Erasmus University Rotterdam
Erasmus School of Economics
Master Thesis Econometrics and Management Science: Business Analytics and Quantitative Marketing.

Decision-Making Process of Football Players: Estimating the Components of the Risk of Pass

Vasileios Kapetanios (616118)

Supervisor:	Schoonees Pieter
Second assessor:	Michel van de Velden
Date final version:	1st April 2024

Abstract

Football is perhaps the most popular team sport globally with the most passionate and loyal fans. Over the past decades, there has been a significant increase in the use of advanced statistics in football for predicting the game outcome, players' performance ratings and even for betting strategies. In this research, we focus on individual performance, and we achieve that by focusing on the passing of football players. We aim to estimate the risk and the components of the pass and understand the decision-making process of football players. For instance, we estimate the effect of the distance of the pass or the effect of the body part of the football player when passing. First, we incorporate the two most common models in classification problems, logit and probit regression. Since we aim for higher predictive performance, we employ state-of-the-art machine learning methods for classification problems as the relevant literature suggests. Namely, we revise the performance of tree-based algorithms, using random forest, which we will evaluate using accuracy measures as described in this thesis. The purpose of this research is to identify the risk factors of passing and provide an insightful tool for predicting the risk of a certain pass based on its characteristics. Given the pass characteristics (set of explanatory variables), the predictive tool will estimate the likelihood of this pass being unsuccessful and its risk components. Consequently, coaches and players will be able to identify situations and playing styles with higher occurrences of unsuccessful passes and, thus will be able to make well-informed decisions about their action or tactical approach to the game.

Keywords— Decision-Making, Risk, Logit, Probit, Bagging, Random Forest

Contents

1 Introduction

Football or soccer, as it is also known, can be considered one of the most popular team sports around the globe with more than five billion fans worldwide (FIFA, 2021). Over the past decades, football analysis has been gaining ground using advanced statistical analysis to inform in and out of the pitch decisions. This rapid development is mainly fostered by access to new kinds of data sets and the development of new methodological tools (Burriel & Buldú, 2021).

Decision-making is a fundamental element of any sport, especially in fast, dynamic team sports such as volleyball, football and basketball. It is the study of identifying and choosing alternatives based on the values and preferences of the decision-maker (agent) (Govindarajan, 2014). More precisely, in sports, the athletes naturally encounter the decision with a higher degree of task familiarity which is firmly correlated with the dynamic nature of this process. In this thesis, we slightly shift our focus on the components of decision-making, by estimating the risk factors of a specific action of football players, their passing. We aim to decompose the features of an unsuccessful pass and estimate the likelihood of a pass being unsuccessful, i.e. the risk of a pass, based on its characteristics.

To narrow our research, and since it is focused on passing, we study the players in possession of the ball. In general, a football player, when in possession of the ball, has three possible options: pass, shoot or dribble (Schelling & Robertson, 2020). In the frame of this research, we focus on passing as this is one of the most effective actions in terms of goal scoring or creating a scoring opportunity (Burriel & Buldú, 2021). For naming convention, from now on in this thesis the terms action and pass are interchangeable.

As mentioned in Schelling and Robertson (2020), there are various components in the decision-making process in sports, such as available information, the cognitive limitations of the decision makers (heuristic and biases), the finite amount of available time to take action, the levels of risk and reward. In connection with this thesis, we observe the risk of a pass based on its outcome as discussed in Section 3. The primary question in this research is to identify and estimate the factors of an unsuccessful pass and consequently understand the decision-making process of a player when passing. To achieve that, we estimate the likelihood of a pass being unsuccessful and we explain the factors that affect the riskiness of the pass. To do so, we assign a binary variable, called outcome, whether the action had a negative result (outcome = 1), i.e. it will lead to a loss of possession. We study the characteristics of the pass, such as its length and angle combined with the characteristics of the football player who committed the pass, such as the execution foot, their body part etc. For instance, we observe that the type of the pass is one of the major factors that increases the riskiness of this action. Namely, when a pass is committed

without first controlling the ball, the so-called no-touch pass, there is a sufficient increase in the risk of that pass.

From this research, we aim to establish an informative method to assess the risk of football players' passing. Possession is considered a key factor of success in modern football, in terms of predicting the winning team. According to Jones, James and Mellalieu (2004), longer possessions, i.e. longer than 3 seconds, are related to a successful team performance. Given that longer possessions are defined as the longer amount of time the team has possession of the ball, we relate successful teams with fewer wrong passes per game, and we further elaborate on that in Section 5. With the term successful team performance, we refer to the evolving score of a team which is defined by all possessions that are categorised as taking place in a goal-scoring opportunity, rather than evaluating the teams based on the game's outcome (Jones et al., 2004). In the past decades, traditionally, goal scoring was established as a performance measure of a team. However, in football analytics, due to its low-scoring nature other novel metrics such as expected goals are being implemented (Mead, O'Hare & McMenemy, 2023).

Taking into consideration the aforementioned metrics for evaluating team performance in soccer, we aim to elaborate on an approach which will asses a player's risk behaviour and we will give insights into the playing style of a team based on their respective risk rates. We intend to identify the factors that affect a pass's risk and estimate its risk based on its outcome. Next, we will classify the players into two categories: risk-averse and risk-takers based on their unsuccessful pass rates as mentioned in Section 5. In other words, we interpret the players with larger percentages of wrong passes as risk-takers, as the likelihood of their pass being risky is greater than those with fewer wrong passes.

Additionally, we perform exploratory analysis on teams' unsuccessful pass rates by aggregating the risk rate per pass. From the results, we observe different segments of teams based on their percentage of unsuccessful passes (outcome = 1). Given the available data, we can relate each pass to the player who commits it. Consequently, coaches can review the characteristics of a player's passes and train them to improve their decision-making.

The most challenging part of this problem lies in the fact that these options need to be identified. Unsuccessful action will be considered the one which leads to a negative outcome for the football player or the team, i.e. loss of possession or turnover, while a successful pass will be considered the one which leads to a positive outcome, i.e. assist to a goal or to a shot on target. In this way, we denote the risk of a pass as the probability of the pass being unsuccessful, i.e. $\mathbb{P}(outcome = 1)$. Consequently, we will estimate the factors of an unsuccessful pass and the level of their effect on its risk.

More precisely, given the binary outcome variable of the estimated variable (unsuccessful or not), we are in the case of a binary classification problem and different predictive models are employed. Initially, we discuss the traditional econometric approaches of a logit and a probit model, studying both their predictive performance and the coefficient estimates of the explanatory variables, in other words, the risk components of a pass. Next, we employ machine learning methods for classification problems, using random forest aiming to achieve better predictive performance. This research must identify the best predictive model which can be used by coaches and athletes to improve their individual and team performance, which is discussed in Section 6.1.

In Section 2, we discuss the relevant literature and the related research that has already been conducted and inspired this thesis. Next, in Section 3 we provide an overview of the data we observe and their quality. In Section 4, we describe the employed models, the evaluation methods we will apply to the selected models and the tuning that needs to be done. Thereafter, in Section 5 we illustrate the results of the employed methods and conduct an explanatory analysis of the observed dataset. The interpretation of the final model will allow us to understand which attributes increase the risk of a pass. Finally, in Section 6, we summarize our findings and discuss how they can be applied by coaches and players. Finally, we elaborate on the limitations of this research as well as the future work which can improve our findings.

The code that was used for this thesis can be found in the repository:
https://github.com/VassilisGitHub/MScThesis.

2 Literature Review

In this Section, we will review the related research on the decision-making process in sports, the evaluation methods of a predictive model's performance and risk assessment in sports analytics and other sciences. In addition, we will elaborate on the effect of play styles and decision-making within a football game. Finally, we will discuss the relevant research that is already in place for the methods that are most commonly applied for prediction models and their evaluation approaches.

2.1 Relevant Research

In economics and psychology literature, there is an extensive study of decision-making behaviour and many theories have been introduced (Mishra, 2014). Mishra (2014) emphasizes the need for interdisciplinary integration in understanding decision-making under risk and incorporating perspectives from biology, economics, and psychology. These disciplines are often correlated in the research's framework and the author highlights the factors which have been ignored in the past in similar research. To create a 'rule of thumb' that allows for quick and efficient decisions the author suggests heuristic approaches. Mostly, these approaches are the product of inductive reasoning from actual patterns of decision behaviour and we employ such an approach since our agents are looking for the best choice at a given time rather than the optimal one.

To further expand our knowledge of risk estimating and to determine the state-of-the-art methods that are currently being implemented, we focus on related research in medical studies and other sports. Injury prevention is a major topic in sports medical science and many researchers estimate the risk factors of several injuries and how can they be estimated (Pasanen et al., 2015; Owoeye, Palacios-Derflingher & Emery, 2018; Read, Oliver, De Ste Croix, Myer & Lloyd, 2018). In their research, Owoeye et al. (2015) aim to investigate risk factors for traumatic non-contact lower extremity injuries in young sport athletes. The authors develop a screening tool for predicting future injury risk by identifying the main predictors of these injuries. In addition, the results of this study will contribute to the optimization of training programs and identify players at risk. Univariate and multivariate regression models were applied to investigate the relation between traumatic non-contact injuries and several intrinsic risk factors for lower extremity injuries. In similar studies, multivariate Poisson regression has been used to examine the risk factors of ankle sprain injuries (Owoeye et al., 2018).

As mentioned in Section 1, we will use the results of this research to identify risk-averse (or not) play patterns. Consequently, this approach will lead us to identify the winning playing style or

in other words the one with the best performance.

In the vast majority of the available literature, the researchers focus on the possession of the ball as a metric of a successful team (Jones et al., 2004; Fernández, Bornn & Cervone, 2019). As Jones et al. (2004) explains, it is intuitively expected that teams with longer possession of the ball create more goal-scoring opportunities. However, after comparing twenty-four matches involving successful and unsuccessful English Premier League teams it was concluded that possession is indeed related to successful performance but it is likely this is due to the differences in individual characteristics (Jones et al., 2004). Hereby, we need to study alternative aspects of the game rather than the possession of the team itself, to establish a successful and accurate performance metric.

This task has proved to be quite challenging for football analysts. Due to its low-scoring nature when compared to other sports, football's uncertainty often influences the result of a match, which led the researchers to deploy novel metrics such as expected goals, also known as 'xG' (Mead et al., 2023). The models described by the authors allow the analysts to quantify how likely it is, for a given shot, to result in a goal. Several features are incorporated in the modelling part of this research paper such as the match importance, player rank metric and Elo rating (Hvattum & Arntzen, 2010). With the aforementioned additions, the authors improve the model's performance to further support the argument that the studied metric is of significant importance for the football community.

Other aspects of the game have been also studied, such as the importance of situational variables, like the quality of opposition, and game period on team performance (Pratas, Volossovitch & P Ferreira, 2012). However, when the author examined the interactive effect between offensive sequences ending in a shot on goal and the situational variables, he did not find any significant evidence for improving the offensive efficiency. As we are interested in the risk of a pass in football, we revise similar studies that examine the effect of passing on expected possession value (Burriel & Buldú, 2021).

2.2 Sports Analytics Literature

Additionally, we consider performance evaluation measures and advanced statistical predictive models as they have been introduced for other sports rather than football.

Predicting the outcome of the NCAA tournament was extensively studied by Kvam and Sokol (2006). The authors' suggested approach was built on previously used methods to predict the tournament's winner such as tournament seedings, polls, ratings and rankings of the teams. In this study, a Markov chain model was used for ranking the teams and a logistic regression

model for calculating the transition probabilities. The underlying Markov chain model denotes one state per team, which intuitively represents the behaviour of a hypothetical voter and the current state of the voter corresponds to the team that the voter now believes to be the best. At each time step, the voter reevaluates their judgement. The prediction outcome appears to be predicting individual game results more accurately than the standard ranking systems. Similar studies on predicting game outcomes in basketball use the naive Bayes classification method in combination with Elo Rating (Miljković, Gajić, Kovačević & Konjović, 2010). This method successfully outperforms its counterparts like traditional linear regression, in terms of prediction accuracy. Furthermore, many machine learning algorithms are by Sarlis and Tjortjis (2020) for individual evaluation of basketball players.

Similarly, researchers aim to predict the result of a tennis match and estimate tennis players' performance (Kovalchik, 2016; McHale & Morton, 2011). In the search for the best method to predict the outcome of a tennis game, Kovalchik (2016) reviews several prediction models for forecasting wins in tennis and they separate those into four main categories, regression-based, point-based, paired comparison and bookmakers consensus model (BCM). In this research, probit and logit models, which belong to the regression-based category, seem to perform better in terms of predictive accuracy. Additionally, McHale and Morton (2011) reviews a logit model to predict the winner of a tennis game and discusses whether the outcome of the research can be applied to betting strategies.

2.3 Prediction Literature

In general, in the field of sports analytics, statistical and probabilistic models are employed to quantify and assess the decision-making process of football players and evaluate individual and team performance. Since we determine whether an action is unsuccessful or not the dependent variable is binary.

Therefore we employ the logit and probit models to estimate the parameters as it was reviewed by Horowitz and Savin (2001). In this paper, the authors review a general approach to the econometric problem of estimating the conditional probability that the outcome variable is 1 as a function of the explanatory variables, which in our case are the pass characteristics. In the logit model framework, the relative utility associated with each alternative is represented in a discrete choice model as described in Section 4.1. Similarly, the probit regression is described in Section 4.2. The maximum likelihood estimator is suggested for the parameter estimation due to its asymptotically efficient properties in large samples. In addition, due to the imbalanced nature of our data, maximum likelihood estimation is preferred.

To improve the predictive accuracy of our problem we employ state-of-the art methods in classification from Machine Learning literature. Namely, the random forest method will be used (Breiman, 1996a, 2001). These approaches are improvements of the naive decision tree algorithms which do not report such accurate results as the aforementioned methods (James, Witten, Hastie, Tibshirani et al., 2013).

As we fit several different models in our research, with probit and logit being used as the benchmark models, we focus on evaluation measures most commonly used in binary classification problems to determine which one will achieve the highest accuracy. Akosa (2017) reviews different techniques for addressing the problem of imbalanced data in classification problems evaluation. This thesis emphasises the importance of choosing the appropriate performance measure when selecting between different classifiers. The suggested metrics are described thoroughly in Section 4.4.

3 Data

A good understanding of the data is essential for accurate analysis. Moreover, data reprocessing is necessary to improve the algorithm's efficiency and performance. This section outlines the available data collected from Statsbomb (StatsBomb, 2023), it gives an overview of the available data set and a description of the important variables. More precisely, in Section 3.1 we describe the provider of our data and the available attributes we observe, in Section 3.2 we provide an overview of the quality of the data we use and, finally, in Section 3.3 we elaborate on the necessary transformations we performed.

3.1 Data Description

We are using the football event data provided by Statsbomb 360 free access data (StatsBomb, 2023). Our data provider is a well-known data platform specialising in collecting quality data for football and creating analytical tools for teams, individuals and betting companies. We are mostly interested in the event data as they are released in Statsbomb 360. Statsbomb 360 is an upgraded data specification of the provider including detailed event data and additional features such as the so-called freeze-frames, which display the location of all the players at the moment of an event.

With the term event data, we refer to a collection of detailed data on various events during a football game. Each event is a specific action, such as a pass, a shot or a substitution and each action has several characteristics or sub-events per action. For instance, for the pass event, we observe, among others, the length or the angle of the pass which will be explained thoroughly below. In addition, we observe general and less dynamic information about a football game such as the playing style of the teams, the outcome of the game etc. Hereby, this dataset contains both team and individual-specific information which will be used to create the risk profile of the players and eventually their team.

To access the free data set offered by Statsbomb we will use the Python package Statsbombpy (StatsBomb, 2023). This package contains data from various soccer competitions such as the FIFA World Cup, UEFA Euro, Champions League and many domestic leagues. The first season of collecting information is back in 1958 and the latest date we could find is 2022.

In this research, we will use the available data for the FIFA World Cup 2022 (male competition). When applying this filter, we get a variety of information like the specific identification of every game played in this tournament. Furthermore, for every unique game, we observe the

home and away team name, home and away team's score, manager and other higher-level information, which would be suitable for an analysis based on team performance. However, as we stated above, we will only utilize the events data available for this competition, which is at a more granular level and will fit our predictive models. Event data sets consist of every single action performed by a player during a match, containing crucial information such as the players involved, or the outcome of the action (Burriel & Buldú, 2021).

The employed data pertains to 111 unique types of events per each of the 64 played games in the competition. A possible drawback of this data set is the sparsity problem we observe since approximately 80% of the data points are null values. With the label null values, we describe a missing value of a certain attribute, either because it is not observed or due to the bad quality of the data. Given the large number of events per game, not all the variables per event are filled in our data set which generates the null (empty) values in our observed data. As we mentioned above, we are interested only in one action of the football players, the passing. In Table 1, we give a summarised description of the available attributes of the pass action and the possible values they can take, where we observe 3 categorical variables. Overall we have a large number of events, 234,652 pass events across all the games of the tournament. In Section 3.2 we elaborate further on the quality of the employed data set.

To have a better understanding of the used data, in Table 1 we provide the description of applicable events for the action of interest, the pass, where N/A denotes an empty value of the respective variable.

Variables	**Description**	**Values**
Length	The length of a pass in yards	Numerical values , i.e. 15.08
Height	Name Specifying the height of the pass	'Ground Pass', 'High Pass', 'Low Pass'
Angle	The angle of the pass in radians. Values between 0 and π, indicating an angle clockwise and with 0 pointing straight ahead.	Numerical values, i.e. 2.30
Body-Part	Name of the body part used to make the pass.	'Drop Kick', 'Head', 'Left/Right Foot', 'No Touch', 'Keeper Arm'
Outcome	Describes the outcome of the pass.	'Incomplete', 'Injury Clearance', 'Out', 'Pass Offside', 'N/A'.
Shot-Assist	Pass was an assist to a shot (not a goal).	TRUE or FALSE
Goal-Assist	Pass was an assist to a goal.	TRUE or FALSE

Table 1: Description of Pass events (variables).

As we observe there are both numerical and categorical variables for the pass events and we would like to distinguish the ones that are most applicable to our study. For instance, we use the 'Shot-Assist', 'Goal-Assist' and 'Outcome' variables to determine the outcome of a pass, and whether it was successful. For this purpose, we introduce the Outcome variable as described below:

- **Outcome 'O_i':** Binary variable with values 1 or 0, denoting whether the pass was unsuccessful (Outcome = 1). We define an unsuccessful pass as the one from which the team loses possession. Hereby, when we observe one of the values {'Incomplete', 'Injury Clearance', 'Out', 'Pass Offside'} for the 'outcome' variable as described in Table 1, we assign the value 1 to 'O_i' for pass $i = 1, \ldots, N$ with N the number of observed pass events.

To get a better understanding of the dataset we provide in Table 2 an example of an unsuccessful and a successful pass from the employed dataset. With the three dots (...), we denote the omitted variables from the sample data, since we only wish to provide an example of the transformation we used to construct the outcome 'O_i' variable. As we observe in Table 2, when

Outcome	**Height**	**Angle**	**Body-Part**	**...**	**Shot-Assist**	**Goal-Assist**
Incomplete	10.05	23.845	Right Foot	...	FALSE	FALSE
N/A	22.46	12.536	Head	...	TRUE	FALSE

Table 2: Sample data of pass events (categorical variables included).

the 'Outcome' variable is not empty, the 'Shot-Assist' and 'Goal-Assist' variables are FALSE and we observe an unsuccessful pass, while on the other hand, when the 'Outcome' variable is empty (N/A), one of the 'Shot-Assist' or 'Goal-Assist' variables is TRUE and we observe a successful pass. Therefore, incorporating the definition of the outcome variable 'O_i, we obtain the following transformed dataset, displayed in Table 3.

Outcome	**Height**	**Angle**	**Body-Part**	**...**	**Shot-Assist**	**Goal-Assist**
1	10.05	23.845	Right Foot	...	FALSE	FALSE
0	22.46	12.536	Head	...	TRUE	FALSE

Table 3: Sample data of pass events with the transformation of the outcome variable (categorical variables included).

It is worth mentioning that this is the initial transformation we perform to the dataset and further transformations will be described in Section 3.3, where we remove the 'Shot-Assist' and 'Goal-Assist' variables and take care of the categorical features.

3.2 Data Quality

In this subsection, we discuss the data quality of our data set. In our case, with the term data quality we refer to the completeness, consistency and fitness for our purpose of the used data sets. The data consists of 234,652 passes over the whole duration of the competition. Our main concern lies in the fact that many of the attributes of our interest are not sufficiently filled in. We illustrate the completeness of the pass variables in Figure 2. With this histogram, we aim to get a better understanding of the sparsity of the employed data set. We observe that the highest level of completeness is approximately 30% for the pass event.

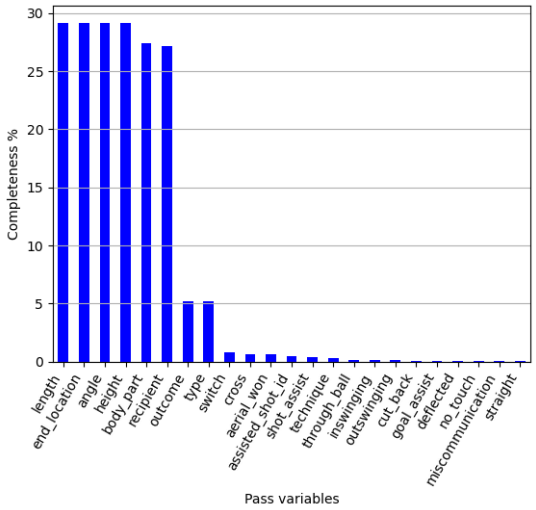

Figure 1: Histogram of completeness % of the Pass attributes.

Hereby, due to data quality and to avoid our dataset containing empty values we drop all the records that are empty and we observe 64,380 pass events without empty values. As mentioned above, we obtained the events from the FIFA 2022 male competition, which has different stages of games as described below:

- **Group Stage:** Consists of 32 teams and 48 games.
- **Round of 16:** Consists of 16 teams and 8 games.
- **Quarter-Finals:** Consists of 8 teams and 4 games.
- **Semi-Finals:** Consists of 4 teams and 2 games.
- **3rd Place:** Consists of 2 teams and 1 game.
- **Final:** Consists of 2 teams and 1 game.

We are interested in the different phases of the competition, as we analyse what are the differences in the completeness percentage of the variables of interest when the importance and attendance of a football game increases as we proceed to later stages. In the histogram displayed in Figure 2, we visualize our findings regarding the data quality of the pass attributes in different stages of the competition, namely for the groups stage and the final game.

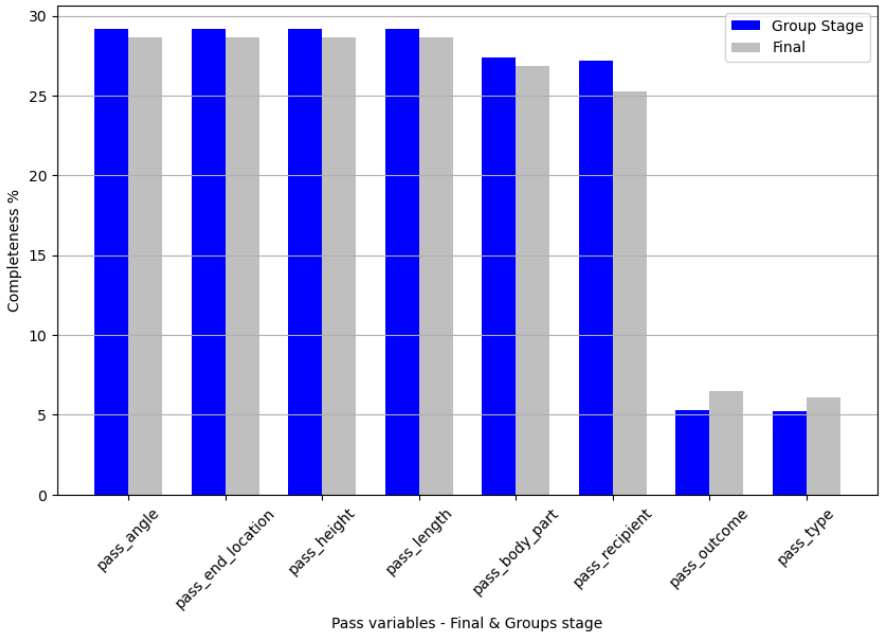

Figure 2: Histogram of completeness % of the Pass attributes in different stages.

In the groups stage, we observe 58,432 pass events (including empty records) with a completeness of approximately 30%, followed by a similar percentage in the competition's final game, approximating a completeness of 30% with 4,407 events. We choose to display the completeness rates of the least important and the most important games in Figure 2, since for the games of the remaining competition's stages, we observe a similar completeness with a range between 25% and 30%. Therefore, we conclude that the different stages of the competition do not differ significantly in the quality of the observed data. However, it is worthwhile to mention that the same set of attributes (angle, end_location, height, length, body_part, recipient) is sufficiently completed in every different stage, leaving type and outcome attributes the next best candidate variables in terms of data quality. Hereby, we conclude that we include all the different stages of the competition in our dataset for the implemented models.

To determine the outcome variable, we remove the empty records and we observe 64,380 pass events, where 10,952 pass events, roughly 17%, are unsuccessful, i.e. the 'Outcome' attribute is equal to 1. The aforementioned 64,380 pass events (unique passes) will be used for our model implementation for the estimation of the outcome variable.

To get a thorough understanding of the explanatory variables of our dataset we plot in Figure 3 the distributions of the two numerical attributes that are included in the set of independent variables. Since the rest of the pass attributes are binary, it would not be of significant importance to include their histograms in our visualization. In Figure 3, we observe the distributions of the 'Length' and 'Angle' attributes.

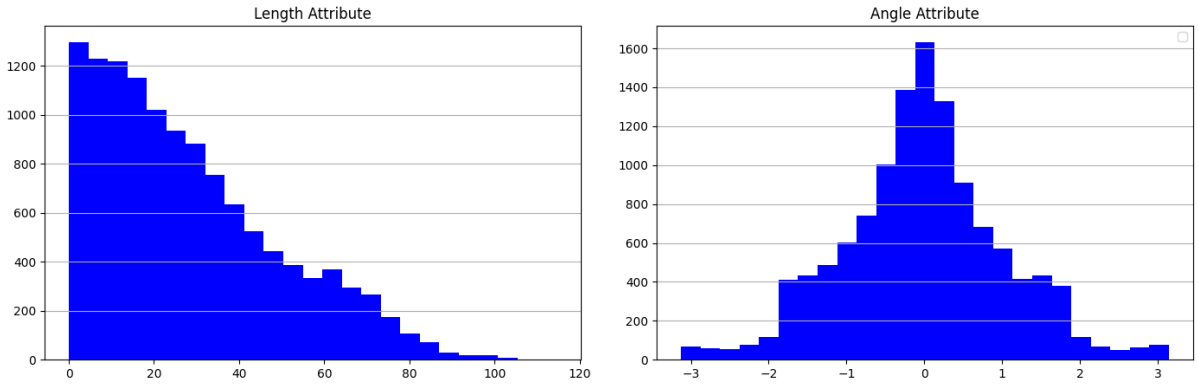

Figure 3: Histogram of the distribution of Length and Angle attribute.

The 'Length' attribute is measured in yards and as we observe from the above histogram, the majority of the passes are between 0 and 20 yards, a distance that covers approximately one fifth of the field. For the 'Angle' attribute, the majority of the observations lie in the range between -1 and 1 ,and thus, we mostly observe passes with relatively low angles.
As mentioned above, we observe both numerical and categorical values and for this reason, we need to further manipulate and transform the data for our model needs, as described in Section 3.3.

3.3 One-Hot Encoding

In this subsection, we elaborate on the transformation technique we use to translate the categorical variables into quantitative ones. In general, when some or all of the input variables are categorical, we need to define how they will be treated so they can be combined with numerical variables (Breiman, 2001). According to Hastie, Tibshirani, Friedman and Friedman (2009) in tree-based algorithms, the partitioning algorithm tends to favour categorical predictors with many levels and the more choices we have, the more likely we can find a good one for the employed data. If we denote by q the number of possible values per category, then for large q the computations become prohibitive. However, with a binary zero or one outcome, the computations simplify (Hastie et al., 2009).
According to Harris and Harris (2015), categorical data can be divided into two groups, which

are nominal (no particular order) and ordinal (with some particular order), and in our case, our input consists of nominal features such as the pass variable 'Body-Part' with possible values {'Drop Kick', 'Head', 'Right Foot', 'Left Foot', 'No Touch', 'Keeper Arm'}. Therefore, we use the One-Hot Encoding scheme for nominal variables. One-Hot Encoding transforms a single variable with N observations and d distinct values, to d binary variables with N observations containing either zero or one. Each observation indicates the presence (1) or absence (0) of the dichotomous binary variable. For instance, the aforementioned 'Body-Part' variable will be transformed into six new binary variables. A possible drawback of this technique lies in the fact that it produces high-dimensional representations increasing the sparsity problem of the data and introducing multicollinearity which negatively affects the logit and probit models.
After the implementation of the One-Hot Encoding transformation algorithm, we obtain five additional attributes, adding up to a total of nine explanatory variables in our dataset. To get a better understanding of the dataset we observed, we provide an example of an unsuccessful pass where we applied One-Hot Encoding in the 'Body-Part' categorical variable. Hereby, in Table 4, we illustrate the One-Hot Encoding transformation of the 'Body-Part' categorical variable from the initial sample data displayed in Table 2. Similarly, the table's columns will expand for the rest of the categorical explanatory variables.

...	**Drop Kick**	**Head**	**Left Foot**	**Right Foot**	**No Touch**	**Keeper Arm**
...	0	0	0	1	0	0
...	0	1	0	0	0	0

Table 4: One-Hot Encoded sample data of pass events (categorical variable 'Body-Part' from Table 2 transformed to numerical, binary).

As explained above, One-Hot Encoding generates binary variables for the categorical features where the sum of the rows per feature, i.e. per the set of generated binary variables per category, is equal to one. This might result in multicollinearity and to avoid that, we drop one of the binary columns per feature (Hastie et al., 2009). The criterion to drop one of the binary columns is the correlation between the explanatory variables. We computed the correlation matrix of the features set and we dropped the variable with the highest correlation per category. As a result, in Table 5, we observe the final transformation of the data as displayed in Table 4, where the column 'Keeper Arm' was dropped.
In this way, we ensure that we avoid collinearity with an intercept and, consequently, we reduce the number of explanatory variables for the logit and probit classification, from eleven parameters to nine.

...	Drop Kick	Head	Left Foot	Right Foot	No Touch
...	0	0	0	1	0
...	0	1	0	0	0

Table 5: One-Hot Encoded sample data of pass events (final transformation, one column per category dropped).

4 Methods

The following sections will give an overview of the econometric methods that will be applied in this research. In Sections 4.1 and 4.2 the logit and the probit models are introduced, for the estimation of the risk of a pass and the estimation of its components. Next in Section 4.3, we elaborate on the machine learning methods. Consequently, we discuss the split of the dataset into the train, test and validation datasets. Finally, evaluation measures are introduced in Section 4.4, to compare the performance of the models and make the the final selection of the applicable model.

4.1 Logit Model

Logistic regression is a linear classification method that models the posterior probabilities of the two classes via linear functions in x, where x is the vector of independent variables, while at the same time ensuring that they sum to one and remain in [0,1] (Hastie et al., 2009). In our case, for the binary setting of our target variable, we want to model the probabilities,

$$\mathbb{P}(O_i = 1 \mid X = x_i) = F(\beta_0 + x_i^T \beta) \tag{1}$$

where β_0 is the intercept, x_i is the vector of K explanatory variables, β is the parameter vector of the K explanatory variables, F is a known function, and $i = 1, \ldots, N$ with N the number of passes. For naming convention, we denote the outcome binary dependent variable as $O_i \in \{0, 1\}$, for $i = 1, \ldots, N$ with N the number of passes. Therefore we have the following specification,

$$O_i = \begin{cases} 1 & \text{when the pass } i \text{ is unsuccessful,} \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

where $i = 1, \ldots, N$ with N the number of passes.

According to Horowitz and Savin (2001), in this parametric approach where the function F is known and the model parameters (β_0, β) are unknown, the linear probability model specifies that the conditional probability is a linear function of $(X = x_i)$. This implies that the probability

may take either negative or greater than one values. Therefore, under the main assumption of the logit model, we set F to be the cumulative logistic distribution function which is symmetrical around 0 as displayed in the graph in Figure 4.

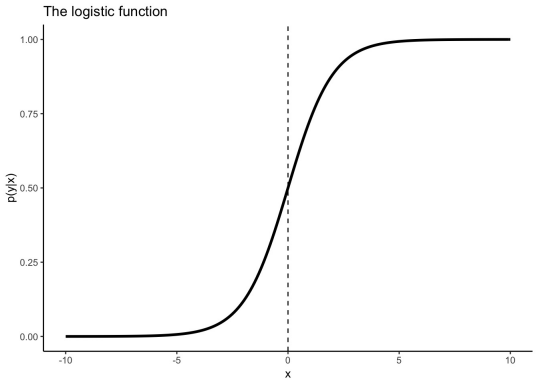

Figure 4: Logistic Function.

Applying the log-odds or logit transformations and the constraint that the target variable should stay between 0 and 1 we transform the aforementioned probabilities in equation (1) into the following equation:

$$\mathbb{P}(O_i = 1 \mid X = x_i) = \frac{\exp(\beta_0 + x_i^T \beta)}{1 + \exp(\beta_0 + x_i^T \beta)} \tag{3}$$

where they clearly sum up to 1, and $i = 1, \ldots, N$.

In this framework, the model parameters are $(\beta_0,\ \beta)$ and need to be estimated. Logistic regression models are usually fit by maximum likelihood, using the conditional distribution of O given X. Since $\mathbb{P}(O_i \mid X = x_i)$ completely specifies the conditional distribution, the multinomial distribution is appropriate. Hereby, we compute the log-likelihood for N observations and by maximizing this function we obtain the estimated parameters. The log-likelihood for N observation is given by

$$l(\theta) = \sum_{i=1}^{N} \log(p(x_i; \theta)) \tag{4}$$

where $p(x_i; \theta) = \mathbb{P}(O_i = 1 \mid X = x_i; \theta)$, where x_i is the vector of explanatory variables, $\theta = \{\beta_0, \beta\}$ is the parameter set and $i = 1, \ldots, N$. To maximize the log-likelihood, we set its derivatives equal to 0 and we obtain the score equations. We will not go into further details regarding the estimation of the model since it is out of the scope of this research.

The output for each action is a probability $\mathbb{P}(O_i = 1 \mid X = x_i)$, which determines the probability of a pass being unsuccessful, in other words, the risk of that pass. It is necessary to set a threshold probability for which any action above is classified as risky and below as not risky. The natural choice and most popular in literature is 0.5 (Manel, Dias & Ormerod, 1999). Nonetheless, this threshold will be subject to the tuning in our algorithms for both logit and probit models.

Denote with τ the different values of the candidate thresholds, where $\tau \in (0.05, 0.95)$ with the increment step equal to 0.05. As discussed later in Section 4.3.3, the tuning of the threshold is performed on the test set. We will evaluate the employed models with the different threshold values and we will make our final selection of the model with the best fit in Section 5.

4.2 Probit Model

In this section, we explore the probit regression to predict the likelihood of the examined action being unsuccessful, similarly to the logit model mentioned in Section 4.1. The baseline assumption for this specification is that the relationship between the independent variable and the binary outcome follows a normal (Gaussian) distribution. Hereby, the model specification as introduced in equations (1),(2),(3) can be written as

$$\mathbb{P}(O_i = 1 \mid X = x_i) = \Phi(\beta_0 + x_i^T \beta), \tag{5}$$

where Φ is the cumulative standard normal distribution with mean 0 and variance 1, whereas x_i is the vector of independent variables, $i = 1, \ldots, N$ with N the number of passes. In line with the logistic regression, we estimate the model parameters by maximum likelihood. To make the distinction visible between the two models we introduce the cumulative distribution graph in Figure 5

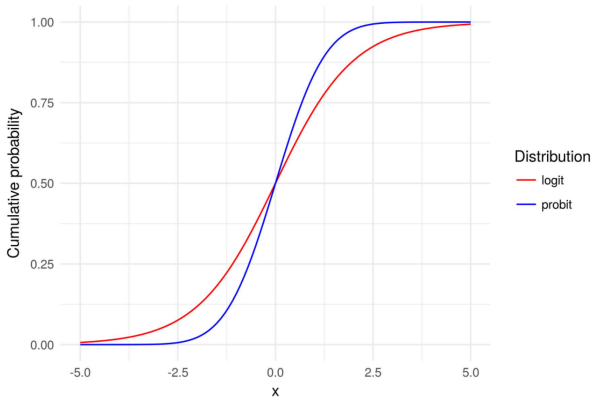

Figure 5: logit and probit Functions.

Overall, the logit model is more robust to outliers as it uses a logistic function while the probit model is more sensitive to outliers. Hereby, we aim to observe if this difference among the two models will guide us to significantly different outcomes.

The main disadvantage of the aforementioned methods in Sections 4.1 and 4.2 lies in the fact that they do not account for non-linear dependencies in the explanatory variables, as their decision criterion is linear (James et al., 2013). In addition, in problems with larger number

of explanatory variables, like in our case, it is likely that non-linear relationships will arise. Given that we encoded the categorical variables into numerical ones as a separate variable we further increased the number of independent variables. To overcome these limitations, we consider non-linear classification methods. Therefore, in the next section, we elaborate further on more complex classification algorithms which will be employed. We expect these methods to account for the non-linear dependencies which are present in our data and provide more accurate predictions. Logit and probit models will be used as the benchmark to compare their performance with the employed machine learning algorithms.

4.3 Tree-Based Methods

In this section, we describe tree-based methods for classification problems. The main reasoning behind the selection of the below classification methods lies in the fact that these are non-parametric algorithms. There are no assumptions made regarding the underlying distribution of the values of the predictor variables (Lewis, 2000). Initially, we give a concise introduction to the decision trees method in Section 4.3.1, followed by a description of bagging and random forest in Sections 4.3.2 and 4.3.3 respectively.

4.3.1 Decision Trees

In this section, we focus on decision trees, which can be split into two categories, regression and classification trees. Regression trees are not in scope for this research and we only focus on classification decision trees, since these are the ones used to predict a qualitative response (James et al., 2013). Decision trees divide the predictor space into several smaller, distinct and non-overlapping regions. This algorithm predicts that each observation belongs to the most commonly occurring class of training observations in the region to which it belongs. To illustrate the latter definition, consider a feature space X and the response variable Y. We split the space into two regions, and model the response by the mean of Y in each region. We choose the variable and split-point to achieve the best fit. Then one or both of these regions are split into two more regions, and this process is continued until some stopping rule is applied (Hastie et al., 2009). The result of this process is a partition into the set R of M distinct regions, with $R = \{R_1, R_2, \ldots, R_M\}$ provided by the function $\hat{f}(X)$ given in Equation (6)

$$\hat{f}(X) = \sum_{m=1}^{M} c_m \mathbb{I}\left\{(X_1, X_2) \in R\right\}, \tag{6}$$

where c_m a constant in region R_m and $m = 1, \ldots, M$, with M the number of total regions.
An important aspect of this method which is also subject to tuning is the splitting rule. According to (James et al., 2013), in the regression setting, we use binary splitting to grow a classification tree and try to find the regions that minimize the residual sum of squares (RSS). However, in the classification setting RSS can not be used as a criterion and we consider its natural alternative, the classification error rate. On the other hand, it has been proved that the classification error rate is not sufficiently sensitive for tree-growing and we prefer another measure, the Gini index as it is defined by Equation (7)

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}), \tag{7}$$

where $\hat{p}_{mk}$ represents the proportion of training observations in the $m-th$ region that are from the $k-th$ class and in our case $k = 2$ since the predicted outcome is either zero or one. Gini index can be interpreted as a measure of variance within the $m-th$ region, and this error measure takes a small value if all the $\hat{p}_{mk}$ are close to zero or one.
In general, one of the main advantages of decision trees is their interpretability. On the other hand, they suffer from high variance and they are not a robust predictive method since a small change in the training data set can alter the predicted outcome. The main reason for this instability is the hierarchical nature of the process: the effect of an error in the top split is propagated down to all of the splits below it (Hastie et al., 2009). In addition, we can categorize the decision trees based on their depth and bias-variance trade-off. Deep trees can have lower bias but higher variance and shallow trees can have lower variance but higher variance.
Another concern about decision trees is overfitting the data, resulting to poor test performance. As described in Equation 6, we partition the feature space into R distinct regions, however, when the number of regions is increasing the tree might be too complex leading to overfitting the data (James et al., 2013). To tackle this problem, another strategy is implemented which results in smaller trees, the so-called pruning. This approach aims to grow a large tree T_0 and then prune it back to obtain a subtree which minimizes the test error rate. According to James et al. (2013), calculating the test error for each subtree would be too cumbersome and an alternative method was suggested. Namely, the cost-complexity pruning considers a sequence of subtrees, instead of the whole set of subtrees, indexed by a nonnegative tuning parameter α. The idea of cost-complexity pruning is to minimize the value displayed in Equation 8, for each value of α which corresponds to a subtree $T \subseteq T_0$

$$\sum_{m=1}^{\Sigma} \sum_{i:x_i \in \mathbb{R}^m} (y_i - \hat{y}_{R_m})^2 + \alpha|T|, \tag{8}$$

where $|T|$ indicates the number of terminal nodes of the tree T, R_m the subset of predictor space corresponding to the $m-th$ terminal node, and $\hat{y}_{R_m}$ is the predicted response associated with R_m, i.e. the mean of the training observations in R_m.

Taking the above restrictions of decision trees into account, alternative classification algorithms need to be considered. For that reason, we introduce ensemble methods to obtain more powerful models. An ensemble method is an approach that combines many simple 'building block' models to obtain a single and potentially very powerful ensemble model. Also known as weak learners, these building block models may lead to mediocre predictive performance on their own (James et al., 2013). In the next sections, we will study the ensemble method for which the building block is a classification tree. Initially, we introduce the procedure of bagging or bootstrap aggregation, which was introduced prior to the random forest by Breiman (1996a). This approach averages predictions over a collection of bootstrap samples for regression problems and applies the majority voting rule for classification to reduce the variance. Consequently, the random forest was built on this idea, improving its performance by decorrelating trees.

4.3.2 Bagging

The reasoning behind bagging lies in the fact that decision trees as discussed in Section 4.3.1 suffer from high variance. For instance, if we randomly split our data set into two parts and fit a decision tree, the results will potentially be significantly different (James et al., 2013). On the other hand, a procedure with low variance will yield similar results if applied repeatedly to distinct data sets. Bagging, also known as bootstrap aggregation, is a procedure to reduce the variance of a statistical learning method.

Bagging of decision trees is the process of constructing B decision trees for each bootstrap sample of size N and then using the majority vote, for classification problems, to assess the predictions. To further understand the implementation of this method we will illustrate it utilizing an example. Suppose we have a set of N observations $Z_1, Z_2, \ldots, Z_N$ each with variance σ^2. Then, the variance of the mean $\bar{Z}$ of the observations is given by σ^2/N. Therefore, averaging a set of observations reduces variance (James et al., 2013). Denote with $\hat{f}(x)$ the resulting prediction for input x. According to Hastie et al. (2009), we average the prediction over a collection of B bootstrap samples, where we define $\hat{f}^{*1}(x)$, $\hat{f}^{*2}(x)$, $\ldots$, $\hat{f}^{*B}(x)$ as the predicted

value for B different bootstrapped training sets. The average of all predictions is the definition of the bagging estimated given in equation (9)

$$\hat{f}_{\text{bag}}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{*b}(x), \tag{9}$$

for $b = 1, 2, \ldots, B$. As mentioned above, in the classification context, for test data we record the class predicted by each of the B trees and take a majority vote. Namely, the average prediction is the most commonly occurring class among the B predictions. One of the main assumptions in bagging is that this method implicitly assumes trees are independent. However, in some cases, the random training datasets derived from the same initial dataset can be highly correlated. Therefore, in the next section, we elaborate on a method which was built on the idea of bagging and it significantly reduces the correlation between the B trees, thus decreasing the variance of the overall estimate.

4.3.3 Random Forest

We will study random forest as they were initially introduced by Breiman (2001). Random forest is a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest (Breiman, 2001). As a result, this method provides significant improvement in classification accuracy, which mainly occurs by growing an ensemble of trees and letting them vote for the most popular class. Random forest can be interpreted as a modification of bagging, as it manages to reduce the correlation among threes by applying random feature selection.

The random forest method builds several forests of decision trees on bootstrapped training samples. But when building these decision trees, each time a split in a tree is considered, a random sample of m predictors is chosen as split candidates from the full set of p predictors. The split is allowed to use only one of those m predictors. A fresh sample of m predictors is taken at each split, and typically we choose $m \approx \sqrt{p}$, that is, the number of predictors considered at each split is approximately equal to the square root of the total number of predictors (James et al., 2013). Note that if $m = p$ then we are in the generalized case of bagging, a special case of random forest. To further illustrate the functionality of this method we provide the random forest Algorithm 1, as it is described by Hastie et al. (2009).

Algorithm: Random Forest - Classification.

1. For $b = 1, \ldots, B$:
   (a) Draw a bootstrap sample Z^* of size N from the training data.
   (b) Grow a random-forest tree T_b to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree until the minimum node size n_{min} is reached.
      i. Select m variables at random from the p variables.
      ii. Pick the best variable/split-point among the m candidates.
      iii. Split the node into two daughter nodes.
2. Output the ensemble of trees $\{T_b\}_1^B$.

To predict a new point x in the classification context:

Let $\hat{C}_b(x)$ be the class prediction of the b-th random-forest tree.
Then,

$$\hat{C}_{\mathrm{RF}}^B(x) = \text{majority vote}\{\hat{C}_b(x)\}_1^B.$$

Algorithm 1: Random Forest - Classification.

At each split, Algorithm 1 considers only a random sample from the whole predictor space for splitting. Consequently, each of the B trees is grown to a different bootstrapped sample. The reasoning behind this algorithm can be explained in the next example. Suppose in the predictor set, there is one very strong predictor along with many other moderately strong predictors. In that case, bagged trees will use the strong predictor in the top split, thus all of the bagged trees will look significantly similar to each other and they will be highly correlated. Therefore, averaging over highly correlated quantities does not cause the same variance reduction as averaging over uncorrelated quantities, as random forest do.

The random forest described in Algorithm 1 has several hyper-parameters that need to be set by the researcher, however, in most cases, it works reasonably well with the default values of these hyper-parameters specified in software packages (Probst, Wright & Boulesteix, 2019). These hyper-parameters control the structure of each tree (node size), the structure and size of the forest (number of trees) as well as its randomness (number of candidate variables) (Probst et al., 2019). In Table 6, in columns 'Hyper-parameter' and 'Description' we provide an overview of the most effective hyper-parameters of the random forest as they are described by Probst et al. (2019) and in column 'Default Values' we provide the default values of the implemented software,

Hyper-parameter	Description	Default Values
Number of trees	The of trees in the forest	500, 1,000
Number of candidates	Number of drawn candidate variables in each split	$\sqrt{p}$
Node size	Minimum number of observations in a terminal node	1
Replacement	Draw observations with or without replacement	TRUE (with replacement)
Splitting rule	Splitting criteria in the nodes	Gini impurity, p-value, random.

Table 6: Description and default values of random forest hyper-parameters.

where p is the number of explanatory variables in the dataset. As we mentioned above bagging is a special case of random forest when the number of candidates is equal to p.

According to Probst et al. (2019), random forest is less tunable than other machine learning methods, in most cases however, there is a small performance gain to be achieved by tuning the hyper-parameters with the largest impact on the performance of the algorithm. For instance, the number of trees will not be tuned as they should be set sufficiently high (Probst & Boulesteix, 2018) and for that reason, we set them at 1,000. Similarly, we prefer the default values for the Replacement and for the Splitting rule for classification, and we fix these values to 'True' and 'Gini impurity' respectively, as introduced by (Breiman, 2001).

Regarding the hyper-parameters we select for tuning, we optimize the number of candidates and the node size. The possible values for the number of candidates are (1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12), where we should mention that the default value is $\sqrt{p} = 3$ and for $p = 12$, we are in the special case of bagging. For node size, the respective set of values is (1, 2, 5, 10, 20, 35, 50, 75), where we are using a large increment on the candidate values since the computation time decreases with an increase in node size without a substantial loss of the algorithm's performance (Probst et al., 2019). To perform the aforementioned tuning, we implement the grid-search approach, where all possible combinations of the tunable hyper-parameters are evaluated. To decide upon the best combination of hyperparameters, we employ the out-of-bag estimate as described in (Breiman, 1996b).

Out-of-bag (OOB) estimate measures the misclassification error utilizing bootstrap aggregation, the so-called bagging. As described in Algorithm 1, when we train the random forest classifier we draw a bootstrap sample from the whole training data, leaving a portion of the data unused in the training of the algorithm. Therefore, these out-of-bag observations will be used to estimate the out-of-bag misclassification error of the classifier. Consequently, we estimate the OOB predictions to define the optimal set of hyperparameters.

After the selection of the optimal hyper-parameter set, we need to evaluate the performance of the employed algorithms and to achieve that, we introduce the evaluation measures in the next section. Moreover, we describe the splitting procedure of the available data to train, test and validate the performance of the employed models.

4.4 Evaluation Measures

In this section, we elaborate on the method we used for the training and tuning of the predictive models and we describe the evaluation measure that will be implemented. Before the models were trained and fitted to get the prediction outcome we applied the so-called 60/40 rule to divide the training, test and validation dataset. The first 60% of the data will be employed for training the predictive models (logit, probit and random forest). Furthermore, we split the remaining 40% of the dataset into two equal parts of 20% each, where the tuning of the threshold value and the hyperparameters will be performed on the first 20% of that dataset, the test set. The latter 20% of the dataset will be the validation set which will be used to evaluate the predictive performance of the selected model based on the measures described in Section 4.4. Other suggestions for the splitting rule are 70/30 or 75/25, however, empirical studies suggest that the initial splitting rule does not drastically affect the algorithm's performance (Gholamy, Kreinovich & Kosheleva, 2018).

The most commonly reported model evaluation metric is predictive accuracy. This metric can be misleading when the data are imbalanced, which is the case in our research. This is the case because more weights are put on the majority class rather than on the minority class, which makes it more difficult for a classifier to perform well on the minority class. An imbalanced set occurs when one class has lower proportions in the data compared to the other class (Akosa, 2017).

Our main goal is to correctly classify the minority class, which in our case are the unsuccessful passes (outcome = 1). This is not always trivial, since the majority of the classifiers tend to favour the majority class and perform poorly on the minority class. Akosa (2017) suggests that many techniques such as down-sampling or up-sampling can improve the performance of the model during parameter tuning even though they introduce bias into the results. Therefore, other evaluation metrics should be considered in addition to the accuracy based on the performance measure of the testing subset.

Hereby, the evaluation measures we use are based on the $TP(True\ Positive)$, $FN(False\ Negative)$, $FP(False\ Positive)$ and $TN(True\ Negative)$ cases. In our case, with two classes we get the following confusion matrix (also known as error matrix) as displayed in Table 7.

For binary classification, the confusion matrix is used to record the correctly and incorrectly predicted classes.

Actual /Predicted	As Positive	As Negative
Positive	TP	FN
Negative	FP	TN

Table 7: Confusion matrix for binary classification

By convention, in imbalanced data we consider the minority class as the positive class while the majority class is considered to be the negative one. In Table 8, we provide an overview of the most commonly accepted performance evaluation measures based on the main assumption that our variables are identical and independently distributed (IID).

Evaluation Measure	**Formula**
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
Misclassification Rate(1-Accuracy)	$\frac{FP+FN}{TP+TN+FP+FN}$
Sensitivity (True Positive Rate)	$\frac{TP}{TP+FN}$
Specificity (True Negative Rate)	$\frac{TN}{TN+FP}$
Precision (Positive Predictive Value)	$\frac{TP}{TP+FP}$

Table 8: Evaluation Measures for Binary Classification.

To get a better understanding of the aforementioned measures, we describe each one of them according to Sokolova, Japkowicz and Szpakowicz(2006) and we elaborate on how appropriate they are in our case:

- **Accuracy:** The most used empirical measure which, however, does not distinguish between the number of correct labels of different classes. It assesses the overall effectiveness of the algorithm by approximating how effective it is by showing the likelihood of the true value of the class label.
- **Misclassification Rate:** The percentage of times the classifier is incorrect.
- **Sensitivity:** Focuses on one class (true positive) and approximates the probability of the positive class being true. It assesses the effectiveness of the algorithm on a single class (the minority class in our case).

- **Specificity:** Similarly with Sensitivity but now with the focus on the negative class.
- **Precision:** Distinguish the correct classification of labels within different classes as it is concentrated on one class. It estimates the predictive value of a label depending on the class for which it is calculated. In other words, it assesses the predictive power of the algorithm.

In general, we are focused on the ability of the algorithms to distinguish classes and avoid misclassification. Therefore, given the aforementioned set of evaluation measures, we will make our selection on the final model to fit our dataset.

5 Results

In this section, we discuss and evaluate the results retrieved from the proposed methods in Section 4 and provide an exploratory analysis of teams' and players' performance. Initially, we observed the number of unsuccessful passes per player and team and we categorize them based on their risk rates. We define the risk rate of a player or a team by dividing the number of unsuccessful passes per total number of passes. For each game, we observe the variable 'tactics', which determines the formation of each team at the beginning of the game. As a result, we compute the rate of unsuccessful passes per different playing styles (formation). Finally, we display the outcome of the employed models in terms of performance and evaluation criteria before we conclude on selecting the best method.

5.1 Exploratory Analysis

Before we proceed to analyse the results of the employed algorithms and their performance, we provide an overview of the risk rates of the players, teams and different formations that were observed during the FIFA 2022 competition. The employed dataset as described in Section 3, allows us to relate every pass event with the player that took this action, the team of that player and the formation that was being used by that team. Consequently, we can identify the unsuccessful passes per player, per team and playing style of each team. This analysis will give us a clear overview of the risk rates of the players and eventually will identify the risky teams and the risky playing styles, as the ones with higher risk percentages. To achieve that, we illustrate the analysis of the teams' and players' results in different subsections. We provide the distribution of the unsuccessful passes per player and per team.

5.1.1 Player Analysis

In this section, we aim to identify the risk-averse and risk-taker players based on their performance and decisions in the studied competition. In this study, we measure the player's performance based on their rate of unsuccessful passes. It is worth mentioning that not all the players play the same amount of games since this competition follows a knock-out process for determining the winner. In total, we observe 680 players and on average each of them makes 343 passes in the duration of the competition. In addition, as it is displayed in Figure 6, we observe that each player makes on average 105 passes per game where we observe that our observations are concentrated in the range of $(50, 175)$.

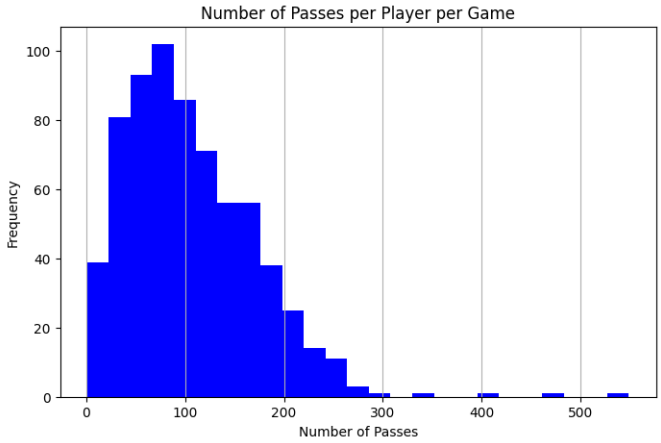

Figure 6: Histogram of number of passes per player.

To further investigate, the passing behaviour of the studied football players we are interested in the distribution of the risk rate of each player. Namely, we calculate the percentage of unsuccessful passes, as it is defined in Section 3, and we illustrate our findings in Figure 7.

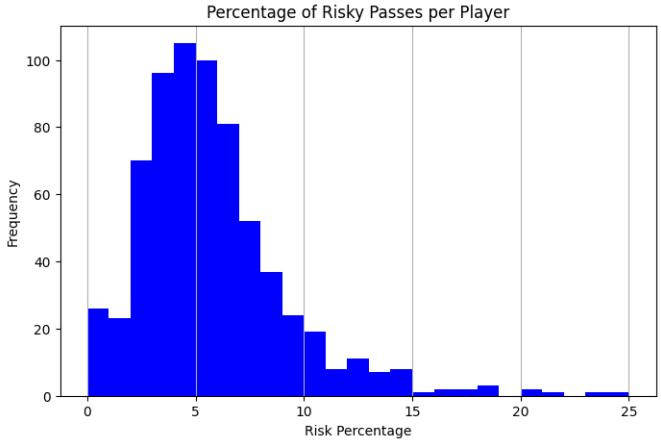

Figure 7: Histogram of risk rate of passes per player.

From Figure 7, we observe that the risk rate of passing is approximately 5.7% per player. There are some outliers with a risk percentage as high as 25%, however, these are unique, exceptional cases which do not alter the results of this study. Next, we focus on similar results on a team level.

5.1.2 Team Analysis

In this section, we aim to describe the risk rate per team and therefore per different formations that are being used throughout the competition and we do so by aggregating the risk rates of passes per team. With this analysis, we provide valuable insights into the team's profiles regarding the riskiness of pass and their formations. In combination with our predictive analysis,

we identify risk-seeking teams or formations based on their passing behaviour.

First, we visualize the number of passes per team per game, to get a clear understanding of the volume of passes. Hereby, in Figure 8, we observe that the vast majority of the teams complete approximately 1,800 passes per game.

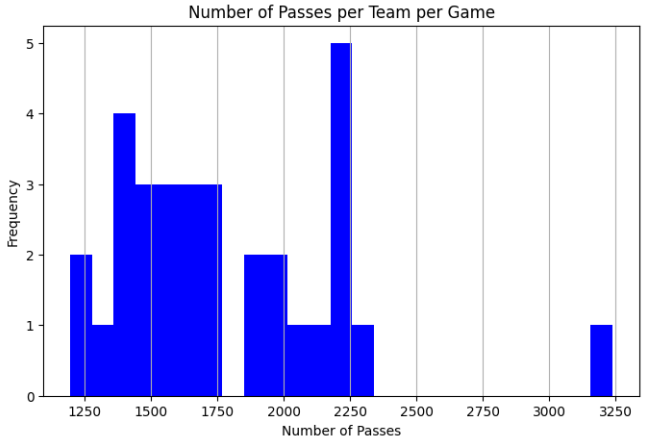

Figure 8: Histogram of Number of Passes per Team.

Hence, we conclude that we observe roughly 3,600 passes per game. It is important to mention in this graph the outlying observation with roughly 3,250 passes. This is the number of passes per game for Spain, which is by far the team with the most passes per game in the competition. It is worth stating that Spain is also the team with the lowest risk rate, roughly 3.1%. Given the history of the playing style of the country and its reputation for quick and accurate passing, our findings confirm that it is the most efficient passing team in the competition. In addition, by speculating the above histogram we can identify three possible groups of teams, one with a relatively low number of passes per game, another one with a moderate number of passes and the outlying observation.

Next, we will elaborate on our results regarding the risk rate of the 32 teams and the formations that have been used. In Figure 9, we observe that the risk rate per team is on average 5.5%, which is close enough to the risk rate per player as it is stated in Section 5.1.1. Similarly, with the volume of passes, we identify three groups of teams, one with a low-risk behavior with a risk rate lower than 4.8%, a group with a moderate risk rate between 5.0% and 6.0% and the group with the riskier teams with rates higher than 6.6% up to 8.0%.

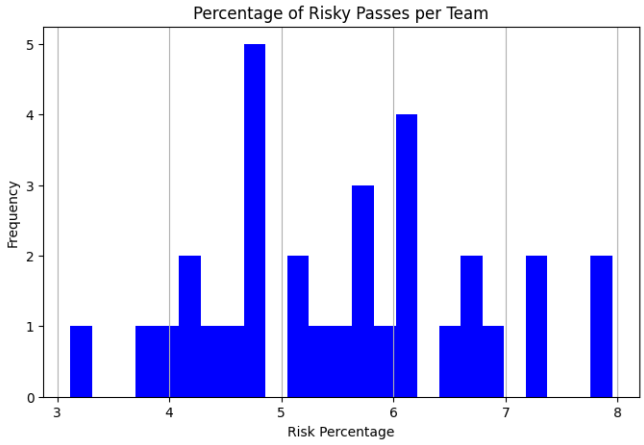

Figure 9: Histogram of Risk Rate of Passes per Team.

Regarding the formations that have been used, '4-2-3-1' and '4-3-3' are the most popular ones, being used in 46 and 41 games respectively. The risk rate for this formation is less than 0.1% and due to the lack of data availability we can not elaborate further.
In the next sections, we discuss the performance of the predictive methods that have been employed. Moreover, we provide the results of the evaluation measures we applied and we suggest the final selection of the most suitable method for this research based on the model comparison we conducted.

5.2 Evaluation Measures and Performance

As described in Sections 1 and 4, we aim to identify the risk components of a pass. Given the available variables and the transformations we performed according to Section 3, we can provide valuable insights regarding the importance and the effect of each of the studied variables. In addition, we perform a predictive analysis since our ultimate goal is to be able to predict whether a team, a player or a formation is risky or not based on their passing behaviour. To achieve that we employed various methods, that are well described in Section 4. Thereafter, in this section, we discuss the results occurring from these algorithms and we evaluate their performance.

5.2.1 Logit and Probit Models

In this section, we discuss the coefficient estimates of the logit and probit regression, as well as the predictive performance of the logit and probit classification models. These predictive models are used as the benchmark methods to be compared with the tree-based model. According to the description of the two methods in Section 4, the binary dependent variable is whether the pass is unsuccessful or not and we estimate the likelihood of the pass being unsuccessful,

i.e. the risk of the pass. In Tables 9 and 10, we display the results of the aforementioned methods.

Logit Regression						
Variables	**Coefficient**	**Std. Error**	**z**	**p - value**	**[0.025**	**0.975]**
Constant	-1.317	0.169	-7.791	0.000***	-1.649	-0.986
Length	-0.005	0.001	-5.669	0.000***	-0.007	-0.003
Angle	-0.013	0.008	-1.533	0.125	-0.030	0.004
Ground Pass	-3.033	0.036	-83.149	0.000***	-3.104	-2.961
Low Pass	-0.800	0.042	-18.847	0.000***	-0.884	-0.718
Drop Kick	2.828	0.249	11.378	0.000***	2.340	3.315
Head	1.327	0.170	7.796	0.000***	0.993	1.661
Left Foot	1.962	0.166	11.788	0.000***	1.636	2.288
No Touch	3.261	0.314	10.385	0.000***	2.646	3.877
Right Foot	1.818	0.166	10.966	0.000***	1.493	2.143

Table 9: Logit Regression Results (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$).

Probit Regression						
Variables	**Coefficient**	**Std. Error**	**z**	**p - value**	**[0.025**	**0.975]**
Constant	-0.679	0.088	-7.727	0.000***	-0.851	-0.506
Length	-0.004	0.001	-7.318	0.000***	-0.005	-0.003
Angle	-0.007	0.005	-1.509	0.131	-0.016	0.002
Ground Pass	-1.771	0.021	-84.959	0.000***	-1.812	-1.730
Low Pass	-0.513	0.026	-19.900	0.000***	-0.563	-0.462
Drop Kick	1.650	0.138	11.913	0.000***	1.378	1.921
Head	0.701	0.089	7.892	0.000***	0.527	0.875
Left Foot	1.097	0.086	12.772	0.000***	0.929	1.266
No Touch	1.795	0.178	10.091	0.000***	1.446	2.144
Right Foot	1.020	0.086	11.920	0.000***	0.852	1.187

Table 10: Probit Regression Results (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$).

In general, for both logit and probit regression, we observe the same set of parameter coefficients have a positive sign, thus these are the variables that increase the risk of a pass. Additionally, in terms of the statistical significance of the estimated variables, only for the 'Angle' variable we report a $p - value$ larger than 0.05. Therefore, on a 5% significance level test, all the variables, except the 'Angle', are statistically significant in both logit and probit regressions. Moreover, we interpret the coefficient estimates as the expected change in the log-odds of a pass being unsuccessful for a unit increase in the corresponding explanatory variable, given that

all the other variables are constant. In the case, of binary explanatory variables, we estimate the expected change in the risk of the pass (i.e. a pass being unsuccessful) when the binary explanatory variable is equal to one.

In Tables 9 and 10, we observe that the variables 'Ground Pass' and 'No Touch' are the ones with the largest effect on the risk of a pass. In other words, when the committed pass is a ground pass there is a decrease in the risk of a pass, whilst when the pass is committed with no touch (direct contact with the ball, without control), there is an increase in the likelihood of the pass being unsuccessful. Given the logit regression estimates (probit regression estimates), if the pass is a 'Ground Pass' there is a decrease in the log-odds of the pass being unsuccessful of $\exp(3.033) = 20.756$ ($\exp(1.771) = 5.876$) or equivalently $1,975\%$ (487%). Similarly, if the pass is a 'No Touch', there is an increase in the log-odds of the pass being unsuccessful of $\exp(3.261) = 26.075$ ($\exp(1.795) = 6.019$) or equivalently $2,507\%$ (501%). Incorporating the standard errors of the aforementioned estimates, we obtain a 95% confidence interval of $\exp(3.261 \pm 2 \times 0.314) = (13.915, 48.867)$, for the 'No Touch' variable in the logit regression approach.

Next, we focus on the prediction performance of the two methods. As mentioned in Section 1, the goal of this study is to offer a tool such that coaches and players can identify a risky player or a risky team based on their pass characteristics. Therefore, the prediction performance of these methods is crucial to our research objectives.

In addition, as discussed in Section 4.1, we perform tuning of the threshold hyper-parameter of the logit and the probit model on the test set and we observe the following accuracy values per different thresholds as illustrated in Figure 10 for the logit model. There is no necessity to report the same figure for the probit model because it is quite similar. Even though the range of the threshold values in Section 4.1 is between 0.05 and 0.95, we only report the accuracy for the threshold values above 0.15, since for the threshold values 0.05 and 0.10, the accuracy was below 20%. Therefore, by visually inspecting Figure 10, we conclude that the default value of threshold $\tau = 0.5$ achieves the highest accuracy and will be our final selection for this hyper-parameter.

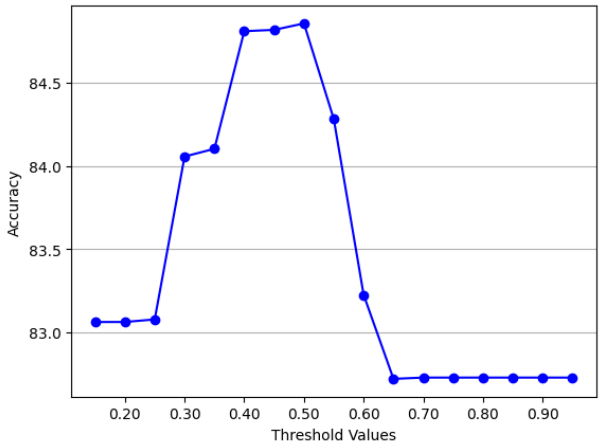

Figure 10: Threshold tuning for logit model.

According to the evaluation measure as discussed in Section 4.4, we display in Table 11 the results for the two classical approaches. Overall, we observe that the predictive performance of

Evaluation Measure	**Logit**	**Probit**
Accuracy	84.86	84.85
Misclassification Rate	15.14	15.14
Sensitivity	45.05	45.19
Specificity	93.17	93.14
Precision	57.95	57.89

Table 11: Logit and probit models evaluation measures.

the two methods is close to each other, with slight differences in Sensitivity, Specification and Precision. In general, this is an expected outcome since relevant research suggests that in the majority of the univariate binary response models, probit and logit yield similar results (Hahn & Soyer, 2005). It is worth mentioning that some differences can be found in the third decimal place, however, it is not in the scope of this research and will not add value to our objectives to study these measures so deeply. In contrast, we maintain this outcome as the benchmark predictive performance to compare it with the respective results of the tree-based model.

5.2.2 Random Forest

As described in Section 4.3, next to the training of the algorithms and the evaluation of their performance, we identify the optimal hyper-parameters, which are the result of the tuning we performed on the test set. As discussed in Section 4.4, there is a variety of metrics to evaluate the performance of the employed ensemble method and initially, the accuracy measure is selected. To illustrate the improvement of the algorithms' predictive power with the optimal

hyper-parameter set, initially, we display in Table 12 the accuracy values without tuning.

Evaluation Measure	**Random Forest**
Accuracy	85.93

Table 12: Random forest evaluation measures (without tuning).

Next, we proceed to the tuning of the hyperparameters for the aforementioned classifier on the test set, using the OOB misclassification error rate as described in Section 4.4. After performing the tuning of the two hyperparameters given the possible candidate values described in Section 4.3.3, we obtain the following set of hyper-parameters for random forest as displayed in Table 13.

Hyper-parameter	**Random Forest**
Number of trees	1000
Number of candidates	10
Node size	50
Replacement	TRUE
Splitting rule	Gini impurity

Table 13: Random Forest Optimal Hyper-parameters (as described in Section 4.3 in Table 6).

As discussed in Section 4.3.3, we maintain the default values for the Splitting rule and the Replacement which are set to 'Gini impurity' (for classification) and 'True' respectively, while we set the number of trees equal to 1000.

To get a better understanding of the effect of the different hyper-parameters values on the algorithm's performance and the reasoning behind our selection, we illustrate in Figure 11 the accuracy values for different numbers of candidates, while we set the rest of the hyper-parameter values to the optimal ones displayed in Table 13.

From Figure 11, we observe that there is an improvement in the model's performance as we increase the number of candidates, where we achieve the highest accuracy when we set it equal to 10. As discussed in Section 4.3.3, random forest increase their accuracy performance by applying random feature selection and Figure 11 verifies this statement.

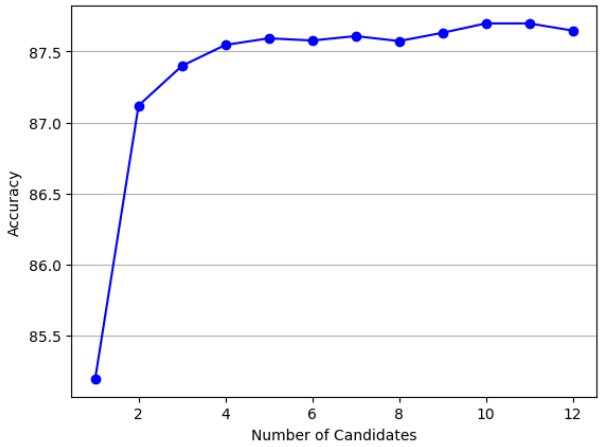

Figure 11: Random forest performance per different number of candidates.

In addition, in Figure 12 we observe the different accuracy values per different node size values. We observe that as we increase the minimum number of observations in a terminal node, there is a substantial increase in the algorithm's predictive performance, in line with the research's results by (Probst et al., 2019). This is a hyper-parameter with a sufficient impact on the algorithm's performance and we set it equal to 50 as it seems to stabilize its effect on the algorithm's performance for this value onwards.

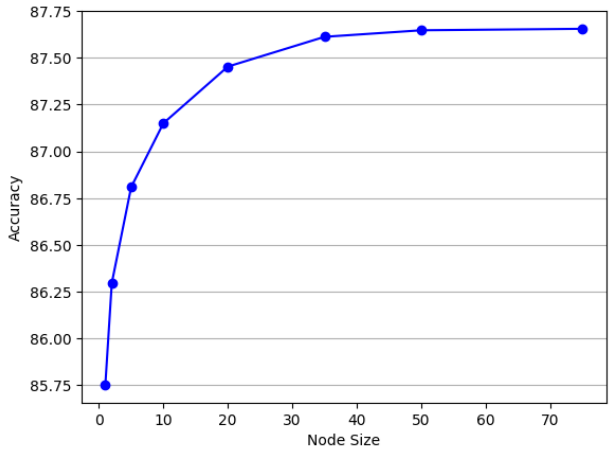

Figure 12: Random forest performance per different node size values.

In line with the logit and probit threshold tuning, we perform a similar tuning for random forest using the same range for the threshold value τ where $\tau \in (0.05, 0.95)$ with the increment step equal to 0.05. However, we do not observe an improvement in the algorithm's performance for threshold values different than the default value ($\tau = 0.50$), thus we do not illustrate the different accuracy values for different thresholds. Hereby, we conclude that due to the increased predictive performance of the classifier after the tuning, the optimal set of optimal hyper-parameters to be used will be the one provided in Table 13.

We obtain the improved performance results, after tuning, in Table 14, where we visualize all the performance metrics as discussed in Section 4.4 for the random forest, with tuning.

Evaluation Measure	**Random Forest**
Accuracy	87.71
Misclassification Rate	12.29
Sensitivity	55.53
Specificity	94.42
Precision	67.52

Table 14: Random forest evaluation measures (with tuning).

Overall, there is a sufficient improvement in the accuracy of the random forest method and the increased values of sensitivity and specificity of the aforementioned method also confirm this. The ultimate goal of this study is to provide an insightful predictive tool for the likelihood of a pass being unsuccessful based on its characteristics, i.e. the risk of the pass based on its characteristics. Thereafter, we study the variable importance methods that have been introduced in the literature to estimate the most effective predictors. Mainly, there are two different approaches for measuring the variable's importance, the Gini criterion and the method based on the mean decrease of the out-of-bag observations (Archer & Kimes, 2008). According to Archer and Kimes (2008), for bootstrapped sampled methods, the out-of-bag observations can be used to calculate the variable importance of the random forest. As a result, we illustrate in Figure 13 the variable importance measures.

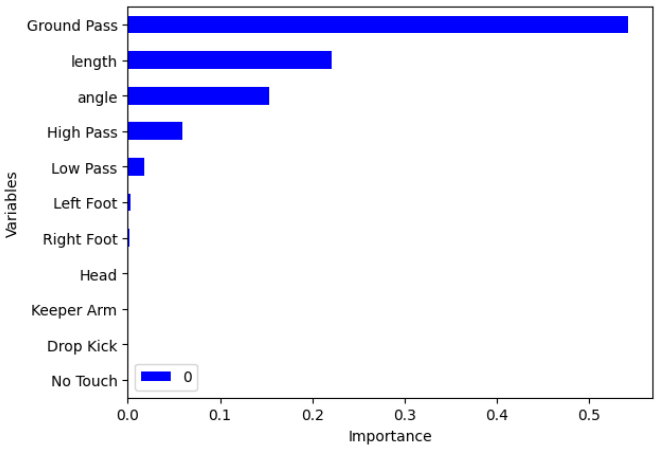

Figure 13: Random Forest - Variable Importance.

From Figure 13, we can identify four variables with sufficient importance in the prediction of a risky pass and these are Length, Angle, Ground Pass and High Pass. We could possibly speculate that Low Pass can be included in the set as well. Given the fact that the random forest

classifier favours the numerical features, it is reasonable that Length and Angle are the ones with the highest importance scores. Moreover, it is reasonable to cite that there are two potential groups of predictors, one could be the pass characteristics in terms of the ball movement, i.e. Length and Angle, and the other group is the technique that is used by the football player. This could be a potential point of view for the agents (coaches, players, etc.) that will interpret the results of this study.

Next, in Section 5.3 we formally compare all the methods and make our final selection.

5.3 Model Comparison and Selection

Given the Tables 11 and 14, we observe the accuracy and precision measures are the ones that stand out and can provide insights for the final selection of the model. As a result, we observe that the random forest classifier is the one with the highest predictive performance and it will be selected as the final model to be implemented. It is worth mentioning, that a final test on the performance of the selected model needs to be made in an unbiased manner.

For this purpose, we make use of the validation set which was described in Section 4. This additional validation set has not been used in the training and the tuning of the model and it allows us to get an independent estimate of the performance of the selected algorithm. We train the model on the 80% of the data which pertains to the original 60%, aggregated with the initial 20% of the test data. Therefore, to measure the predictive performance of the selected model in an unbiased manner, we test it on the last 20% of the data, the validation set, as described in Section 4.3.3.

In our case, the random forest classifier, with the optimal hyper-parameters set resulting from the tuning in Section 5, achieved the following performance measures displayed in Table 15, when it fits the validation set.

Evaluation Measure	**Random Forest**
Accuracy	87.67
Misclassification Rate	12.33
Sensitivity	55.21
Specificity	94.28
Precision	66.26

Table 15: Random Forest Evaluation Measures on the Validation Set.

Therefore, we observe that the selected method performs well since its evaluation measures

are close to the ones observed in Table 14, and we conclude with our selection of a random forest classifier.

6 Conclusion

In this section, we summarize the findings of the performed study, which will provide the answers to the initial research question. Furthermore, we discuss the limitations of this thesis and the topics that could be potential subjects of future research and improvement.

6.1 Summary

This thesis focuses on estimating the risk of a pass and the behaviour of football players based on their risk rates. Additionally, it identifies the risk components of an unsuccessful pass. Consequently, understanding the decision-making process of football players by evaluating their risk profile and the components of a risky pass is an essential part of this study's outcome. To achieve that, we identified the most impactful risk factors of a pass, as described in Section 5.
We performed an exploratory analysis to get a better understanding of the risk behaviour of the teams and players in the studied competition. Based on the findings in Section 5, we observe that we can categorize the players and the teams based on their risk percentages. For the teams' analysis, we aggregate the passes per player per team to estimate their overall risk rate, however, we do not apply the predictive models on a team level, since this research focuses on individual's passing.
In connection with the above reasoning, we studied the player's risk rate on passing and we established a connection with the number of passes they make per game. Our findings suggest that players with a higher number of passes achieve higher risk percentages during each game, which negatively affects their performance. Intuitively this outcome makes sense and, theoretically, football players can estimate the risk factors of a pass and improve their knowledge of which passes will yield a higher risk by using the provided predictive tool. Consequently, they will reduce the amount of these passes per game and eventually, they will achieve lower risk rates. Since an unsuccessful pass leads to a loss of the team's possession, this translates to turnover for the player who committed the pass, thus, it has a direct negative impact on their performance. As described in Section 5, we provide the set of the most effective variables for an unsuccessful pass, i.e. the risk components. Therefore, football players will use these results to understand what are the risk factors of a pass and how they can adjust their game to be in a position to improve their decision-making process.
Additionally, as discussed in Section 1, this thesis can assist football coaches in preparing their game approach and improving their team's performance. Based on the explanatory analysis in Section 5.1, the coaches are in a position to evaluate the risk profile of a team or a player.

Intuitively, most professional coaches have adequate knowledge of their players' capabilities and restrictions. However, this research provides a measurable approach to the riskiness of each player based on their passing behaviour.

For this purpose, we introduced a predictive model which can accurately estimate the likelihood of a pass being unsuccessful, i.e. the risk of the pass. This is a handy tool for coaches since they can estimate the risk of a pass by providing data to the predictive model. The provided data will consist of pass characteristics, which will be used as the explanatory variables for the predictive model. Therefore, coaches can generate a sample set of passes by creating a matrix like the one that appears in Table 4, and they can simulate game situations that require specific characteristics for passes. For instance, according to the results in Section 5, formations which require more passes performed by the head of the football player or without controlling the ball (No Touch), will achieve higher risk rates than formations with more secure ground passes passes. As a result, coaches will estimate the risk of these passes and they will make adjustments until they reach the desired level of risk percentages. In other words, before trying a new formation or a playing pattern, which requires a specific set of pass characteristics, coaches are capable of estimating the risk of this approach beforehand. This is a major advantage compared to the traditional coaching decision-making that is currently in place, where coaches change their approach when they observe a team's performance after many games or hours of training.

Initially, to reach the aforementioned results, we studied classical approaches for classification problems, such as the logit and probit models as described in Section 4. Overall, these approaches offer a comprehensive interpretation of the fit of the data and predictive accuracy of approximately 85%. One of their advantages is that through the estimated coefficients of the parameters, we obtain an elaborate understanding of the importance of each risk component. In other words, we can fulfil the research's purpose by identifying the most important risk factors and providing an informative estimate of their effect on the likelihood of a pass being unsuccessful.

However, since we aim for a higher predictive accuracy and a model with fewer assumptions we studied the ensemble machine learning approaches for classification. Namely, we examine the performance of tree-based algorithms, focusing on the random forest and a special case of random forest, the bagging. The latter algorithm combines weak learners and applies the majority rule to yield an accurate prediction as thoroughly described in Section 4. Consequently, random forest, which can be considered as an extension of bagging, de-correlates the decision trees by constructing a multitude of decision trees at training time. Therefore, we expect a better performance in terms of predictive accuracy compared to the classical approaches. Indeed,

according to Section 5, the machine learning method yields a higher accuracy rate and better fit to the data reaching an accuracy of 88%, thus this is the preferred method. To validate this result, in Section 5, we evaluate the accuracy of this algorithm in the validation set, achieving a predictive performance of 87.67%, which is sufficiently close to the initial measure.
Consequently, we are in a position to estimate the risk of a pass and to identify and estimate its risk factors. However, there are some limitations on these results related to the nature of the available data we have, which will be explained in Section 6.2. Additionally, we discuss potential improvements in this study including additional models to be taken into consideration and possible ways of expanding the available set of explanatory variables.

6.2 Limitations and Future Research

First of all, the limitations of this research lie in the nature of the data and the variety of potential parameters we could consider. In other words, we estimate the likelihood of a pass being unsuccessful based on its characteristics as described in Section 3. However, it would be quite beneficial to consider other factors that are player-specific, such as the age of the player, their years of experience in professional football, their strong foot combined with their execution foot and a fatigue measure for each player, i.e. the total number of minutes played in the game.
Besides the pass variables, another limitation that comes along with the employed dataset is the small number of games in the competition, thus we can not compute an accurate win ratio for every team. Given the fact that several teams play only three games with the larger amount of games for a team being seven, it is not informative to calculate the win ratio for each team. It would add value to our results if the win ratio could be related to the risk rate of every team. Hereby, if we could increase the number of games per team we would relate the predictive model with the win ratio of every team giving more insights to the coaches. Furthermore, while the importance of actions and the movement of all players in the field is not underestimated, for the sake of this research off-ball movement is not taken into account. Intuitively, the decisions made by the players with the ball will have a larger impact on the game's outcome, rather than those who are moving without the ball. However, an additional aspect of the employed dataset could be to consider the off-ball movement of the football player before receiving the pass. As a result, the cooperation and communication of the two players would be also evaluated from the model.
Data attributes like the ones that are mentioned above will allow us to account for the unobserved factors that affect the performance and risk behaviour of a football player. According to Mishra (2014), each individual makes decisions based on the maximization of their expected

utility function. Expected utility describes the relationship between the expected value of an action and the perceived utility. At the elite level, coaches and athletes appear to consistently make good decisions in situations that are highly temporally constrained. In our case, we do not capture the perceived utility for each player and therefore we do not reach the maximum accuracy of our predictive models. As Mishra (2014) mentioned, the ability to assess all options a player faces when in possession of the ball can provide insights into the cognitive processes involved in decision-making and inform strategies for improving decision-making skills and players' performance, as well as developing game tactics.

Taking into consideration the above, it would be interesting and insightful for future researchers to consider predictive models that take into account latent factors. i.e. the perceived utility, of each football player. The study that has been conducted by Joseph, Fenton and Neil (2006), where the researcher compares machine learning techniques to a Bayesian network approach that predicts football game results, can be a starting point for incorporating the Bayesian approach. Moreover, relevant research of state-of-the-art Bayesian models are described by Ryan, Drovandi, McGree and Pettitt (2016). When we account for latent factors in the employed model, we increase the accuracy of our algorithm and potentially can capture the unobserved characteristics of the risk profile of a football player. Therefore, we believe it would be greatly beneficial as a future step of this thesis to include the aforementioned approaches in the estimated models.

As a final conclusion, we can state that the risk factors of a football player's pass are close to the intuitive expectations of the researcher. In other words, the type of the pass (High, low, ground pass) and the type of the execution of the pass, such as the execution foot or controlling the ball before passing, are among the most highly effective factors of an unsuccessful pass, as stated in Section 5. We can conclude that the machine learning method we employed, achieved a higher predictive performance compared to the traditional classification approaches, a result which aligns with the initial expectations of the researcher described in Section 4. Finally, the selected model can be used as an informative tool by the players and coaches to elevate their performance and achieve lower risk rates, and eventually, that model can be combined with the improvements we suggest in the final section of this research.

References

Akosa, J. (2017). Predictive accuracy: A misleading performance measure for highly imbalanced data. *Proceedings of the SAS Global Forum*, *12*, 1–4.

Archer, K. J. & Kimes, R. V. (2008). Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis*, *52*(4), 2249–2260.

Breiman, L. (1996a). Bagging predictors. *Machine Learning*, *24*, 123–140.

Breiman, L. (1996b). Out-of-bag estimation.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Burriel, B. & Buldú, J. M. (2021). The quest for the right pass: Quantifying player's decision making. In *StatsBomb Innovation in Football Conference, London, United Kingdom.*

Fernández, J., Bornn, L. & Cervone, D. (2019). Decomposing the immeasurable sport: A deep learning expected possession value framework for soccer. In *13th MIT Sloan Sports Analytics Conference.*

FIFA. (2021). *The football landscape.* Retrieved from `https://publications.fifa.com/en/vision-report-2021/the-football-landscape/`

Gholamy, A., Kreinovich, V. & Kosheleva, O. (2018). Why 70/30 or 80/20 relation between training and testing sets: A pedagogical explanation.

Govindarajan, M. (2014). Decision Making Methods. In *Encyclopedia of Business Analytics and Optimization* (p. 690-695). IGI Global.

Hahn, E. D. & Soyer, R. (2005). Probit and logit models: Differences in the multivariate realm. *The Journal of the Royal Statistical Society, Series B*, *67*, 1–12.

Harris, S. & Harris, D. (2015). *Digital design and computer architecture.* Morgan Kaufmann.

Hastie, T., Tibshirani, R., Friedman, J. H. & Friedman, J. H. (2009). *The elements of*

statistical learning: data mining, inference, and prediction (Vol. 2). Springer.

Horowitz, J. L. & Savin, N. (2001). Binary Response Models: Logits, Probits and Semiparametrics. *Journal of Economic Perspectives*, *15*(4), 43–56.

Hvattum, L. M. & Arntzen, H. (2010). Using ELO ratings for match result prediction in association football. *International Journal of Forecasting*, *26*(3), 460–470.

James, G., Witten, D., Hastie, T., Tibshirani, R. et al. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

Jones, P., James, N. & Mellalieu, S. D. (2004). Possession as a performance indicator in soccer. *International Journal of Performance Analysis in Sport*, *4*(1), 98–102.

Joseph, A., Fenton, N. E. & Neil, M. (2006). Predicting football results using Bayesian Nets and other Machine Learning techniques. *Knowledge-Based Systems*, *19*(7), 544–553.

Kovalchik, S. A. (2016). Searching for the GOAT of tennis win prediction. *Journal of Quantitative Analysis in Sports*, *12*(3), 127–138.

Kvam, P. & Sokol, J. S. (2006). A logistic regression/Markov chain model for NCAA basketball. *Naval Research Logistics (NRL)*, *53*(8), 788–803.

Lewis, R. J. (2000). An introduction to classification and regression tree (CART) analysis. In *In Annual meeting of the society for academic emergency medicine in San Francisco, California* (Vol. 14). San Francisco, CA, USA: Department of Emergency Medicine Harbor-UCLA Medical Center Torrance.

Manel, S., Dias, J.-M. & Ormerod, S. J. (1999). Comparing discriminant analysis, neural networks and logistic regression for predicting species distributions: a case study with a Himalayan river bird. *Ecological Modelling*, *120*(2-3), 337–347.

McHale, I. & Morton, A. (2011). A Bradley-Terry type model for forecasting tennis match results. *International Journal of Forecasting*, *27*(2), 619–630.

Mead, J., O'Hare, A. & McMenemy, P. (2023). Expected goals in football: Improving model performance and demonstrating value. *Plos One*, *18*(4), 1-29.

Miljković, D., Gajić, L., Kovačević, A. & Konjović, Z. (2010). The use of data mining for basketball matches outcomes prediction. In *IEEE 8th International Symposium on Intelligent Systems and Informatics* (pp. 309–312). IEEE.

Mishra, S. (2014). Decision-making under risk: Integrating perspectives from biology, economics, and psychology. *Personality and Social Psychology Review*, *18*(3), 280–307.

Owoeye, O. B., Palacios-Derflingher, L. M. & Emery, C. A. (2018). Prevention of ankle sprain injuries in youth soccer and basketball: effectiveness of a neuromuscular training program and examining risk factors. *Clinical Journal of Sport Medicine*, *28*(4), 325–331.

Pasanen, K., Rossi, M. T., Parkkari, J., Heinonen, A., Steffen, K., Myklebust et al. (2015). Predictors of lower extremity injuries in team sports (PROFITS-study): a study protocol. *BMJ Open Sport & Exercise Medicine*, *1*(1), 76–80.

Pratas, J., Volossovitch, A. & P Ferreira, A. (2012). The effect of situational variables on teams' performance in offensive sequences ending in a shot on goal. A case study. *The Open Sports Sciences Journal*, *5*(1).

Probst, P. & Boulesteix, A.-L. (2018). To tune or not to tune the number of trees in random forest. *Journal of Machine Learning Research*, *18*(181), 1–18.

Probst, P., Wright, M. N. & Boulesteix, A.-L. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *9*(3), e1301.

Read, P. J., Oliver, J. L., De Ste Croix, M., Myer, G. D. & Lloyd, R. S. (2018). A prospective investigation to evaluate risk factors for lower extremity injury risk in male youth soccer players. *Scandinavian Journal of Medicine & Science in Sports*, *28*(3), 1244–1251.

Ryan, E. G., Drovandi, C. C., McGree, J. M. & Pettitt, A. N. (2016). A review of modern computational algorithms for Bayesian optimal design. *International Statistical Review*, *84*(1), 128–154.

Sarlis, V. & Tjortjis, C. (2020). Sports analytics—Evaluation of basketball players and team performance. *Information Systems*, *93*, 101562.

Schelling, X. & Robertson, S. (2020). A development framework for decision support systems in high-performance sport. *International Journal of Computer Science in Sport*, *19*(1), 1–23.

Sokolova, M., Japkowicz, N. & Szpakowicz, S. (2006). Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. In *Australasian joint conference on artificial intelligence* (pp. 1015–1021). Springer.

StatsBomb. (2023). *statsbombpy.* GitHub. Retrieved from `https://https://github.com/statsbomb/statsbombpy/tree/master`