although 3189 players were successfully matched during the process. This means that over 99.98% of the instances were matched correctly to their respective positional data. It could be argued that the remaining 19 players should be added manually but the lost data should not hamper the models and this could be added at a later date.

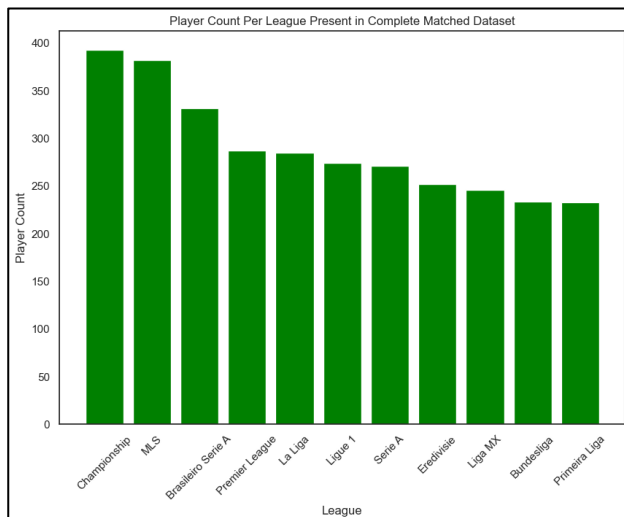


Figure 4: Player Count of Matched Player/Position in Complete Dataset

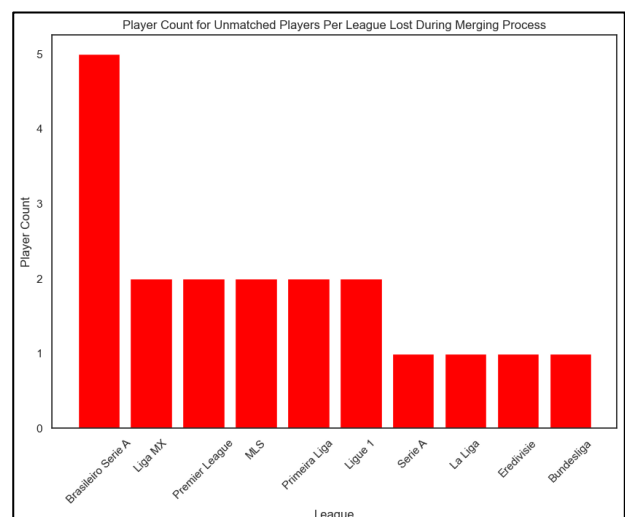


Figure 3: Player Count for Unmatched Players From Original Dataset Lost During Merging Process

4.0 Methodology

The methodology which was followed throughout the course of this project is the Knowledge Discovery in Databases (KDD) data mining model. KDD is a programmed and analytical approach to model data from a database to extract useful and applicable 'knowledge'. This methodology is used in tasks related to machine learning which aim to use grammatical and analytical procedures to retrieve knowledge from a database for use in

the real world.

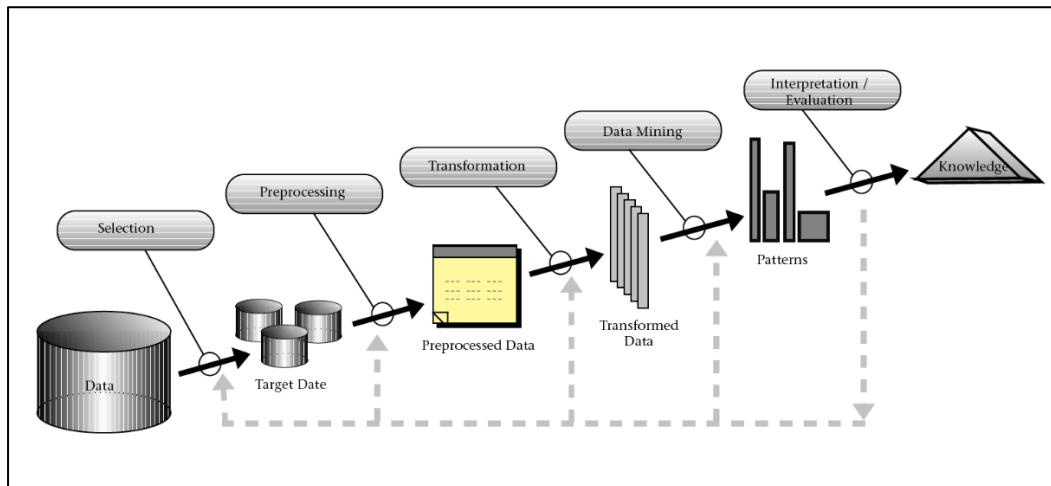


Figure 5: KDD Methodology Steps [19]

The figure above shows 5 key steps involved in the KDD methodology:

- **Selection:** The data which has been collected for the purpose of the project must be analysed and the relevant data must be selected and segregated into meaningful sets based on availability, accessibility, importance, and quality. These parameters are critical for data mining because they make the base for it and will affect what kinds of data models are formed.
- **Cleaning and Pre-processing:** The step of pre-processing and cleaning involves identifying and dealing with missing or noisy data as well as removing redundant or low-quality data from the data set in order to improve the reliability of the data and its effectiveness.
- **Transformation:** The step of transformation of data prepares the data to be fed to various data mining algorithms. Hence, the data needs to be in consolidated and aggregate forms. Various popular techniques are used for this purpose depending on the data and on what data mining algorithms shall be used.
- **Data Mining:** The step of Data Mining is the most important step of the KDD methodology. This is where algorithms are used to identify and extract meaningful patterns from the transformed data, which help in prediction models.
- **Interpretation / Evaluation:** This step is taken to analyse the results of the models employed in the previous step. Any modifications should be made if necessary to the previous steps to try and improve the accuracy of the models based on the results retrieved from this step. If the results retrieved from this step are acceptable for the purpose of the project, the findings can then be documented and presented as necessary.

4.1 Selection

This section of the project details the process of identifying the most relevant data features for the task. A sufficient period of time was devoted at the beginning of the project to

identify and collect relevant player data which could be used for the purpose of creating an analysis of methods to make predictions of players positions on the field and identify similar players at a high accuracy.

Two datasets were identified during this period which were suitable for the task at hand.

Features which were selected to be included for use in the models were strictly chosen on parameters relating to a player's intent and tendencies during a football match. This was done to more accurately profile players based on what they try to achieve rather than the outcome of their actions. This point and reasoning will be expanded on in the analysis section of the report.

A Manual Test dataset was extracted from the complete dataset and stored for manual testing for the various algorithms to validate against overfitting and to show the generality of the models created. This manual test dataset will not be seen by the models during the training or validation processes. The sample size chosen contained 100 players from various outfield positions and leagues chosen at random by the script.

4.2 Cleaning and Pre-processing

After the data collection and data selection phases of the project were completed, it was then time to take on the next very important step of the project which was to clean and pre-process the data to make it fit for the purpose of the project and to make sure the data could be computationally readable and suitable for the proposed models.

Various tasks were carried out during this process such as the renaming of columns and standardizing of both numeric and categorical problems to aid with generality.

4.3 Transformation

Principal Component Analysis: The selected features of the data were normalized using Min-Max scaling. This is to normalize the data so that it can be reduced to a set number of components using **Principal Component Analysis (PCA)**. PCA is a method of dimensionality-reduction which is often used for large data sets which contain a large number of features that need to be reduced for computational purposes.

Reducing the number of variables of a data set naturally comes at the expense of accuracy, but the trick in dimensionality reduction is to trade a little accuracy for simplicity. Due to smaller data sets being easier to explore and visualize and make analysing data much easier and faster for machine learning algorithms without extraneous variables to process.

Training and Validation Dataset Split: The data was then split into training and validation datasets so that the models could be trained on one set of data and then the accuracy of the trained model could then be tested on unseen data to identify the model's accuracy on making predictions on unseen data.

4.4 Data Mining / Information Retrieval

A number of different data mining and statistical tests were tested throughout the work on this project in the hopes of identifying the most accurate solution to the problem at hand.

Predictive Models: Two predictive models were tested and performed on the components obtained from the PCA analysis.

- K-Means Clustering:** The K-Means clustering algorithm is a popular unsupervised machine learning algorithm. This means that it does not require any pre-defined labels to apply to a dataset. This algorithm was applied to and tested on the raw data in the dataset to analyse if the data would naturally cluster into positional clusters without the need for labels to be applied. This was merely used for preliminary analysis and there was never the intention for it to be used as a final model.

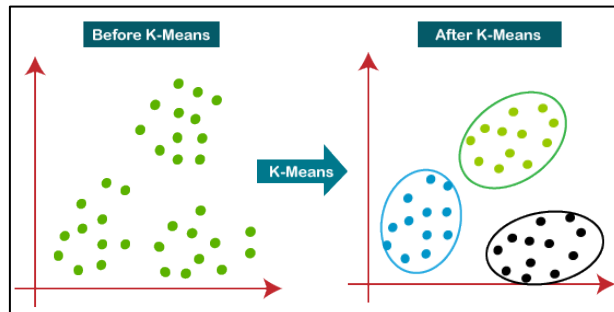


Figure 6: K-Means Clustering by Javatpoint [20]

- K-Nearest Neighbour (KNN):** The K-Nearest Neighbour clustering machine learning model was tested and applied to the components in the hopes of predicting a player's position category and sub-position with as accuracy as possible. Players were clustered using the model and grouped into clusters with similar components created by the PCA. The algorithm then makes calculations on the training and validation data to add the validation data to the clusters with the closest data points in a vector space. These predictions were then analysed to view how accurate the model was in reference to the labels added by the position's dataset.

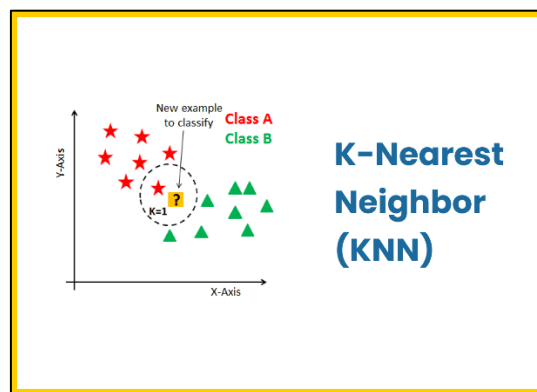


Figure 7: K-Nearest Neighbour Algorithm [21]

- Support Vector Machine (SVM):** The Support Vector Machine, machine learning model was also tested and applied to the components created by the PCA. The principle behind SVM models is to find an optimal plane in a high-dimensional feature space which is used to separate different classes with the maximum margin. This method is then used to identify a decision boundary which maximises the distance between the support vectors which are the data points closest to those decision boundaries. The groups created by this high-level

vector space using the training data can be used to create predictions on the unseen validation data to make predictions on the players positions.

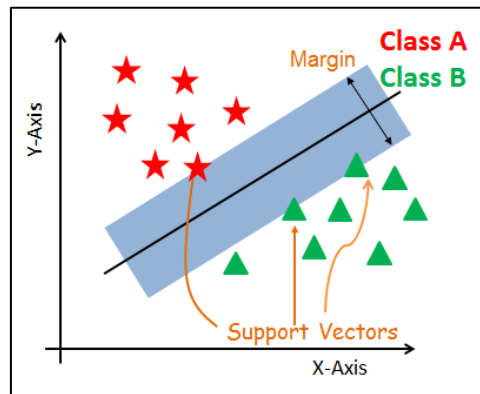


Figure 8: Visualizing SVC (Data camp) [22]

Transfer Learning: Transfer Learning is a powerful technique in machine learning that leverages knowledge learned from one task to improve performance on another related task. This technique has gained significant popularity in Computer Vision and Image Recognition but has also been successfully applied to various other domains such as Natural Language Processing, Speech Recognition, Recommendation Systems, and much more. Using the SVM model created, the data is plotted to a high-level vector space which can then be used to apply the two previous machine learning methods in the hopes of reaching a more accurate result.

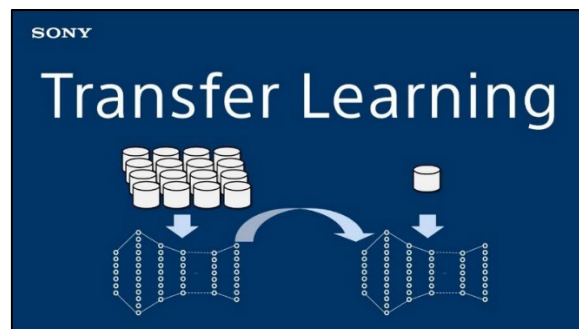


Figure 9: Transfer Learning [23]

Similar Player Analysis: Using the dataset created from the multi-component PCA and the high-level vector space created by the SVM model as comparisons, calculations are made to identify most similar players in the dataset. Using a *Euclidean Distance* calculation function, a target player can be inputted and the function will return a list of the most similar players to the target. This can be applied to both datasets for comparison purposes.

	player	value
0	Cristiano Ronaldo	0.0000
1	Ciro Immobile	0.0724
2	Aleksandar Mitrović	0.1026
3	Sehrou Guirassy	0.1063
4	Kévin Gameiro	0.1183
5	Steven Fletcher	0.1191
6	Sébastien Haller	0.1193
7	Ian Gonzalez Nieto	0.1439
8	Fernando Aristeguieta	0.1439
9	Sasa Kalajdzic	0.1440
10	Andrea Belotti	0.1450
11	Kacper Przybylko	0.1493

Figure 10: Example Similar Player Output (Cristiano Ronaldo)

Targeted Questionnaire: Since there is no mathematical way of determining the accuracy of the Euclidean Distance calculations on both datasets, help was enlisted by 20 participants who claim to have knowledge of football. Each of the 20 participants were asked to submit the 10 players who's playstyle they understand best and would be able to recognise similar players of. The chosen players were then ran through the Euclidean Distance function and a list of 10 players were outputted in the form of a table with the outputs of the 10 most similar players for each player. The participants were then asked which table they thought contained players which were most accurate to their target players.

T-Test: The results of the questionnaire were applied to a T-Test to understand if it can be said that there is a statistical difference between the number of participants which agree that the proposed *improved* method is more accurate than the original model which was created.

Dashboarding: An interactive dashboard was created using the proposed '*improved*' model so that users could test the methods for themselves. The dashboard contains a number of customizable inputs which a user can change to adapt the output they are looking for. The dashboard currently takes inputs such as the target players name in which the user is looking to identify similar player, a league filter which will filter the outputted players if the user only wishes to see players who play in certain leagues, a nationality filter if the user only wishes to identify players of certain nationality,

and age filters if the user is searching for a player of a certain age range, such as under 23.

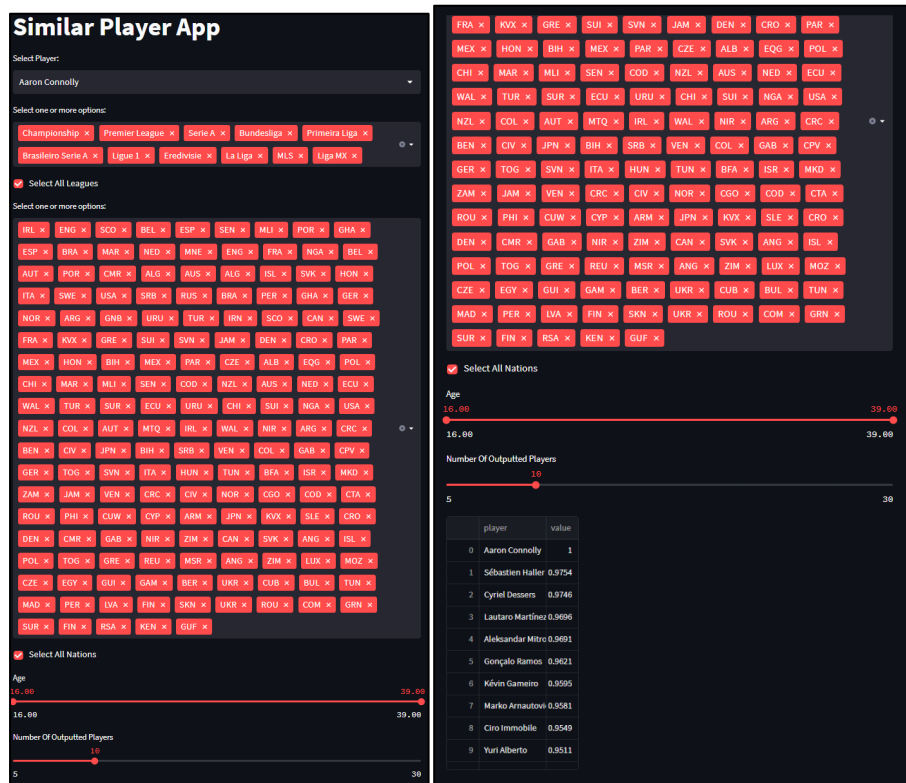


Figure 11: Example Dashboard Output

5.0 Analysis

5.1 Selection

As already stated in the methodology section of this report, the decision was made to only select features for this project which revolved around a player's on-pitch **intent** and tendencies.

Intent:

In this step it was decided that only features showing a player's intention should be included in the dataset. For example, 'Passes_Per90', a data point detailing the number of passes a player attempted per 90 minutes, was a feature that was chosen as the player chose to pass the ball this number of times. However, 'Pass_CompletionPct_Per90', a data point which states the outcome of the passes played as a percentage completed per 90 minutes, should not be included as it relies on the outcome rather than the intent of the player. This is done to remove outcome from the data and focus solely on the player's intentions. A player that tends to pass a lot in matches will be reflected more accurately in the number of passes they attempt in match as opposed to their to the percentage of their attempted passes which they complete per match.

Player	Shots_Per90	Shots_On_Target_Per90
Player1	2.2	0.5
Player2	0.2	0.8

Table 1: Example Player Attribute Values

To demonstrate this point, we look at the figure above. It shows that Player2 has more 'Shots_On_Target_Per90'. Which if taken at face value, would indicate that Player2 shoots more often in match. However, if we look at the intent stat being the 'Shots_Per90', it can be seen that Player1 shoots much more often per 90 minutes than Player2.

The use of only *intent* data was proposed by Mark Carey in his article titled 'Assessing midfielder similarities using machine learning' [15] where the researcher looked to do a similar analysis on the similarity between midfielders in the top 5 footballing leagues.

The data was also filtered using certain parameters to exclude certain groups of players from the dataset. Firstly, the analysis was strictly limited to **outfield players alone**. This meant that all Goalkeepers in the dataset were filtered out during the selection process. This was because the aim of the analysis was to solely analyse outfield players in Europe's top leagues as the assumption is that there aren't many disparities in general attributes for goalkeepers that could be applied to outfield players to create a meaningful analysis. The dataset was also filtered to only include players who had played a minimum of 10 90s across the 2021/2022 season. This meant that players who did not achieve a significant amount of game time would not be included in the analysis. This would ensure that a small sample of a player with cameos in a small number of games throughout the season would not disrupt the weightings towards players with a consistent number of minutes played throughout the season.

5.2 Cleaning and Pre-processing

Once the data had been collected and retrieved it was time to deal with missing or noisy data. The columns needed to be cleaned and standardized as part of the web scraping process so that they could be made machine readable.

A number of functions were created to standardize columns names as well as converting columns from string types to numeric. String columns were cleaned to remove any unnecessary text from cells so that they could be grouped easily and computationally.

A number of duplicated player data was identified and removed so that it did not hinder the performance of the machine learning models. This duplicated data was produced in a case where a player made appearances for two clubs, having transferred to the second club during the season. This would mean that they generated data as one club's player, and another set of data as a second club's player. Both data was combined in this case to create just one profile for every single player.

With the positional data being in categorical form, it was necessary for each position to be converted to a numeric value to make the data more computationally readable. Each positional string was mapped to numeric value that the models could identify during training and testing stages to make predictions on the data.

```
# Hard code encoding mappings
position_category_encoding = {
    'Defender' : 0,
    'Midfield' : 1,
    'Attack' : 2
}

sub_position_encoding = {
    'Full Back' : 0,
    'Centre-Back' : 1,

    'Defensive Midfield' : 2,
    'Central Midfield' : 3,
    'Attacking Midfield' : 4,

    'Wide Midfield' : 5,

    'Winger' : 6,

    'Striker' : 7
}
```

Figure 12: Positional Mapping for Each Position in the Dataset

5.3 Transformation

Principal Component Analysis: The selected features of the data were normalized using Min-Max scaling. This is to normalize the data so that it can be reduced to a set number of components using **Principal Component Analysis (PCA)**. PCA is a method of dimensionality-reduction which is often used for large data sets which contain a large number of features that need to be reduced for computational purposes.

Reducing the number of variables of a data set naturally comes at the expense of accuracy, but the trick in dimensionality reduction is to trade a little accuracy for simplicity. Due to smaller data sets being easier to explore and visualize and make analysing data much easier and faster for machine learning algorithms without extraneous variables to process. An article from builtin.com titled, ‘

A Step-by-Step Explanation of Principal Component Analysis (PCA)’ [24], an article written by builtin.com, details the reasoning and use cases of PCA in machine learning.

A Univariate Optimization Function was created to calculate the minimum number of components needed to explain 95% of the variance in the dataset. This is done to understand what is lowest number of components the chosen features can be reduced to without sacrificing key information contained in the dataset.

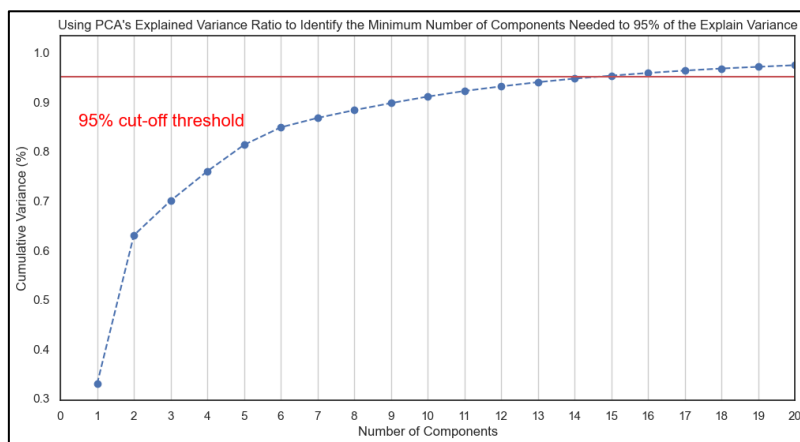


Figure 13: Calculating Minimum Number of Components Necessary to Explain 95% of the Data's Variance

The chart above plots the percentage values of the cumulative variance explained for each component. The industry standard cut off is set at 95% and the red line in the chart denotes the cut off point on the chart. It can be seen from the chart that the optimal value is 14 components which was used for this analysis.

Training and Validation Split: Once the number of components were identified it was then necessary to split the dataset into training and validation datasets. This is done so that the models can be trained on one set of data and then the performance of the model can be validated on unseen data so that an accurate evaluation can be made on the model's performance. With this in mind, a 70%-30% training to validation split was chosen as the ratio to be used for this analysis. This is generally an industry standard practice and can be seen to be applied across all machine learning disciplines, except in special cases.

5.4 Data Mining / Information Retrieval

1) K-Means Clustering: A K-Means Clustering model was fitted to the 2 components dataset. The optimum number of clusters was calculated using 2 methods:

- 1) **Elbow Method:** The elbow method iterates through a selected range of k clusters and calculates the sum of squared differences between each cluster in the given model. The values are then plotted as a line plot showing a curve as the sum of squared distances of the clusters decreases as the number of clusters increases. The optimum number of clusters should be chosen at the 'elbow' of the curve. The last point of the curve before it starts to smoothen.

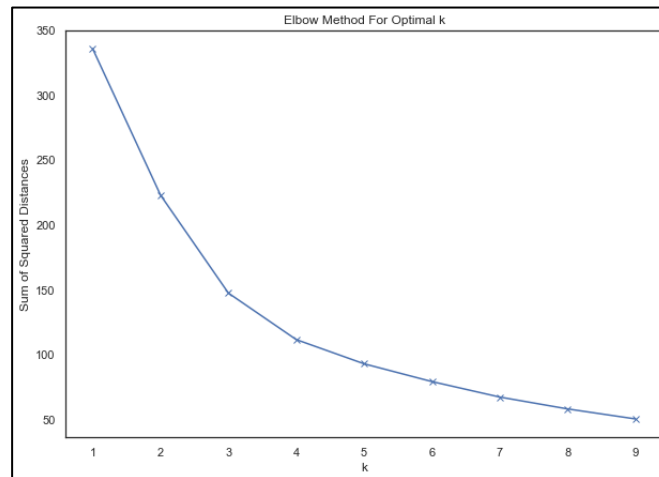


Figure 14: Elbow Method Identifying Optimal k Value

It can be seen from the figure that the optimum number of clusters that should be chosen for this data is at the elbow which is 3 clusters.

2) **Silhouette Method:** The silhouette method is widely regarded as a more accurate method for choosing the optimum number of clusters for complicated datasets. Similar to the elbow method, the silhouette method iterates through a selected range of k clusters but instead calculates silhouette scores for each cluster in the given model. The values are then plotted as a line plot showing a curve as the silhouette values for each cluster decreases as the number of clusters increases. The optimum number of clusters should be chosen as the highest silhouette value for the clusters.

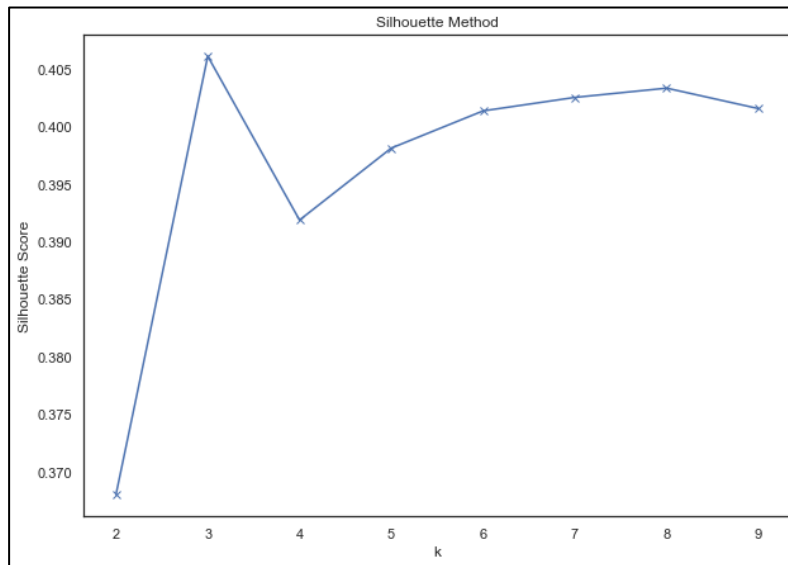


Figure 15: Silhouette Method Identifying Optimal k Value

It can be seen from the figure that the optimum number of clusters that should be chosen for this data is at the high point of the line chart which is 3 clusters. This aligns with the result of the elbow method also.

Entire K-Means Clustering Scatter Plot:

This figure shows the plotting of the player data in a 2-dimensional scatter plot using the K means clustering algorithm. This plot can be analysed to show the distribution of the attributes of the data and analyse at a high level, the differences between players, which can then be used to base further, more complicated algorithms.

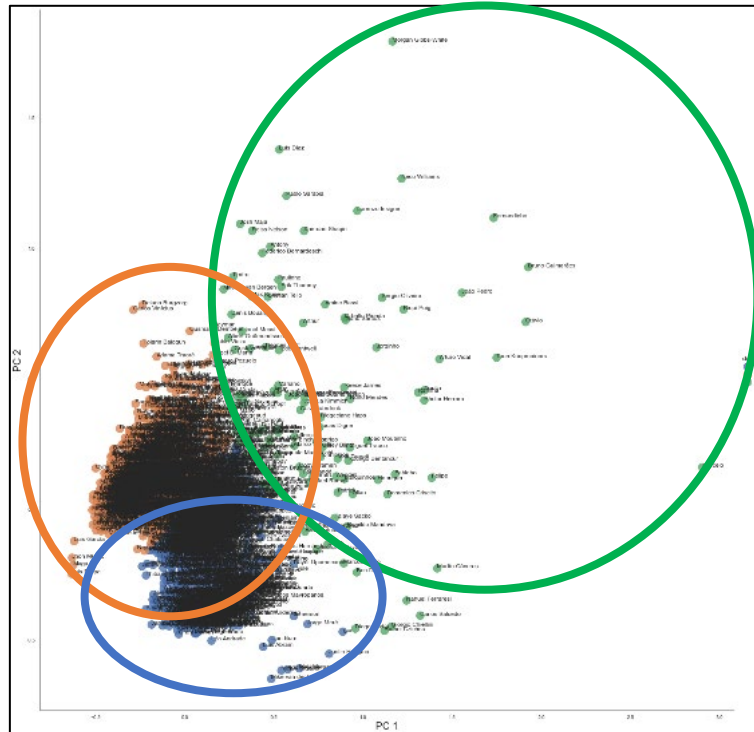


Figure 16: K-Means Clustering Chart

Blue: The blue section of the K Means Clustering plot appears to contain players related to defensive positions such as centre backs and full backs. These players traditionally tend not to shoot or dribble as much as more attacking players and contain mostly high defensive stats such as tackles and interceptions. This could explain why they are lower in the plot than other players.

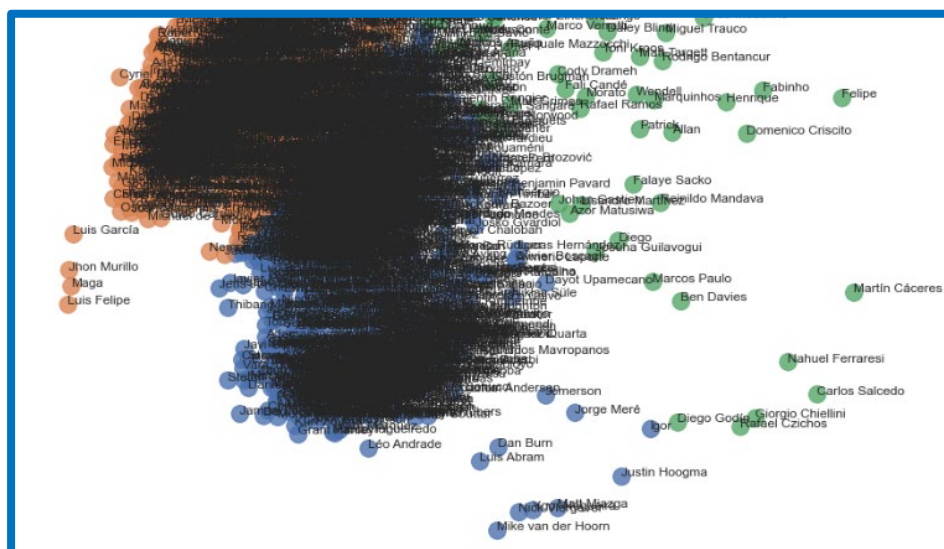


Figure 17: Blue Circled Section of K-Means Clustering Plot

Orange: The orange section of the K Means Clustering plot appears to contain players related to Midfield positions. These players traditionally tend to be more creative than defenders but with some defensive attributes, having slightly more shots and dribbles but not as attacking as attackers. This could explain why they are at the midpoint of the plot.

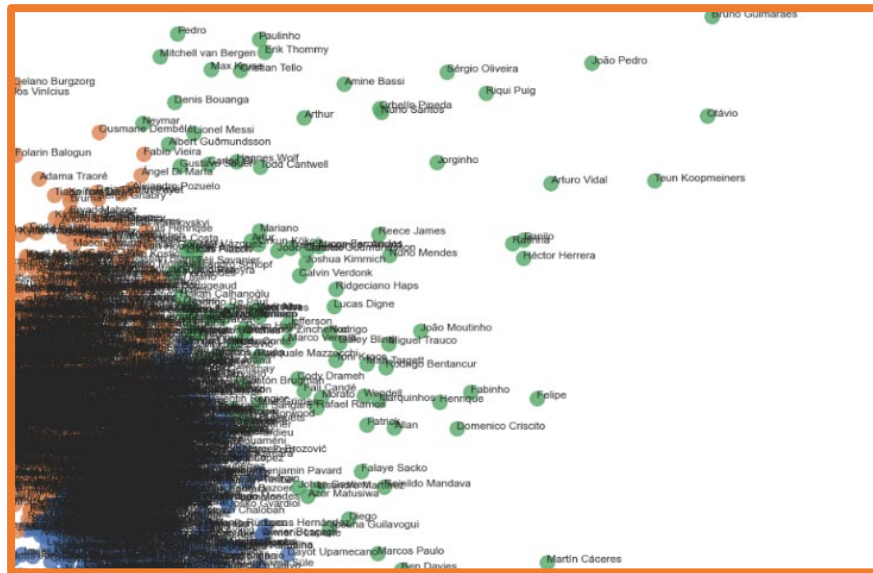


Figure 18: Orange Circled Section of K-Means Clustering Plot

Green: The green section of the K Means Clustering plot appears to contain players related to attacking positions. These players traditionally tend to be very attacking and will focus mainly on creating chances through passes and dribbles as well as taking a lot of shots. This could explain why they are at the higher point of the plot.



Figure 19: Green Circled Section of K-Means Clustering Plot

2) K-Nearest Neighbour Clustering: K-Nearest Neighbour (KNN) is a supervised machine learning algorithm which means that it relies on labelled input data to learn a function that produces an appropriate output when given new unlabelled data.

Training data in the form of X data and Y data are given to the model. X data being the numeric features and Y being the labels which in this case are the position labels converted to numeric values for machine readability.

The model plots the given training data points in a vector space. Once the testing data is fed to the trained model, the model will identify the nearest neighbours to the data in the vector space using Euclidean distance and assign it the same label as its neighbour. The model will do this until each test data point is assigned a label known as a prediction.

The figure below shows an example plot showing KNN on a 2-dimensional plane.

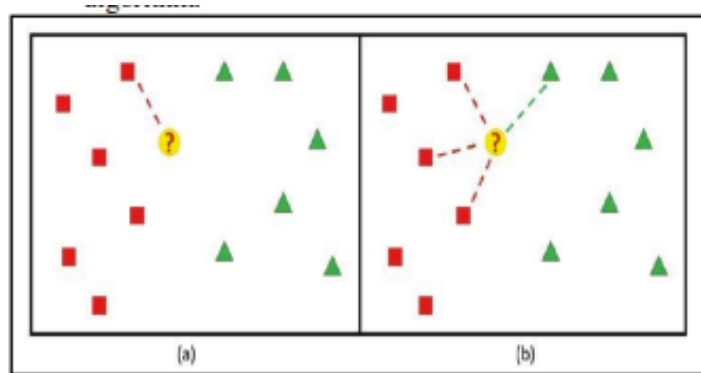


Figure 20: Presenting the Workings of the KNN Algorithm [15]

A univariate optimization function created to identify the most suitable values for k-neighbours for the KNN model. This means the model chooses which cluster a new data point applies to based on a set number of closest neighbours. The function iterates through a selected range of k-neighbours and calculates the both the accuracy and error rate for each value of k. The values are then plotted as a line plot and the highest value for accuracy and lowest value for error rate are chosen as the most suitable value for k.

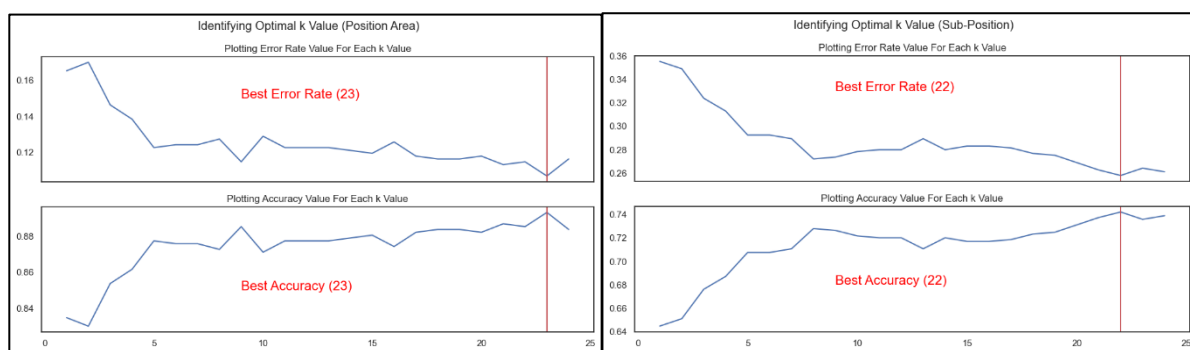


Figure 21: Plotting Error Rate and Accuracy to Identify Optimal k Value

It can be seen from the charts above, plotting the error rate and accuracy for 'Position Area' data points, that 23 is the most suitable value for k in this instance and that plotting the error rate and accuracy for 'Sub Position' data points, that 22 is the most suitable value for k in this instance.

3) Support Vector Machine (SVM): Support Vector Machines are a popular machine learning algorithm used in both classification and regression problems. The underlying principle behind SVM is to find an optimal hyperplane in a high-dimensional feature space that separates different classes with the maximum margin. In other words, SVM aims to identify a decision boundary that maximizes the distance between the support vectors, which are data points closest to the decision boundary. This margin ensures better generalization and robustness of the model. SVM achieves this by transforming the input data into a higher-dimensional space using a kernel function. In this transformed space, SVM finds the hyperplane that best separates the classes by solving an optimization problem. The key idea is to map the data into a feature space where it is linearly separable, even if the original data may not be. SVMs are versatile and can handle both linearly separable and non-linearly separable data by using different kernel functions.

SVM models take a parameter known as the C parameter which is crucial to the model in helping to control the trade-off between a low training error and minimizing the complexity of the decision boundary which in turn makes it easier for the model to generalize on unseen data and prevents overfitting on the training data.

A Univariate Optimization Function was again used to identify the ideal value for C to apply to the model during training. Just like what was used for the KNN optimization the same method is used to identify the optimal value for C for the Support Vector Machine Models. The function iterates through a selected range and calculates both the error rate and accuracy for the function with C as the value. The function then identifies the value for C which returns both the lowest error rate and highest accuracy within the range which can then be used to train the model.

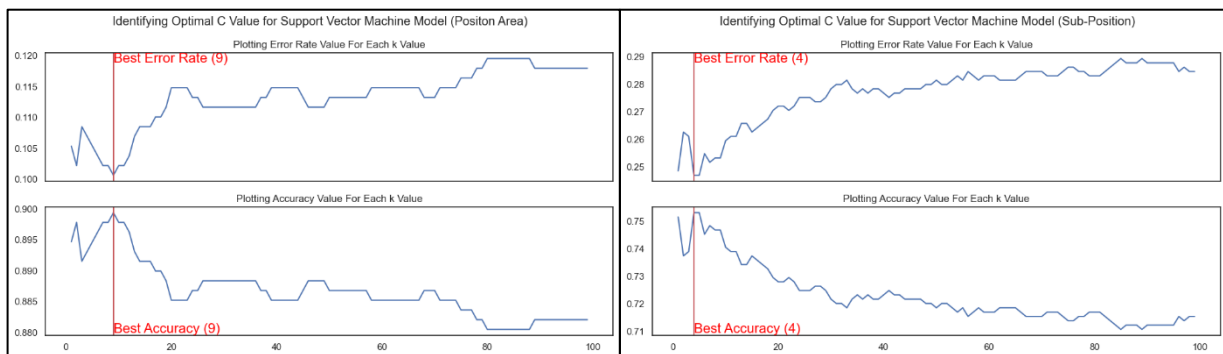


Figure 22: Charts Plotting Univariate Optimization Function Results to Identify Optimal Value for C for Support Vector Machine Models

The charts above which are plotting the error rate and accuracies for each C value from 1-100 for both the Position Area and Sub Position labels. It can be seen clearly in the charts what values for C the function identifies which can then be applied to the model during training for each target label set.

The algorithm was again applied to the X and Y training datasets in order to train the model to make predictions on player's positions using the component dataset.

4) Transfer Learning: A Transfer Learning Approach was tested seeking to improve on the results already gained from both models. Transfer Learning involves leveraging information obtained from a previous model and using the information to create a high-level vector space where the same models can be applied in the hopes of achieving more accurate results.

In the case of this task, the SVM model, which was trained and validated in the last section, is applied to the entire dataset. The model creates a vector space through calculating the decimal probability weighting to each position for each data point. This information is then used to create a new high-level dataset which the earlier models can then be applied to in the hopes of achieving a higher accuracy. In this case, the SVM model which was used to create the probability calculations to create the high-level vector space used a C value of 4. This model was trained and validated on the original component training and validation datasets and was then applied to the combined dataset to output the high level vector space.

5) Similar Players Analysis using Euclidian Distance: Making use of the components that were created using PCA in the transformation step of this project, the components were relayed to a data frame where the player's name labels were added to the corresponding values for each player. A calculation of *Euclidean Distance* is then used to identify similar players to a given input. A function which was created which takes the users input and retrieves the 10 most similar players to the given input player.

Euclidean Distance is a formula which was created to identify the distance between 2 points in a vector space. The formula is much simpler when working in a 2-dimensional space but in the case of using 14 components the calculation becomes much larger.

$$d = \text{sqrt}((x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2, \dots)$$

Equation 1: Euclidean Distance Equation

This function calculates the Euclidean Distance between the inputted player and every other player in the data frame and returns a data frame filtered to return the 10 highest values.

An example output of similarity function with 'Lionel Messi' as input can be seen in the figure.

	player	value
0	Lionel Messi	0.0000
1	Neymar	0.4195
2	Darwin Quintero	0.4205
3	Thiago Almada	0.4213
4	Alejandro Pozuelo	0.4432
5	Max Kruse	0.4791
6	Fabio Vieira	0.5242
7	Luis Montes	0.5248
8	Luciano Acosta	0.5281
9	Jamiro Monteiro	0.5485
10	Sergio Canales	0.5662
11	Luis Alberto	0.5801

Figure 23: Top 10 Similar Players for Lionel Messi Using PCA Components as Input

The function was created to calculate the most similar players using *Euclidean Distance* based on both the original components vector space and also for the high-level vector space which was created using Transfer Learning in the previous step. This meant that 2 outputs could be retrieved and could be analysed to understand which vector space was deemed to return the most accurate similar players. It was noticed that there was a large difference between the outputs in both vector space embeddings.

6) Targeted Questionnaire: Through some extensive research, it was established that there was no mathematical method to evaluate which of the outputs between the two embeddings was most accurate with regard to the most similar players analysis. Therefore, it became evident that the involvement of knowledgeable individuals well-versed in the world of football was imperative to decide which output was most accurate.

With this in mind, a questionnaire was created to pitch to 20 participants who claimed to be well-versed in football to choose 10 football players of whom they thought they understood their style of play best and would be able to identify whether proposed players would be similar to their chosen player. Once this questionnaire was completed by the participants, outputs were created containing two columns for each selected player. One column containing the ten most similar players deemed using the Euclidean Distance calculation on the reduced component dataset, and ten most similar players chosen using the Euclidean Distance calculation on the high-level dataset created using Transfer Learning.

The participants were then asked, without knowing which model was which, to choose which of the two columns were most accurate in their opinion. The number of choices for both models were noted to a spreadsheet to be analysed using statistical tests to understand if there was a statistical difference between the choices for the two vector

spaces. A T-test statistical test was applied to the participants choices in order to assess the significance of differences between the two choices. A t-test is used to evaluate whether the means of two independent samples are significantly different from each other. It is particularly useful when working with continuous numerical data and seeking to draw inferences about population parameters based on sample data. By calculating the t-statistic, which quantifies the difference between the sample means relative to the variability within each group, the t-test provides a statistical measure of the likelihood that any observed differences are not merely due to random chance. The resulting p-value from the t-test helps researchers make informed decisions about whether to reject or fail to reject the null hypothesis, which posits that there is no significant difference between the two groups. When properly applied and interpreted, the t-test serves as a robust tool in hypothesis testing, aiding researchers in drawing reliable conclusions and advancing scientific knowledge in various academic disciplines.

7) Dashboarding: The most accurate vector space was taken as a dataset and hosted on an interactive dashboard using Python code with the help of the Streamlit library. This interactive dashboard is proposed to be used as a recruitment tool wherein the user can select target player and the dashboard will retrieve the players most similar to the user's input. This dashboard contains a number of useful filters that can be applied to the user's output to help the user to specify a specific group this similar player should be linked to. These filters include League, Nationality, and Age.

6.0 Results

K-Means Clustering: K-Means clustering was used in this project as a form of early preliminary analysis. Plotting on a 2-dimensional plane using 2 component PCA and showing the spread of the data points on the plot was done to understand and show that the features can be used to show various play styles of the players included in the dataset. The fact that the plot shows more attacking players higher in the plot and defensive players towards the bottom of the plot shows that there are distinguishable differences between the data points in the dataset.

The information retrieved from this preliminary method was used to develop more complicated models to identify a metric to quantify the differences between various players and to gauge the accuracy of machine learning algorithms.

K-Nearest Neighbour and Support Vector Machine Algorithms: The KNN and SVM models were used to make the classifications of the positional data relating to each data point. The models would cluster the data to make predictions on which positional cluster the validation data would be classified as.

The models were ran once each on the original component vector space, and then again on the high-level vector space created with the help of transfer learning using the SVM model.

The results can be seen presented as follows:

Predictive Model Results:

Original KNN Position Area Results:

Classification Report:				
	precision	recall	f1-score	support
Defender	0.93	0.96	0.95	249
Midfield	0.87	0.84	0.86	212
Attack	0.86	0.86	0.86	175
accuracy			0.89	636
macro avg	0.89	0.89	0.89	636
weighted avg	0.89	0.89	0.89	636
Accuracy: 0.8930817610062893				

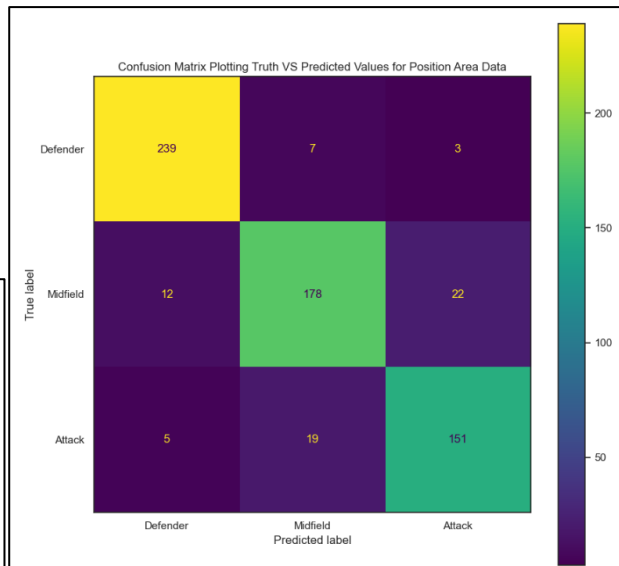


Figure 24: Original KNN Area Results

Original KNN Sub Position Results:

Classification Report:				
	precision	recall	f1-score	support
Full Back	0.80	0.89	0.84	106
Centre-Back	0.93	0.95	0.94	143
Defensive Midfield	0.52	0.55	0.53	58
Central Midfield	0.64	0.58	0.61	105
Attacking Midfield	0.35	0.20	0.25	41
Wide Midfield	0.00	0.00	0.00	8
Winger	0.65	0.75	0.69	80
Striker	0.84	0.85	0.84	95
accuracy			0.74	636
macro avg	0.59	0.60	0.59	636
weighted avg	0.72	0.74	0.73	636
Accuracy: 0.7421383647798742				

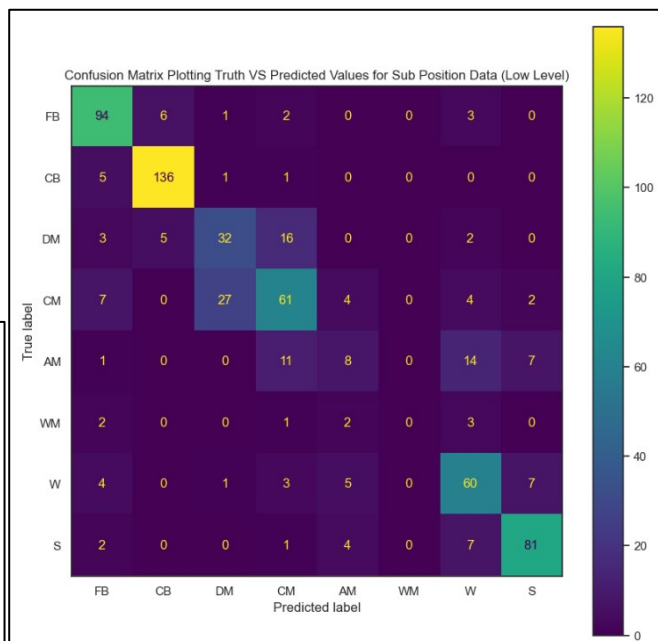


Figure 25: Original KNN Sub Position Results

Original SVM Position Area Results:

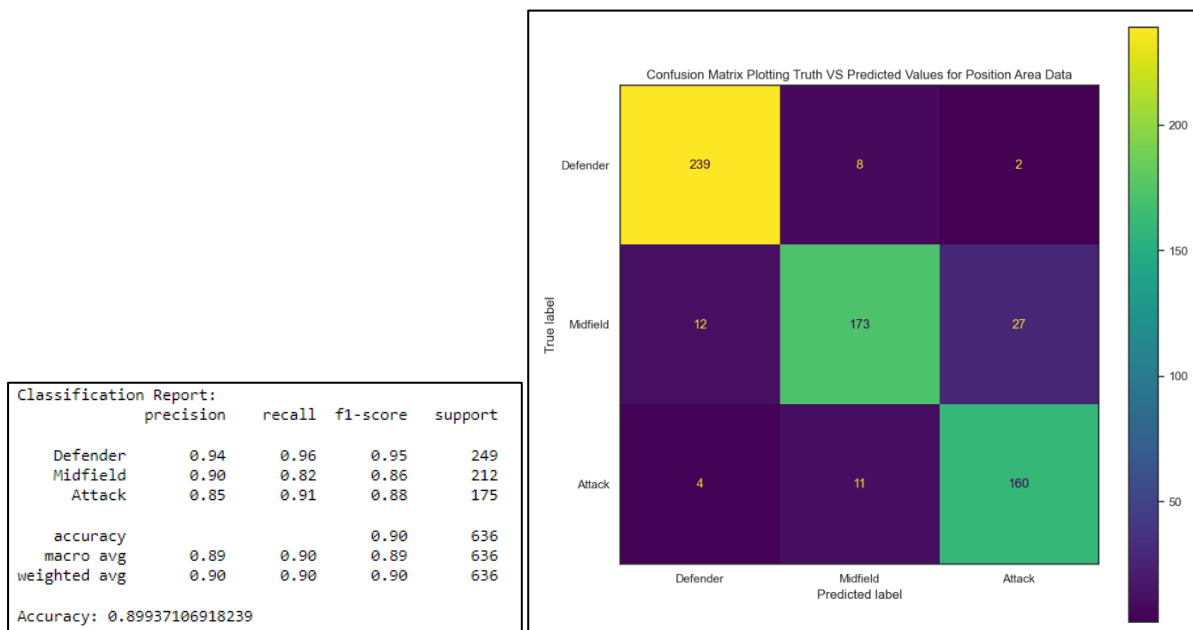


Figure 26: Original SVM Position Area Results

Original SVM Sub Position Results:

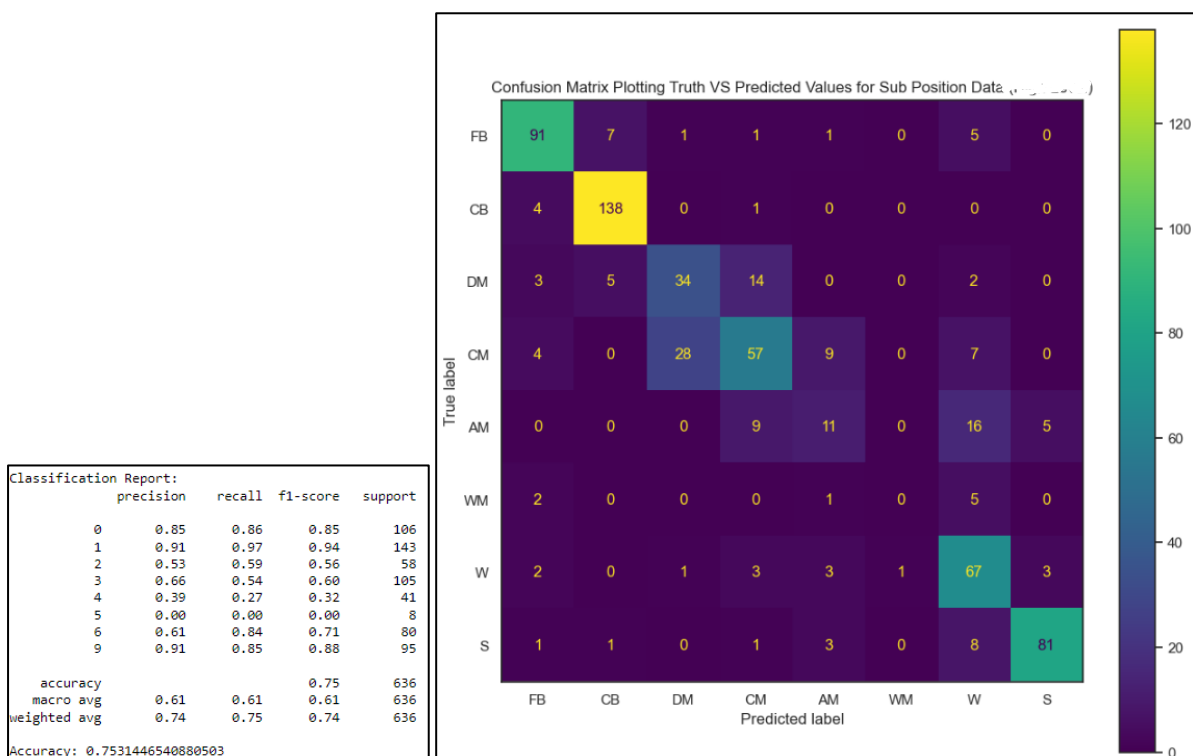


Figure 27: Original SVM Sub Position Results

High-Level KNN Position Area Results:

Classification Report:				
	precision	recall	f1-score	support
Defender	0.94	0.96	0.95	249
Midfield	0.87	0.82	0.84	212
Attack	0.82	0.86	0.84	175
accuracy			0.89	636
macro avg	0.88	0.88	0.88	636
weighted avg	0.89	0.89	0.88	636

Accuracy: 0.8852201257861635

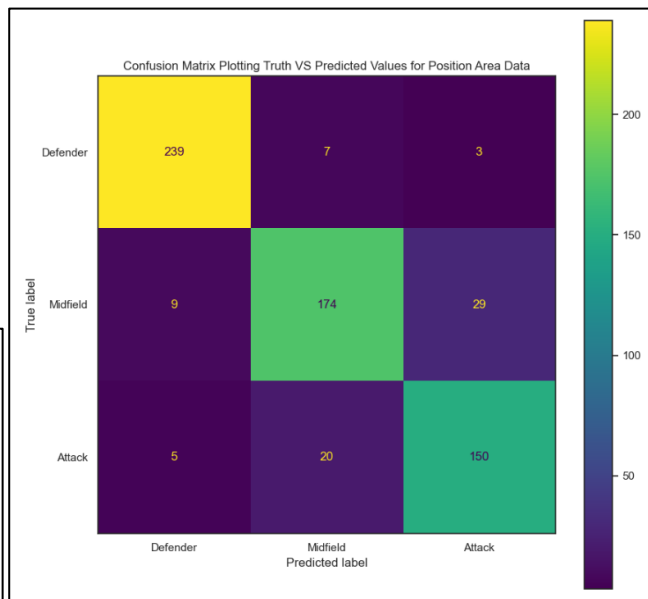


Figure 28: High-Level KNN Position Area Results

High-Level KNN Sub Position Results:

Classification Report:				
	precision	recall	f1-score	support
Full Back	0.84	0.83	0.83	106
Centre-Back	0.91	0.97	0.94	143
Defensive Midfield	0.52	0.57	0.55	58
Central Midfield	0.64	0.54	0.59	105
Attacking Midfield	0.45	0.34	0.39	41
Wide Midfield	0.00	0.00	0.00	8
Winger	0.60	0.80	0.68	80
Striker	0.91	0.84	0.87	95
accuracy			0.75	636
macro avg	0.61	0.61	0.61	636
weighted avg	0.74	0.75	0.74	636

Accuracy: 0.7452830188679245

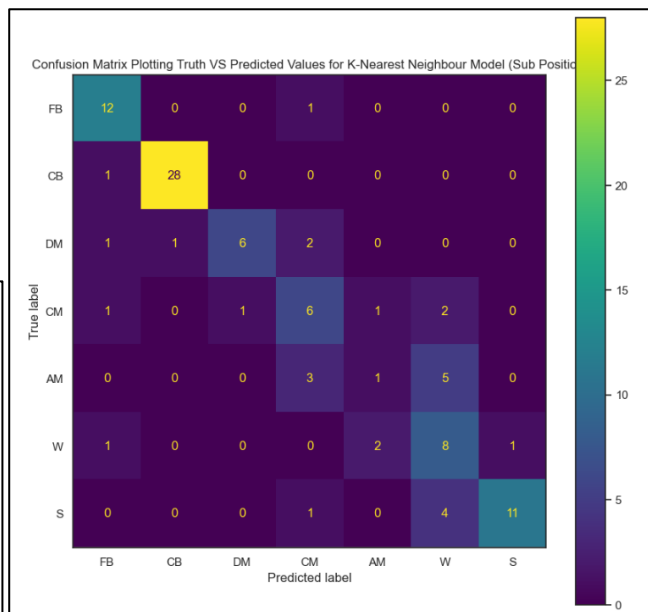


Figure 29: High-Level KNN Sub Position Results

High-Level SVM Sub Position Results:

Classification Report:				
	precision	recall	f1-score	support
0	0.82	0.84	0.83	106
1	0.91	0.97	0.94	143
2	0.54	0.57	0.55	58
3	0.63	0.54	0.58	105
4	0.46	0.29	0.36	41
5	0.00	0.00	0.00	8
6	0.62	0.82	0.71	80
9	0.91	0.85	0.88	95
accuracy			0.75	636
macro avg	0.61	0.61	0.61	636
weighted avg	0.74	0.75	0.74	636
Accuracy: 0.7484276729559748				

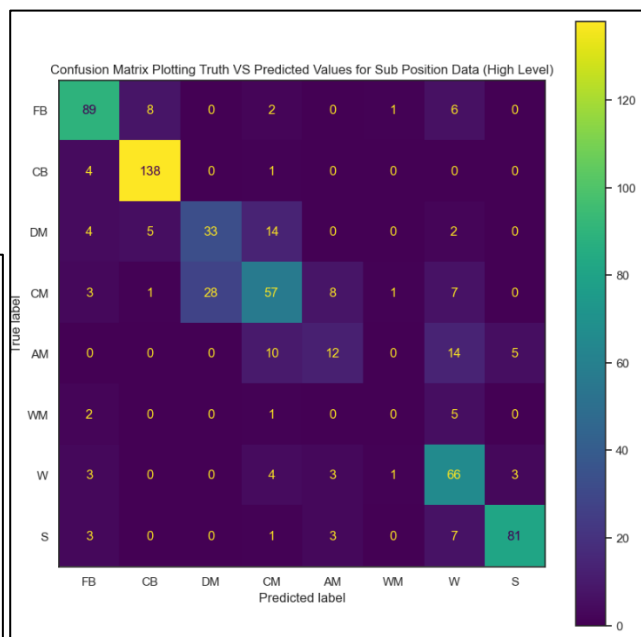


Figure 30: High-Level SVM Sub Position Results

Test Data: Original KNN Position Area Results:

Classification Report:				
	precision	recall	f1-score	support
Defender	0.91	0.93	0.92	42
Midfield	0.83	0.67	0.74	30
Attack	0.82	0.96	0.89	28
accuracy			0.86	100
macro avg	0.85	0.85	0.85	100
weighted avg	0.86	0.86	0.86	100
Accuracy: 0.86				

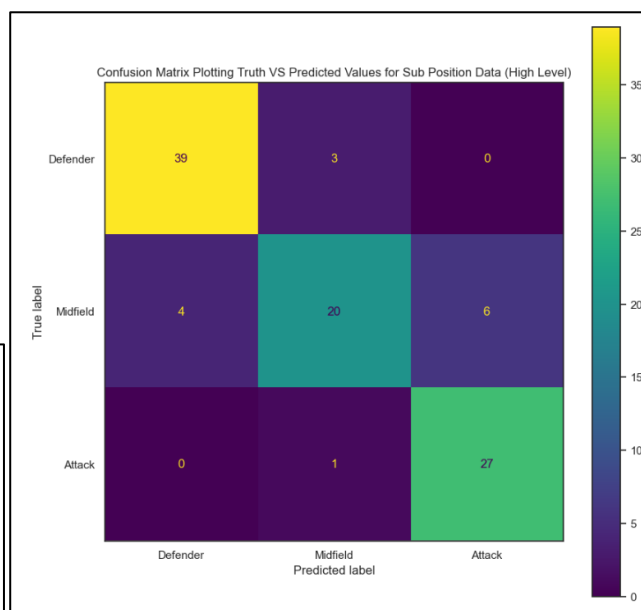


Figure 31: Test Data: Original KNN Position Area Results

Test Data: Original KNN Sub Position Results:

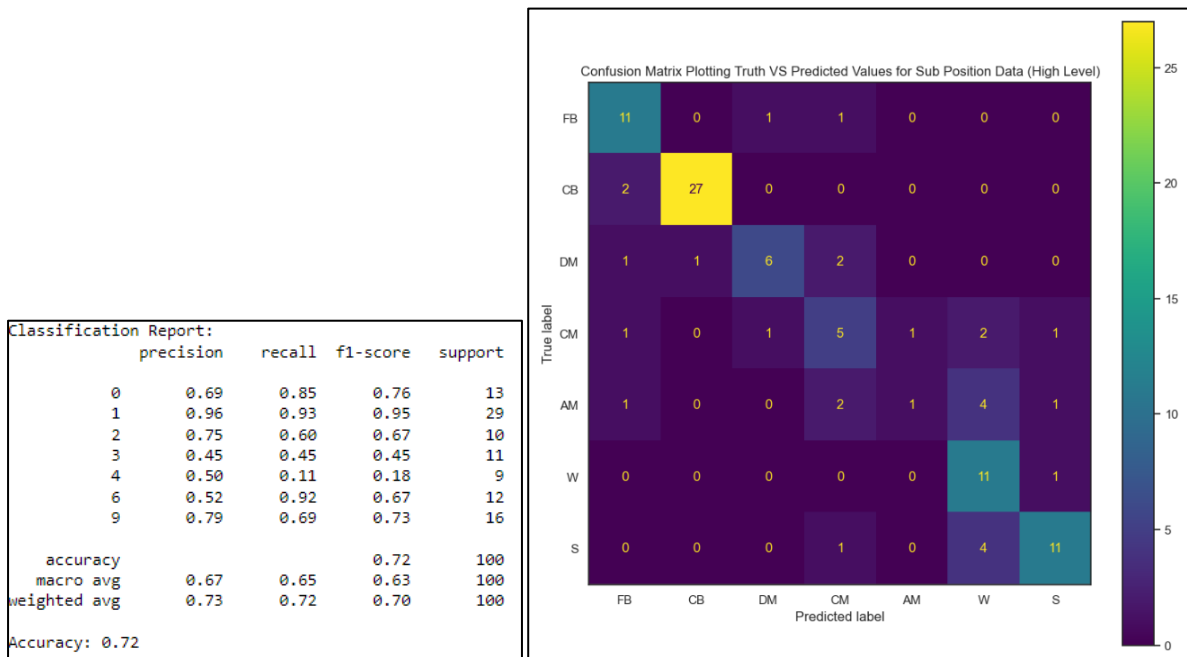


Figure 32: Test Data: Original KNN Sub Position Results

Test Data: High-Level KNN Position Area Results:

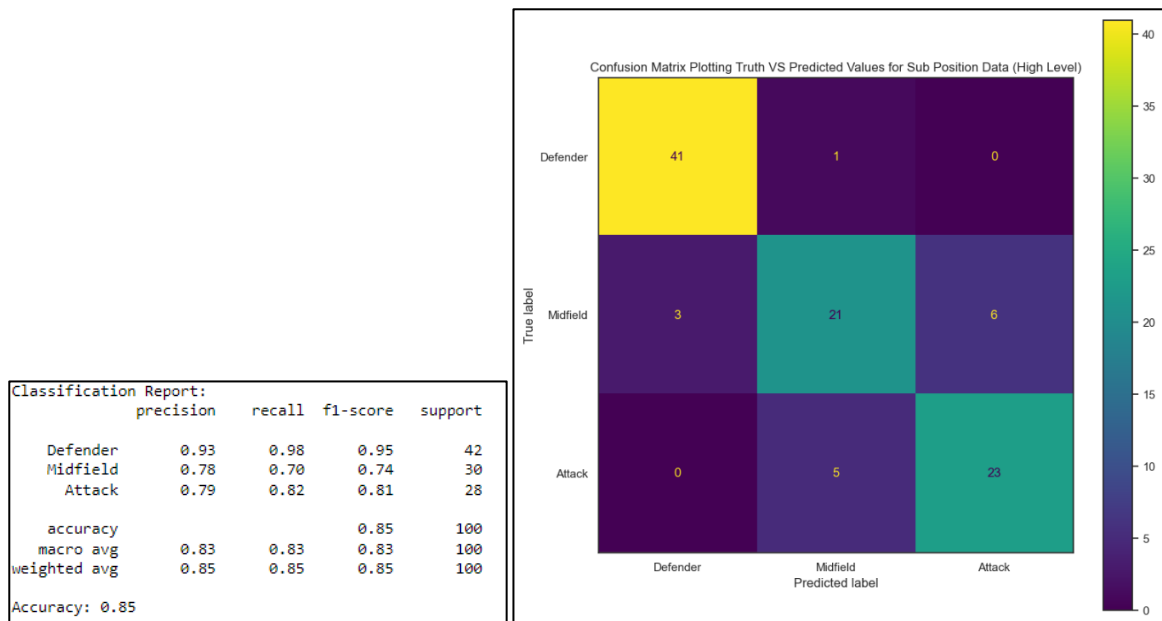


Figure 33: Test Data: High-Level KNN Position Area Results

Test Data: High-Level KNN Sub Position Results:

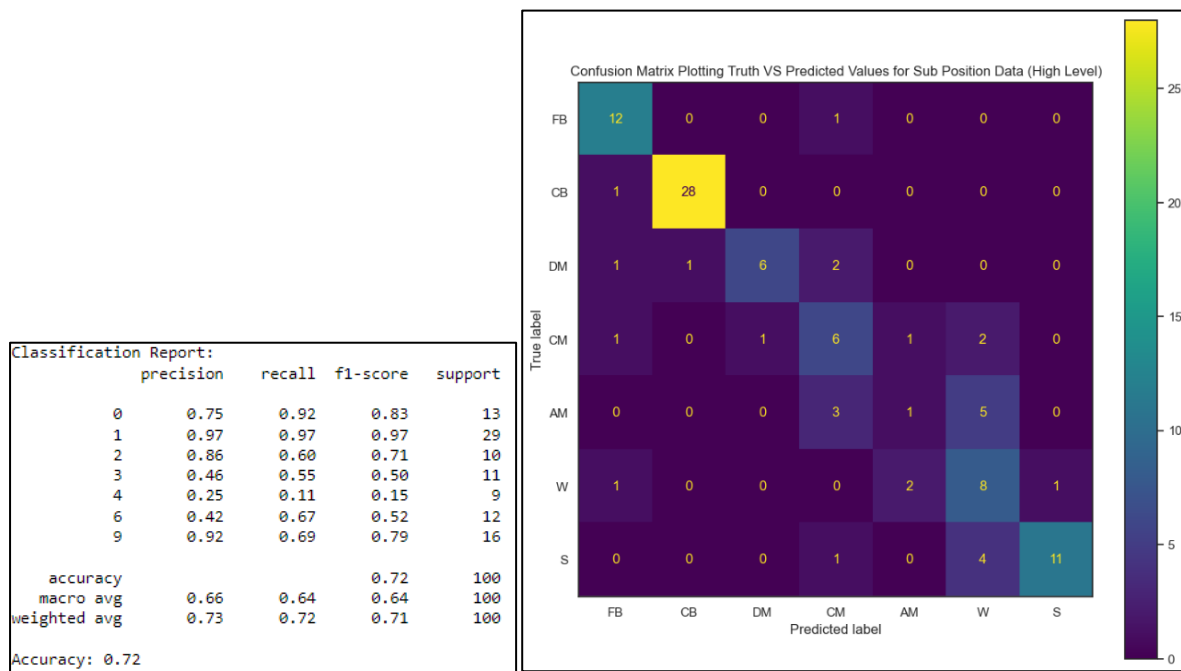


Figure 34: Test Data: High-Level KNN Sub Position Results

Results Breakdown:

Original Component Vector Space Results:

Labels	Model	Accuracy (%)	Scatter
Position Area	KNN	89.31	
Sub Position	KNN	74.21	66.57
Position Area	SVM	89.94	
Sub Position	SVM	75.31	71.75

Table 2: Machine Learning Model Results Breakdowns - Original Vector Space (Accuracy Scores)

High-level Vector Space:

Labels	Model	Accuracy (%)	Scatter
Position Area	KNN	88.52	
Sub Position	KNN	74.53	25.07
Position Area	SVM	87.25	
Sub Position	SVM	74.84	27.29

Table 3: Machine Learning Model Results Breakdowns - High-Level Vector Space (Accuracy Scores)

The results retrieved from each iteration of the models can be seen in the table above. It can be seen that The Support Vector Machine performs best on the original vector space data which is why it that particular model is suitable to be used during the transfer learning

process to create the high-level dataset which is later used in the hopes of gathering more accurate results in terms of the most similar player identification model.

It can be seen that an accuracy score of 89.94% is the highest achievable accuracy from the models for the Position Area labels and that 75.31% is the highest achievable accuracy for the Sub Position labels which are both returned by the Support vector Machine Algorithm.

Furthermore, when examining the scatter of the clusters returned by each algorithm, it is notable that the K-Nearest Neighbour algorithm exhibited the tightest scatter, 25.07, when applied to the high-level dataset. This suggests that the vector space representation used in this study has the potential to yield higher accuracy results with the similar player identification algorithm.

Targeted Questionnaire: For the purpose of the questionnaire, 20 targeted participants who were chosen based on their knowledge of the football world, were asked to choose 10 players each of whom they felt they understood their playstyle based. The participants would then fill out a form detailing their 10 chosen players along with which leagues they wished to be included in their output.

The participants were then returned a table for each of their chosen players which were created using the Euclidean Distance Similar Players Identification model. Each table contained two columns. One column contained the top ten players deemed most accurate to their chosen player based on the original components dataset. Whereas, the other model contained the top ten players deemed most accurate by the high-level dataset created using transfer learning using the Support Vector Machine Model created earlier. The participants were then asked to choose which of the two columns they thought was most accurate to their chosen target player. The participants were asked to repeat this choice for each of their chosen target players. This meant that the results returned would be 10 answers from each participant with 20 participants which resulted in 200 answers being returned.

To analyse the answers which were returned by the participants, the choices were encoded as a 0 if the participant did not choose the improved model and a 1 if they did. A t-test was created in order to established if the participants choices were statistically different to the mean value which would be expected of 0.5 which relates to the expected mean value for a list of 200 0 and 1s.

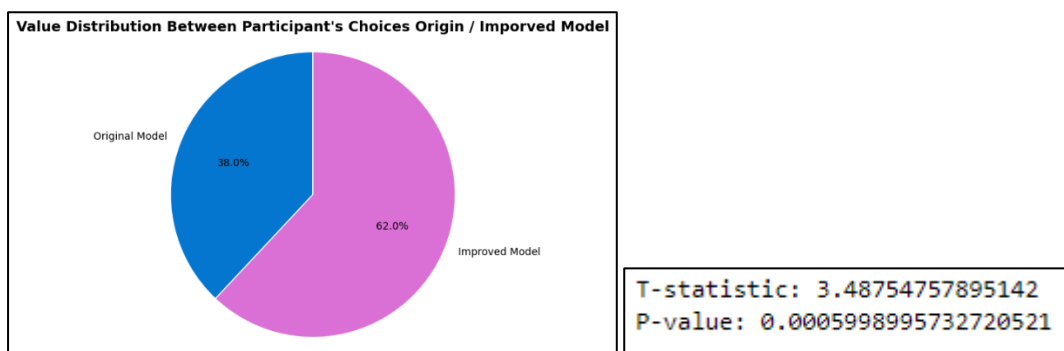


Figure 35: Questionnaire Results Breakdown

It can be seen from the figures above that 62% of the 200 participant choices related to the improved model with 38% relating to the original model. Through the use of a t-test it was established that a T-statistic of 3.49 which would indicate that there is a significant difference between the sample mean and the null hypothesis mean, with the null hypothesis being that there is no statistical difference between the participants choices. The P-value returned shows a very low number of 0.0006 which means that the observed difference is unlikely to occur by chance alone, and we can reject the null hypothesis in favor of an alternative hypothesis which would be that the participants found the improved model more accurate than the original model.

7.0 Conclusions

In conclusion, this project aimed to explore different machine learning algorithms for similar player identification in football data. Through experimentation and analysis it was identified that the Support Vector Machine models which were created during the analysis performed exceptionally well on the original vector space data. It achieved the highest accuracy scores of 89.94% for Position Area labels and 75.31% for Sub Position labels, making it a suitable choice for the transfer learning process to create a high-level dataset.

Additionally, when applying the K-Nearest Neighbour models on the high-level dataset, it was observed that the algorithms contained the tightest scatter of clusters of any of the models on either of the vector spaces, which would indicate the potential for increased accuracy in similar player identification. This highlights the effectiveness of the vector space representation used in this study and backs up the proposed methodology to improve the similar player identification algorithm.

These findings could have significant implications for football analysis and player scouting and could be expanded on and used to aid scouts and professional clubs at the highest level of the sport. By accurately identifying similar players, teams and coaches can make informed decisions regarding player recruitment, strategic planning, and performance optimization. Moreover, the success of the SVM algorithm in the original vector space data and the potential of the high-level dataset created through transfer learning showcase the importance of algorithm selection and feature engineering in similar player identification tasks.

As with any research project, there are limitations to consider. The results presented in this study are based on a specific dataset and may not generalize to all leagues or players. The data which was used in this analysis relates to data pulled from one specific year and only relates to the most elite leagues in the football world. This means that the models which were created for the purpose of this study may not as easily generalize to fit data potentially from a lower skilled league. Further exploration with different datasets and sports domains would provide a more comprehensive understanding of the algorithms' performance. Additionally, the evaluation metrics used in this study focused on accuracy alone, and it would be valuable to assess other performance measures such as precision, recall, and F1-score to gain a more nuanced perspective.

In conclusion, this project contributes to the advancement of similar player identification techniques and offers valuable insights for football analytics and player scouting. The

findings highlight the effectiveness of the SVM algorithm and the potential of transfer learning with the creation of a high-level dataset to be used throughout the process of recruitment and scouting and the analysis itself could help to create scouting tools which may expand upon methods suggested in this report.

8.0 Further Development or Research

Based on the results and identified limitations of this project, there are several avenues which could be suggested with regard to future research and development on the subjects and methods discussed throughout this report.

1) Dataset Expansion: As mentioned in earlier sections of the report, the current study analyses only a very specific dataset from an elite level of football which is limited to a snapshot of data gathered within one year. Potential proposed future work that could be carried out to expand upon the methods detailed in this report could involve incorporating data from a wider range of leagues specifically seeking to include lower-tier leagues and also different seasons, to increase the generalizability of the models. This would provide a more comprehensive understanding of the algorithms' performance across various levels of football.

2) Feature Engineering: While the current study focused on the vector space representation of players, future research could explore additional features to improve the accuracy of similar player identification. To mention more features that could be taken into account in future work, it may be beneficial to creating a more accurate algorithm for identifying similar players to incorporate more features in the analysis such physical traits, playing style, or to even go as far as incorporating video analysis data. The inclusion of such features could provide a more holistic representation of player similarity.

3) Evaluation Metrics: With accuracy being the main evaluation metrics which was focused on throughout the course of this particular study, it may be beneficial for future work to consider the other metrics such as *precision*, *recall*, and *F1-score*. These metrics provide a more comprehensive assessment of model performance and can help evaluate the trade-off between different evaluation measures.

By pursuing these future directions, researchers and practitioners can further advance the field of similar player identification in football. These developments have the potential to revolutionize player scouting, talent identification, and team performance analysis, providing valuable insights and support for decision-making processes in the football industry.

9.0 References

- [1] 'Predicting Job Performance: The Moneyball Factor'. Accessed: May 09, 2023. [Online]. Available: https://faculty.wharton.upenn.edu/wp-content/uploads/2012/05/Moneyball-Foresight_1.pdf
- [2] 'THE BEAUTIFUL (COMPUTER) GAME: HOW DATA SCIENCE WILL REVOLUTIONIZE THE WORLD'S MOST POPULAR SPORT'. Accessed: May 09, 2023. [Online]. Available: https://repositories.lib.utexas.edu/bitstream/handle/2152/116332/Takvorian_Thesis_The%20Beautiful%20Computer%20Game_2021.pdf?sequence=2&isAllowed=y
- [3] 'Football Statistics and History', *FBref.com*. <https://fbref.com/en/> (accessed May 09, 2023).
- [4] 'Football transfers, rumours, market values and news'. <https://www.transfermarkt.com/> (accessed May 09, 2023).
- [5] 'pandas - Python Data Analysis Library'. <https://pandas.pydata.org/> (accessed May 09, 2023).
- [6] 'NumPy'. <https://numpy.org/> (accessed May 09, 2023).
- [7] M. Waskom, 'seaborn: statistical data visualization', *JOSS*, vol. 6, no. 60, p. 3021, Apr. 2021, doi: 10.21105/joss.03021.
- [8] 'Matplotlib — Visualization with Python'. <https://matplotlib.org/> (accessed May 09, 2023).
- [9] 'Plotly: Low-Code Data App Development'. <https://plotly.com/> (accessed May 09, 2023).
- [10] 'scikit-learn: machine learning in Python — scikit-learn 1.2.0 documentation'. <https://scikit-learn.org/stable/> (accessed Dec. 20, 2022).
- [11] 'Requests: HTTP for Humans™ — Requests 2.28.1 documentation'. <https://requests.readthedocs.io/en/latest/> (accessed Dec. 20, 2022).
- [12] 'Beautiful Soup Documentation — Beautiful Soup 4.9.0 documentation'. <https://www.crummy.com/software/BeautifulSoup/bs4/doc/> (accessed Dec. 20, 2022).
- [13] 'Streamlit • A faster way to build and share data apps', Jan. 14, 2021. <https://streamlit.io/> (accessed May 09, 2023).
- [14] Y. Li, S. Zong, Y. Shen, Z. Pu, M.-Á. Gómez, and Y. Cui, 'Characterizing player's playing styles based on player vectors for each playing position in the Chinese Football Super League', *Journal of Sports Sciences*, vol. 40, no. 14, pp. 1629–1640, Jul. 2022, doi: 10.1080/02640414.2022.2096771.
- [15] CAREYANALYTICS, 'Assessing midfielder similarities using machine learning', *Carey Analytics*, May 26, 2020. <https://careyanalytics.wordpress.com/2020/05/26/assessing-midfielder-similarities-using-unsupervised-machine-learning/> (accessed May 10, 2023).
- [16] 'In-game behaviour analysis of football players using machine learning techniques based on player statistics'. <https://journals.sagepub.com/doi/epub/10.1177/1747954120959762> (accessed May 14, 2023).
- [17] J. Fan, J. Lee, and Y. Lee, 'A Transfer Learning Architecture Based on a Support Vector Machine for Histopathology Image Classification', *Applied Sciences*, vol. 11, no. 14, Art. no. 14, Jan. 2021, doi: 10.3390/app11146380.
- [18] J. Zivkovic, 'worldfootballR_data'. Apr. 26, 2023. Accessed: May 09, 2023. [Online]. Available: https://github.com/JaseZiv/worldfootballR_data/blob/665084c8422024a374c009f0799c25e6ead4aa89/raw-data/fbref-tm-player-mapping/output/initial-match/fbref_to_tm_up_to_20-21.csv
- [19] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, 'The KDD process for extracting useful knowledge from volumes of data', *Commun. ACM*, vol. 39, no. 11, pp. 27–34, Nov. 1996, doi: 10.1145/240455.240464.
- [20] 'K-Means Clustering Algorithm - Javatpoint'. <https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning> (accessed May 12, 2023).
- [21] 'Yuk Kenali Apa itu Algoritma K-Nearest Neighbors (KNN)', Sep. 16, 2022. <https://www.trivusi.web.id/2022/06/algoritma-knn.html> (accessed May 12, 2023).
- [22] 'Scikit-learn SVM Tutorial with Python (Support Vector Machines)'. <https://www.datacamp.com/tutorial/svm-classification-scikit-learn-python> (accessed May 12, 2023).

- [23] *Transfer Learning in Deep Learning - Introduction to Deep Learning*, (Sep. 12, 2022). Accessed: May 12, 2023. [Online Video]. Available: <https://www.youtube.com/watch?v=pFQ2qVqsLvg>
- [24] 'Principal Component Analysis (PCA) Explained | Built In'. <https://builtin.com/data-science/step-step-explanation-principal-component-analysis> (accessed May 10, 2023).

10.0 Appendices

This section should contain information that is supplementary to the main body of the report.

10.1. Project Proposal



National College of Ireland

Project Proposal

Identifying Similar Players Across the World's Top Football Leagues Through Profiling Based on Performance Data

30/10/2022

BSc in Data Science

2022/2023

Thomas Cannon

X19405504

X19405504@student.ncirl.ie

Contents

1.0	Objectives.....	42
2.0	Background	42
3.0	State of the Art.....	43
4.0	Data	43
5.0	Methodology & Analysis	44
6.0	Technical Details	46
7.0	Project Plan	46
8.0	References	48

11.0 Objectives

The aim for this project is to retrieve data from FBREF a 3rd party football performance data source for analysis and to try and predict a player's position while also identifying similar players based on the performance data. A number of machine learning methods and technologies will be analysed to identify the most accurate method of predicting positions as well as the similarities. Data from FBREF [1] relating to the most recent season of football from the top 11 leagues around the world to be combined in a dataset which contains over 100 statistics pertaining to each individual players performance.

It is hoped that learnings from this project could then be used in the form of application which could be used to leveraged to identify talent from other leagues or even applied to data from lower-level leagues to identify lesser-known players. This system could be used by professional football clubs and scouts to aid in the process of scouting and analysing transfer prospects.

This project will make use primarily of Python code to scrape, compile, analyse, model and to evaluate the findings from the project which will be detailed in a comprehensive report.

12.0 Background

Statistics and data are tools which are used constantly by large and medium sized businesses and companies. These companies leverage data on a daily basis to make important business decisions. Sales strategies, marketing campaigns, user experience, services and product improvements are examples of areas which companies leverage data to make decisions in these areas.

These companies have spent thousands and sometimes even millions to create infrastructures to store, handle and gain insights from data they collect.

However, with the growth of money spent by sports teams, it may come as a surprise to many people that the use of data in sport is only a relatively new addition and still in it's infancy.

It is widely regarded that Billy Beane, who was General Manager of the American Baseball team Oakland Athletic in 2002, was the first to leverage the analysis of players previous performance data in the recruitment process for a sports team. Billy established metrics in which he would base the recruitment of players on and retrieved huge backlash from

baseball fans and experts because he would not sign big names for the team and instead sign players who had performed at a high level which usually went unnoticed.

Data in sport has progressed from this time and football teams have started to create large data teams which are used in their recruitment and on-field *strategies*.

With my passion for data and football I decided to combine the two and to work on a project to create a tool which is used to profile similar players based on previous performances.

13.0 State of the Art

Profiling football players is an extremely important part of the recruitment process for all football clubs. If a club is looking to fill a certain position in their team, they want to ensure the player they sign is the right player to fill the role required.

For years, managers and club staff identified these profiles through what is widely known in football as the 'eye test'. An experienced scout would be commissioned to go and watch matches in various leagues and to find players to fit the role through watching them play. Nowadays, teams have transitioned to reviewing previous performance data to identify these players before going to watch to narrow down the player pool in which they are reviewing. With this in mind, there have been a number of articles written by analysis websites and freelance analysts in which they have created projects to identify these roles.

[An article by AnalyticsFC \[2\]](#) summarised the problem with Conor Gallagher's recent departure from Crystal Palace. They wrote about seeking to identify a replacement who plays a similar role to Gallagher at Crystal Palace.

[Various articles by Mark Carey \[3, 4, 5\]](#) outline a project he worked on between 2018-2020 where he created a machine learning model and dashboard to identify similar midfielders based on profiling of similar metrics. This project hopes to improve on these projects to provide more up to date data and also to provide profiling for all positions rather than just limited to Mark's analysis of midfielders while also expanding the player pool to a much larger collection of footballing leagues

14.0 Data

The data necessary for a project of this nature requires a large number of rows/instances as well as a sufficient number of metrics which are relevant to the objective of the project. I believe that the dataset which will be used for this project satisfies these requirements.

The dataset is made up of tabular data which include 3,000 rows/instances in the form of players who play in Europe's top five leagues. These rows also contain around 140+ player metrics across a number of selected seasons. These metrics pertain to all aspects of the game. Metrics are related to Standard Stats, Goalkeeping Advanced, Goalkeeping, Shooting, Passing, Pass Types, Goal and Shot Creation, Defensive Actions and Possession.

The data which is intended to be used in this dataset is from the football statistics website FBREF [1]. This website is a football data website which contains data collected for a host of various leagues around the world. This data is open source and can be scraped using web scraping tools with permission from the company.

The project will make use of Python code containing scraping packages such as **Requests** and **Beautiful Soup** to scrape and compile performance data from the top five football leagues for a number of previous seasons to create a web application which can be used by a user to identify similar players in the top five football leagues based on profiles which will be created using selected metrics from previous performances.

Permission was received for the use of this data from FBREF as per the email which is linked in the ethics report. The email states that the project has permission for the use of the data from the website as long as there is citation and credit given to the website for the data and that the project is used solely for academic purposes.

These metrics will be compiled in to one large Pandas data frame which will be used for the bulk of the data analysis.

The majority of the metrics in this dataset are not relevant to every position. For example, advanced goalkeeping is not a relevant metric for a forward. For this reason, a preliminary analysis will need to be conducted on the data to analyse the most relevant metrics for each part of the project.

It is proposed to also select features which only relate to a player's intent, such as attempted passes, attempted shots, attempted dribbles etc. rather than the outcome of the decision. It is thought that this is a more accurate measurement of a player's profile and tendencies as opposed to including outcome data such as goals and assists.

15.0 Methodology & Analysis

The methodology which was followed throughout the course of this project is the **Knowledge Discovery in Databases (KDD)** data mining model. KDD is a programmed and analytical approach to model data from a database to extract useful and applicable 'knowledge'. This methodology is used in tasks related to machine learning which aim to use grammatical and analytical procedures to retrieve knowledge from a database for use in

the real world.

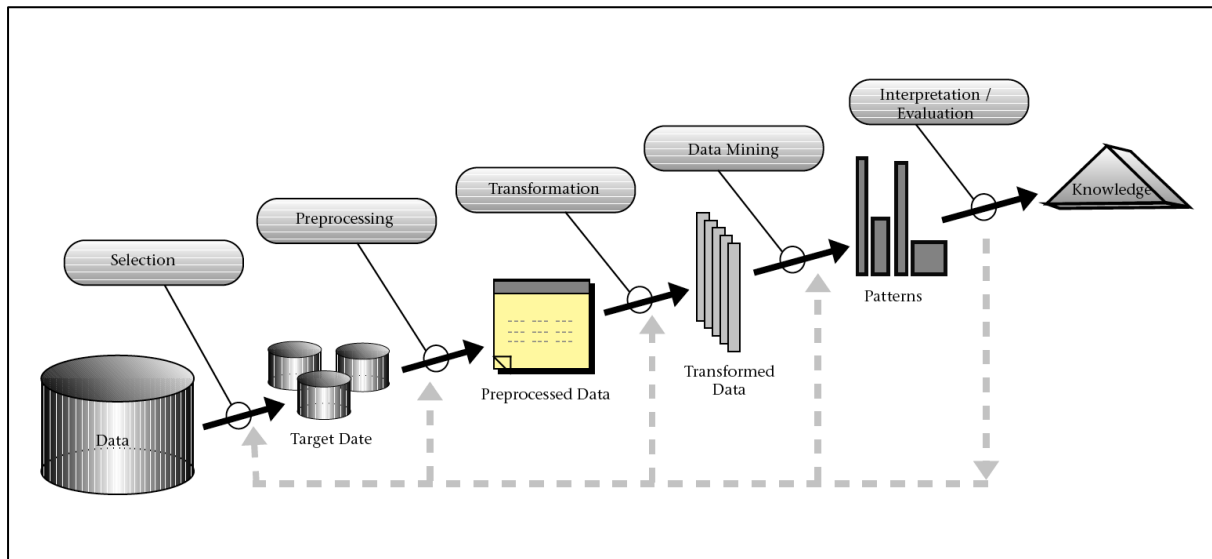


Figure 36: KDD Methodology Steps

The figure above shows 5 key steps involved in the KDD methodology:

- **Selection:** The data which has been collected for the purpose of the project first needs to be analysed and the relevant data must be selected and segregated into meaningful sets based on availability, accessibility importance and quality. These parameters are critical for data mining because they make the base for it and will affect what kinds of data models are formed.
- **Cleaning and Pre-processing:** The step of pre-processing and cleaning involves identifying and dealing with missing or noisy data as well as removing redundant or low-quality data from the data set in order to improve the reliability of the data and its effectiveness.
- **Transformation:** The step of transformation of data prepares the data to be fed to various data mining algorithms. Hence, the data needs to be in consolidated and aggregate forms. Various popular techniques are used for this purpose depending on the data and on what data mining algorithms shall be used.
- **Data Mining:** The step of Data Mining is the most important step of the KDD methodology. This is where algorithms are used to identify and extract meaningful patterns from the transformed data, which help in prediction models.
- **Interpretation / Evaluation:** This step is taken to analyse the results of the models employed in the previous step. Any modifications should be made if necessary to the previous steps to try and improve the accuracy of the models based on the results retrieved from this step. If the results retrieved from this step are acceptable for the purpose of the project, the findings can then be documented and presented as necessary.

16.0 Technical Details

There are a number of processes that will need to be implemented during the course of this project. Python code will need written to automatically scrape the code from the website and write it to a Pandas data frame. Python code will then be used to clean the data frame. It has already been identified that the data which will be scraped from the source is not clean. Columns must be re-named and standardised and in some cases new columns will need to be created with mathematical equations to standardise the data for the purpose of the project.

The data must then be analysed to identify any missing values, any problems or inconsistencies in values or any outliers which cause potential risk to the analysis.

It is then intended to make use of dimensionality reduction to reduce a collection of specific metrics to a reduced axis plane. It is intended that **Principal Component Analysis (PCA)** will be the dimensionality reduction method to be used in this project. The metrics will be analysed and selected based on the preliminary analysis to identify the most important metrics for use in the PCA.

A **K-Nearest Neighbour (KNN)** algorithm will then be used to select the nearest point on the reduced axis plane to identify the most similar player to the target player statistically. Through this, the model will identify players with similar statistical profiles from the dataset.

It is intended to expand the analysis to use other more complex methods such as transfer learning and artificial neural networks to attempt to increase the accuracy of the predictions.

17.0 Project Plan

With this proposal due on 30/10/2022 (week 7), I propose to start work on the project as soon as feedback is received. There are 7 weeks between the due date for this proposal and the due date for the Preliminary Analysis/Mid-Point Review 20/12/2022 (week 14). In this space in time there needs to be three steps and activities to be completed in this time.

Data Collection (week 8-10): As stated previously in earlier sections of this proposal, the project will make use of Python code containing scraping packages such as **Requests** and **Beautiful Soup** to scrape and compile performance data from the top five football leagues for a number of previous seasons to create a web application which can be used by a user to identify similar players in the top five football leagues based on profiles which will be created using selected metrics from previous performances.

It is expected that this process will take around 2 of the allocated 7 weeks for this portion of the project.

Data Cleaning (week 10-12): The data must be cleaned and prepared in a way that it can be inputted in to a model which will be created later. The data should also be cleaned to create an analysis on the data to identify metrics which are most relevant to be used in a model. It

has already been identified that the data which will be scraped from the source is not clean. Columns must be re-named and standardised and in some cases new columns will need to be created with mathematical equations to standardise the data for the purpose of the project.

It is expected that this process will also take around 2 of the allocated 7 weeks for this portion of the project.

Preliminary Analysis(week 12-15): A Preliminary Analysis will need to be completed to identify key metrics which are important for the model as well as removing metrics from the dataset which will not be important to the model. Some visualisations will be created during this Preliminary Analysis to visualise the data easier and to explain the findings.

It is expected that this process will also take around 3 of the allocated 7 weeks for this portion of the project. An extra week will be allocated to this process when compared to the other 2 processes to be completed in this period as it is a key part of the project and it is important that sufficient time is given to provide a thorough analysis of the dataset.

This will lead the project to the mid-point review along with the Preliminary Analysis. The work that will have been completed in this time will then be submitted for review.

This will take the time frame to week 15 where the work on the remaining parts of the project will be completed.

Principle Component Analysis (PCA) (week 15-20): It is then intended to make use of dimensionality reduction to reduce a collection of specific metrics to a reduced axis plane. It is intended that **Principal Component Analysis (PCA)** will be the dimensionality reduction method to be used in this project. The metrics will be analysed and selected based on the preliminary analysis to identify the most important metrics for use in the PCA.

It is expected that this process will also take around 6 of the allocated 21 weeks for this portion of the project.

Modelling (week 20-26): A **K-Nearest Neighbour (KNN)** algorithm will then be used to select the nearest point on the reduced axis plane to identify the most similar player to the target player statistically. Through this, the model will identify players with similar statistical profiles from the dataset.

It is expected that this process will also take around 6 of the allocated 21 weeks for this portion of the project.

Experimentation and Evaluation of More Complex Models (week 27-32): It is intended to spend this time experimenting and analysing the performance of more complex models which have been created to try to increase the accuracy of predictions. The models will also be evaluated through the use of a survey presented to around 10 participants who will be asked to choose which outputs for the models appear to be most accurate from their own knowledge of the sport.

18.0 References

- [1] <https://fbref.com/en/>
- [2] <https://analyticsfc.co.uk/blog/2022/04/19/the-key-to-consistency-tracking-footballers-statistical-profiles-across-seasons/>
- [3] <https://careyanalytics.wordpress.com/2020/05/26/assessing-midfielder-similarities-using-unsupervised-machine-learning/>
- [4] <https://careyanalytics.wordpress.com/2020/04/30/a-closer-look-into-european-midfielder-playing-styles/>
- [5] <https://careyanalytics.wordpress.com/2018/02/22/quantifying-player-profiles-the-evolution-of-the-full-back/>

18.1. Ethics Approval Application (only if required)

Appendix II

Journal of Statistics Education: http://jse.amstat.org/jse_users.htm

JSE Copyright and Usage Policy

Unlike other American Statistical Association journals, the Journal of Statistics Education (JSE) does not require authors to transfer copyright for the published material to JSE. Authors maintain copyright of published material. Because copyright is not transferred from the author, permission to use materials published by JSE remains with the author. Therefore, to use published material from a JSE article the requesting person must get approval from the author.

National College of Ireland

**Ethical Guidelines and Procedures for Research involving
Human Participants**



SEPTEMBER 2017

18.2. Reflective Journals

19.0 Supervision & Reflection Template

Student Name	Thomas Cannon
Student Number	X19405504
Course	BSc Data Science
Supervisor	William Clifford

20.0

21.0 Month: October

What?

The work I have completed for my project during this month is the planning and mapping out of tasks which will need to be completed during my project. I completed and uploaded the official written proposal for my project on Wednesday 26th of October. This proposal and plan will help to guide me through the project as I complete the work over the next months.

So What?

The work I completed this week will be a very important step of the project. It is imperative to any project to first establish a plan and to map out the steps which will be taken during the duration of the project to create timeframes which you expect work to be completed. From this starting point, I can now start to experiment with my dataset and start to work on the preliminary analysis and assess any cleaning which needs to be processed on the data. From that point I can then proceed to creating my machine learning model.

Now What?

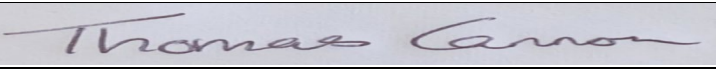
What can you do to address outstanding challenges?
Once I have met with my project supervisor and we have discussed my idea more thoroughly, I can then start work on these challenges and hopefully progress more in to the technical side of work of the project now that the planning phase has been completed.

Student Signature

Supervision & Reflection Template

Student Name	Thomas Cannon
Student Number	X19405504
Course	BSc Data Science
Supervisor	William Clifford

Month: November

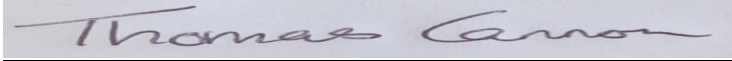
What? The work I have completed for my project during this month is the preliminary analysis of the datasets. Originally the plan was to only make use of the player performance data from FBREF, but it was identified that more data was required for the project in the form of players specific positions which can be retrieved in the form of a Transfermarkt dataset. The two datasets were merged on using the common column of player names.	
So What? The preliminary analysis and identification of another dataset which will be useful to the project will be very helpful to create accurate machine learning models. These datasets will add more information to add to the datasets which will increase the quality of the data and hopefully increase precision of the clustering models.	
Now What? In the coming weeks I will start to build clustering classification models to predict the position of players in the dataset which will in turn identify the importance of features to different positions/roles.	
Student Signature	

Supervision & Reflection Template

Student Name	Thomas Cannon
Student Number	X19405504
Course	BSc Data Science
Supervisor	William Clifford

Month: December

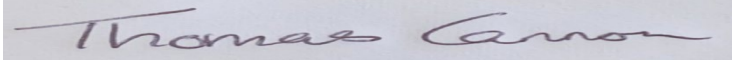
What? The work I have completed for my project during this month is to create a prediction model using a Support Vector Machine. The performance of this model will be compared to the performance of the K-Nearest Neighbour model.	
So What? It is helpful to trial and test as many models as possible for the purpose of this analysis to ensure that a model with the best performance is chosen to make accurate predictions	
Now What? I will wait to hear the grade and feedback of the work which I have completed so far and hope to use the feedback to make changes to the plan if needed and to progress with the rest of the project for the next semester.	

Student Signature	

Supervision & Reflection Template

Student Name	Thomas Cannon
Student Number	X19405504
Course	BSc Data Science
Supervisor	William Clifford

Month: January

What?	
The work I have completed for my project during this month is working on creating an accurate Support Vector Machine model in hopes to make accurate predictions on a players position category and subposition.	
So What?	
This model will add to the K-Nearest Neighbour Classifier which was created previously and their performances can be compared to identify the most accurate model for the purpose of the project.	
Now What?	
This Support Vector Machine will be used to identify the probabilities for each player for each position to create a high level vector space to use similarity identification methods to identify the most similar players.	
Student Signature	

Supervision & Reflection Template

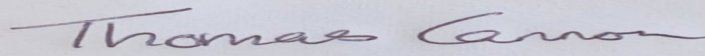
Student Name	Thomas Cannon
Student Number	X19405504
Course	BSc Data Science
Supervisor	William Clifford

Month: February

What?	
The work I have completed for my project during this month is the creation of the dashboard on which the models will be used in the background. Some testing was done on some designs to identify the most desirable layout for said dashboard. The application takes a user input in the form of a player and returns the most similar players to the input based on the parameters set by the user.	
So What?	
This interactive dashboard will bring a real world use the analysis and could be used for the purpose of profiling most similar players.	
Now What?	

From this point more work will be carried out to improve the performance of the background models as well as analysis into the most important features used in the Principle Component Analysis.

Student Signature



Supervision & Reflection Template

Student Name	Thomas Cannon
Student Number	X19405504
Course	BSc Data Science
Supervisor	William Clifford

Month: February

What?

The work I have completed for my project during this month is the completion of the preliminary analysis of the datasets. I then spent the rest of the month writing the mid-point submission report which I submitted for grading.

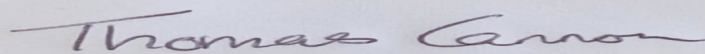
So What?

This mid-point submission is an important deliverable and holds a large weighting of the marks for the overall project.

Now What?

I will wait to hear the grade and feedback of the work which I have completed so far and hope to use the feedback to make changes to the plan if needed and to progress with the rest of the project for the next semester.

Student Signature



Supervision & Reflection Template

Student Name	Thomas Cannon
Student Number	X19405504
Course	BSc Data Science
Supervisor	William Clifford

Month: March

What?

The work I have completed for my project during this month is improving on the Support Vector Machine Classification model to create a high level vector space on which to run more clustering models in the hopes of improving on previous results.

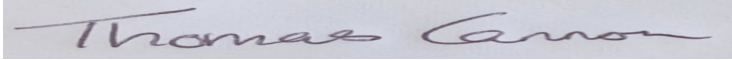
So What?

This is a key step of the project as it seeks to improve on results gained from the original models. With the high-level vector space created it is hoped that the clustering models will be able to achieve a better accuracy than the raw data points.

Now What?

Testing must now be completed with the new vector space to see what accuracy can be achieved with the new data points.

Student Signature



Supervision & Reflection Template

Student Name	Thomas Cannon
Student Number	X19405504
Course	BSc Data Science
Supervisor	William Clifford

Month: April

What?

The work I have completed for my project during this month was the administration work for the questionnaire portion of the project. 20 participants were contacted and invited to fill out a questionnaire where each participant chooses 10 football players from the worlds top 11 footballing leagues that they feel they are most familiar with, to be then asked questions based on the accuracies of the 2 models to understand which the participants feel is more accurate.

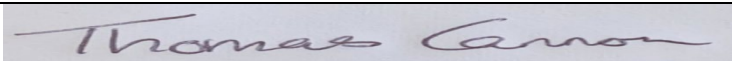
So What?

This feedback from participants will be important to the project to understand if people who are interested in football can identify the 'improved' model simply by using the eye test to determine their opinions on players similarities.

Now What?

The feedback retrieved from this questionnaire will be analysed using statistical tests to determine if there is a significant difference between the participants choosing a certain model over the other. This can be used in the results section of the report and will be a key part in understanding if the newer improved model is more accurate or not.

Student Signature



21.1. Other materials used

Any other reference material used in the project for example evaluation surveys etc.

Google Form: Targeted Questionnaire:

Final Year Project: Beyond the Eye Test:

Improving Football Recruitment Through The Use Of Clustering And Support Vector Machines

This project aims to create the most accurate model for identifying most similar players based on performance data collected from league games from the top 12 footballing leagues during the 2021/22 season. Participants will receive 2 outputted tables of 5 players related to each selected

player. One of an original model and a proposed improved model. Participants must choose which output table they feel most accurately chooses players similar to their selected player.

Please select the 10 players who you understand their playstyle and tendencies best. Do not select players you may be confused about, such as position or role.

Note: Players are not deemed similar based on overall skill or ability but in game tendencies and playstyle. For example, players with tendencies to pass more than they will shoot will be weighted more similarly.

Please select the leagues in which you wish the algorithm to choose players from for the output. Please only select leagues you are familiar with as this algorithm will return players most similar to the chosen target player based on this input. **At least 1 league MUST be ticked below:**

* Indicates required question

1. Participants Name (This is required to return the tables at a later date and will * not be stored or used beyond that point):
2. Which leagues would you like to be included in your output? *

Check all that apply.

- Premier League (England)
- La Liga (Spain)
- Ligue 1 (France)
- Serie A (Italy)
- Bundesliga (Germany)
- Championship (England)
- Major League Soccer (USA) Eredivisie (Netherlands)
- Primeira Liga (Portugal)
- Liga MX (Mexico)

Campeonato Brasileiro Série A (Brazil)

Player Selection

Please select the 10 players whos role and tendencies you understand the best. You will be asked to judge whether other players are similar or not in role and tendencies to your selections.

3. Player 1: *

4. Player 2: *

5. Player 3: *

6. Player 4: *

7. Player 5: *

8. Player 6: *

9. Player 7: *

10. Player 8: *

11. Player 9: *

12. Player 10: *

This content is neither created nor endorsed by Google.

Google Forms

Participant Permission Form template:

Final Year Project: Beyond the Eye Test: Improving Football Recruitment Through The Use Of Clustering And Support Vector Machines

Plain Language Statement and Consent Form

You are being invited to take part in a research study. Before you decide it is important for you to understand why the research is being done and what it will involve. Please take time to read the following information carefully and discuss it with others if you wish. Please don't hesitate to get in contact if there is anything that is not clear or if you would like more information. Take time to decide whether or not you wish to take part.

This project aims to use machine learning methods and analysis to identify similarities between professional football players in the world's top 12 footballing leagues. Two models were created in the hopes of finding the most accurate model for this purpose.

The purpose of this questionnaire is to gain participants feedback on the accuracy of the algorithms created during the course of this final year project. It is pitched to participants to decide, from their background knowledge of the game, which algorithm is the most accurate in identifying similar players.

This project is being undertaken as part of a Final Year Thesis Project for the BSc in Data Science programme at the National College of Ireland.

Willing participants will be given a questionnaire to gain feedback on which algorithm they have decided is most accurate in achieving the goal.

It is up to you to decide whether or not you wish to take part. If you give your permission, you are still free to withdraw at any time and without giving a reason.

All information which is collected during the course of the research will be kept strictly confidential. However, it should be noted that no participants' personal information will be gathered or stored throughout the course of the project.

The data from this research study will be anonymised and securely stored on the researchers' NCI OneDrive in passphrase protected folders for secondary analysis.

The results are due to be published in 2023. In the publication, the data will be anonymised, and no individual participant will be identifiable from the research.

This project has been reviewed by the NCI Ethics Committee. If you (participants) have any concerns regarding the conduct of the research project, and/or you would like to discuss this further, please email x19405505@student.ncirl.ie to speak with the researcher.

By signing and returning this consent form you are indicating your agreement with the following statements:

I give my consent to participate in this research study.

- I understand that even if I agree to participate now, I can withdraw at any time or refuse to answer any question without any consequences of any kind.
- I have had the purpose and nature of the study explained to me in writing and I have had the opportunity to ask questions about the study.
- I understand that I will not benefit directly from participating in this research.
- I understand that all information provided for this study will be treated confidentially.
- I understand that signed consent forms and the anonymised data will be retained by the researchers on the NCI Staff network in passphrase protected folders.
- I understand that I am free to contact any of the people involved in the research to seek further clarification and information.
- I understand that I give permission for material/data to be stored for possible future research unrelated to the current study without further consent being required but only if the research is approved by a Research Ethics Committee.

Signature-----

Print name-----