National College of Ireland

BSC in Data Science

Computing Project (BSHCSD4)

2022/2023

Daragh Carroll

X19477436

X19477436@student.ncirl.ie

Classification of Goalkeeper Dives

Data Analysis Report

Contents

Table of tables:

Table of figures:

Executive Summary

Performance analysis plays a major role in helping professional footballers compete at the highest level. Through technological advances such as GPS trackers, data relating to player performance during matches and training sessions can be gathered and clubs now need a way of leveraging this data. One such area relates to the analysis and classification of goalkeeper actions, to create individual training programmes and improve performance.

The goal of this study is to investigate whether machine learning techniques can be a used to assist in classifying goalkeeper actions, firstly as dives/non-dives and then based on the nature of those dive e.g. (Left High Dive).

A dataset built on GPS tracker data and video footage from goalkeeper training sessions, at multiple clubs across Europe and America, together with goalkeeper characteristics and weather data, which had been manually classified provided a rich dataset on which to train and test machine learning models.

A Crisp-DM approach was implemented to train, test and validate a comprehensive set of classification models including, logistic regression, k nearest neighbours, decision trees, support vector classification and random forest. Each model was run iteratively on the dataset with different levels of pre processing applied. Cross validation was used to validate the models.

The study found that the best model for binary classification of goalkeeper actions was the random forest, with minimal pre-processing techniques applied to the dataset (Cross-Val = 0.86, Accuracy = 0.84, Precision = 0.84, Recall = 0.84, F1-score = 0.84). For the subsequent multi-class classification of those dives by type the random forest was the best performing model, with minimal pre-processing techniques applied to the dataset (Cross-Val = 0.81, Accuracy = 0.81, Precision = 0.83, Recall = 0.82, F1-score = 0.81). The results show that machine learning techniques can be successfully implemented on goalkeeper performance data, with high levels of accuracy.

1) Introduction

1.1. Background

Performance analysis tracking in sports has grown significantly in recent years, thanks to advances in technology and the resulting increased availability of data. In the past, performance analysis in sports was limited to simple summary descriptive statistics and charts, such as counting goals or tracking the number of passes. However, sports teams and organizations now have access to a vast array of tools for capturing, tracking and analysing performances. One major area of growth has been the use of wearable technology, such as GPS tracking devices, heart rate monitors, and accelerometers.

GPS trackers are now widely used across a variety of sports, including football, rugby, athletics, and cycling. They allow coaches and athletes to track a wide range of performance metrics and provide real-time data on athletes' physical performance, including distance covered, speed, and intensity of effort. This data can be used to identify areas for improvement, optimize training programmes (creating individual targeted training / recovery programmes), and monitor athletes' recovery.

One of the key developments in GPS tracking technology has been the increasing accuracy of the devices. Early GPS trackers had relatively low accuracy, leading in some instances to unreliable data. However, modern GPS devices now have much higher levels of accuracy, allowing coaches and analysts to get a more detailed understanding of an athlete's performance. Now that teams have an abundance of highly accurate data available on their athletes, they wish to leverage the data to gain a deeper understanding of their players performances and make data-driven decisions that can help improve performance and prevent injuries. The use of data science and, in particular machine learning techniques in interpreting data is growing rapidly.

Examples of studies where machine learning techniques have been applied include: A study by Al-Ashida, M. (Al-Asadi, 2018), into the use of machine learning algorithms to create intelligent decision support systems for team management. In this study the author attempts to develop a method to select a teams optimum starting 11 as well as identifying each players preferred available position. The author found that he could accurately classify the general role of each player (defensive or attacking) but struggled to classify their position due to a “lack of available data”. With the large volume of highly accurate data available to clubs now this may be an approach that could be developed further.

A similar study conducted by Cwiklinski, B et al. (Ćwiklinski et al., 2021), investigated whether machine learning approaches can be implemented to help clubs with team management and transfer strategies. The authors identified multiple parameters for player assessment and defined three measures of a successful transfer. With the use of random forest models and pre-processing they were able to develop a system with ‘promising results’ that could be used to ‘support a scout or a team manager in the process of transfer planning.’.

These studies show that the high volumes of data currently available to teams can be leveraged in a wide variety of ways to help improve the clubs. For these reasons I believe that the current study into the potential use of machine learning methods on GPS tracker data is highly interesting and holds a lot of promise.

This study utilizes work done during my time on work placement at STATSports. While there, I was given an insight into the applications of machine learning in sport and how teams use GPS performance analysis trackers to make data driven decisions about their players. One of my tasks during this placement was to compile and manually classify data from goalkeeping sessions at a variety of clubs. The output from this task was a large database containing various metrics on goalkeeper actions and the classification of actions as dives/ non dives and then subsequent classification based on the nature of those dive. I was able to run some basic modelling techniques at the end of my placement on the dataset but felt that the natural next step for the data was a comprehensive analysis using a variety of machine learning classification models. STATSports permitted me to use this dataset, which has formed the basis for this analysis, I hope that the results of my study will be beneficial to them.

1.2. Aims

As stated previously in this report, the aim for this study is to assess whether classification models can be a useful tool for goalkeeper performance analysis within football. In order to achieve this goal, I have chosen to apply machine learning classification techniques to GPS tracker data from goalkeeping sessions. By training a variety of models on a dataset comprised of metrics gathered by the STATSports Apex GPS tracker the study will attempt to create two highly accurate classification models I for the binary classification of goalkeeping actions (Dive/ Not Dive) and ii) the multi-class classification of the type of dives made by goalkeepers e.g. High Right Dive Body. In order to assist in achieving the study aim I broke it down into four research questions:

1. Can machine learning and GPS tracker data be used to classify goalkeeper actions into dives and not dives?
2. Can machine learning and GPS tracker data be used to classify goalkeeper dives based on their direction, height and landing?
3. Can pre-processing methods such as scaling, feature selection, and Principal Component Analysis be used to improve these models?
4. Can additional data related to the sessions (weather and keeper characteristics) improve the models accuracy?

Research questions provide clarity on what we are trying to achieve and help us more clearly define the study. If successful, the study will provide teams with an efficient way of quantifying a goalkeepers session performance and identifying necessary areas of improvement. For example, if after a match the club can identify the areas where their goalkeeper was most comfortable at stopping shots through the frequency of each type

of dive in a session then they will be able to assess where to focus their training in order to make more effective dives which they made less of or conceded from.

Once the research questions had been defined, I was next able to identify research objectives which can be stepped through in order to answer questions and achieve goals. These objectives are to:

1. Create a classification model to classify goalkeeper actions gathered by STATSports GPS trackers and additional data (weather and personal) as dives and not dives.
2. Create a classification model to classify the identified dives by type based on their direction, height, and landing.
3. Implement pre-processing techniques such as scaling, feature selection, and Principal Component Analysis, to check for improvements in models.
4. Investigate whether including additional features, relating to goalkeeper characteristics and weather data can improve the model's accuracy.

By achieving these research objectives, I will be able to determine whether my study was successful in reaching its main goals.

1.3. Technology

The study makes use of a variety of technologies for all stages of the development including data collection, pre-processing, modelling, evaluation, and potentially creating a dashboard. Below I have described each of these technologies and their purpose in this study:

- STATSports Apex Athlete Series(*STATSports, 2023*):
  This is a wearable GPS tracking device designed specifically for athletes. The device is worn on the athlete's body and collects data on their performance during training and competition. It uses advanced algorithms and built-in features such as an accelerometer (measures acceleration) and gyroscope (measures rate of rotation) to track a wide range of metrics. These includes distance covered, speed, acceleration, deceleration, heart rate, jump height, jump load, and ground contact time. The device uses a multi-constellation GPS receiver to track the athlete's movement with high precision. This allows for accurate tracking of metrics which are essential for reliably monitoring an athlete's performance. The GPS trackers from STATSports were used to collect and track the data that formed the basis for this analysis from a variety of goalkeepers.

*Figure 1: STATSports Apex GPS Tracker (*STATSports, 2023*)*

- STATSports Apex Coach Series App *(STATSports, 2023*):
  Along with the wearable GPS tracking devices STATSports provides an app that allows athletes to view a variety of key metrics on their sessions data. The app provides multiple features such as live player tracking, real-time data analysis, video analysis and tagging which allows the user to tag actions that occurred during their sessions. These features allowed for the collected data in this study to be manually classified by mapping to available video footage to prepare it for use in training and testing classification models.

*Figure 2: STATSports Apex Coaching App (*STATSports, 2023*)*

- Microsoft Excel:
  After extracting the collected data from the GPS tracker excel was then used to safely compile and store the dataset. It was used as the data files extracted from the

Apex app came in a .csv format due to the application's large storage capacity, easy of use and flexible.

- Python:
  As the goal of this study is to create machine learning classification models, I chose to use python as the coding language to conduct this analysis. This is because python is a flexible and robust coding language that can be used for web development, data analysis, scientific computing, and machine learning. Its features make it the most efficient coding language for this study as it means it can be fully developed and deployed through python. As well as this Python has a vast collection of open-source libraries and frameworks which provide an extensive set of tools and algorithms for building machine learning models. These libraries simplify the development process, reducing the time and effort required for efficient and effective data preparation, model building, and evaluation.

- Pandas:
  The Pandas library is a versatile library used for data manipulation and analysis. It was used throughout the study to read in the data and apply different pre-processing methods, to prepare the data for modelling, such as handling missing data, handling duplicates, and performing descriptive statistics.

- Matplotlib:
  Matplotlib provides users with a wide variety of plotting tools and options to create high-quality charts, graphs, and other types of visualizations. It was implemented throughout the study to clearly display the structure of the dataset, the results from pre-processing methods and the accuracy of each model.

- Scikit-learn:
  This library is key to running machine learning tasks, such as classification, regression, clustering, and dimensionality reduction. It is a comprehensive machine learning library providing tools for data pre-processing, feature selection, model selection, and model evaluation. It was implemented throughout this study as it works seamlessly with the other libraries that have been necessary and makes it possible to perform complex data analysis.

1.4. Structure

This paper is organised into seven main chapters. This first chapter provides an introduction to the topic, including the research question and hypothesis and an overview of the study and document. The second chapter Data is where I describe the creation of my dataset along with additional data used and the structure of my data files. In the third chapter I present a review of the relevant literature on the topic,

highlighting previous studies that have investigated similar topics in football and goalkeeping. Chapter four identifies the methodology employed in this study and describes the steps involved. The fifth chapter describes the techniques and models used in the study including pre-processing, data analysis and modelling. The sixth chapter presents the results of the study, including statistical analyses of the data and a discussion of the findings. Finally, the seventh chapter concludes the paper by summarizing the main findings, discussing the implications for future research, and suggesting practical applications of the research in the world of football analysis.

2) Data

Next, I will describe the data that formed the basis for this study. This data consisted of the following elements:

1. STATSports data (GPS and Video)
2. Goalkeeper Characteristics
3. Weather Data

Each of these are detailed below.

1) STATSports data (GPS and video)

As previously stated, the idea for this project arose from work that I undertook during my time on work placement at STATSports. One of my task's during placement was to build a dataset based on GPS and Video data that involved manually classifying goalkeeper actions, detected by the STATSports tracker and identified using the corresponding video footage. I was permitted to carry forward this dataset to provide training and testing data for this project which aims to build and test machine learning classification models.

Construction of dataset through manual classification:

The raw data consisted of metrics recorded by GPS trackers during goalkeeper training sessions at multiple clubs across Europe and America, including Hamburger SV in Germany, Arsenal FC in England and Colorado Rapids in the USA. The raw data from each training session contained 26 different metrics gathered by the STATSports Apex Athlete Series tracker. This tracker uses advanced algorithms and built-in features such as an accelerometer and gyroscope to gather a wide variety of relevant data including the following metrics:

- Start/End Time:

  This is a timestamp of when each action began and ended during the session. They use time zones local to where the data was collected so this had to be taken into account as the data training sessions gathered occurred in different time zones.

These metrics provide a means of calculating the duration of each action, which was itself used by the built-in algorithms to calculate some of the additional metrics below.

- Peak Acceleration (x/y/z):

  These metrics refer to the maximum acceleration experienced by the goalkeeper in each of the horizontal, vertical and depth directions - axes (x/y/z) - during actions. They are calculate using the rate of change of velocity in each direction during the action with respect to the time taken. These can be used to identify sudden changes in speed or direction.

- Mean Acceleration (x/y/z):

  These measure the average acceleration in every direction experienced by the goalkeeper during each action. These provide information on the intensity and consistency of their movement during an action.

- Peak Gyro (x/y/z):

  These metrics measure the maximum angular velocity or rotation rate around each of the axes (x/y/z). These are important values for dives as they can indicate rapid changes in direction or orientation of the goalkeeper, so it helps to detect when a dive occurs. These are measured in units of degrees per second.

- Mean Gyro (x/y/z):

  These give the average rotation rate around each axis and can be useful in detecting imbalances in the user's movements. These are also measure in units of degrees per second.

- Peak/Mean Mag:

  These metrics provide the maximum and average values for the user's magnitude of acceleration. This means it can measure the total acceleration of the user in all directions using the following formula.

  Equation 1: Magnitude of Acceleration

  $$Mag_{Acclerometer} = \sqrt{{Acc_x}^2 + {Acc_y}^2 + {Acc_z}^2}$$

  These can be useful to detect the acceleration of the user in every direction and help track their movement and any rapid breaking or acceleration.

- Peak/Mean Speed:

  These are the maximum and average speeds that the user is traveling at during an action. These are helpful in identifying diving actions, as a goalkeeper's increased

speed will indicate when they are attempting to quickly close the distance to the ball and make a save.

- Peak Acceleration:

  This is measured in meters per second squared (m/s^2) and measures the highest rate of change of velocity in all directions. It can be used to measure how quickly the goalkeeper is accelerating during the dive, which can be an indicator of the action's explosiveness.

The task was to use these metrics to manually classify goalkeeper movements based on corresponding video footage that had been provided by each club for their training sessions.

The GPS Tracker constantly gathers metrics on all of the goalkeeper's actions during a session and separates them into instances. Using the STATSports Apex app I was able to access this data and corresponding video footage. The app allowed me to combine the video footage and goalkeeper data for a given session and align their time stamps. I was then able to decipher which actions in a given session corresponded to different types of movement to identify where dives occurred, the type of dive, and then classify each dive within the app.

The classifications I chose were based on 3 factors, the direction of the goalkeeper, the height of their jump and their form of landing. Each dive was assigned to one of 14 different classes as shown in the table below.

Class	Type	Side	Landing
1	High Dive	Right	Body
2			Feet
3		Left	Body
4			Feet
5	Low Dive	Right	Body
6			Feet
7		Left	Body
8			Feet
9	High Catch		Body
10			Feet
11	Block		Body
12			Feet
13	Centre		Body
14			Feet

Figure 3: Dive Classifications

Once all actions were classified within the Apex app using video footage the next step was to extract the data into .csv files, one for each session. Unfortunately, the app doesn't allow

you too extract both the data and classifications together and classifications had to be re-entered into the .csv files.

This process involved adding two new attributes in the datasets (.csv files), the first identified each action identified as a dive/not a dive (1/0) and the second recorded the dive classification based on direction, height and landing (as in figure 1). The requirement to re-enter the classifications was unfortunate, time consuming and potentially error prone. Randomly selected portions of the data were continually double checked by another member of the STATSports team to minimize the potential error and provide confidence in the manual classification of the data, but due to the large scale of the datasets it wasn't possible to check the complete dataset. Although I am confident in my work this is a potential area for errors to occur that could have a negative impact on the classification models accuracy. Having been fully classified, all the data .csv files were combined to create a single dataset, which now forms the basis of this study. This was done using a python function that utilised lambdas to identify each .csv file within a folder, store their data one at a time into an empty data frame and then create a new .csv file containing all of the data. Below is a summary of the final dataset and starting point for this study.

Dataset Name	***Size (MB)***	***Format***	***No. Attributes***	***No. Instances***	***No. Sessions***	***No. Goalkeepers***	***No. Teams***
All_Dives	1.79	.csv	29	6693	20	9	3

Table 1: Dataset Summary

Additional Data:

In order to increase the potential accuracy and complexity of the analysis I felt it was important to build in other potentially relevant factors including characteristics of individual goalkeepers and weather details, where available these are detailed below.

2) Goalkeeper Characteristics

Using information provided in the GPS Tracker dataset I was able to identify each goalkeeper involved in the study. With this information I gathered relevant characteristics on the goalkeeper's, from each club's website, that may influence their actions and potentially improve the models in their classification. These included each goalkeeper's age and height at the time of each session. The data was added to the .csv files before being incorporated into the dataset using excel functionalities. After this all identifying data relating to the keepers was removed to anonymize them.

3) Weather Data

The weather conditions during a training session can have a major impact on a goalkeeper's performance, in many ways, such as rain and fog impeding the goalkeeper's vision and reaction time. In order to see if additional information on the weather during sessions would help to classify the goalkeeper's actions and dives, I sourced in-depth metrics from

each session. These included the temperature, humidity, wind speed and a variety of other data. This additional data for my analysis was sourced from" Visual Crossings" (Historical Weather Data & Weather Forecast Data | Visual Crossing, 2023) weather data service which provided a csv file containing data from each training ground's location on the date of each session. Using Python functions, I was able to merge this data with my original dataset to assist in a more complex analysis.

From being tasked with compiling this dataset in its entirety including the manual classification of each goalkeeper's actions I gain an in depth and complete understanding of the datasets structure, and content. After my placement concluded it felt that the next obvious step for my work was to see if I could implement the machine learning techniques, I learn through my programming modules, to see if the process could be achieved accurately using machine learning. Being able to work on as a part of my studies entire data lifecycle was of huge benefit to me and my analysis.

Ethical considerations:

A key consideration of any data science project is that the study data is collected, processed, and used ethically, in order to achieve this, I applied some important ethics principles.

The most crucial principle followed was ensuring I had informed consent to use the data in this study. I obtained consent from a representative at STATSports to use this data within my study after providing them with details of my project proposal. In addition to this all the data was gathered by multiple clubs, who all gave permission to STATSports to use their data for research purposes, so it could be used within this study.

Ensuring the privacy of the individuals involved was another major ethical principal that needed to be implemented. To achieve this, all data used in the study was anonymised, with any personal information such as goalkeeper's names being removed from the dataset.

Finally, I felt it was important to consider potential bias within the dataset and attempt to prevent it. The data was gathered from multiple leagues, clubs and goalkeepers making the dataset diverse and as representative as possible.

In order to determine the appropriate methodology, algorithms, pre-processing and testing approaches, to my investigation into the use of machine learning algorithms in goalkeeper performance classification I looked into previous studies, below is a summary of some of the studies that informed my choices.

4) Literature Review

> "In football, the goalkeeper has a unique and critical role in the team which requires timed and explosive adjustments of body, speed, position and orientation in response to a stimulus." (Ibrahim, R et al., 2019). Making it critical to make use of all available information and techniques to help goalkeepers maximise their performance.

Past research has mainly focused on the psychological aspects of goalkeeping tasks, with a focus on perception and cognitive skills used when diving during penalty kicks. (Ibrahim, R et al., 2019) Other studies have also tried to incorporate basic analysis of numeric data; however the availability of performance data is leading to an increasing focus on technical and physical characteristics and skills of goalkeepers.

In their paper Ibrahim, R. et al., (Ibrahim, R et al., 2019), investigated the contribution of starting position, linear and angular momentum and legs to end-performance in goalkeeper dives. They used kinetics and full body kinematics from ten elite goalkeepers diving to save high and low balls at both sides of the goals. They conducted their analysis of the data collected in Mathlab and SPSS, with the data being presented as the mean ± standard deviation between the subjects who participated. This allowed the authors to analyse significant movements which could help coaches in determining appropriate training programmes for goalkeepers and adapt their approach to "highlight lateral skills, to both sides of the body".

In another study (Szwarc et al., 2010) the authors looked at cognitive models of goalkeeper's actions in soccer based on observations made from some matches at Euro 2008. The purpose was to try and identify efficiencies in the goalkeeper's actions. The study focused on two aims, identifying the most frequently performed actions in offensive and defensive phases, by a goalkeeper, and subsequently analysing their effectiveness and reliability. An observational method was used in this study, with the play of both goalkeepers in seven cup games being observed, recorded and analysed. The study found that most goalkeeper actions are aimed at either taking control of the pitch or retaining ball possession with only a small number of offensive actions focused on creating goal scoring opportunities. Defensive actions are typically individual and most reliable when catching the ball. The study concluded that the efficiency model of goalkeeper's actions are useful in creating models of play, in lower level competitions to help improve the effectiveness of their game play.

"The use of new technology that allows the coaching staff to record data from all players is extremely important, thus allowing for the design of specific training programmes according to the physical demands on each position" (Di Salvo et al., 2008).

This study's objective was to analyse goalkeeper activities throughout a match to identify the distances covered at different velocities in 1st and 2nd halves and to analyse the time and frequency of these activities, which for goalkeeping should be short distances. Data was collected from 62 goalkeepers across 28 teams in the premier league using cameras to capture the goalkeeper's movement across 109 matches, data was collected on distance covered at different speeds and also the number of activities.

Their system used specifically developed software to obtain the time spent at different intensities and the frequency of occurrence of these activities. The authors used SPSS to carry out a statistical analysis of the data, with summary statistics including - averages and standard deviations. Bi variate correlations were examined using

Pearson's product momentum correlations and the mean differences were analysed using student t-tests.

Their results were tested using a variety of metrics and they concluded that there were no real differences identified between the distances covered in the two halves and that goalkeeper's walked during 73% of the match with 2% at high intensity. Goalkeeper's physical activities are not as great as other players but high-intensity actions are very decisive in the final results of the match.

The ever increasing availability of performance data combined with new technologies is leading to an increased focus on the use of more advanced statistical and machine learning algorithms.

In a recent study (Hosp BW, 2021) the authors presented a model for the recognition of goalkeepers expertise when making decisions, in build-up game situations by means of machine learning algorithms, that rely solely on eye movements. The study used VR headsets with integrated high speed eye trackers. The steam VR framework was used to create highly immersive realistic scenes using footage from German league games. These scenes were displayed to 12 experienced players to assess the perceptual skills of these players in an optimized manner. Using gaze data their experience was classified using a support vector machine model and then validated in 2 steps, cross-validation and leave-out validation. The results showed that eye movements contain highly informative features and enable classification of goalkeepers between experienced, interim and novice.

As mentioned previously, an earlier study (Al-Asadi, 2018), focused on using machine learning algorithms to build intelligent decision support systems for team management. Models were built on player skills (technical, physical and psychological) and aimed to identify the best available squad, with positions. In order to help with identifying player positions and predicting their progression(in terms of dribbling) the author introduce data from the football videogame FIFA. To achieve his study goal the author applied a suite of machine learning techniques including linear and logistic regression, random forest, neural networks and k Nearest neighbours. He also employed a recursive feature elimination algorithm (with REF and PCA) to reduce the data dimensionality. The author found that he could accurately classify player position in a binary classification (defensive or attacking) but struggled to classify their exact position with any model due to a "lack of available data". The random forest model proved to have the highest accuracy in this study at predicting a players progression in dribbling skills.

A study by Cwiklinski, B et al. (Ćwiklinski et al., 2021), investigated whether machine learning approaches can be implemented to help clubs with team management and transfer strategies. The data used was on match statistics, transfers and the transfer market, and which were drawn from a number of different websites to build a database. The authors also defined three different measures of a "successful" transfer. A train test split was applied to the data as well as 3-fold cross validation which was used to prove

the lack of dependency on the dataset. A number of machine learning techniques were applied to this data including Random forest, Naïve Bayes and an Adaboost. With the use of random forest models and pre-processing they were able to develop a system with "promising results" that could be used to "support a scout or a team manager in the process of transfer planning.".

In his study to utilize machine learning techniques in football predictions Duarte, R. (Duarte, 2022) investigated whether machine learning can accurately gauge the winner of a football match between two Premier League teams. Three research objectives were identified:

1) create a model that can predict the result of any given Premier League game using standard in-game statistics,

2) improve the model by incorporating additional statistics and psychological factors,

3) assess whether the best-performing model from Objective 2 can be used to predict the results of the most recent Premier League season.

A Crisp-DM methodology was used to provide structure to his analysis of eight different models the kNN model was found to be the most accurate. The first objective was viewed as highly successful, with an accuracy metric of approximately 85% for the kNN model. Objective 2 was also deemed successful, with some models seeing an increase in accuracy and others seeing a decrease, but the kNN model still performed the best with an accuracy of 75%. Objective 3 was also a success, as the kNN model achieved an accuracy of 72.37% across the whole season.

The author's final model accurately predicted the close fight at the top of the league, who would be relegated and in what order they would be relegated in, and it successfully deduced who would finish where and what other close battles would be fought. The author concluded that a machine learning model can accurately gauge the winner between two Premier League teams. However, it is important to note that football is still a game of chance, and the accuracy of the final model was 75%.

Another study by Herbinet, C (Herbinet, 2018) focused on predicting scores looked at using an alternative data analysis method to predict the results. Typically most score prediction models use the number of goals scored and conceded by teams as the base values to train their models but this can be unreliable due to a large random factor in scoring a goal. A method of using more complex data generated by a combination of regression and classification algorithms in order to evaluate a team's performance was proposed. A metric called "expected goals" was introduced, it calculates the number of goals a team should have scored based on the likelihood of their shots and chances created resulting in a goal. Along with this the algorithms also calculated the 'ELO rating', which is a method of rating how well a team has been performing. These metrics were used as inputs for the studies classification model (for match outcomes) and regression model (for match scores).

The classification model had a predictive accuracy of 0.511 and a F1 score of 0.382, meaning it correctly predicted the outcome of a match in more than 50% of games. When using actual goals instead of expected goals metrics, the accuracy dropped to 0.496 and the F1 score became 0.361. The regression model had an RMSE value of 1.153, MAE value of 0.861. Therefore, the model's predictions were on average 0.861 goals away from the reality. Similar to the classification model when using actual goals instead of expected goals, the performance worsened. When compared to other benchmark models such as bookmakers models the studies performed just as well, returning a similar accuracy and F1 score for both models. Overall, the study showed that using expected goals in the models generates better predictive performance.

A study's by Mainsh. S, et al looked at "the prediction of football player performance through machine learning and deep learning algorithms" (Manish. et al., 2021). The studies aim was to improve the player's performance prediction through the use of machine learning and deep learning algorithms by considering previous session data, strength, and weaknesses of players.

In this research paper the authors implemented and compared machine learning and deep learning algorithms to predict the player's performance. Taking into account the player's position, they built models for the four different positions (forward, midfield, defence and keeper). The study used multiple regression, neural network, Xgboost regression, and support vector regression models in order to determine which would be best suited for this task. The results of the study showed that Multiple Regression outperformed the other models returning a higher accuracy and lower rate of error across all 4 positions.

The final study described (Herold, M, 2019), involved a critical review of applications of machine learning across a variety of studies, and also discussed the current challenges and possible future directions. It critically reviewed thirty-one different papers linked to machine learning analysis on professional football data. These papers looked at a variety of ways to use machine learning on attacking play data such as, evaluating a team's strengths and weaknesses through pattern recognition, or quantifying the relation between performances and success based on goal difference through logistic regression and classification. The study found that "The quantitative analysis that machine learning offers is beyond the scope of observational analysis" as it can provide richer observational data, however they also found that it currently lacks "practicality and adoptability" for coaches and teams. They concluded that the use of machine learning approaches that included computer scientists working along side sports scientists competent in interpreting the value of the machine learning outputs, are essential to fully leverage data, technology and the power of machine learning.

The above exploration of the literature guided the approach in this study into whether classification models can be a useful tool for goalkeeper performance analysis within football.

5) Methodology

Having built the dataset, the next step in achieving the aims and objectives of this study was to determine an appropriate methodology. Two aspects that I felt were important for my choice were an ability to build in the business context of my problem and also to work iteratively making changes/ potential improvements as required. It was clear from the literature (Azevedo & Santos, 2008), (Dåderman & Rosander, 2018) and previous studies that the more common methodologies employed in data mining projects include, KDD, SEMMA, and Crisp-DM.

KDD – "Knowledge Discovery in Databases", is a data mining technique that's goal is to assist the user in discovering useful information from a collection of data. It is an iterative process that allows for repeated improvement of data mining and modelling to help achieve project goals.

SEMMA – "Sample, Explore, Modify, Model, Assess", is another data mining and machine learning project approach that consists of five steps, it is described as a toolset for carrying out the core tasks of data mining. Unlike other techniques SEMMA focuses mostly on data management and the modelling aspects of data mining.

Crisp-DM – "Cross Industry Standard Process for Data Mining", is currently the most used framework due to its clear structure, with well defined steps, and its iterative nature of implementation. As an approach it originates from a business perspective providing users with clear steps to set out their projects business goals, and the process to achieve them.

A comparison of these methodologies conducted by Azevedo, A. and Santos, M.F. (Azevedo & Santos, 2008), found them to be very similar data mining methodologies with comparable steps. They argued that both 'Semma and Crisp-DM can be viewed as an implementation of the KDD process' achieving the studied goal through a similar procedure. However Crisp-DM is more iterative in nature and uses additional business understanding and deployment steps to guide users in how data mining can be applied in practice.

Table 2: Summary of the correspondences between KDD, SEMMA and CRISP-DM (Azevedo & Santos, 2008)

KDD	SEMMA	CRISP-DM
Pre KDD	----------	Business Understanding
Selection	Sample	Data Understanding
Pre processing	Explore	
Transformation	Modify	Data preparation
Data mining	Model	Modelling
Interpretation/Evaluation	Assessment	Evaluation
Post KDD	----------	Deployment

Given the applied nature of my study which has the objective of developing a model that can be implemented to help football clubs with performance analysis I decided that Crisp-

DM was the best approach. It also importantly allows for iterative improvement of the models which I required in implementing different pre-processing procedures.

Crisp-DM Procedure:

*Figure 4: Crisp-DM Methodology Steps (*Data Scientists and the Practice of Data Science | CustomerThink*, 2015)*

The Crisp-DM methodology is made up of 6 steps, I have described these below, along with their application within this study:

- **Business Understanding**, the first stage of this process focuses on gaining a clear understanding of the projects aims and requirements from a business perspective, then using this knowledge to determine the studies data mining goals and create a project plan to achieve them.

  As stated earlier, the main aim of this study is to assess whether classification models can be a useful tool for player performance analysis. From a business perspective this will be achieved by developing a model that provides clubs with a way to leverage their goalkeeper data to help make data driven decisions about individuals and the team. These includes helping to identify a goalkeeper's areas of weakness and tailor their training programme to help them address these problems.

- **Data Understanding**, this phase involves identifying, collecting and analysing the datasets necessary to accomplish the projects goals.

  Within the data section of this report, I have clearly described how the dataset used for my analysis was identified and collected. After compiling my complete dataset and introducing additional external data I then ran a preliminary analysis on the dataset using descriptive statistics to better understand the distribution of the data

and the frequency of each dive. I also created some simple visualisations and charts to display the results from this analysis.

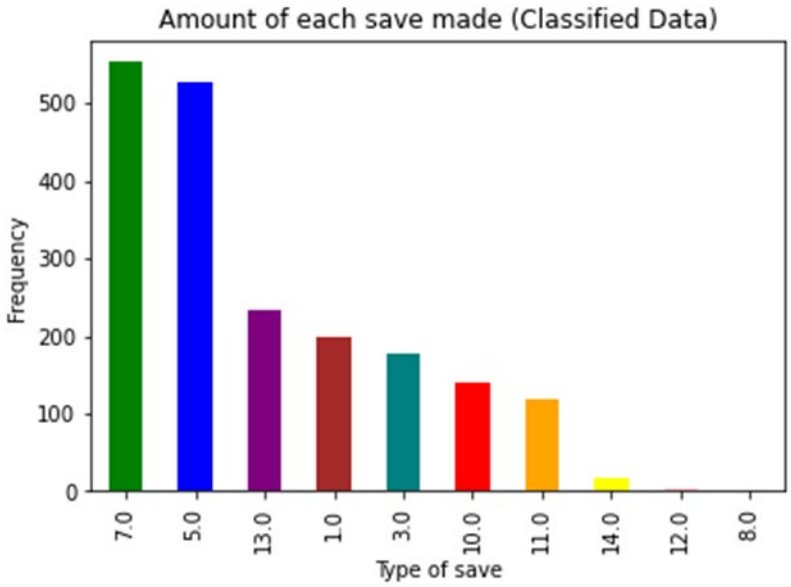

Figure 5: Bar Chart on Frequency of Each Dive

In order to gain a better perspective of the data and its structure I plotted the frequency of each classification within the dataset. From the figure above I was able to learn that a specific 7 of the original 14 classes were common across the different training sessions with the rest occurring less than 50 times and many classes didn't occur at all either due to their unlikely nature or that they were unable to occur. Low body dives to the right (class 5) and left (class 7) occurred more than twice as much as any other dive type and certain dive types never took place across any of the training sessions such as High Right Feet (class 2) and High Catch Body (class 9). These insights into the training sessions provided me with a clearer idea of the specific actions in goalkeeping, for example when a keeper makes a dive to the left or right it almost never results in the keeper landing on their feet.

After reviewing the results of the preliminary data analysis and assessing the dataset it was clear that the original 14 classes of dives was too broad and included dives that were highly uncommon, or unable to occur such as low dives that result in the keeper landing on their feet. This led to me restructuring the classes involved in the study to only include the 7 most common dives. Data relating to any of the outlying dives were removed from the dataset, this was done to avoid them skewing the models and negatively affect their accuracy. The table and figure below provide a clearer view of the remaining classifications of dives within the dataset, as well as the areas in which each class of dives occurs.

Table 3: New Dive Classifications

Class	Dives	Side	Landing
1	High Dive	Right	Body
3		Left	Body
5	Low Dive	Right	Body
7		Left	Body
10	High Catch		Feet
11	Centre		Body
13	Block		Body

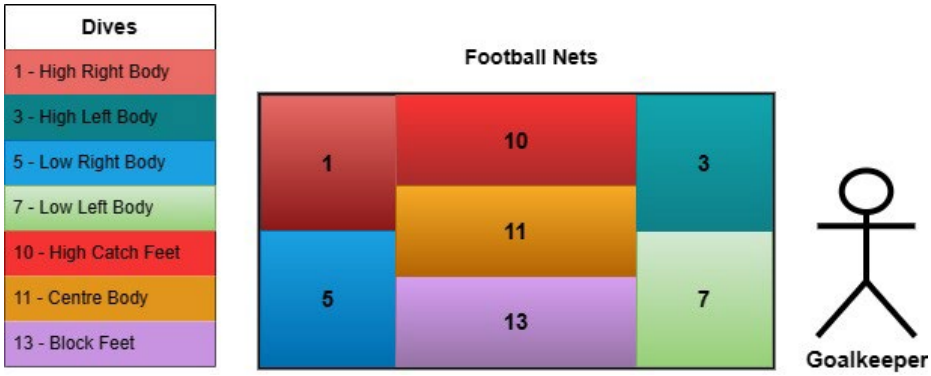

Figure 6: Area in the Nets of each type of save.

- **Data Preparation**, this step details the entire process involved in preparing the initial raw dataset for modelling. It includes steps such as cleaning, reformatting, and extracting data.

  I employed a number of different pre-processing techniques in order to prepare my data for modelling. This was done with a view to improving the accuracy and reducing the computational power necessary to model the data and answer the research questions. The pre-processing techniques used in this study were:

  1. **Standardization** – A technique to reduce the scale of the datasets features and prepare it for further pre-processing.
  2. **Univariate selection** – An approach for feature selection to identify the variables with the lowest variance.
  3. **Extra Tree Classifier** – A modelling approach for feature selection to identify the best features for classifying the data based on their predictive power.
  4. **Principal Component Analysis** – An algorithm for feature extraction to reduce the dimensionality of the dataset.

- **Modelling**, here a variety of modelling techniques are selected and applied to the dataset. It is important that each model is assessed, and their parameters are calibrated to produce optimum results.

For this stage of the analysis, I chose to run a number of different models on my dataset. Based on a review of the literature I selected five of the more common classification models. I ran each model three times on datasets with different levels of pre-processing (Minimal/Scaling & Feature Selection/Principal component analysis) using techniques described above, in order to build a comprehensive set of models and identify the best approach to take for this study. These models included:

1. **Logistic Regression –** Uses logistic functions to classify binary and multiclass data.
2. **K-Nearest Neighbours –** Classifies each data point based on its K number of closest data points (Neighbours).
3. **Decision Trees –** This method creates a tree like structure from the training data where the internal nodes are features, the branches are rules for classifying each data point and the leaf nodes are the classification results.
4. **Support Vector Classifier –** that seeks to find the optimal hyperplane to separate different classes of data points.
5. **Random Forest –** Uses bagging techniques to extract subsets of data, builds a forest decision tree and aggregates their predictions to classify data points.

- **Evaluation**, this stage involves thoroughly evaluating each model and the steps taken to construct the model in order to ensure it achieves the studies objectives.

  To properly evaluate each model and the effects of the different pre-processing methods implemented, output from the fifteen different approaches was fully analysed and comparisons made. The most accurate and efficient model for goalkeeper dive classification, trained and tested on the dataset was then identified.

  The performance of each model was assessed using two methods, cross validation scoring and standard model testing, these returned F1-scores which could be used to measure and compare accuracy.

- **Deployment**, the final phase is to create a way for users to access the model and its results. In order to achieve this within my study the aim is to create an interactive dashboard that will allow users to upload their own data from goalkeeping sessions and dashboard will run the most accurate model on their data and classify each action and the resulting dives. The dashboard will provide users with visuals detailing the frequency of dives made during each session and other interesting statistics.

The next part describes how the pre-processing and machine learning techniques were applied to the dataset.

6) Analysis

Having collected the data necessary for this study, compiled it into a single dataset and run a preliminary analysis on its structure and content I was able to begin my full analysis. At this stage objectives one and two were to build two machine learning classification models based on my dataset, the first was for the binary classification of goalkeeper actions into Dive/Not Dive (1/0) and the second was a multi-class classification of each dive into one of seven classes. Based on my research into similar studies to my own I decided to apply a comprehensive set of machine learning classification models in order to identify the most appropriate method. Each of the models were trained on data frames with different levels of pre-processing (minimal/ scaling and feature selection/ feature extraction), once this was completed cross-validation was used to validate the models as well as estimate their accuracy and the train_test_split module was used to test the validity of the models on unseen data.

7.1. Pre-processing

Pre-processing can have a significant effect on the accuracy of machine learning models. I decided to employ multiple methods of pre-processing to the dataset and then train and test each model for each stage of pre-processing in order to identify whether they improved or impeded the model's accuracy and efficiency. The dataset was pre-processed in three different steps:

1. Feature Scaling
2. Feature Reduction
3. Feature Extraction

1) **Feature Scaling**, this technique is used to transform the scale of features within the dataset and to bring all features to a similar scale, to avoid any bias on the analysis or modelling. For this study scaling was applied using standardization, which is done by subtracting the mean of each variable from each observation and then dividing by the standard deviation of the variable. This transforms the data so that it has a mean of zero and a standard deviation of one. It is important when comparing variables with different units of measurement or when some variables have much larger values than others.

2) **Feature Selection**, the goal of feature selection is to improve the accuracy, efficiency and performance of machine learning models by identifying and keeping the most informative and relevant features, while reducing the size and complexity of the dataset. It's an important part of pre-processing data for modelling, as it can help prevent the models from overfitting on the training data which can occur when too many features are included. This can result in the models memorizing training data features and prevent it from identifying patterns, to allow any new testing data to be accurately classified. For this study I chose to use two feature selection methods to ensure all the important features were identified and retained. The approaches used were Univariate selection, and Extra Tree Classifier.

Univariate selection, is a technique used to identify the best features for modelling by evaluating each feature independently based on their relationship with the output variables. The Anova F-test was used in this study to compare the variance between the means of the feature values for each class in the target variables (binary and multi-class) to the variance within each class. This produced a score for each feature, which represents the relationship between the feature and the target variable. The features with the highest score were selected these are displayed in the tables below:

Table 4: Univariate 10 Best feature (Binary Models)

	Specs	Score
4	Mean Acc y	2780.608778
13	Peak Mag	1997.877594
15	Peak Gyro Mag	1702.773561
1	Peak Acc x	1540.606899
5	Peak Acc z	1343.306705
16	Mean Gyro Mag	1129.597215
9	Peak Gyro y	1108.309314
7	Peak Gyro x	1053.471994
11	Peak Gyro z	1015.975057
20	Peak Acceleration (m/s2)	978.142397

Table 5:Univariate 10 Best feature (Multi-Class Models)

	Specs	Score
2	Mean Acc x	361.636877
1	Peak Acc x	306.475797
12	Mean Gyro z	208.938959
4	Mean Acc y	116.803360
19	Mag Speed	109.076901
17	Peak Speed	98.881743
18	Mean Speed	98.881743
5	Peak Acc z	95.404293
9	Peak Gyro y	93.768706
13	Peak Mag	87.930894

From these tables it's clear that the target value from both analyses are dependent on very different factors within the dataset as they only have five factors in common. As well as this we can see that no factors introduced in the additional data (goalkeeper characteristics and weather) are in either table, therefore they seem to have had minor influence on the classification of both target variables.

Extra Tree Classifier, is a technique for feature selection that uses multiple tree models on randomly selected subsets of data from each feature to identify the best features for high accuracy modelling with low variance. For this study I used a built-in class to build these models and find each features importance in classifying the dives. The 10 best features are displayed in the graphs below.

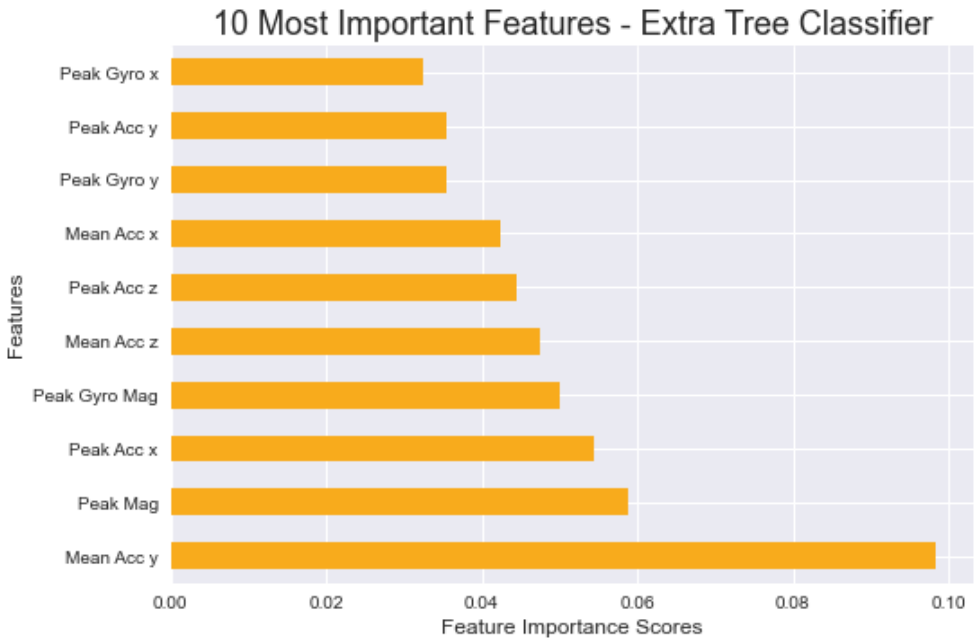

Table 6: Extra Tree Classifier 10 Best Features (Binary Models)

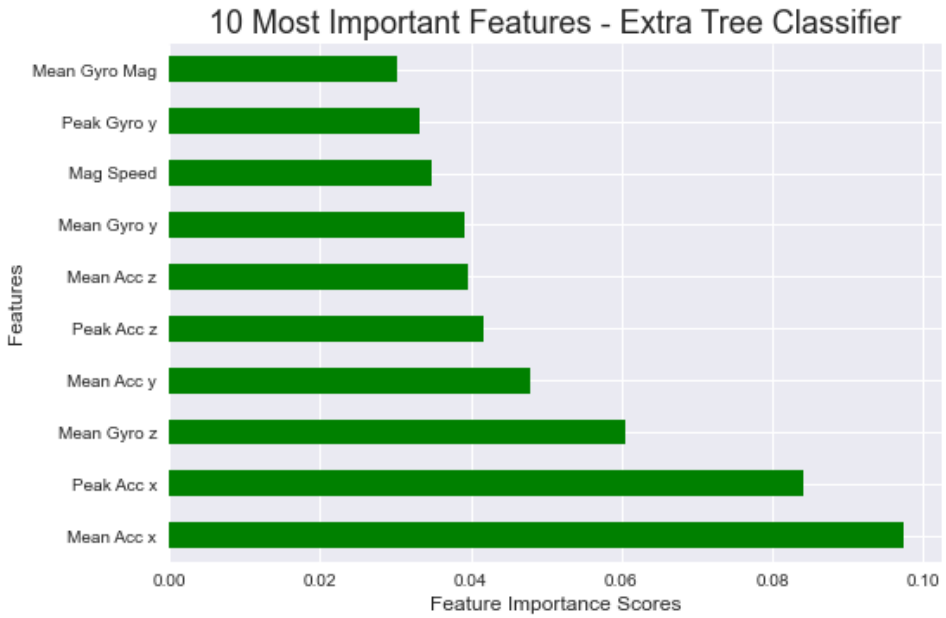

Table 7: Extra Tree Classifier 10 Best Features (Multi-class Models)

From the results of these tables, it is clear that the 5 most important features identified when using both techniques were the same. The remaining features differed between both methods and provided a variety of features of high importance to classifying the target variables. From the findings of both methods, I was able to identify the most informative and relevant features in classifying the dives with relation to their predictive power and the statistical significance of their relationship to the target variables. I chose to use a combination of both methods results and remove any other features in an attempt to improve the accuracy of my studies models.

3) **Feature extraction**, these techniques involve transforming the datasets original features into a new set of features that capture the most important information in the dataset. To achieve this, I used Principal component analysis (PCA) to reduce the dimensionality of my features while retaining the important information. PCA is an unsupervised learning method that examines the relationships between the different attributes in a dataset and identify patterns that it can preserve while reducing the dimensionality of the data. This technique creates a new set of variables smaller than the original set that retains most of the original information. This allows the models to achieve similar accuracy results more efficient with a lower computational time.

   In order to determine the optimum number of principal components to use for both my binary and multi-class target variables I applied the elbow method and also identified each components silhouette score. See the graphs below.

   The **elbow method** works by running PCA on the dataset for a range of principle components, the percentage variance explained by each those components is then calculated and plotted on a line chart. Finally, the point of inflection (the elbow) in the graph is identified, this point shows the number of components where the rate of increase in variance explained begins to level off. This point shows the optimum number of components to use, as any additional components will only capture a marginal increase in variance, likely to only contain noise and unimportant features.

Figure 7: Elbow Method for Optimal K (Binary Classification)

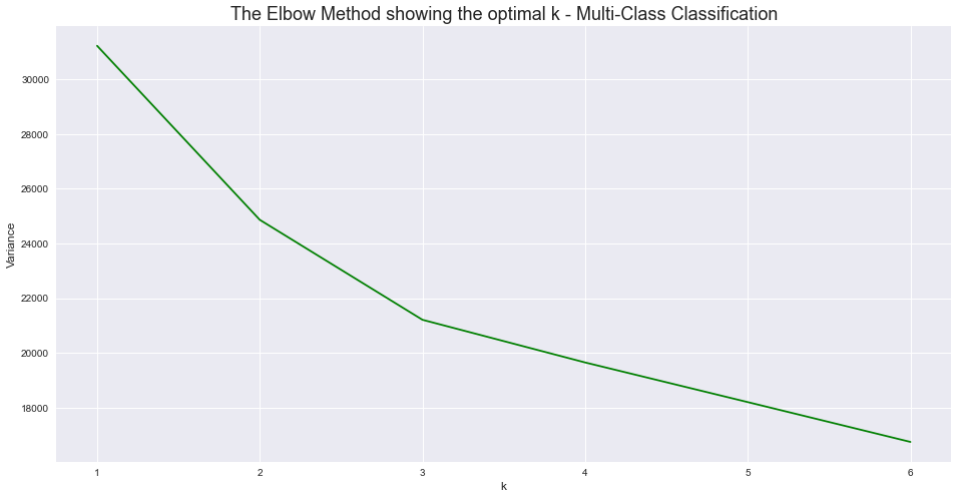

Figure 7: Elbow Method for Optimal K (Multi-Class Classification)

Silhouette scores, evaluate the quality of clustering in machine learning by measuring how similar each data point is to its own cluster compared to other clusters. These can be applied to PCA as it is able to identify the number of principal components with the maximum quality of data clustering. The process works by first running PCA on the dataset for a range of principle components, the data is then clustered for each number of PCA's using the K-means algorithm. The silhouette score is then calculated for each of these clusters and the scores were plotted on a line chart along with the component number, the component with the highest silhouette score has the highest quality of clustering in its data.

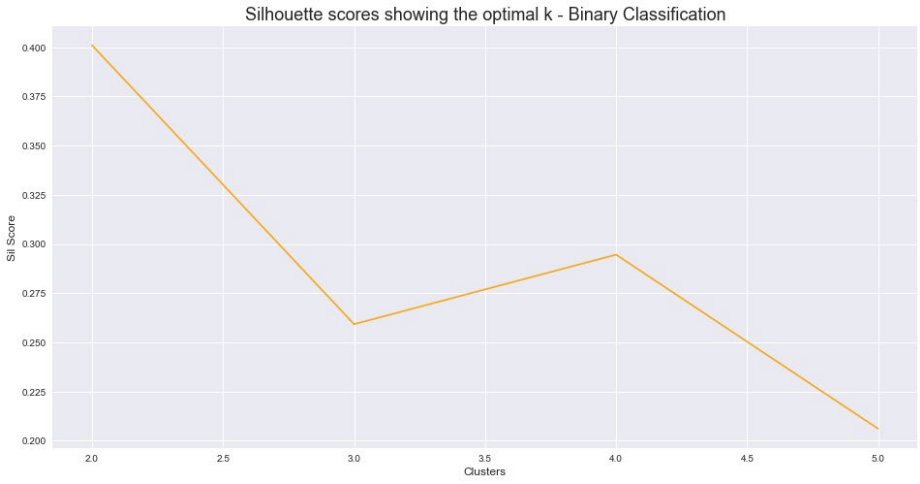

Figure 8: Silhouette Scores Chart for Optimal K (Binary Classification)

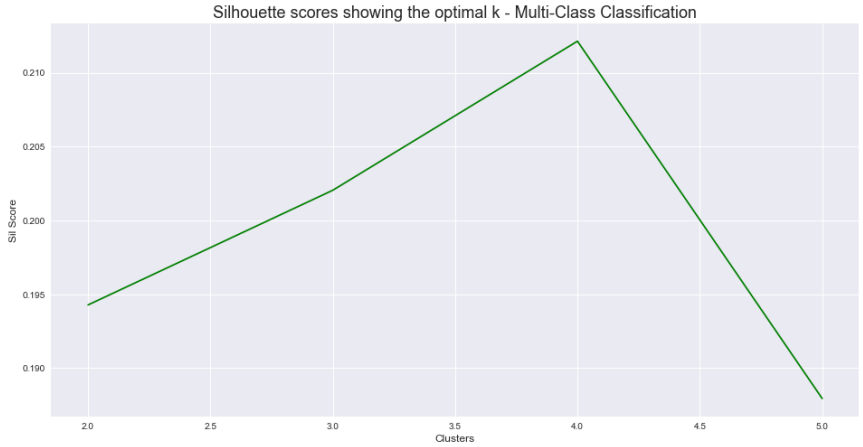

Figure 9: Silhouette Scores Chart for Optimal K (Multi-Class Classification)

From applying both the Elbow method and calculating the silhouette scores it was clear that for the binary classification, identifying if the action is a dive, the optimum number of components was 2. The results of both methods on the multi class classification, identifying the type of dive, wasn't as clear as the optimum number of components was different between both methods (Elbow = 3, Silhouette = 4). In order to identify which produced better results I ran the models on the dataset with each number of components and found that 3 was the optimum value for the multi class classification models as it returned the highest accuracies.

7.2. Modelling

After applying each of the three methods above to prepare the dataset I then moved on to building and testing models for both the binary and multi class classification tasks. I selected a variety of models that were used in similar academic studies which I had previously assessed and used them to develop my classification models. These models were iteratively applied to the dataset at different levels of pre-processing to achieve a clear view of which techniques provided the highest accuracy from each model. 5 models were selected from the research papers reviewed, these included:

1. Logistic Regression
2. K-Nearest Neighbours
3. Decision Tree
4. Support Vector Classifier
5. Random Forest

1) **Logistic Regression**,
   This is a supervised machine learning algorithm used to predict the probability of an event by matching the data on a logistic curve. The model uses the logistic function to calculate the probability of each data point belonging to class 1 or 0. Each data point is classified based on which point it has a higher probability of belonging to.

2) **K-Nearest Neighbour (kNN)**,
kNN works by finding the k number of closest data points (neighbours) in the training set to the input data points, and then predicts the class of each input data point based on the majority vote or the average of the k neighbours. The algorithm follows the following steps to classify data:
    1. Determine number of K.
    2. Calculate distance between data points (Euclidean or Manhattan)
    3. Determine K-nn minimum distance
    4. collect category Y values of nearest neighbours
    5. Calculate the predicted class or value of the input data point based on the majority vote or average of the k-neighbours.

3) **Decision Tree**,
Every Decision tree is comprised of 3 similar elements. A root node which is the base of every tree and is made up of the entire data set. From this a series of questions are asked, based on the data set being used. These resulting nodes branch off and produce decision nodes. At each node the model asks itself, what feature will allow the data to be split in such a way that the resulting groups are as different from each other as possible, while the data items of each group are as similar to each other as possible? This process continues until reaching the final leaf or terminal node. The model's aim is to find the best method of splitting data to derive actionable and useable information.

4) **Support Vector Classifier (SVC)**,
SVC's work by laying the dataset on a hyper-plane and creating boundaries that separates the data into partitions on both sides. Each row of data becomes a data item that is plotted on this created plane. The model performs classification by finding the hyper-plane that best differentiates each of the classes. This method is similar in principle to kNN as it splits data points based on similar features, but it also prioritises accuracy. Should a data point contain features that closely resemble another classification the model will choose accuracy over splitting data points evenly.

5) **Random Forrest**,
This functions similarly to Decision Trees, discussed above. However, it utilizes many Decision Trees to come to its prediction, using trees to build a forest, whereby the answer that most trees predict is the answer the model will return. The concept of this machine learning algorithm is that a large number of relatively uncorrelated models working as a committee will outperform any of the individual models within that committee. It differs to decisions trees as instead of training each model on the entire dataset it uses a technique called bagging to select random features from the dataset. This allows each model to remain relatively uncorrelated and produce an ensemble of predictions which are more powerful and accurate than any singular prediction.

After training each of these models on the original dataset and recording their results from testing it was then important to tune each model's hyper-parameters in order to return the highest accuracy possible. To achieve this a combination of cross-validation and gridsearch was used to evaluate the models with different parameters and identify the optimum values for each model. GridsearchCV uses the following steps:

1. Define a grid of hyperparameters to search over.
2. Split the data into training and test sets.
3. Fit the model with each set of hyperparameters.
4. Select the best set of hyperparameters.

From implementing this process I was able to tune each of the model's parameters to the optimum level and significantly improve each of their accuracies when testing the original dataset. For example, when constructing the K nearest neighbours' model on the PCA data the gridsearch function results showed that in order to optimize the model it should be set to use 17 neighbours to classify each action. This led to the final step of the analysis which consisted of training the improved models on the dataset at each stage of its pre-processing and comparing the final accuracy results from testing the models.

7) Results

7.1. Testing

The next phase of this analysis was to test each of the models and identify which returned the highest classification accuracy results. In order to achieve this, I applied a number of different testing and validation methods and metrics to each of the models. Each technique applied is described below.

- **Cross-Validation Testing:**

Cross validation is a method applied when training a model, in order to avoid overfitting, which is the situation when the model fits the training data very well but cannot generalise to data that has not been seen before. K-fold cross validation was used for this study as it was successfully implemented in a previously analysed research paper in the literature review (). The process of cross validation uses the following steps:

1. Split the total dataset into k small subsets (k folds).
2. The model will be trained by k-1 subsets of the total subsets.
3. The model will be tested by remaining one data subset and get the scoring of the currently trained model.
4. The testing dataset of k folds is then changed until every subset is regarded as the testing dataset for one time.
5. Finally, there will be k scoring of the model, and the average value of k scoring is the final result of the models cross-validation.

From running the cross-validation with different values of k and comparing the results I found that five was the optimum value for k (number of subsets)

- **Accuracy:**

Accuracy is a simple metric that allows us to understand the performance of our classification model by seeing what proportion of instances it has correctly predicted. It uses the following steps to calculate a model's accuracy:

1. Get predictions from your model.
2. Calculate the number of correct predictions.
3. Divide it by the total prediction number.
4. And analyse the obtained value.

Equation 2: Accuracy (Accuracy, 2023)

$$Accuracy = \frac{TrueNegatives + TruePositive}{TruePositive + FalsePositive + TrueNegative + FalseNegative}$$

The accuracy is always between 0 and 1, and better performance is achieved for higher accuracy.

- **Precision:**

Precision is a measure of how many of the positive predictions made are correct (true positives). The formula for it is:

Equation 3: Precision (Accuracy, 2023)

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

- **Recall:**

Recall is a measure of how many of the positive cases the classifier correctly predicted, over all the positive cases in the data. The formula for it is:

Equation 4: Recall (Accuracy, 2023)

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$= \frac{True\ Positive}{Total\ Actual\ Positive}$$

- **F1-Score:**

F1-Score is a measure combining both precision and recall. This metric calculates the harmonic mean of precision and recall and provides a balance between the two metrics. It is calculated using the following formula:

Equation 5: F1-Score (Accuracy, 2023)

$$F1 = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

7.2. Model Results

Table 6: Binary Classification of Goalkeeper Actions Results

Variant	Measure	Log Reg	kNN	Decision Tree	SVC	Random Forest
Minimal Pre-Processing	Cross-Val	0.81	0.81	0.82	0.82	0.86
	Accuracy	0.81	0.80	0.79	0.81	0.84
	Precision	0.80	0.79	0.82	0.80	0.84
	Recall	0.81	0.80	0.79	0.81	0.84
	F1-score	0.80	0.79	0.79	0.81	0.84
Scaling and Feature Selection	Cross-Val	0.84	0.85	0.83	0.85	0.85
	Accuracy	0.82	0.82	0.80	0.82	0.83
	Precision	0.81	0.82	0.82	0.82	0.82
	Recall	0.81	0.82	0.80	0.82	0.82
	F1-score	0.82	0.82	0.81	0.82	0.82
Principal Component Analysis	Cross-Val	0.81	0.82	0.80	0.83	0.81
	Accuracy	0.79	0.79	0.76	0.80	0.78
	Precision	0.78	0.79	0.79	0.80	0.80
	Recall	0.79	0.79	0.76	0.80	0.78
	F1-score	0.77	0.79	0.77	0.80	0.79

In the table above I have displayed a variety of testing and validation accuracy metrics for each model built on the data at three different levels of pre-processing, this allowed me to conduct a comprehensive comparison of each models results and the effects of each form of pre-processing. The results in this table related to the binary classification models that were built to classify every action detected by the GPS tracker into one of two categories Dive and Not Dive.

It is clear from the table that while both the Support Vector Classification model and Random Forest model had promising results with scaling and feature selection applied the most accurate model built during this study was the Random Forest model with minimal pre-processing. This method returned high levels of accuracy (0.84) when classifying the actions while having similarly high results for the F1-score (0.84) meaning that it was able to correctly identify most of the positive cases while minimizing the number of false positives.

There are claims that the random forest model can "perform well 'out-of-the-box', with no tuning or feature selection needed, even with so-called high-dimensional data"[] so this may be a factor in why the random forest performed at its best with minimal pre-processing applied.

From the table we can see the effect each pre-processing method had on each of the models. The table clearly shows that by applying scaling and feature selection to the dataset it resulted in a slight increase in most of the model's accuracy, this could be due to the lack of noise in the dataset following feature selection or the change in data's scaling from standardization. The table also shows that by applying Principal Component Analysis to the dataset it caused a decrease in accuracy in all of the model's, this could have been caused by an accidental loss of important information when reducing the data's dimensionality.

Table 7: Multi-Class Classification of Goalkeeper Dives Results

Variant	Measure	Log Reg	kNN	Decision Tree	SVC	Random Forest
Minimal Pre-Processing	Cross-Val	0.67	0.64	0.68	0.70	0.81
	Accuracy	0.66	0.63	0.67	0.68	0.82
	Precision	0.66	0.64	0.69	0.67	0.83
	Recall	0.66	0.63	0.67	0.68	0.82
	F1-score	0.65	0.63	0.68	0.67	0.81
Scaling and Feature Selection	Cross-Val	0.73	0.78	0.68	0.79	0.78
	Accuracy	0.74	0.77	0.66	0.77	0.78
	Precision	0.74	0.77	0.71	0.78	0.79
	Recall	0.74	0.77	0.66	0.77	0.78
	F1-score	0.73	0.77	0.67	0.77	0.78
Principal Component Analysis	Cross-Val	0.68	0.72	0.63	0.72	0.71
	Accuracy	0.67	0.71	0.62	0.70	0.78
	Precision	0.67	0.72	0.68	0.71	0.71
	Recall	0.67	0.71	0.62	0.70	0.78
	F1-score	0.66	0.71	0.63	0.70	0.79

Similar to the previous table, I have displayed the results from each multi-class model's metrics above. The results in this table were for the multi-class classification models that were built to classify every dive the goalkeeper makes into one of seven categories based on the direction of the dive, the height of the dive, and the keepers landing.

From the table I can see that with scaling and feature selection applied the k-Nearest Neighbours (kNN), Support Vector Classification (SVC) and Random Forest models all had promising results but the model that returned the highest classification accuracy was the Random Forest with minimal pre-processing applied to the dataset. This model was able to accurately classify goalkeeper dives (accuracy = 0.82) and correctly identify the majority of the positive cases with a small number of false positives (F1-score = 0.81).

The table also displays the results from the different levels of pre-processing applied to the dataset prior to modelled. Unlike the previous models these results showed that scaling and feature selection had an unexpected effect on the Logistic Regression, k-Nearest Neighbours (kNN) and Support Vector Classification (SVC) models. This pre-processing method increased each of their accuracies and F1-scores by between 8% to 14%, these results were a surprise and although I am unsure what caused them, I believe standardization may have played a factor in the dramatic improvement as a similar change didn't occur in the decision tree and random forest models which aren't effected by feature scaling.

8) Conclusions

A Crisp-DM methodology was taken to frame this study, and achieve the following research objectives:

1) Create a classification model to classify goalkeeper actions gathered by STATSports GPS trackers and additional data (weather and personal) as dives and not dives.

   The best model for binary classification of goalkeeper actions was found to be the random forest, with minimal pre-processing techniques applied to the dataset (Cross-Val = 0.86, Accuracy = 0.84, Precision = 0.84, Recall = 0.84, F1-score = 0.84).

2) Create a classification model to classify the identified dives by type based on their direction, height, and landing.

   For the multi-class classification of dives by type, the random forest was the best performing model, with minimal pre-processing techniques applied to the dataset (Cross-Val = 0.81, Accuracy = 0.81, Precision = 0.83, Recall = 0.82, F1-score = 0.81).

3) Implementing pre-processing techniques such as scaling, feature selection, and Principal Component Analysis, to check for improvements in models.

   The application of pre-processing techniques resulted in an increased accuracy in many of the models implemented however the application did not appear to improve the accuracy of the random forest model which gave the best overall results for both classification models.

   For the binary classification model, the application of scaling and feature selection techniques to the dataset resulted in a slight improvement in most model's accuracy, except random forest. This could be due to a reduction in noise and or a change in scaling. The results also showed that applying Principal Component Analysis to the dataset resulted in a decrease in accuracy in all of the models. This could have been caused by an accidental loss of important information when reducing the data's dimensionality.
   Increases, were also observed in the multi-class classification models however these were much larger with an increase of between 8% to 14%, these results were unexpected and although I am unsure what caused them, I believe standardization may

have played a factor in the dramatic improvement as a similar change didn't occur in the decision tree and random forest models which aren't affected by feature scaling.

4) Investigate whether including additional features relating to goalkeeper characteristics and weather data can improve the model's accuracy.

   None of the additional factors (goalkeeper characteristics and weather) were selected following feature selection methods, therefore they seem to have had minor influence on the classification of both target variables. However, the models identified with the best results was Random Forest without feature selection or PCA, and therefore contained these factors. This is an aspect that I would like to further investigate to understand and potentially identify additional contributing factors, which may improve models.

9) Further Development or Research

Having access to such a rich dataset to train and test models was hugely beneficial leading to models that while useful in themselves, can be further developed in a number of ways.

- **Increased dataset**
  The models would benefit from training on a more extensive dataset potentially involving data from other leagues.
- **Real time data modelling**
  The ultimate aim would be to build a classification model that could handle data streamed from STATSports GPS trackers in real time. I would hope that learning from this study could inform the development of such a model.
- **Further investigate the impact of factors on model accuracy**, potentially adding additional external data such as psychological factors.
- **Incorporation of the model within the STATSports app** to provide teams with a mechanism for automated classification of actions during a goalkeeping session.

10) References

1) *STATSports*. (n.d.). Retrieved 14 May 2023, from https://eu.shop.statsports.com/

2) Szwarc, A., Lipińska, P., & Chamera, M. (2010). The Efficiency Model of Goalkeeper's Actions in Soccer. *Baltic Journal of Health and Physical Activity*, *2*. https://doi.org/10.2478/v10131-0013-x

3) *STATSports*. (n.d.). Retrieved 14 May 2023, from https://eu.shop.statsports.com/

4) *Soccer goalkeeper expertise identification based on eye movements | PLOS ONE*. (n.d.). Retrieved 14 May 2023, from https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0251070

5) Ibrahim, R., Kingma, I., de Boode, V. A., Faber, G. S., & van Dieën, J. H. (2019). Kinematic and kinetic analysis of the goalkeeper's diving save in football. Journal of Sports Sciences, 37(3), 313–321. https://doi.org/10.1080/02640414.2018.1499413

6) Azevedo, A., & Santos, M. F. (2008). KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW. *IADIS*, *European Conference on Data Mining 2008, Amsterdam, The Netherlands*, 182--185.

7) Historical Weather Data & Weather Forecast Data | Visual Crossing. (n.d.). Retrieved 14 May 2023, from https://www.visualcrossing.com/weather-data

8) DÅDERMAN, A., & ROSANDER, S. (2018). Evaluating Frameworks for Implementing Machine Learning in Signal Processing A Comparative Study of CRISP-DM, SEMMA and KDD. EXAMENSARBETE INOM TEKNIK, GRUNDNIVÅ, 15 HP(STOCKHOLM, SVERIGE 2018), 1–43.

9) Al-Asadi, M. (2018). Decision support system for a football team management by using machine learning techniques. https://doi.org/10.13140/RG.2.2.17433.75367

10) Data Scientists and the Practice of Data Science | CustomerThink. (n.d.). Retrieved 14 May 2023, from https://customerthink.com/data-scientists-and-the-practice-of-data-science/

11) Di Salvo, V., Benito, P. J., Calderón, F. J., Di Salvo, M., & Pigozzi, F. (2008). Activity profile of elite goalkeepers during football match-play. The Journal of Sports Medicine and Physical Fitness, 48(4), 443–446.

12) Accuracy. (n.d.). Hasty.Ai. Retrieved 14 May 2023, from https://hasty.ai/docs/mp-wiki/metrics/accuracy

13) Duarte, R. (2022). *Utilizing machine learning techniques in Football prediction*.

14) Herbinet, C. (2018). *Predicting Football Results Using Machine Learning Techniques*.

15) Manish., S., Bhagat, V., & Pramila, R. (2021). Prediction of Football Players Performance using Machine Learning and Deep Learning Algorithms. *2021 2nd International Conference for Emerging Technology (INCET)*, 1–5. https://doi.org/10.1109/INCET51464.2021.9456424

Appendices

National College of Ireland

Project Proposal

The Classification of Goalkeepers Dives

20/10/2022

BSC in Data Science

Computing Project (BSHCSD4)

2021/2022

Daragh Carroll

X19477436

X19477436@student.ncirl.ie

Contents

16) Objectives

The aim for my project is to build a machine learning classification model using data gathered by professional football clubs in order to classify the different types of dives made by goalkeepers. The end goal is to have a reliable model that is highly accurate which STATSports customers can use to analyse their goalkeepers performances during a match or training session. Once this model has been developed and tested my aim is to build a web application where I can deploy the model for public use, this way it can be easily accessed by STATSports elite clients and their individual consumers.

17) Background

Over the past two decades football has adapted and improved in many different ways due to the introduction and influence of sport science and statistics. They have revolutionised the sport and had a drastic affect on how coaches and players approach every aspect of the modern game from tactics and training to rehabilitation and recovery. This is a rapidly growing industry and one that earlier this year I got the opportunity to be a part of. I was lucky enough to spend my six month work placement in 3rd year at a company called STATSports, they are the global leaders in GPS sports performance analysis and have award-winning technology that allows them and their consumers to accurately gather data.

here I learned a lot about the influence of sport science on professional football and how data science is becoming one of the most important tools to analyse athletes performances. One of the projects I got to be involved in was developing a dataset compiled of different actions made by goal keepers throughout training sessions. Once a substantial dataset had been created I then got to develop some basic machine learning models in order to see if this data could be used to classify the goal keepers actions into different categories of dives. Although I ran out of time to develop a highly accurate model I was very impressed by the results I did get. This is why for my final year project I wish to continue using the data I compiled in order to build a reliable model that STATSports can provide to their consumers to give them more metrics to analyse a keepers performance.

18) State of the Art

One of the most important parts of putting together a high level project is identifying and analysing similar projects. This allows you to understand how others have approached the topic and provides clarity in how to structure your own project. From searching for similar projects I have found a number of research projects analysing goal keepers actions. One

analysis looked at the efficiency of goalkeepers actions during the 2008 euros in Portugal while another used machine learning to analyse how a goalkeeper plays in the build up to an attack. These reports were interesting and informative but I found nothing relating directly to the analysis that I plan on conducting. One of the differences I identified between my analysis and any similar work was that there are very few projects looking at the action of diving and differentiating between different dives. This is mainly due to the limited availability of data on this topic. The main difference between my proposed work and similar projects is the data being used. The majority of the work done on this topic has used data gathered through observing matches and taking note of the different actions. This means that there was limited access to in depth and accurate data. The dataset that I have compiled was gathered using the most accurate gps trackers in the world and provides me with access to metrics that were unavailable during other analysis on this topic such as the goalkeepers peak and mean values for their gyroscope, acceleration and speed. Due to this I believe that my project is extremely unique.

19) Data

When building a Machine Learning model a well-prepared training dataset drives the quality of your Machine Learning model and its effectiveness in fulfilling its goal. It is important to have a high quality dataset, This means that the data needs to be relevant to the models goal and have feature representation as well as being gathered from a reliable source. If the dataset has all of these features it will help to build a model that will provide accurate results which you can be confident in trusting. As well as this the dataset should contain a high volume of instances and attributes as the more data a model has access to the more accurate it will be.

Therefore, in order to achieve my goal of building a classification model that can classify dives made by a goalkeeper I will need access to live data from professional goalkeepers. The dataset will need to have a classification of the different dives a goalkeeper can make as well as a selection of metrics relevant to the action of diving. As well as this the data needs to be gathered by an accurate and reliable source and needs to be collected from a variety of goalkeepers across multiple different sessions.

As I previously stated in this proposal during my 6-month work placement I was provided with access to data gathered by multiple professional football teams using STATSports FIFA approved GPS trackers. These trackers compile highly accurate data for various metrics based on the sport. Using this data and video footage of the training sessions I was able to manually classify every action made by the goalkeepers into 10 different categories based on the direction, height & landing, for example a ‘High Right Dive Body’. The dataset that I created contains over 6600 instances and 26 different attributes such as the players gyroscope and their acceleration.

Due to the fact that I have manually classified each instance in the dataset I know that it accurately reflects the goalkeepers actions as well as this it does not including any null values or major outliers. The dataset will still need to be cleaned as there are too many

unnecessary variables with data not relating to the goal of the project. Once cleaned the dataset that I have assembled will hopefully provide all the information I need to build a complex and accurate model for classifying dives.

20) Methodology & Analysis

The end goal of my project is to have a publicly accessible tool that can be used by STATSports consumers and high profile clients to classify and analyse the dives made during a goalkeeper session. Due to these factors I have decided to use Crisp-DM as my methodology. This will allow me to analyse the business purpose of the model for STATSports as well as deploying it online for their clientele to use. Another benefit to using Crisp-DM as my methodology is that it is flexible and agile so it will allow me to move back and forth between phases and make alterations to improve my models.

The first stage in my analysis of the project will be to run a preliminary analysis and gather descriptive statistics on my dataset in order to gain a better understanding of the data. With this analysis I will be able to clearly interpret meaningful information about the raw dataset and visualise it. Using what I learn from the descriptive statistics I will be able to alter my dataset to only include relevant and important information relating to the projects end goal.

Using the Crisp-DM Methodology structure I will break my project down into different stages, The first will be business understanding.

- **Business Understanding:**
  Here I will attempt to achieve a clear understanding of my projects objectives. In order to identify the success criteria for my model I will first need to understand the needs of the customers from a business perspective.
- **Data Understanding:**
  During this step I will discuss how I compiled my dataset and describe its structure and each of the different metrics. I will also look at the preliminary analysis of the dataset and its results. From this information I will produce visualisations for the data and identify its structure as well as its quality.
- **Data Preparation:**
  Once I have completed my analysis of the data I will then proceed to clean the dataset by identifying which data is unnecessary to my analysis and removing it. The aim of this step will be to get the data ready to be modelled
- **Modelling:**
  Here I will select the models that I think will suit my dataset best and produce the most accurate results. I will then train and test these models using cross validation in order to use the entirety of the dataset. Once the models have been built I will calculate their accuracy.
- **Evaluation:**

The purpose of this step will be to evaluate the accuracy of my models and to compare them. This will allow me to identify if I need to make any adjustments to the dataset or the models parameters and eventually result in a model I can deploy.

- **Deployment:**
  The final step will be to deploy my model for public use potentially through an online dashboard. I will also plan the monitoring and maintenance of this model.

21) Technical Details

As previously stated the aim for my project is to build a machine learning model using the dataset I have gathered that can classify saves made by a goalkeeper into different categories. I also aim to deploy this model onto a website where STATSports customers can use the model on their own data. In order to complete this project and develop a highly accurate and reliable model I will have to build multiple different models and compare their accuracy. The machine learning algorithms that I am currently considering using are a Random Forrest model and a Neural Network model.

Developing the web application in order to deploy my model is another important technical development in this project. After researching different methods in order to complete this goal I believe that Flask is the best option. It is an application that I briefly learned about during my placement and it allows you to connect Python to HTML/CSS to create websites able to run Python.

The other applications I plan on using during my project are:

- Excel (Storing Dataset)
- Tableau (Visualisations)
- SPSS (Data Analysis)
- Spyder (Running Python Code)

22) Project Plan

The goal that has been set for our mid point report is to have the preliminary analysis on the data completed. Therefore in order to reach this goal within the next 6 weeks I will be following the following project plan to try and have as much information as possible in my midway report.

- Week 1: Identify 6/7 projects that analyse similar topics to my project. From this I will be able to structure my project and select methods to use for analysing and modelling my data based off methods used by others.
- Week 2: Gather background information for my project. This will include things like the use of data science in sport and its impact as well as the development and change of goalkeepers and how machine learning can influence it
- Week 3: Clean my dataset, In order to do this I will use methods that I have previously used as well as finding different methods used in similar projects for thoroughly cleaning and preparing data for modelling.

- Week 4: Run a preliminary analysis on the cleaned dataset. This will include analysing the data's descriptive statistics in order to summarize the datasets structure and contents.
- Week 5: Create visualisations of the results from the preliminary analysis in order to make the information clear and easy to understand.
- Week 6: finish writing the midway report and make any necessary additional alterations to the dataset and/or preliminary analysis.