Universiteit
Leiden
The Netherlands

Opleiding Informatica

Analyzing Offensive Player- and Team Performance in Soccer Using Position Data

Lars Tijssen

Supervisors:
L.A. Meerhoff & A.J. Knobbe

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

05/12/2018

Abstract

As Cruyff said: "You should always ensure that you score one goal more than the opponent.". The number of goals, however, does not always reflect the relationships between the teams on the field. In addition, scoring a goal is a rare occurrence. Therefore, coaches are looking for other techniques to measure the offensive player- and team-performance, and so does the KNVB. Link's dangerousity is a good method for measuring this, according to the KNVB's analyst expert. Therefore, this study has two purposes: (1) to investigate the offensive player- and team-performance of the Dutch national soccer team and (2) to validate Link's dangerousity method. This has been done by transforming the data according to Link's dangerousity method using a step-by-step Python pipeline. The position data is linked to the event data (e.g., shots, goals) to determine the outcome of an attack. The aggregated features per attack are used for subgroup discovery to find interesting patterns in the data in relation to the outcome of an attack.

On the basis of the results of this research, it can be concluded that Link's dangerousity is a valid method for measuring player- and team-performance. However, we have investigated that the maximum dangerousity score without the Density feature performs even better.

Contents

Chapter 1

Introduction

Over the last years there has been an exponential growth in the availability of data in general. As a result of technological advancements, data is becoming more prevalent in sports too, resulting in the emergence of Sports Analytics [1]. In soccer, data analysis is also becoming more and more accepted and so its use at the Royal Dutch Football Association (KNVB) is also increasing. The current research is done in collaboration with the Performance & Innovation department of the KNVB.

The KNVB mainly uses video analysis to analyze the matches of the Dutch national soccer team. Before the match, the video analyst analyzes all recent games of the Dutch national soccer team and its opponent. During the match, the video analyst provides a half time report with the most important moments of the first half. The KNVB aims to support the video analysis with data to increase the objectivity.

The KNVB already uses data to support the performance evaluation of the Dutch national soccer team. Most of it is physical data of the players during the match and training, like the players' heart rate. Systematically analyzing tactical patterns of play is not yet common practice. As a first step, data scientists attempt to reconstruct the coach's view with data. As data analysis has not yet been widely accepted in soccer, this may also make it easier for the coaching staff to implement these innovative analyses. In a later stage, the use of data may even provide unexpected insights that the coaching staff had not considered.

At the moment, the KNVB wants to have more insight in the offensive performance of individual players and the team during attacks in matches. Traditional statistics like possession of the ball, shots on goal, tackle and pass rates are not sufficient for this, because they do not cover the entire load of the team's offensive performance. Previously, Link and colleagues have proposed a measure that captures the level of danger players exert during an attack, aptly named the "dangerousity". Link and colleagues argue that their dangerousity method [2] can be seen as a better method to measure offensive player- and team performance. It is more reliable than the traditional performance indicators, but it is more difficult to interpret. For our purpose, it is useful because dangerousity can evaluate players on individual level, but is also describing the offensive team performance. Therefore, this method is implemented during this research.

Link's dangerousity is a measure of how dangerous (i.e., likely to score) a player is for every point in time at which that player is in possession of the ball. This method describes dangerousity on the hand of four components: Zone, Control, Pressure and Density. *Zone* is the position of the player with ball in the final third (i.e., the last 34 meters of the field). *Control* is the ball control of the player with ball, calculated using the relative speed of player and ball. *Pressure* is exerted by the defender who is in a certain range of the player with ball. *Density* is divided into two components: Shot Density and Pass Density. *Shot Density* is the chance of blocking a shot for a defender. *Pass Density* is the chance of intercepting a pass or cross.

In this research, the reliability of Link's dangerousity as an indicator for offensive player- and team-performance is examined, because the parameters of the algorithms are not publicly available and the method has only been tested in the highest German soccer league (Bundesliga), so it may not be generalizable.

Therefore, the main research question is: Can Link's dangerousity be used to meaningfully analyze offensive player- and team-performance? So the goal of this research is to quantify the contribution of players and the team during attacks, and to validate Link's dangerousity method.

Chapter 2

Related Work

In recent years there have been several studies where position data was used for. Not only in soccer, but also in other sports, such as basketball, hockey and rugby [3]. In the early years mainly physical parameters (e.g., speed, distance covered) were analyzed, but in recent years more and more tactical analysis is done on the basis of position data.

Position data describes where the players (and the ball) are located on the field at a specific point in time. This data can be obtained through video analysis or tracking sensors. We work with position data that is obtained through video analysis, because it has the advantage that it can track all the players and the ball (unlike sensor technologies, which cannot track the ball). In addition, it is relatively inexpensive and portable, because only a couple of cameras are needed, so it is easy to place the cameras in any stadium before the match. The data from each stadium is not always equally reliable, due to backlight from the setting sun. Large stadiums do not suffer from this. Tracking sensors are generally more reliable, but they are more expensive, require both teams to wear a sensor and cannot track the ball.

The acquisition of the position data is provided by STATS [4], which is a commercial sports data and technology company. They analyze more than 100,000 games a year in different sports. The KNVB uses the SportVU soccer playing tracking algorithm from STATS. This soccer tracking algorithm has been developed using some of the best capture, aggregation and analytical technologies [5]. SportVU uses computer vision to extract the position data of all the players and the ball from tactically placed video cameras around the pitch. Figure 2.1 shows an setup of the cameras.

In this research, we use a combination of position data and event data. The acquisition of event data is done manually during the match. This is a commonly used method and is done by highly trained analysts. An advantage of this method is that the event data captures information which cannot directly be derived from the position data, like the type of event (e.g., pass, shot) or even the foot the player passes with. This makes the combination of position data and event data valuable. A disadvantage is that the acquisition is done by humans. The difference between a pass and a shot is sometimes hard to see, which means that there could be different interpretations between analysts, despite they are highly trained. Therefore, the event data is more subjective than the position data. The event data is provided by Opta [6], which is an international sports analytics company.

There are several studies that use a combination of position data and event data, like Memmert and colleagues [1]. Their study examines the differences in variability in inter-team distance (distance between two teams' centroids) during the 30 seconds prior to a critical event (goal attempts and goals). The team centroid represents the mean position of all outfield players. It is also possible to determine centroids per line (e.g., defenders, midfielders and forwards), which is a more accurate method to capture the players' movement behavior. In the current research, we also look at the critical period before a goal attempt or goal, where the duration of the period is variable, depending on the length of the attack.

In the critical period we look at Link's dangerousity score [2], because dangerousity identifies four important offensive components according to the domain expert (the KNVB's analyst expert) involved in this project. These four components are (1) the zone in which the player with ball is located, (2) the ball control that the player has, (3) the pressure exerted by the defense and (4) the density of the defense. A full explanation of these features can be found in Appendix A. It is important to check the validity of Link's dangerousity, because

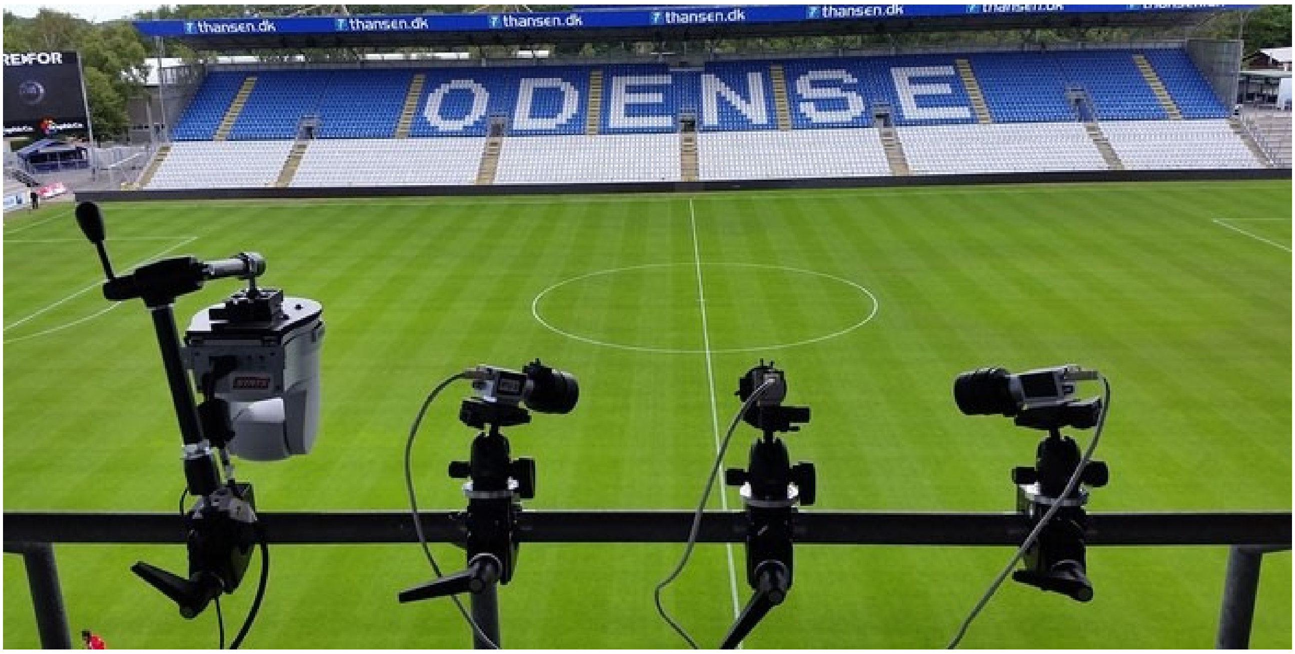


Figure 2.1: setup of the cameras to obtain position data through video analysis (adapted from https://www.stats.com/sportvu-football/).

the constants of the algorithms are not publicly available and Link only tested dangerousity in an specific setting, namely the Bundesliga. Therefore, it may not be generalizable to other competitions.

To be able to test the validity of Link's dangerousity, we use a similar method to that of Stein and colleagues [3]. They analyze the current approach of data in team sports in general. Their approach consists of data modeling, data mining, information visualization and visual analytics. *Data modeling* is about giving structure to the problem of sport analysis. Stein and colleagues describe two main approaches: (1) domain-specific modeling and (2) data-driven or explorative modeling. *Domain-specific modeling* is based on theories from Sport Science, which look at the relationship between actions and outcomes. *Data-driven modeling* typically not assumes previous knowledge about the domain, but obtains its insights directly from the data. The two approaches often go hand-in-hand, as in the current research. We have the expectation that Link's dangerousity says something relevant with respect to the outcome of an attack, but we also use data to arrive at new insights.

Data mining includes a selection, preprocessing, transformation and interpretation phase [8]. These steps are necessary before the actual data mining algorithm can be applied. *Data selection* is done to collect the correct data to be able to solve the problem described in the data modeling phase. Section 3.1 describes which data we use in this research. *Data preprocessing* includes removing noise out of the data and strategies for handling missing data fields for example, as described in sections 3.3 and 3.4.1. *Transformation* of the data is done to find useful features that represent the data depending on the goal of the task. The goal of this research is to quantify the contribution of players and the team during attacks, and to validate Link's dangerousity method. So in the transformation phase we try to replicate the dangerousity method as well as possible, which can be found in sections 3.4.2, 3.4.3 and 3.4.4.

The implementation of the steps mentioned above is done by using an existing Python pipeline [7], which allows for a direct comparison of "fingerprints" of tactical behavior (see section 3.4). A fingerprint then refers to a specific combination of features derived from the position data that describes the characteristics of a certain playing style. For the current work, this pipeline has been adjusted to create a system that analyzes dangerousity per player per match. At the end of the transformation phase, features are aggregated per attack, such as the maximum dangerousity score during an attack. These aggregated features are the output of the pipeline.

After the aggregation of the features, the data is ready for data mining. There are several data mining techniques that can be applied to the outcome of the pipeline. By applying a data mining algorithm, we will discover whether there is a relation between the dangerousity score and the outcome of an attack. Fayyad and colleagues group data mining techniques in six general categories [8]: classification, regression, clustering, summarization, dependency modeling and change and deviation detection. *Classification* is a technique whereby items are divided into predefined classes on the basis of a rule. When *Regression* is applied then the data items are mapped to real-value prediction variables. With *Clustering*, the dataset is divided into a finite number of categories. *Summarization* includes methods for finding compact descriptions for a subset of the data. *Dependency Modeling* is a technique whereby significant dependencies between variables are found. *Change and deviation detection* includes the detection of significant changes in the data from previously measured or normative values.

Data mining techniques can be applied from two different perspectives: predictive induction and descriptive induction [9]. The goal of predictive induction is to predict the outcome variable. Classification and regression can be seen as predictive induction. The goal of descriptive induction is to find interesting patterns in the data. Clustering, summarization, dependency modelling and change and deviation detection can be seen as descriptive induction.

A technique that is not mentioned by Fayyad and colleagues [8] is subgroup discovery. Subgroup discovery is a combination of predictive and descriptive induction and its goal is to describe relations between independent variables (subgroups) and the target. We apply subgroup discovery in this research, because we want to discover interesting patterns in the data with respect to the outcome of an attack. The other data mining techniques are less sufficient for this. In addition, with subgroup discovery many features can be examined without having to select in advance. This gives us the possibility to determine which features contribute the most to a successful outcome of an attack.

After the data mining phase, information visualization is applied to get a visual representation of the data. *Information visualization* has three main tasks: exploration, hypotheses validation, and hypotheses generation [3]. As an explorative information visualization technique we use some temporal visualizations to show how the dangerousity scores change over time (see sections 3.4.5, 4.3.1 and 4.3.2). For hypotheses validation we use boxplots and ROC curves, which can be found in chapter 4. Hypotheses generation as an information visualization technique is not used in this research.

Visual Analytics is the combination of data mining and information visualization. This allows experts to use their domain knowledge in the analysis process by applying interactive and controllable data mining methods with immediate visual feedback of the results. This can be applied to the results of the information visualization by the domain experts of the KNVB, in a Business Intelligence tool for example. This is beyond the scope of this study and is therefore not implemented.

In short, we follow the steps described by Stein and colleagues [3] in a pipeline [7] in which Link's dangerousity method [2] is applied. We examine Link's dangerousity method, because it is not complete and it has not been tested in a setting outside the Bundesliga. Additionally, we are validating the method in an event-based approach, which is also not done before.

Chapter 3

Methods

This chapter describes the method of approach for analyzing offensive player- and team-performance based on the dangerousity score. First, we give an overview of the data to understand how we can obtain interesting insights from this data. Secondly, to study offensive sequences we need to define a formal definition of an attack. Thirdly, the software program Inmotio does the first preprocessing step to check the X and Y coordinates, and to calculate the first features. Fourth, we give a description of the Python pipeline with his preprocessing, spatial aggregation, event selection, temporal aggregation and visualization phase. The dangerousity feature and its underlying components are computed per attack in the pipeline, so that the data is ready for doing experiments. Finally, we give an introduction to the experiments that are done in chapter 4, including subgroup discovery.

3.1 Data Overview

Two types of data are used in this research: position data and event data. Position data is obtained through video analysis from the company STATS. The data consists of a timestamp, X coordinate, Y coordinate, player identification, shirt number, team name, speed and a Boolean for ball possession. A more detailed description can be found in Table 3.1. The data is available in CSV files per half match. A sample of the data can be found in Table 3.2.

Data	Description
Timestamp	A timestamp with a measurement frequency of 10 Hz. (0.1 seconds).
X coordinate	Representing the length of the field, so from -52.5 to 52.5 meters.
Y coordinate	Representing the width of the field, so from -52.5 to 52.5 meters.
PlayerID	A unique identifier per player.
Shirt	Shirt number of the player during the match. For the opponents of the Dutch national soccer team this number is used as a unique identifier if the players of the opponent do not have a PlayerID.
Team Name	Netherlands for example.
Speed	Speed in meters per second.
Ball possession	A Boolean variable to determine which player is in possession of the ball. A player has possession of the ball if his distance to the ball is less than 1.5 meters, and one of the following conditions is true: 1. Ball is moving a certain distance during possession. 2. Ball is changing direction. The incoming direction differs 10 with the outgoing direction. 3. Ball is gaining speed. The acceleration of the ball is at sending time greater than 5 meters per second. 4. Ball has stopped moving. Speed is less than 0.5 meters per second. Filtering the data with a weighted Gaussian algorithm with a sensitivity of 85%.

Table 3.1: Detailed description of the position data available in CSV files.

Timestamp	X	Y	Speed	Dist to closest home	Dist to closest visitor	Shirt	PlrID	In Ball-pos	Name
12400	-22.825	9.225	3.53	11.316	8.092	17	0	0	Latvia
12400	-18.215	-2.296	4.93	12.408	4.337	14	0	0	Latvia
12400	-2.665	6.33	3.13	8.845	7.997	16	0	0	Latvia
12400	-37.848	0.125	4.79	17.565	4.194	10	0	0	Latvia
12400	-14.673	-19.048	4.59	17.123	8.336	15	0	0	Latvia
12400	-41.132	2.733	1.38	4.194	18.327	0	1233	1	Netherlands
12400	-10.425	27.872	2.26	13.577	15.961	0	1214	0	Netherlands
12400	-36.874	20.559	5.8	15.049	18.327	20	20	0	Netherlands
12400	-31.628	-14.717	2.01	16.093	19.871	28	28	0	Netherlands
12400	-15.402	-27.352	2.44	8.336	20.565	0	1209	0	Netherlands

Table 3.2: Sample of the position data from a CSV file.

Event data describes all the different actions during a match. The acquisition of this data is done manually during the match by highly trained Opta analysts [6]. Event data is split into ball events and match events. The data is available in XML files per half match and can be related to the position data by using the timestamp. A detailed description of the event data used in this research can be found in Table 3.3. The event data is used to determine the outcome of an attack, which is needed to do the experiments in Chapter 4. If there is no shot on target, shot off target or goal during an attack, then the attack leads to no shot.

Ball or Match event	Event	Description
Ball	Shot on target	Any attempt at shooting that would reach the goal if it was not blocked by the goalkeeper or if the ball touches the post or cross bar.
Ball	Shot off target	Any attempt at shooting that does not hit the goal, post or crossbar and is not blocked by the goalkeeper.
Match	Goal	Awarded when the whole of the ball crosses the whole of the goal-line.

Table 3.3: Detailed description of the event data available in XML files.

3.2 Attack

To determine the dangerousity of a player during an attack, a definition for attack has to be given. In close collaboration with the domain expert (the KNVB's analyst expert) involved in this project, the start and end times of an attack are defined. An attack starts when all of the following statements are true:

1. The player is in possession of the ball for 0.5 seconds in the final third.
2. The ball is on the field.
3. The ball is moving.
4. The previous attack has ended.

An attack ends if one of the following statements is true:

1. The ball passes the center line.
2. The ball goes out for a goal-kick or a throw-in.
3. The attacking team does not have the ball for more than five seconds.
4. The defending team shoots the ball out of the final third.
5. The ball is not moving for more than five seconds (the referee may have stopped the game temporarily).
6. It is half or full time.

This definition does not cover all possession sequences. It is thus possible that a goal is scored without it being considered as an attack according to this definition (for example penalties).

3.3 Preprocessing in Inmotio

The STATS position data is loaded into Inmotio software. This program does the first preprocessing of the data. It calculates the following features: speed, acceleration, ball possession and distance to closest home/opponent. Ball possession is determined on the basis of the definition in Table 3.1.

The software of Inmotio always adjusts the field dimensions to 105 by 68 meters. So the X coordinates have values from -52.5 to 52.5 meters, and the Y coordinates have values of -34 to 34 meters. For example, if the actual size of the field is 110 x 75 meters (maximum size according to the FIFA [10]) and the player has coordinates (40,20) the new coordinates are: X = 40 / 55 * 52.5 = 38.18 and Y = 20 / 37.5 * 34 = 18.13. The adaption of the field dimensions has the advantage that for each match the side- and goal line have the same X and Y coordinates. A disadvantage is that the actual distances between players can only be calculated if the actual field dimensions are known.

3.4 Pipeline

The preprocessing, transformation and visualization of the data is done by using an existing python pipeline of the Leiden University [7]. The transformation phase exists of a spatial aggregation, event selection and temporal aggregation part. The pipeline is initially made to do research to the tactical differences between the Netherlands and Brazil. Therefore, it is well suited to use for our tactical analyses. Code has been added to this pipeline to create the features needed for this research. The steps that have been taken to do a good analysis can be found in this section.

3.4.1 Preprocessing

In the preprocessing phase the position data is loaded from the CSV file into a pandas dataframe, so it can easily be used and adjusted in all phases of the pipeline. Several actions are done to clean up the position data. First of all, the column headers are checked to determine if all the necessary data is available. Secondly, the rows are checked to see if they have the right data type. Thirdly, the Player Identifiers of the players and ball are checked and set, so that every player has a unique identifier. As shown in Table 3.2 the players of the opponent do not always have correct Player IDs; some of them are 0. To give them a unique identifier their shirt numbers are multiplied by -1. The shirt numbers per team are unique and the Player IDs of the Netherlands are always positive numbers. The ball always has shirt number 1, Player ID 0 and is not part of a team. Fourth, the referees are thrown out of the dataframe, because they are not interesting for these research purposes. Fifth, rows that have no team value, and rows where the X and Y coordinates are 0 or empty are omitted. This may be missing data or data from players who are no longer on the field (i.e., they are substituted), making this data useless. Finally, extreme X and Y values are omitted. The X coordinates need to be in the range -52.5 to 52.5 and the Y coordinates in the range -34 to 34, because the field size is adjusted to 105 by 68 meters (see section 3.3). Rows with X and Y values far out of this range are omitted.

After the position data is loaded and checked, the event data is loaded from the XML file. Only the columns containing the time, NumAmisco, ball event code and match event code are loaded. The time is needed to link the event data with the position data. NumAmisco is a unique number per player per match to determine which player of which team is occurring in the event. The ball and match event code indicates what kind of event occurs, which is needed to determine the outcome of an attack.

3.4.2 Spatial Aggregation

To rigorously explore the patterns in the data we derive new features from the raw X and Y coordination. This requires the data to be aggregated spatially. For this research, we want to know if the dangerousity score of Link [2] has a relationship to the attack outcome (target). Therefore the features of Link (Zone, Control, Pressure and Density) and its underlying components are implemented as well as possible.

The player in possession of the ball gets a value for *Zone* based on how dangerous he is from his current position on the field. In general, a position closer to the goal scores higher. Ball *Control* is calculated for the player with ball based on the average relative speed of ball and player. High relative speed leads to a low control and low relative speed to a high control. *Pressure* is exerted from the defenders who are in a certain range from the player with ball. There are four different pressure zones based on the distance from the player with ball to the defender and the angle they have to the centre of the goal. Every pressure zone has its own weighting factor. The *Density* from the defense is determined on the basis of two components: Shot Density and Pass Density. *Shot Density* is the chance of blocking a shot, and *Pass Density* is the chance of intercepting a pass or cross.

A combination of the four Link features is called "dangerousity" which is calculated according to the following formula:

$$Dangerousity = Zone\left(1 - \frac{1 - Control + Pressure + Density}{3}\right)$$

Dangerousity is only determined for the player in possession of the ball in the final third (i.e., the last 34 meters of the field), so not for every timestamp and not for every player. Dangerousity always have a value between 0 and 1, where 0 means no danger and 1 means very dangerous.

We have also implemented a combination of the Link features with one or two features set to 1, so that we can examine how the relation is between these features and the outcome of an attack (see chapter 4). For example, the dangerousity score consisting of Zone, Control and Pressure (DA(ZCP)), with the Density feature set to 1. When the score for Zone for the player with ball is high (i.e., the player is close to the goal), he has a good control over the ball (i.e., the relative speed between player and ball is low) and the pressure of the defensive team is low (i.e., there are few or no defenders between the player with ball and the goal), the Zone-Control-Pressure dangerousity score (DA(ZCP)) is high. A full list of the features can be found in Table B.1 and Table B.2 in Appendix B.

3.4.3 Event Selection

In the event selection part the start and end times of an attack are computed according to the definition of an attack, described in section 3.2. After that, the attacks are labelled with the outcome of an attack. This is necessary to determine the contribution of dangerousity to an attack, at which a higher dangerousity score leads to a better attack outcome. An attack can lead to no shot (label 0), a shot off target (label 1), a shot on target (label 2) or a goal being scored (label 3). The attacks are labelled by using the annotated event data, described in section 3.1.

3.4.4 Temporal Aggregation

The spatial aggregates are summarized in the temporal aggregation part of the pipeline to reduce them to a single value, so it can be linked to the outcome of an attack and used for subgroup discovery (see section 3.5.1). The standard deviation, average, minimum and maximum values of the features during an attack are computed. Because of this, the amount of features grows exponentially. Additionally, features that have been computed at the player level are aggregated into one team level feature. We take the average, minimum and maximum of all the aggregated player values to obtain such a team measure (e.g., the average of the average distance to the ball of all players of one team). This leads to a total of 376 features. For this research, the values have been calculated for the full duration of an attack, according to the definition in section 3.2.

3.4.5 Player & Team Reports

In the visualization part of the pipeline player and team reports are generated. For pragmatic reasons, we took the maximum dangerousity score per five seconds. After that, we take the sum of all these scores per fifteen minutes, because the typical analysis of the KNVB occur in that timeframe so the KNVB can analyze the progressions during a match. The sum is taken, because the dangerousity score is not determined for every second, but only for players in possession of the ball in the final third, as described in section 3.2. Suppose we take the average, then a team that has been very dangerous for only a short period can achieve a high score, while a team that has had multiple dangerous attacks and a few less dangerous attacks achieves a lower score. With the graphical representation of this score, the KNVB gains insight into the offensive player- and team-performance. This could be valuable information for the coach.

3.5 Analysis

In the analysis we examine whether dangerousity can be used to meaningfully analyze offensive player- and team-performance. This is the case if there is a relation between the dangerousity score and our target; the outcome of an attack. Our target attribute has four different outcomes: no shot, shot off target, shot on target and goal.

We will first provide a descriptive analysis to show the distribution of the features. A summarization of the dataset is given, it is tested whether the aggregated dangerousity scores are correlated to the outcome of an attack and which one has the strongest correlation. For this, the Spearman‘s rank-order correlation is used. Boxplots and histograms are given for the one with the strongest correlation. We want to show that a high dangerousity score leads to a better outcome of an attack.

Secondly, the relationships between the dangerousity scores of the different attack outcomes are also interesting. Before we test that, the dangerousity scores for the four attack outcomes are tested for normality. If all the four attack outcomes are normally distributed, a One-way ANOVA is done, otherwise a Kruskal-Wallis H test is done. After that, a Mann-Whitney U test is done to test if the distribution of the dangerousity scores in the groups differ from each other.

Finally, we want to know if there are other interesting features that contribute to a good outcome of an attack and perhaps even better describe the offensive player- and team-performance. After the temporal aggregation of the data, the dataset contains 2929 attacks and 376 features per attack. With this amount of data, subgroup discovery is a good method to find interesting deviations in the data.

3.5.1 Subgroup Discovery

Subgroup discovery is an exploratory data mining technique which scans the data without much prior focus and find unusual parts of the data. This is done to find out if there are interesting subgroups in the dataset which meet a specific rule. A subgroup is a part of the dataset that show a significant deviation in the distribution of the target attribute. Subgroup discovery can handle binary, discrete or numeric target attributes, but our target attribute has an ordinal scale (there is a certain order in the attribute, e.g., a goal is better than no shot). So our target is converted into three binary targets, namely: no shot, shot off target and shot on target (including goals). The subgroup discovery is done by using Cortana [11]. Cortana is a Data Mining Tool for discovering local patterns in data. It supports multiple data types, contains multiple quality measures, includes statistical validation of mining results and provides a graphical presentation of results. Before we perform the subgroup discovery, a quality measure must be selected. A quality measure determines when the deviation in the distribution of the target attribute is significantly different in the subgroup than in the rest of the dataset. Cortana contains all usual quality measures. We will use the Weighted Relative Accuracy (WRAcc) measure:

$$WRAcc(S,T) = p(ST) - P(S) \cdot P(T)$$

where S is the subgroup and T the target. As the definition shows, this measure is a balance between coverage and unexpectedness.

As a search strategy, the default beam search method of Cortana is used with the search width set to 100. Beam search is a heuristic search algorithm with a predetermined number of paths, called the search width. Only the best 100 paths are kept as candidates according to the best first search algorithm. The validation of the subgroups is done by using swap-randomization [12]. This technique replaces the target column with a random permutation of itself so that all the relations between the attributes and the target disappear. Then the subgroup algorithm is run on the resulting dataset using WRAcc as a quality measure. This process is repeated 100 times. After that, a threshold is calculated to distinguish statistical significant results and accidental findings.

The subgroup discovery is performed on search depth 1 and 2, which refers to the number of features included in the rules to define the subgroups. So for search depth 1 there is only one feature defining the subgroup. For example, if the dangerousity score is higher than 0.7, the percentage of attacks that lead to a shot on goal is considerably larger in the subgroup than in the whole dataset.

To measure the quality of a set of subgroups, the subgroups are plotted in the Receiver Operating Characteristic (ROC) space expressed in its False Positive Rate (FPR) and True Positive Rate (TPR). The FPR represents the fraction of the negative examples that occur in the subgroup. The TPR represents the positive examples. The accuracy of the test is measured with the Area Under Curve of the ROC. An AUC near 1.0 means a perfect test; almost perfect subgroups are found. An AUC of 0.5 means a worthless test; no significant subgroups are found.

Chapter 4

Experiments

To make sense of the data, we first need to describe how the features relate to the target value. Depending on which of Link's components seems most promising, we can perform a subgroup discovery to discover the patterns that best describe the outcome of an attack. All in all, this analysis will give us an idea of the meaningfulness of the features in our specific population.

4.1 Descriptive Statistics

The dataset contains 2929 attacks from 31 matches of the Dutch national soccer team. The target contains four possible outcomes of an attack: no shot (label 0), a shot off target (label 1), a shot on target (label 2) or a goal being scored (label 3). Eighty-three of these attacks leads to a goal (3%), 183 to a shot on target (6%), 340 to a shot off target (12%) and 2323 to nothing (79%).

There were 97 goals scored in the 31 matches, but only 83 goals were labelled correct (86%). Eight goals are not labelled due to incomplete ball possession data. Six goals are not labelled because of very short ball possession (shorter than 0.5 seconds) at the beginning of an attack, two of them were penalties.

For this research, we want to know if the dangerousity score of Link [2] has a relationship to the attack outcome (target). A test for correlation is done to check if there is a relationship between the dangerousity score and the attack outcome. To do so, the Spearman's rank-order correlation is used, because the target is an ordinal scale and the data is not normally distributed. The significance level is set at $\alpha = 0.05$. The following hypothesis is tested:

Ho: there is no association between the dangerousity score and the outcome of an attack.

Ha: there is an association between the dangerousity score and the outcome of an attack.

In Table 4.1 the temporal aggregated dangerousity features with its correlation are shown.

Feature	**Correlation**	**p-value**
DA_{max}	0.42329	<0.001
DA_{avg}	0.38877	<0.001
DA_{std}	0.38134	<0.001
DA_{min}	0.17998	<0.001

Table 4.1: Spearman's rank-order correlation of the temporal aggregated dangerousity scores with the outcome of an attack.

Results of the Spearman's rank-order correlation indicated that there was a significant positive association between the maximum dangerousity score and the outcome of an attack, $(rs(2927) = 0.423, p < 0.001)$. The same applies to the other cases, so Ho is rejected and we can conclude that there is an association between the dangerousity score and the attack outcome.

The maximum, average and standard deviation of the dangerousity score are moderately correlated with the attack outcome. The minimum dangerousity score has a weak correlation. From now on, we take the maximum dangerousity score for plotting purposes, because it has the strongest correlation with the attack outcome. The boxplots of the maximum dangerousity scores per attack outcome can be found in Figure 4.1. It can be deduced from this that there is a relationship between the maximum dangerousity score and the outcome of an attack. In general, a higher dangerousity score leads to a better outcome of an attack.

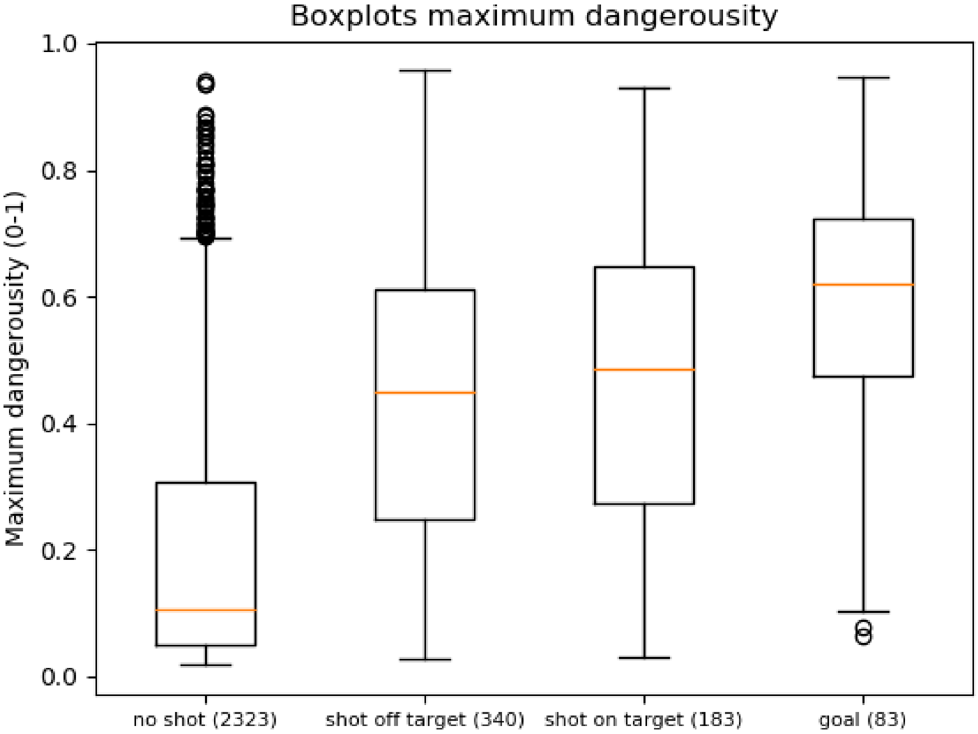


Figure 4.1: Boxplots of the maximum dangerousity scores per attack outcome. In parentheses the number of attacks per outcome.

In Figure 4.2 a histogram of the maximum dangerousity scores can be found. It can be concluded that most attacks lead to a low maximum dangerousity score. Figure 4.3 shows that the distribution of the maximum dangerousity scores is very different depending on the attack outcome. Most attacks that lead to nothing have a maximum dangerousity score of 0.2 or lower, while most attacks that lead to a goal have a maximum dangerousity score of more than 0.7.

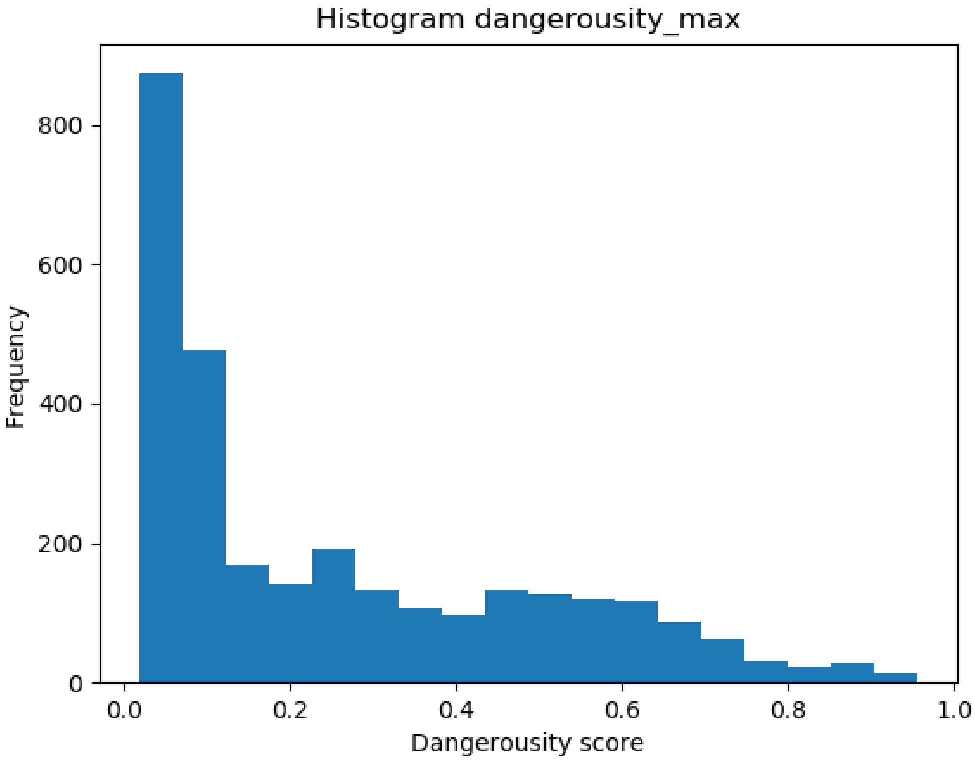


Figure 4.2: Histogram of the maximum dangerousity scores.

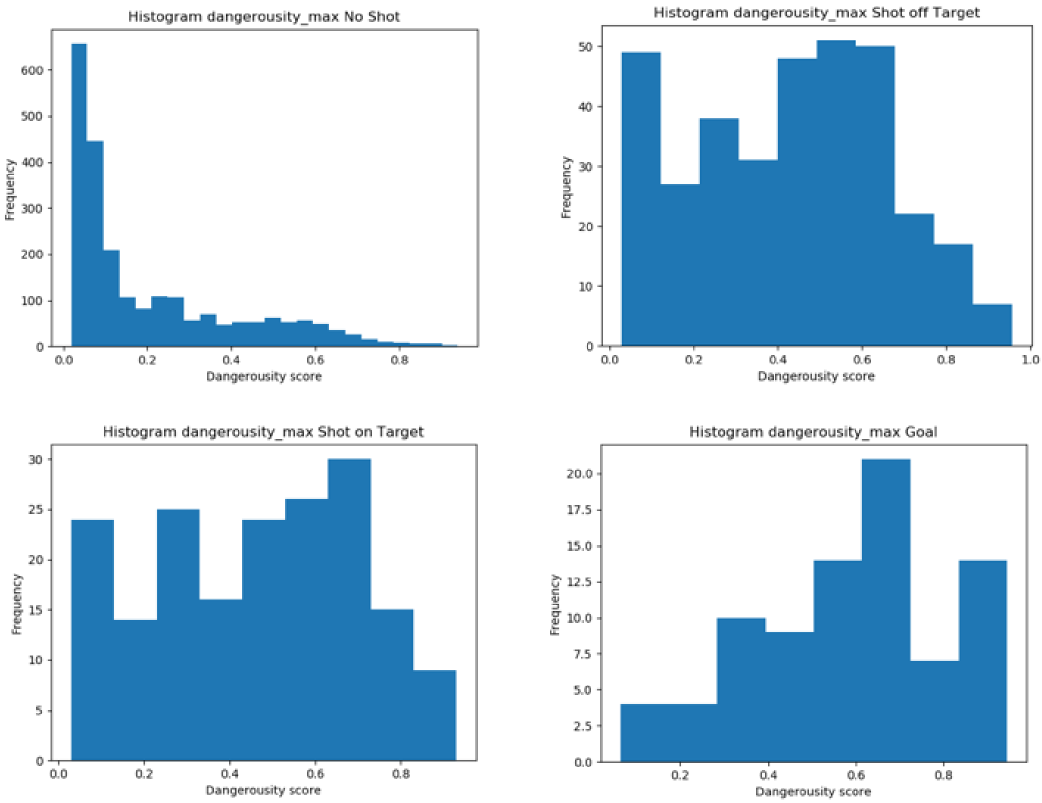


Figure 4.3: Histograms of the maximum dangerousity scores for the four different outcomes of an attack. Top left the dangerousity scores which leads to no shot, top right scores which leads to a shot off target, bottom left the scores which leads to a shot on target and bottom right the scores which leads to a goal.

4.2 Attack Outcomes

It is now interesting to know if there is a difference in the maximum dangerousity scores between the four outcomes of an attack. Before we test that, we need to know if the four groups come from a normal distribution. This is done via a test for normal distribution, which is based on D'Agostino and Pearson's test that combines skewness and kurtosis to produce an omnibus test for normality. The hypothesis below is tested for significance for all four groups separately ($\alpha = 0.05$):

Ho: the sample comes from a normal distribution.

Ha: the sample does not come from a normal distribution.

The results can be found in Table 4.2.

Target	Mean	Standard Deviation	Median	Skewness	Kurtosis	p-value	Normal distribution?
All	0.256	0.235	0.154	0.897	-0.360	< 0.001	No
No shot	0.203	0.203	0.107	1.246	0.539	< 0.001	No
Shot off target	0.428	0.231	0.448	-0.018	-0.918	< 0.001	No
Shot on target	0.460	0.239	0.486	-0.112	-1.035	< 0.001	No
Goal	0.589	0.224	0.620	-0.384	-0.482	0.222	Yes

Table 4.2: Descriptive statistics per attack outcome.

We can say that only "goal" does come from a normal distribution, with a skewness of -0.384 and a kurtosis of -0.482. For "no shot", "shot on target" and "shot off target" the p-value is less than 0.001, so Ho is rejected and we can say that these groups does not come from a normal distribution.

A Kruskal-Wallis H test is used to see if there is significant difference between the groups, because not all groups have a normal distribution. The following hypothesis is tested for significance:

Ho: there is no difference in the maximum dangerousity score between the outcomes of an attack.

Ha: there is a difference in the maximum dangerousity score between the outcomes of an attack.

The Kruskal-Wallis H test showed that there was a statistically significant difference in the dangerousity scores between the outcomes of an attack, ($H(3) = 529, p < 0.001$), with a mean dangerousity score of 0.203 for "no shot", 0.428 for "shot off target", 0.460 for "shot on target" and 0.589 for "goal".

Now we know that there is a significant difference between the groups, we want to know if there is a significant difference between all the groups. So we do a post hoc analysis with the Mann-Whitney U test to find out if the distribution of the maximum dangerousity scores is different between two groups. The hypothesis below is tested for significance:

Ho: the distribution of the maximum dangerousity scores for the two groups is not different.

Ha: the distribution of the maximum dangerousity scores for the two groups is different.

The results of the test can be found in Table 4.3.

Test	Accept / Reject	Conclusion
Ho: the distribution of “no shot“ and “shot off target“ is not different.	Reject Ho	The Mann-Whitney U test indicated that there is no significant difference in the distribution of the maximum dangerousity score between “no shot“ ($Mdn = 0.107$) and “shot off target“ ($Mdn = 0.448$), ($U = 175625, p < 0.001$).
Ho: the distribution of “shot off target“ and “shot on target“ is not different.	Accept Ho	The Mann-Whitney U test indicated that there is no significant difference in the distribution of the maximum dangerousity score between “shot off target“ ($Mdn = 0.448$) and “shot on target“ ($Mdn = 0.486$), ($U = 28613, p = 0.065$).
Ho: the distribution of “shot on target“ and “goal“ is not different.	Reject Ho	The Mann-Whitney U test indicated that there is no significant difference in the distribution of the maximum dangerousity score between “shot on target“ ($Mdn = 0.486$) and “goal“ ($Mdn = 0.620$), ($U = 5354, p < 0.001$).

Table 4.3: Results of the Mann-Whitney U test to test on equal distribution between two groups.

4.3 Player & Team Reports

Now we know that the dangerousity score reflects something relevant about the outcome of an attack, we show some results on the player and team level in section 4.3.1 and 4.3.2. The dangerousity score is only determined for the player with ball in the final third. The maximum score per five seconds has been taken, after that we take the sum per fifteen minutes, according to section 3.4.5.

4.3.1 Dangerousity per Team

We analyze the first half between Team AAA and Team BBB on its offensive performance. Team BBB won the game with 3-0, with this score already on the scoreboard at half time. They scored one goal in the first fifteen minutes and two goals in the last fifteen minutes of the first half. This cannot directly be deduced from the dangerousity score, but when we look at the dangerousity scores of Team BBB more closely, we see that they were much more dangerous than Team AAA in the last part of the first half, resulting in two goals (see Figure 4.4).

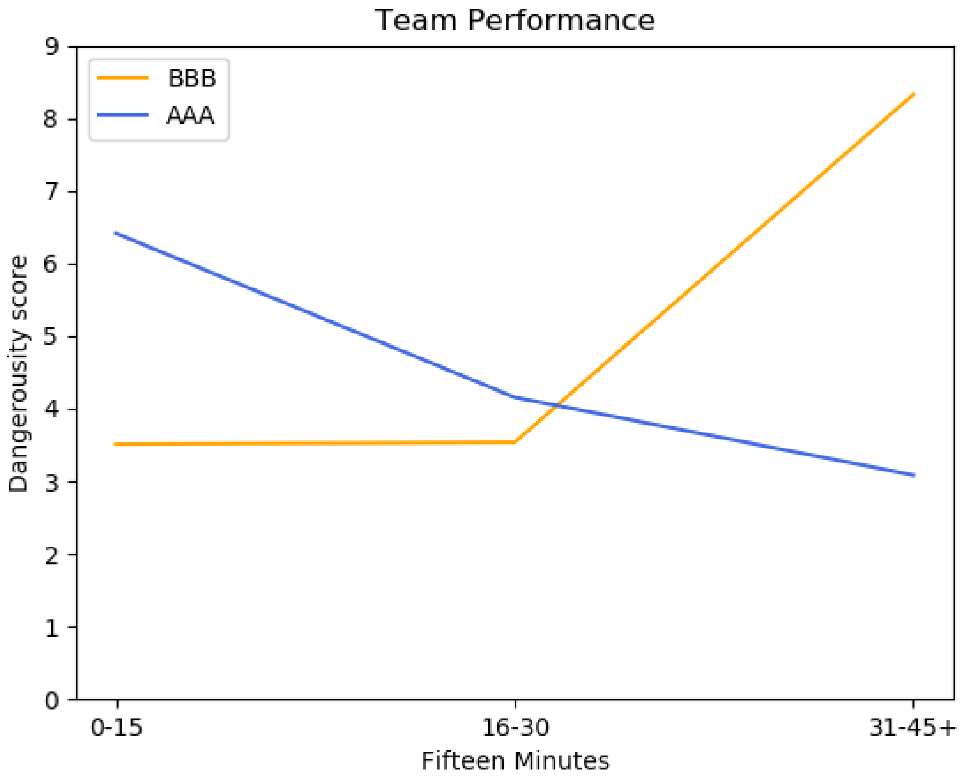


Figure 4.4: Offensive team performance based on the dangerousity score of Team AAA and Team BBB during the first half of the match. The dangerousity scores on a single timestamp vary from 0 to 1. The maximum score per player per five seconds has been taken, after that the sum per team per fifteen minutes has been taken.

In the first fifteen minutes, Team AAA (59.7 seconds) spent much more time with attacks than Team BBB (35.7 seconds), resulting in an almost twice higher dangerousity score. In the second fifteen minutes, Team AAA (45.0 seconds) was more dangerous than Team BBB (105.2 seconds), but they were engaged in attacks for far less time. This is because Team BBB had three attacks that lasted longer than 18 seconds, but were not dangerous. They mostly played the ball around between the center line and the beginning of the final third. In the last fifteen minutes of the first half, Team BBB (58.7 seconds) were much more dangerous than Team AAA (70.9 seconds), but they were attacking for a shorter period than Team AAA.

The statistics of the first half between Team AAA and Team BBB can be found in Table 4.4. If we look at the number of goals, attempts and shots on target, Team BBB seemed much stronger. If we look at ball possession, Team AAA seemed to be the better team. If we look at the total dangerousity score, we see that the teams were in balance with each other. The table also shows that Team BBB had a higher total dangerousity score, but a lower dangerousity score per second. All in all, we can conclude that both teams were almost equally dangerous, but that Team BBB was much more effective than Team AAA by scoring three goals.

Team	**Team AAA**	**Team BBB**
Goals	0	3
Ball possession	60%	40%
Attempts	2	6
Shots on target	2	4
Total dangerousity score	14.61	15.75
Seconds in final third	175.6	199.6
Dangerousity per second	0.083	0.079

Table 4.4: Statistics of the first half between Team AAA and Team BBB.

4.3.2 Dangerousity per Player

We can also look at the dangerousity score at player level. Figure 4.5 shows the performance of all the players of Team BBB during the first half against Team AAA. Player 1266 made the biggest contribution to the attacks of Team BBB during the last fifteen minutes of the first half, according to his dangerousity score. This player has therefore been very dangerous, but did not score a goal and did not give an assist in the last fifteen minutes, so the dangerousity score reflects more than only goals and assists do. Player 1182 scored a goal in the first 15 minutes. His dangerousity scores over the first half are very constant. Player 1241 and 1267 scored a goal in the last 15 minutes of the first half.

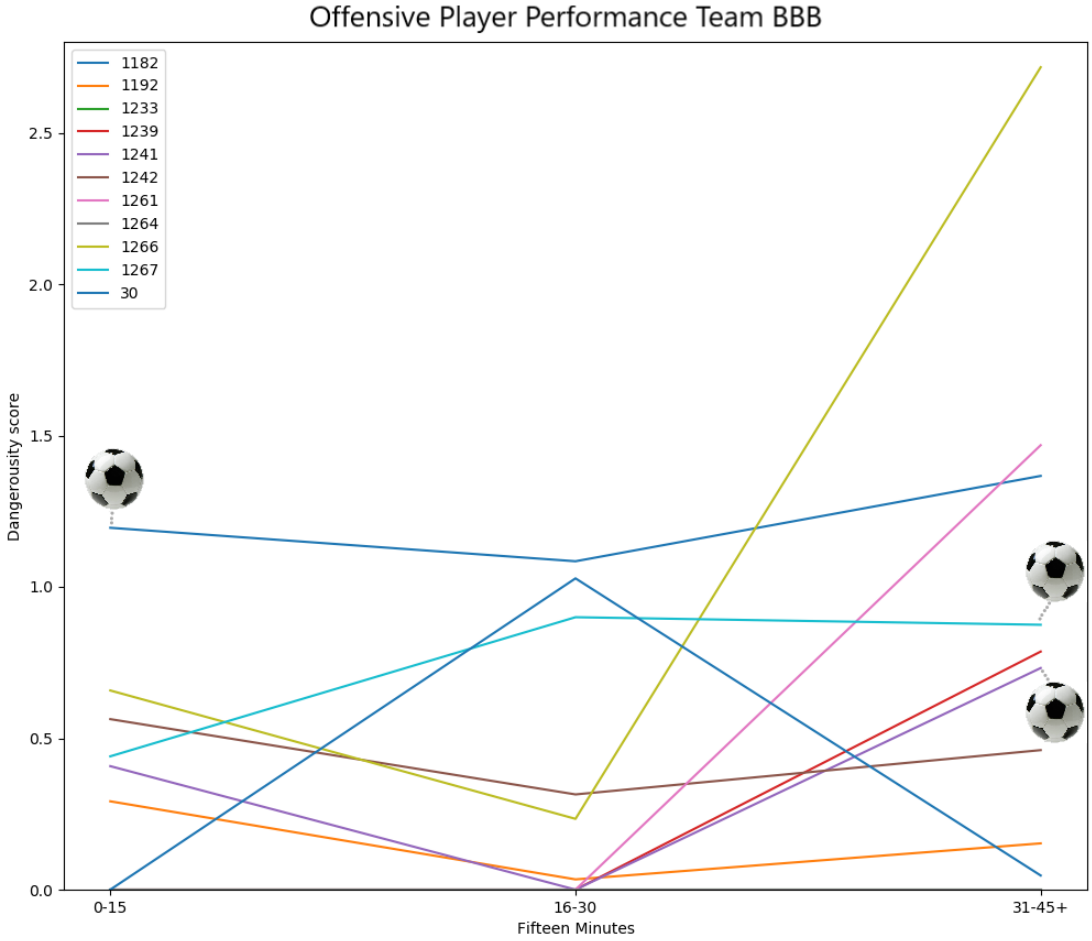


Figure 4.5: Offensive player performance based on the dangerousity score of Team BBB during the first half against Team AAA. Player 1182 scored in the first 15 minutes, player 1241 and 1267 in the last 15 minutes.

In Table 4.5 the dangerousity scores of all players during the first half of Team AAA against Team BBB can be found. Player 1239 has one of the lowest total dangerousity scores, but he has the highest dangerousity score per second, so in the 0.7 seconds that he had possession of the ball in the final third, he was very dangerous. This could be valuable information for the coach.

TeamID	PlayerID	0-15	16-30	31-45+	Total	Seconds in Final Third	Dangerousity per Second
Team BBB	1182	1.20	1.08	1.37	3.65	50.0	0.07
Team BBB	1266	0.66	0.23	2.72	3.61	10.7	0.34
Team AAA	1619	2.11	0.91	0.16	3.19	15.9	0.20
Team AAA	1618	0.80	1.40	0.67	2.87	38.0	0.08
Team BBB	1267	0.44	0.90	0.87	2.21	13.8	0.16
Team AAA	1612	1.92	0.00	0.03	1.95	3.4	0.57
Team AAA	1616	0.13	1.20	0.32	1.65	16.1	0.10
Team BBB	1261	0.00	0.00	1.47	1.47	2.2	0.67
Team AAA	1614	1.23	0.16	0.00	1.39	6.2	0.22
Team BBB	1242	0.56	0.31	0.46	1.34	31.2	0.04
Team AAA	1611	0.10	0.48	0.74	1.32	25.8	0.05
Team BBB	1241	0.41	0.00	0.73	1.14	3.4	0.34
Team BBB	30	0.00	1.03	0.05	1.07	13.1	0.08
Team AAA	1617	0.03	0.00	0.98	1.01	4.6	0.22
Team AAA	1613	0.50	0.06	0.31	0.87	22.9	0.04
Team BBB	1239	0.00	0.00	0.79	0.79	0.7	1.12
Team BBB	1192	0.29	0.03	0.15	0.48	6.2	0.08
Team AAA	1615	0.28	0.05	0.03	0.36	2.8	0.13
Team AAA	1610	0	0	0	0		
Team BBB	1233	0	0	0	0		
Team BBB	1264	0	0	0	0		
Team AAA	1609	0	0	0	0		

Table 4.5: Dangerousity scores of all the players during the first half of Team AAA against Team BBB, including the number of seconds the player is in possession of the ball in the final third and the dangerousity per second.

4.4 Subgroup Discovery

The subgroup discovery is done according to section 3.5.1. We perform a subgroup discovery at depth 1 in section 4.4.1 and at depth 2 in section 4.4.2. Depths higher than 2 are mostly too difficult to interpret. We look at three different binary targets: no shot, shot off target and shot on target (including goal).

4.4.1 Subgroup Discovery at Depth 1

At refinement depth 1, there is one feature that defines the rule for the subgroup. Swap-randomization is used to compute a threshold value per subgroup discovery setting to determine if a subgroup is significant. With a WRAcc of above 0.014 "no shot" is significant at 5%. For "shot off target" the threshold value is 0.011 and for "shot on target" this value is 0.010.

For the target "no shot", a maximum Zone-Control-Pressure dangerousity score ($DA_{max}(ZCP)$) below 0.2286 is the most important, with a WRAcc of 0.082, a coverage of 1831 (62.5%) and a probability of 92.4%. For the target "shot off target", a minimum distance of 23.7 meters or smaller is the most predictive, with a WRAcc of 0.042, a coverage of 1099 (37.5%) and a probability of 22.8%. In Table 4.6 the first ten subgroups for "shot on target" are shown. They are ranked according to the quality measure, WRAcc. The first subgroup is defined by a maximum Zone-Control-Pressure dangerousity score ($DA_{max}(ZCP)$) higher than 0.2286. The coverage is 1099 (37.5%) with a probability of 19.8%, which means that nearly 20% of the attacks in this subgroup lead to a shot on target. In the whole dataset this is only 9%.

The maximum Zone-Control-Pressure dangerousity score ($DA_{max}(ZCP)$) is hard to interpret, but this feature is the most predictive for the outcome of an attack, at refinement depth 1. When the score for Zone for the player with ball is high (i.e., the player is close to the goal), he has a good control over the ball (i.e., the relative speed between player and ball is low) and the pressure of the defensive team is low (i.e., there are few or no defenders between the player with ball and the goal) during an attack, the maximum Zone-Control-Pressure dangerousity score ($DA_{max}(ZCP)$) is high. Other, easier to interpret features in the top 10 are $distToGoal_{min}$

and $centrality_{max}$. They are both a bit obvious, but when the minimum distance to the goal is less than 23.7 meters or the maximum centrality is above 0.57 (where 0 is the side line and 1 the middle of the field) the chance of shooting on target is bigger. A full list of the features can be found in Table B.1 and Table B.2 in Appendix B.

Nr	Coverage	Quality	Probability	Positives	Conditions
1	1099	0.04001	0.19745	217	$DA_{max}(ZCP) >= 0.229$
2	1117	0.03980	0.19517	218	$DA_{max}(ZP) >= 0.233$
3	1141	0.03905	0.19106	218	$Zone_{max} >= 0.45$
4	1099	0.03899	0.19472	214	$DA_{max}(ZPD) >= 0.284$
5	1099	0.03865	0.19381	213	$DA_{max}(ZC) >= 0.142$
6	1099	0.03830	0.19290	212	$DA_{max} >= 0.273$
7	1099	0.03796	0.19199	211	$DA_{avg\_pTpP\_avg}(ZCP) >= 0.102$
8	1099	0.03796	0.19199	211	$distToGoal_{min} <= 23.716$
9	1099	0.03762	0.19108	210	$DA_{avg\_pTpP\_avg}(ZC) >= 0.062$
10	1465	0.03754	0.16587	243	$centrality_{max} >= 0.570$

Table 4.6: Outcome of the Subgroup Discovery in Cortana for the target “shot on target“ at refinement depth 1. $pTpP$ refers to the features that are first aggregated at player-level and then at team-level.

The area under the Receiver Operating Characteristic (ROC) curve is a good metric to compare the performance of classifiers. The ROC curve of “shot on target“ at refinement depth 1 can be found in Figure 4.6, with an AUC of 0.77. Each subgroup represent a point in ROC space, expressed in its false positive rate (FPR) and true positive rate (TPR). The closer the subgroups are to the diagonal, the more random they are. The AUC of “no shot“ is 0.786 and the AUC of “shot off target“ 0.741. Now we can say that the classifier for “no shot“ is the best one.

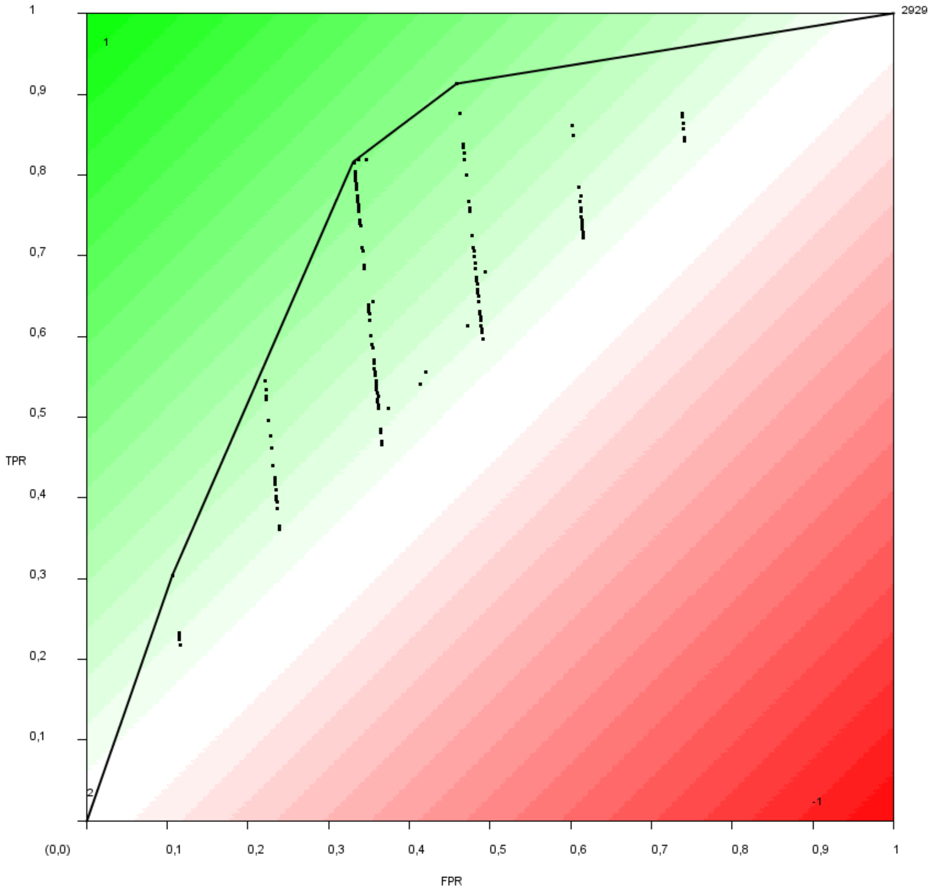


Figure 4.6: Receiver Operating Characteristic curve for the target "shot on target" at refinement depth 1. The Area Under Curve is 0.77.

For all three targets, Zone-Control-Pressure dangerousity score ($DA_{max}(ZCP)$) is (one of) the most important features. In Figure 4.7, boxplots of this feature are shown per attack outcome. A higher $DA_{max}(ZCP)$ score leads in most cases to a better outcome of the attack.

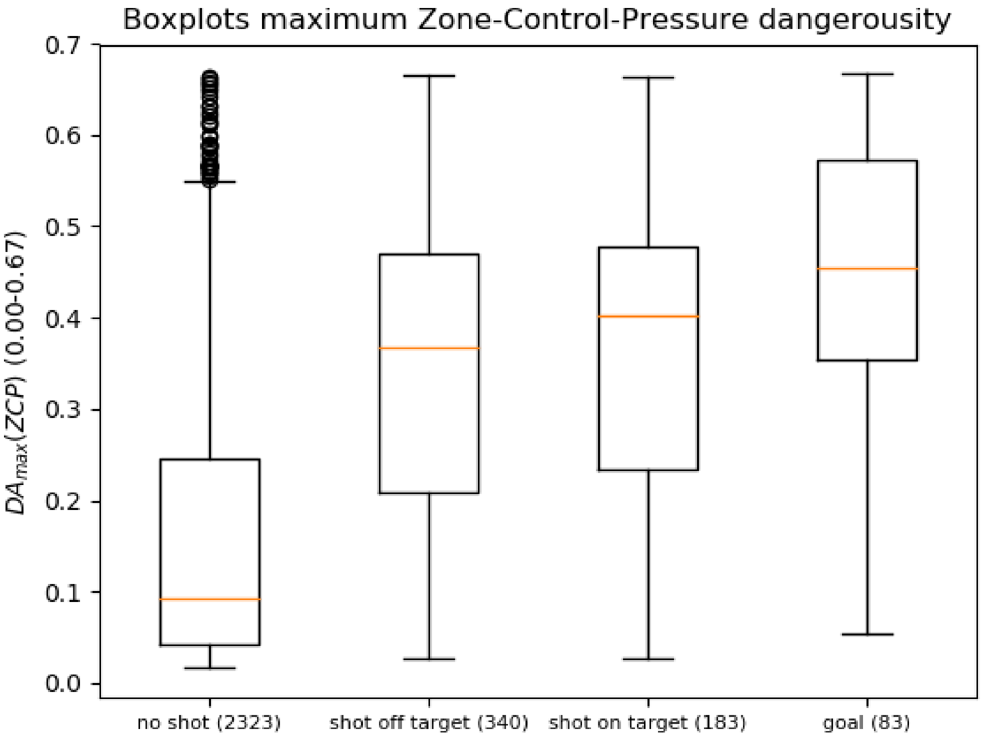


Figure 4.7: Boxplots of the maximum Zone-Control-Pressure dangerousity score ($DA_{max}(ZCP)$) per attack outcome. The value per attack is between 0.00 and 0.67, because the Density feature is set to 1. In parentheses the number of attacks per outcome.

The full Cortana output at refinement depth 1 can be found in Figure C.1, C.2 and C.3 in Appendix C.

4.4.2 Subgroup Discovery at Depth 2

At refinement depth 2, a combination of two features describes the rule of the subgroup. We look again to three targets, as mentioned in the previous section. With a WRAcc of above 0.018 “no shot“ is significant at 5%. For “shot off target“ the threshold value is 0.015 and for “shot on target“ this is 0.013.

For the target “no shot“, the subgroup with the highest WRAcc is when there is a maximum Zone-Pressure dangerousity score ($DA_{max}(ZP)$) below 0.23 in combination with the average (at team-level) of the average (at player-level) angle to goal of all players of the attacking team of 23.49 or higher. This subgroup is somewhat difficult to interpret. If the score for zone in which the player in possession is located is low (i.e. the player is far from the goal) and the pressure of the defensive team is high (i.e. there are many defenders between the player in possession and the goal), and the average (at team-level) of the average (at player-level) angle to the goal of the players with ball of the attacking team is no greater than 57.94, the chance that it does not lead to a shot is greater. The WRAcc of this subgroup is 0.083, the coverage 1620 (55.3%) and the probability 94.4%. For “shot off target“, a maximum centrality of 0.57 or higher in combination with a maximum Zone-Pressure dangerousity score ($DA_{max}(ZP)$) of 0.13 or higher can be seen as the best subgroup. The WRAcc of this subgroup is 0.046, the coverage 1106 (37.8%) and the probability 24.1%. In Table 4.7 the outcome of the subgroup discovery for “shot on target“ can be found. The full Cortana output at refinement depth 2 can be found in Figure C.4, C.5 and C.6 in Appendix C.

Nr	Coverage	Quality	Probability	Positives	Conditions
1	962	0,042552	0,220374	212	$DA_{max}(ZCP) >= 0.229$ AND $angleToGoal_{avg\_pTpP\_avg} <= 57.942$
2	1099	0,042401	0,203822	224	$centrality_{max} >= 0.570$ AND $DA_{avg\_pTpP\_avg}(ZCP) >= 0.065$
3	1099	0,042401	0,203822	224	$centrality_{max} >= 0.570$ AND $DA_{std}(ZP) >= 0.012$
4	1099	0,04206	0,202912	223	$centrality_{max} >= 0.570$ AND $DA_{avg}(ZPD) >= 0.071$
5	1099	0,04206	0,202912	223	$centrality_{max} >= 0.570$ AND $DA_{max}(ZPD) >= 0.154$
6	1099	0,04206	0,202912	223	$centrality_{max} >= 0.570$ AND $DA_{std}(ZPD) >= 0.014$
7	1099	0,04206	0,202912	223	$centrality_{max} >= 0.570$ AND $DA_{avg} >= 0.068$
8	1099	0,04206	0,202912	223	$centrality_{max} >= 0.570$ AND $DA_{max} >= 0.150$
9	978	0,042056	0,216769	212	$DA_{max}(ZP) >= 0.233$ AND $angleToGoal_{avg\_pTpP\_avg} <= 58.457$
10	916	0,04193	0,224891	206	$centrality_{max} >= 0.570$ AND $DA_{max}(ZCP) >= 0.222$

Table 4.7: Outcome of the Subgroup Discovery in Cortana for the target "shot on target" at refinement depth 2. $pTpP$ refers to the features that are first aggregated at player-level and then at team-level.

The ROC curve of "shot on target" at refinement depth 2 can be found in Figure 4.8, with an AUC of 0.807. The test is therefore very accurate. The AUC of "no shot" is 0.805 and for "shot off target" the AUC is 0.767. We can conclude that the classifier for "shot on target" is the best one.

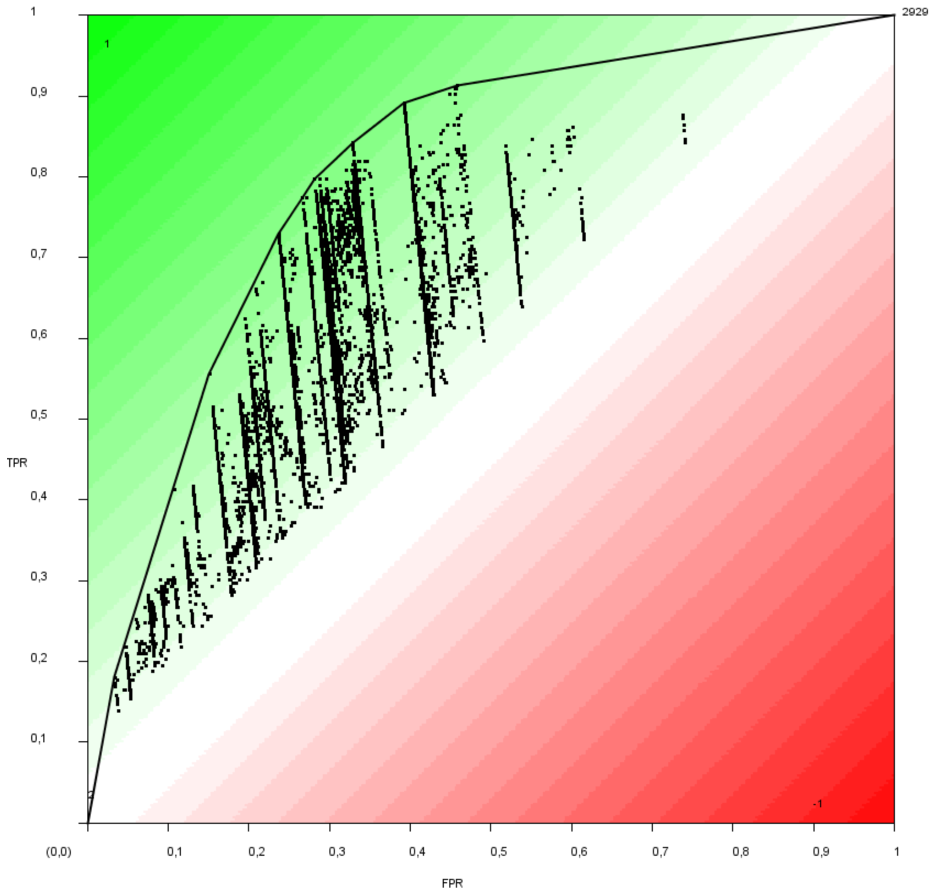


Figure 4.8: Receiver Operating Characteristic curve for the target "shot on target" at refinement depth 2, with an Area Under Curve of 0.807.

Chapter 5

Discussion

The experiments in chapter 4 have shown that the maximum dangerousity score has the strongest correlation with the outcome of an attack (no shot, shot off target, shot on target and goal). There is a moderate correlation that is well represented by boxplots. Afterwards, we have compared the attack outcomes with each other to see if there is a difference in the maximum dangerousity scores between them. From the test for normality follows that only "goal" does come from a normal distribution. Therefore we have performed a Kruskal-Wallis H test, which showed that there was a statistically significant difference in the dangerousity scores between the outcomes of an attack. Therefore, we have repeatedly compared two attack outcomes with each other using a Mann-Whitney U test, which showed that only "shot off target" and "shot on target" have no significant difference in the distribution of the maximum dangerousity score. From these test follows that Link's dangerousity is a good method to measure the offensive performance of individual players and the team. By plotting the dangerousity scores in graphs, the coach gets a good impression of the offensive performances of players and the team.

The experiments have also shown that by using subgroup discovery interesting patterns appear in a dataset with a large number of aggregated features. We have performed the subgroup discovery in three different settings. It shows that the maximum Zone-Control-Pressure dangerousity score ($DA_{max}(ZCP)$) during an attack comes up best. So we can say that when the player is in a good position (i.e., close to the goal), has good ball control and there is little pressure from the defense, the chance of a good result of an attack (i.e., shot or goal) is higher. The value for the maximum Zone-Control-Pressure dangerousity score is rather difficult to interpret. Therefore, future work must prove the practical application of our research by creating an interactive visualization tool to plot the values during an attack in combination with the positions of the players on the field. Visual analytics could be a useful method for doing this, for example by using a Business Intelligence tool. In such a BI tool the KNVB has various possibilities to visualize and analyze the data. Other possibilities in such a tool are comparing the dangerousity scores from one match to another, comparing the dangerousity scores from one player in multiple matches, and comparing the dangerousity scores over a certain period (e.g., per coach) to analyze the offensive performance in this period.

In contrast with Link, we discovered that the density feature adds nothing to the dangerousity feature. This may be due to the fact that we did not have the right parameters, we tested it in an event-based setting with aggregated features or because we tested dangerousity in a different environment, namely matches of the Dutch national soccer team instead of the Bundesliga. Future research must prove this.

Other future research should show whether adjusting the length of the critical period leads to better results. We have now aggregated over the full duration of an attack, but maybe only the last 10 seconds before a critical event (e.g., shot or goal) take place are interesting. In addition, other features could be added that may be related to the outcome of an attack. Examples are the movement dynamics of the players and the ball, the direction in which the players are looking, their position in relation to the ball, the extent to which teammates are available and different individual skills [2]. Moreover, the performance results for dangerousity have to be compared with the coach's view on offensive performance. Only if they match, the use of this tool can be a success.

Chapter 6

Conclusions

In this research, we give an answer on the following question: Can Link's dangerousity be used to meaningfully analyze offensive player- and team-performance? Dangerousity is meaningfully if it says something relevant about the outcome of an attack. Therefore, we have matched the aggregated (min, max, std, avg) dangerousity score and its underlying components (Zone, Control, Pressure and Density) to the attack outcome (no shot, shot off target, shot on target and goal). From this, it follows that the dangerousity method of Link is a good measure to analyze the offensive performance of individual players and the team, but it is not the best. Overall the maximum Zone-Control-Pressure dangerousity score ($DA_{max}(ZCP)$) during an attack comes up best, with the Density feature excluded. We can therefore conclude that when the player with ball is in a good position (i.e., close to the goal), has a good control over the ball (i.e., the relative speed between player and ball is low) and the pressure from the defense is low (there are a few or no defenders between the player and goal) that the chance that the attack leads to a good result is greater (i.e. shot or goal).

Unfortunately, we were not able to replicate the dangerousity method of Link exactly, because the algorithms are not publicly available. Therefore, it may be that with the right parameters the model scores even better. For now, we can say that it can be a useful tool to gain insight into a characteristic that is difficult to measure objectively. Future work should examine how these objective measures match with the coach's views on the performance.

Bibliography

[1] Daniel Memmert and Robert Rein, Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science, Springerplus. 2016; 5(1): 1410.

[2] Link D, Lang S, Seidenschwarz P, (2016) Real Time Quantification of Dangerousity in Football Using Spatiotemporal Tracking Data. PLoS ONE 11(12): e0168768. doi:10.1371/journal. pone.0168768

[3] Manuel Stein , Halldr Janetzko, Daniel Seebacher, Alexander Jger, Manuel Nagel, Jrgen Hlsch, Sven Kosub, Tobias Schreck, Daniel A. Keim and Michael Grossniklaus, How to Make Sense of Team Sport Data: From Acquisition to Data Modeling and Research Aspects, MDPI, 2016.

[4] https://www.stats.com/about/ on 03-06-2018

[5] https://www.stats.com/sportvu-football/ on 03-06-2018

[6] https://www.optasports.com/about/the-opta-difference/ on 08-11-2018

[7] Meerhoff, L.A., de Leeuw, A.-W., Goes, F., & Knobbe, A. (Under review). Mining Soccer Data: Discovering patterns of tactics in tracking data.

[8] Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. From data mining to knowledge discovery in databases. AI Mag. 1996, 17, 37.

[9] Francisco Herrera, Cristbal Jos Carmona, Pedro Gonzlez, Mara Jos del Jesus, An overview on subgroup discovery: foundations and applications, 2010.

[10] https://img.fifa.com/image/upload/datdz0pms85gbnqy4j3k.pdf on 12-11-2018, page 7.

[11] Marvin Meeng, Arno Knobber, Flexible Enrichment with Cortana Software Demo, 2011

[12] Wouter Duivesteijn, Arno Knobbe, Exploiting False Discoveries Statistical Validation of Patterns and Quality Measures in Subgroup Discovery, 2011 11th IEEE International Conference on Data Mining.

Appendices

Appendix A

Dangerousity

Dangerousity (DA) can be described as the chance of scoring a goal for every moment in time a player is in possession of the ball during an attack. An attack starts when a player walks into the last 35 meters of the field (final third) with the ball and ends when the opposite team retrieves ball possession or when the ball passes the center line.

Dangerousity is based on four components: Zone, Control, Pressure and Density, which will be discussed in the next sections. Zone and Control are attacking components, so they increase Dangerousity. Pressure and Density are defending components, so they decrease dangerousity. The values for the four components are in a range between 0 (low) and 1 (high).

Dangerousity is calculated for every moment in time (t) with the following formula:

$$DA(t) = ZO(t) * \left(1 - \frac{1 - CO(t) + PR(t) + DE(t)}{k_1}\right)$$

A.1 Zone

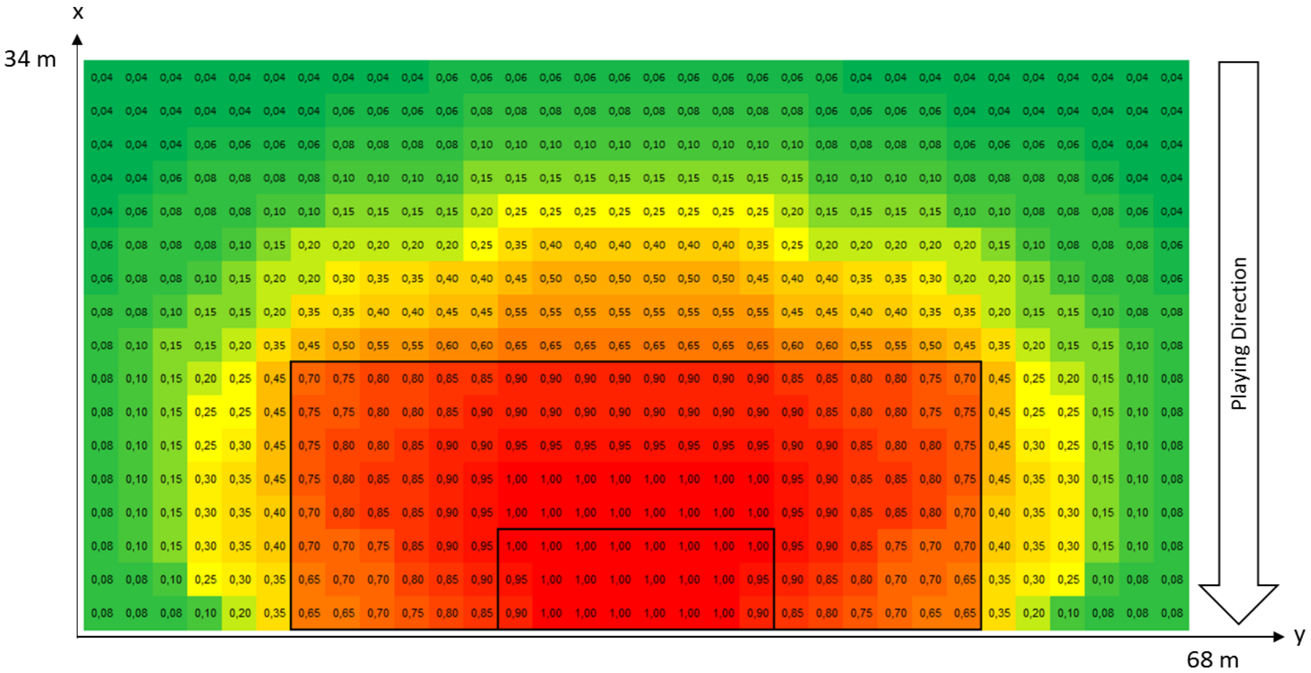
Zone (ZO) is a value for the dangerousity only based on the position of the player on the field. Link determines for every player who is in possession of the ball in the last 34 meters of the field a value for Zone. In this research the players who are in the last 35 meters to the goal get a value for Zone, because this can be qualified as the final third of the field.

The values for Zone can be found in Figure A.1. Link made some assumptions to evaluate the position on the field for the player with ball. First, the danger rises if a player is more central and closer to the goal. Second, if a player walks into the penalty area the danger rises, because of the chance of a penalty kick. Third, there is a area in front of the goal where the danger does not increase any further. Fourth, if a player is in an sharp angle to the goal the danger decreases. Fifth, the danger arises on the side of the penalty area because of the chance of a cross with little risk of offside.

The implementation of Zone is done by simply putting all the values for every 1 by 1 meter in a CSV file. This CSV file is loaded into an array in Python. For every player which is in possession of the ball in the final third (last 35 meters) the X and Y coordinates are determined and the value of zone is read from the array.

A.2 Control

The ball control (CO) of a player is estimated by means of the average relative speed between player and ball (v_{rel}). It is assumed that a player has a high Control when the relative speed between player and ball is low, for example as a player is dribbling with the ball. A player has low Control when the relative speed is high, for



0,04	0,04	0,04	0,04	0,04	0,04	0,04	0,04	0,04	0,04	0,06	0,06	0,06	0,06	0,06	0,06	0,06	0,06	0,06	0,06	0,06	0,06	0,04	0,04	0,04	0,04	0,04	0,04	0,04	0,04	0,04	0,04
0,04	0,04	0,04	0,04	0,04	0,04	0,04	0,06	0,06	0,06	0,06	0,08	0,08	0,08	0,08	0,08	0,08	0,08	0,08	0,08	0,08	0,06	0,06	0,06	0,06	0,04	0,04	0,04	0,04	0,04	0,04	0,04
0,04	0,04	0,04	0,06	0,06	0,06	0,06	0,08	0,08	0,08	0,08	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,08	0,08	0,08	0,08	0,06	0,06	0,06	0,06	0,04	0,04	0,04
0,04	0,04	0,06	0,08	0,08	0,08	0,08	0,10	0,10	0,10	0,10	0,15	0,15	0,15	0,15	0,15	0,15	0,15	0,15	0,15	0,15	0,10	0,10	0,10	0,10	0,08	0,08	0,08	0,08	0,06	0,04	0,04
0,04	0,06	0,08	0,08	0,08	0,10	0,10	0,15	0,15	0,15	0,15	0,20	0,25	0,25	0,25	0,25	0,25	0,25	0,25	0,25	0,20	0,15	0,15	0,15	0,15	0,10	0,10	0,08	0,08	0,08	0,06	0,04
0,06	0,08	0,08	0,08	0,10	0,15	0,20	0,20	0,20	0,20	0,20	0,25	0,35	0,40	0,40	0,40	0,40	0,40	0,40	0,35	0,25	0,20	0,20	0,20	0,20	0,20	0,15	0,10	0,08	0,08	0,08	0,06
0,06	0,08	0,08	0,10	0,15	0,20	0,20	0,30	0,35	0,35	0,40	0,40	0,45	0,50	0,50	0,50	0,50	0,50	0,50	0,45	0,40	0,40	0,35	0,35	0,30	0,20	0,20	0,15	0,10	0,08	0,08	0,06
0,08	0,08	0,10	0,15	0,15	0,20	0,35	0,35	0,40	0,40	0,45	0,45	0,55	0,55	0,55	0,55	0,55	0,55	0,55	0,55	0,45	0,45	0,40	0,40	0,35	0,35	0,20	0,15	0,15	0,10	0,08	0,08
0,08	0,10	0,15	0,15	0,20	0,35	0,45	0,50	0,55	0,55	0,60	0,60	0,65	0,65	0,65	0,65	0,65	0,65	0,65	0,65	0,60	0,60	0,55	0,55	0,50	0,45	0,35	0,20	0,15	0,15	0,10	0,08
0,08	0,10	0,15	0,20	0,25	0,45	0,70	0,75	0,80	0,80	0,85	0,85	0,90	0,90	0,90	0,90	0,90	0,90	0,90	0,90	0,85	0,85	0,80	0,80	0,75	0,70	0,45	0,25	0,20	0,15	0,10	0,08
0,08	0,10	0,15	0,25	0,25	0,45	0,75	0,75	0,80	0,80	0,85	0,90	0,90	0,90	0,90	0,90	0,90	0,90	0,90	0,90	0,90	0,85	0,80	0,80	0,75	0,75	0,45	0,25	0,25	0,15	0,10	0,08
0,08	0,10	0,15	0,25	0,30	0,45	0,75	0,80	0,80	0,85	0,90	0,90	0,95	0,95	0,95	0,95	0,95	0,95	0,95	0,95	0,90	0,90	0,85	0,80	0,80	0,75	0,45	0,30	0,25	0,15	0,10	0,08
0,08	0,10	0,15	0,30	0,35	0,45	0,75	0,80	0,85	0,85	0,90	0,95	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	0,95	0,90	0,85	0,85	0,80	0,75	0,45	0,35	0,30	0,15	0,10	0,08
0,08	0,10	0,15	0,30	0,35	0,40	0,70	0,80	0,85	0,85	0,90	0,95	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	0,95	0,90	0,85	0,85	0,80	0,70	0,40	0,35	0,30	0,15	0,10	0,08
0,08	0,10	0,15	0,30	0,35	0,40	0,70	0,70	0,75	0,85	0,90	0,95	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	0,95	0,90	0,85	0,75	0,70	0,70	0,40	0,35	0,30	0,15	0,10	0,08
0,08	0,08	0,10	0,25	0,30	0,35	0,65	0,70	0,70	0,80	0,85	0,90	0,95	1,00	1,00	1,00	1,00	1,00	1,00	0,95	0,90	0,85	0,80	0,70	0,70	0,65	0,35	0,30	0,25	0,10	0,08	0,08
0,08	0,08	0,08	0,10	0,20	0,35	0,65	0,65	0,70	0,75	0,80	0,85	0,90	1,00	1,00	1,00	1,00	1,00	1,00	0,90	0,85	0,80	0,75	0,70	0,65	0,65	0,35	0,20	0,10	0,08	0,08	0,08

Figure A.1: **Values for Zone.** Retrieved from Real Time Quantification of Dangerousity in Football Using Spatiotemporal Tracking Data, Link et al. (2017), by PLoS ONE.

example if the player has brief contact with the ball when shooting on goal. Control is calculated with the following formula:

$$CO(v_{rel}) = 1 - k_2 * {v_{rel}}^2$$

In the paper of Link k_1 is not defined, but when v_{rel} is above 25 m/s, Control is equal to 0. k_2 can now be calculated: $k_2 = \frac{1}{25^2} = 0.0016$.

A.3 Pressure

A defender (D) exerts Pressure (PR) when he is at a certain distance (d_D) from the player with ball (P). Pressure is divided into four different sub-areas called the Pressure Zone (PZ). The sub-area to which a defender belongs is the result of the distance and the angle of the player with ball, defender and goal. The four sub-areas with its values are:

- High Pressure Zone: the defender has a distance shorter than 1 meter to the player with ball. Value: 10.
- Head-On Zone: the defender is between the player and the goal and has a distance from 1 to 4 meters to the player with ball. Value: 8.
- Lateral Zone: the defender is on the side of the player with ball and has a distance from 1 to 3 meters to the player with ball. Value: 4.
- Hind Zone: the player is behind the player with ball and has a distance from 1 to 2 meters to the player with ball. Value: 2.

A visual representation of Pressure can be found in Figure A.2.

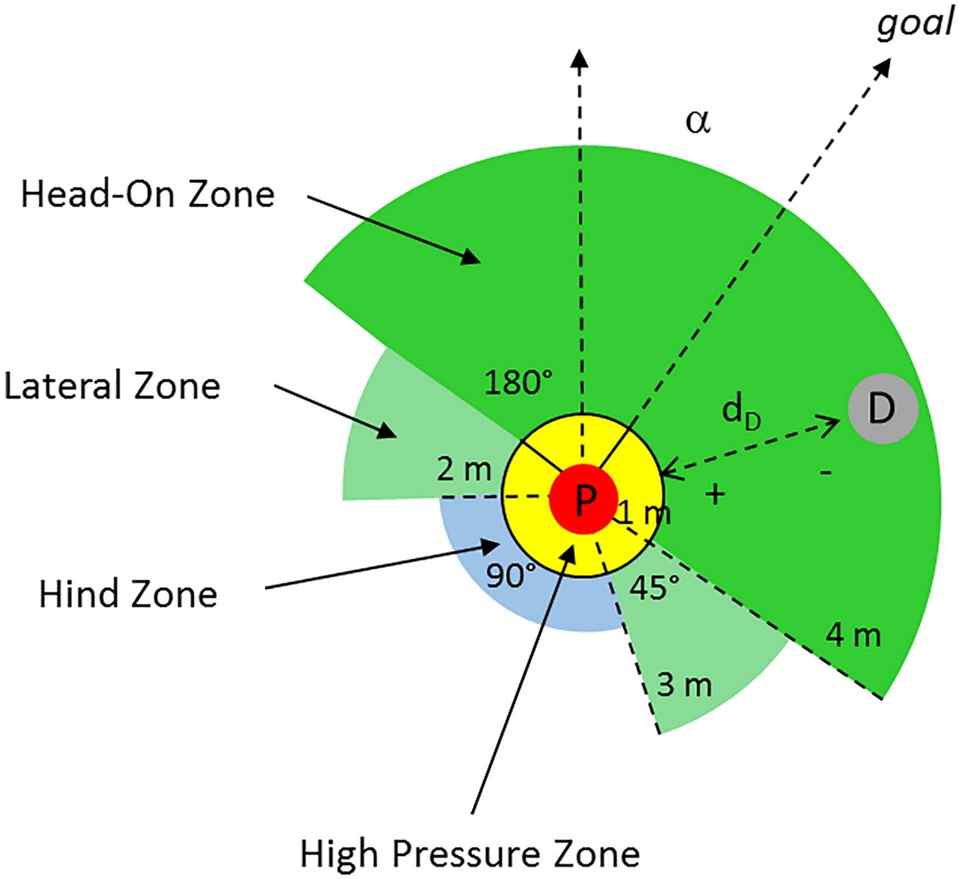


Figure A.2: **The four sub-areas of Pressure.** Defender (D) exerts pressure on player with ball (P) depending on his distance and angle to P. Retrieved from Real Time Quantification of Dangerousity in Football Using Spatiotemporal Tracking Data, Link et al. (2017), by PLoS ONE.

For every player which is in possession of the ball in the final third, the distance to every defender and the angle between the player with ball (P), defender (D) and goal (G) is calculated. To calculate the distance between the player with ball and the defender the Euclidean distance is used:

$$Dist(P, D) = \sqrt{(P_X - D_X)^2 + (P_Y - D_Y)^2}$$

To calculate the angle (α) between the player with ball, defender and goal the law of cosines is used:

$$\alpha = cos^{-1}\left(\frac{Dist(P, D)^2 + Dist(P, G)^2 + Dist(D, G)^2}{2 * Dist(P, D) * Dist(P, G)}\right)$$

where cos^{-1} is the arcus cosinus function; the inverse cosinus function.

Based on the angle and the distance, the sub-area (r_{ZO}) a player belongs to is determined. The pressure for every individual defender is then calculated by the following formula:

$$PR_{D_i}(d_{D_i}, \alpha) = 1 - \frac{d_{D_i}}{r_{ZO}(\alpha)}$$

where d_{Di} is the distance between player with ball and defender (equal to $Dist(P, D)$).

Every defender who is in the Pressure Zone increases the Pressure. So the total pressure on the player with ball is calculated by the following formula:

$$PR(x) = 1 - e^{-k_3 x}, wherex = \sum \forall D_i insidePZPR_{Di}$$

A.4 Density

Density is divided into two components: Shot Density (SD) and Pass Density (PD). SD is the chance for a defender of blocking a shot. A defender is able to block a shot when he is in the Blocking Zone (BZ): the zone between the player in possession and the goal (Figure A.3). The BZ starts two meters next to the player in possession and ends ten meters from the center of the goal on both sides; 2.68 meters next to the goalposts.

The SD for a single defender is calculated by means of the distance between the player in possession and the defender (d_D), and between the player in possession and the goal (d_G):

$$SD_D(d_D, d_G) = 1 - \frac{d_D}{d_G}$$

The SD for the player in possession is also logarithmically determined with the sum of all the defenders in the BZ:

$$SD(x) = 1 - e^{-k_4 x}, wherex = \sum \forall D_i insideBZSD_{Di}$$

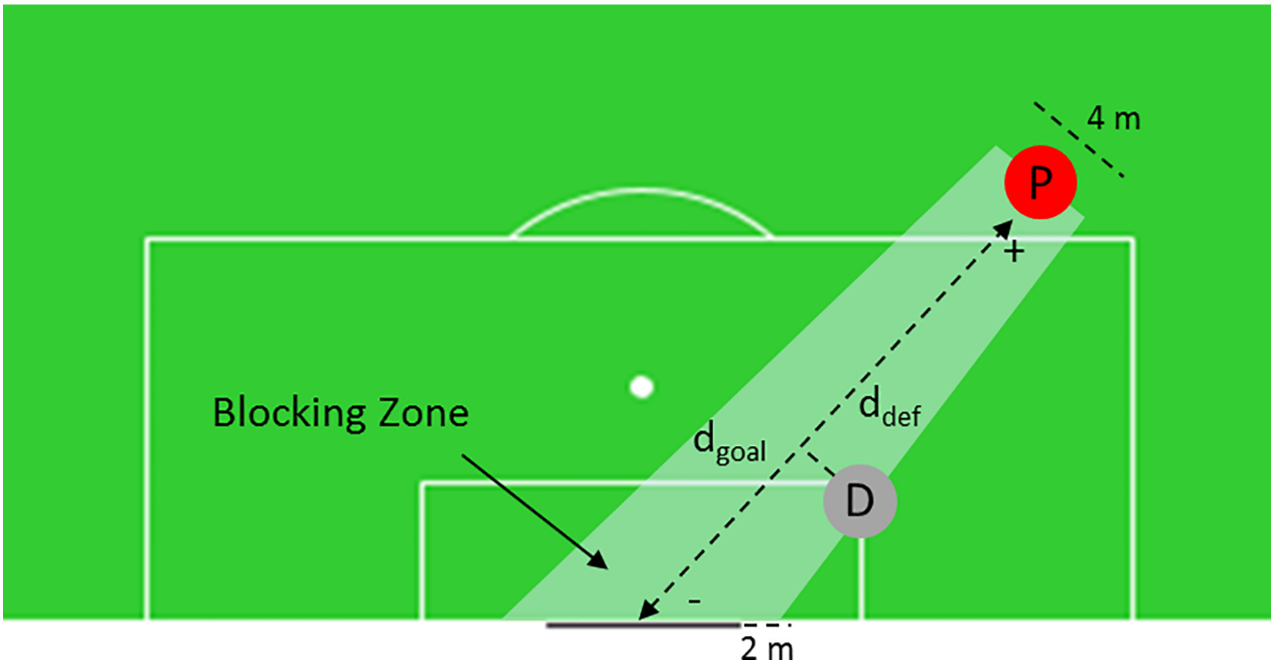


Figure A.3: **Blocking Zone to determine Shot Density.** A defender creates Shot Density when he is in the Blocking Zone: the zone between the player in possession of the ball (P) and the goal, where d_{def} is the distance between P and the defender (D), and d_{goal} is the distance between P and the goal. Retrieved from Real Time Quantification of Dangerousity in Football Using Spatiotemporal Tracking Data, Link et al. (2017), by PLoS ONE.

Pass Density (PD) is the chance of intercepting an offensive pass or cross. This chance increases when the defenders have a majority against the attackers. Majority (M) is defined as the difference between defenders and attackers in the Interception Zone (IZ): the zone between the player in possession (P), the goal and 11 meters in front of the goal (Figure A.4). In Figure A.4 the Majority is 1 because there are four defenders and three attackers in the IZ (excluding P).

An arcus tangent function (tan^{-1}) is used to calculate the PD:

$$PD(M) = 0.5 + \frac{tan^{-1}(k_5 M)}{\pi}$$

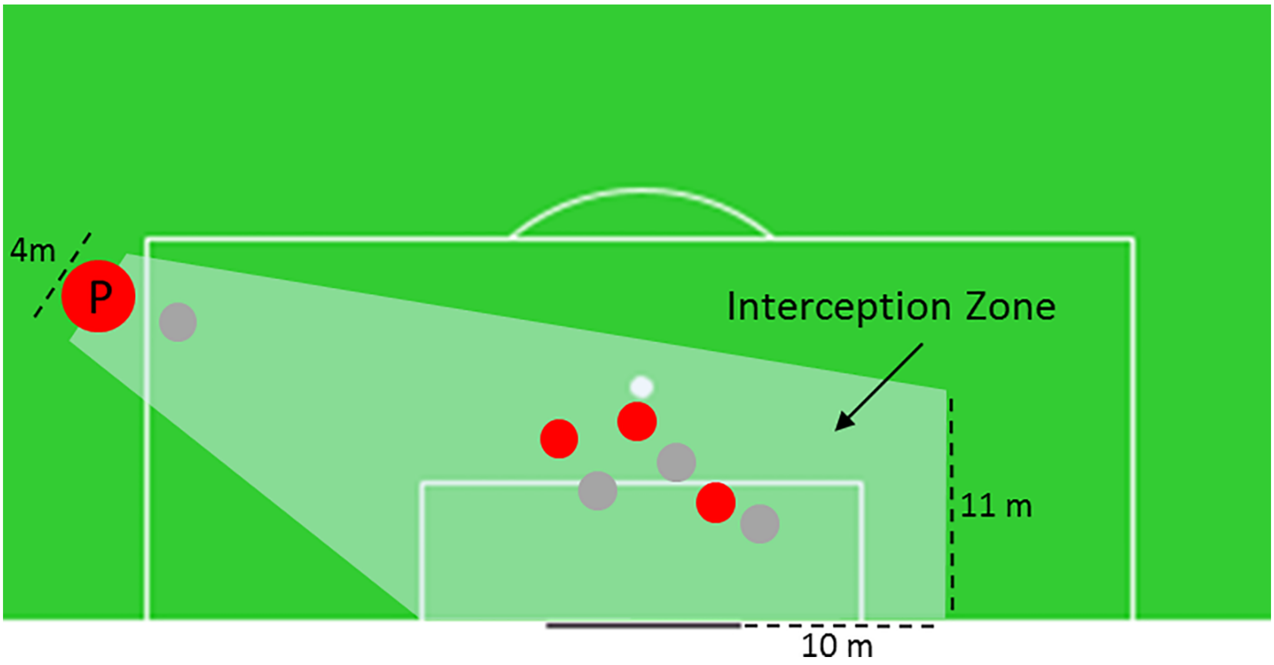


Figure A.4: **Interception Zone to determine Pass Density.** The difference between attackers (red dots) and defenders (grey dots) is called the Majority, where P is the player in possession of the ball. Retrieved from Real Time Quantification of Dangerousity in Football Using Spatiotemporal Tracking Data, Link et al. (2017), by PLoS ONE.

The contribution of Shot Density and Pass Density to the total Density depends on the centrality of the player in possession of the ball. When a player is in front of the goal it is more likely that he shoots on goal than give a pass or cross. When a player has a sharp angle to the goal, it is more likely to give a pass or cross than shooting on goal. So when the centrality (C) of P is high the contribution of SD is higher than the contribution of PD. This leads to the following formula:

$$DE(C) = C * SD + (1 - C) * PD$$

Appendix B

Feature List

Feature	Description
Dist to closest home	Distance to closest player of the home team.
Dist to closest visitor	Distance to closest player of the visitors team.
Speed	Speed in meters per second.
Link_Zone	Value of player with ball in the final third (last 34 meters).
InZone	Boolean to decide if the player with ball is in the final third (last 34 meters)?
OpponentsHalf	Boolean to decide if the player with ball is on the half of the opponent.
InPenaltyArea	Boolean to decide if the player with ball is in the penalty area.
Link_Control	Ball control of player with ball.
VelRelToBall	Relative velocity between player and ball.
VelRelToBallSquared	Square of relative velocity between player and ball.
Link_Pressure	Pressure on player with ball.
MinDistToDef	Distance to closest defender for player with ball.
AvgDistToDef2	Average distance to the two closest defenders for player with ball.
AvgDistToDef3	Average distance to the three closest defenders for player with ball.
AngleInPossDefGoal	Angle between player with ball, defender and goal for the closest defender.
AngleInPossDefGoal2	Average angle between player with ball, defender and goal for the two closest defenders.
AngleInPossDefGoal3	Average angle between player with ball, defender and goal for the three closest defenders.
Link_Density	Density for player with ball.
Link_SDPlayerWithBall	Shot Density for player with ball.
Link_PDPlayerWithBall	Pass Density for player with ball.
AngleToGoal	Angle to goal for player with ball.
Majority	Difference between the number of defenders and attackers within the Interception Zone.
Centrality	Centrality for the player with ball. Formula: 1 abs(Y_coor) / (fieldwidth / 2).
DistToGoal	Distance from player with ball to the goal.

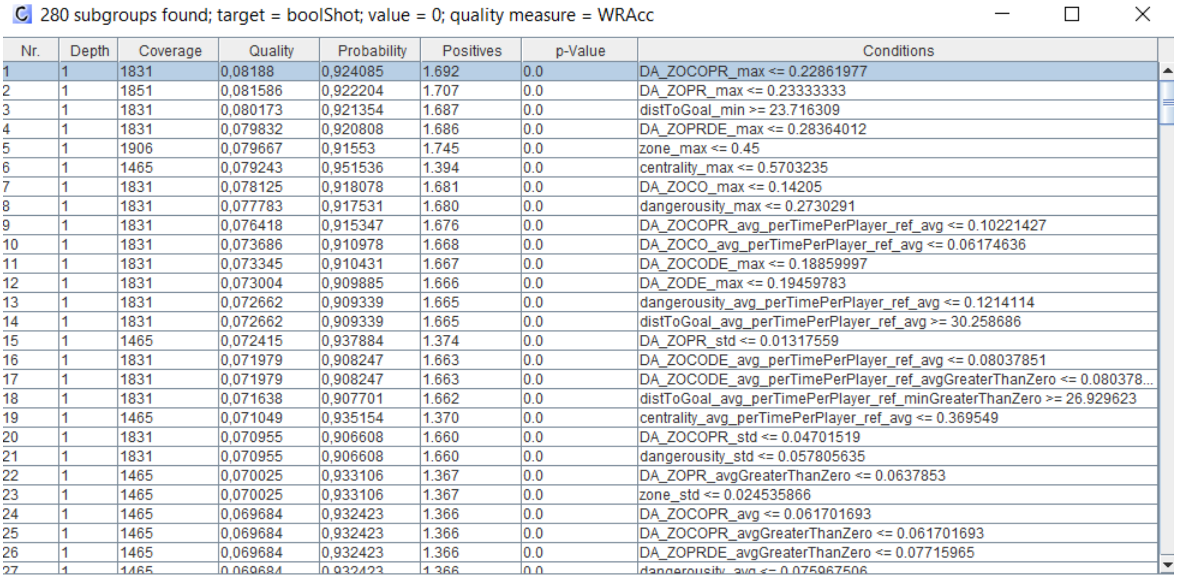
Table B.1: Feature list of all the components to come to the dangerousity score.

Feature	**Description**
Dangerousity	The dangerousity score is a combination of Link_Zone, Link_Control, Link_Pressure and Link_Density. Formula: Dangerousity = Zone * (1 (1 Control + Pressure + Density) / 3).
DA(ZCP)	Dangerousity score consisting of Zone, Control and Pressure, with Density set to 1.
DA(ZCD)	Dangerousity score consisting of Zone, Control and Density, with Pressure set to 1.
DA(ZPD)	Dangerousity score consisting of Zone, Pressure and Density, with Control set to 1.
DA(CPD)	Dangerousity score consisting of Control, Pressure and Density, with Zone set to 1.
DA(ZC)	Dangerousity score consisting of Zone and Control, with Pressure and Density set to 1.
DA(ZP)	Dangerousity score consisting of Zone and Pressure, with Control and Density set to 1.
DA(ZD)	Dangerousity score consisting of Zone and Density, with Control and Pressure set to 1.
DA(CP)	Dangerousity score consisting of Control and Pressure, with Zone and Density set to 1.
DA(CD)	Dangerousity score consisting of Control and Density, with Zone and Pressure set to 1.
DA(PD)	Dangerousity score consisting of Pressure and Density, with Zone and Control set to 1.

Table B.2: Feature list of the dangerousity score and dangerousity scores where a certain feature is made constant.

Appendix C

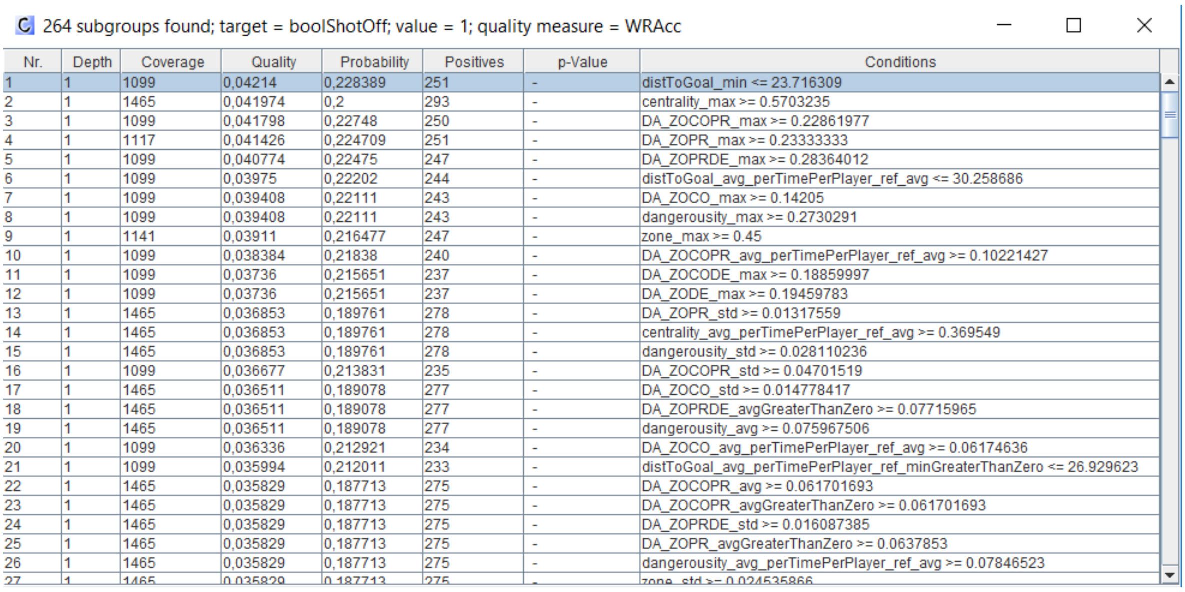
Cortana Output



280 subgroups found; target = boolShot; value = 0; quality measure = WRAcc

Nr.	Depth	Coverage	Quality	Probability	Positives	p-Value	Conditions
1	1	1831	0,08188	0,924085	1.692	0.0	DA_ZOCOPR_max <= 0.22861977
2	1	1851	0,081586	0,922204	1.707	0.0	DA_ZOPR_max <= 0.23333333
3	1	1831	0,080173	0,921354	1.687	0.0	distToGoal_min >= 23.716309
4	1	1831	0,079832	0,920808	1.686	0.0	DA_ZOPRDE_max <= 0.28364012
5	1	1906	0,079667	0,91553	1.745	0.0	zone_max <= 0.45
6	1	1465	0,079243	0,951536	1.394	0.0	centrality_max <= 0.5703235
7	1	1831	0,078125	0,918078	1.681	0.0	DA_ZOCO_max <= 0.14205
8	1	1831	0,077783	0,917531	1.680	0.0	dangerousity_max <= 0.2730291
9	1	1831	0,076418	0,915347	1.676	0.0	DA_ZOCOPR_avg_perTimePerPlayer_ref_avg <= 0.10221427
10	1	1831	0,073686	0,910978	1.668	0.0	DA_ZOCO_avg_perTimePerPlayer_ref_avg <= 0.06174636
11	1	1831	0,073345	0,910431	1.667	0.0	DA_ZOCODE_max <= 0.18859997
12	1	1831	0,073004	0,909885	1.666	0.0	DA_ZODE_max <= 0.19459783
13	1	1831	0,072662	0,909339	1.665	0.0	dangerousity_avg_perTimePerPlayer_ref_avg <= 0.1214114
14	1	1831	0,072662	0,909339	1.665	0.0	distToGoal_avg_perTimePerPlayer_ref_avg >= 30.258686
15	1	1465	0,072415	0,937884	1.374	0.0	DA_ZOPR_std <= 0.01317559
16	1	1831	0,071979	0,908247	1.663	0.0	DA_ZOCODE_avg_perTimePerPlayer_ref_avg <= 0.08037851
17	1	1831	0,071979	0,908247	1.663	0.0	DA_ZOCODE_avg_perTimePerPlayer_ref_avgGreaterThanZero <= 0.080378...
18	1	1831	0,071638	0,907701	1.662	0.0	distToGoal_avg_perTimePerPlayer_ref_minGreaterThanZero >= 26.929623
19	1	1465	0,071049	0,935154	1.370	0.0	centrality_avg_perTimePerPlayer_ref_avg <= 0.369549
20	1	1831	0,070955	0,906608	1.660	0.0	DA_ZOCOPR_std <= 0.04701519
21	1	1831	0,070955	0,906608	1.660	0.0	dangerousity_std <= 0.057805635
22	1	1465	0,070025	0,933106	1.367	0.0	DA_ZOPR_avgGreaterThanZero <= 0.0637853
23	1	1465	0,070025	0,933106	1.367	0.0	zone_std <= 0.024535866
24	1	1465	0,069684	0,932423	1.366	0.0	DA_ZOCOPR_avg <= 0.061701693
25	1	1465	0,069684	0,932423	1.366	0.0	DA_ZOCOPR_avgGreaterThanZero <= 0.061701693
26	1	1465	0,069684	0,932423	1.366	0.0	DA_ZOPRDE_avgGreaterThanZero <= 0.07715965
27	1	1465	0,069684	0,932423	1.366	0.0	dangerousity_avg <= 0.075967506

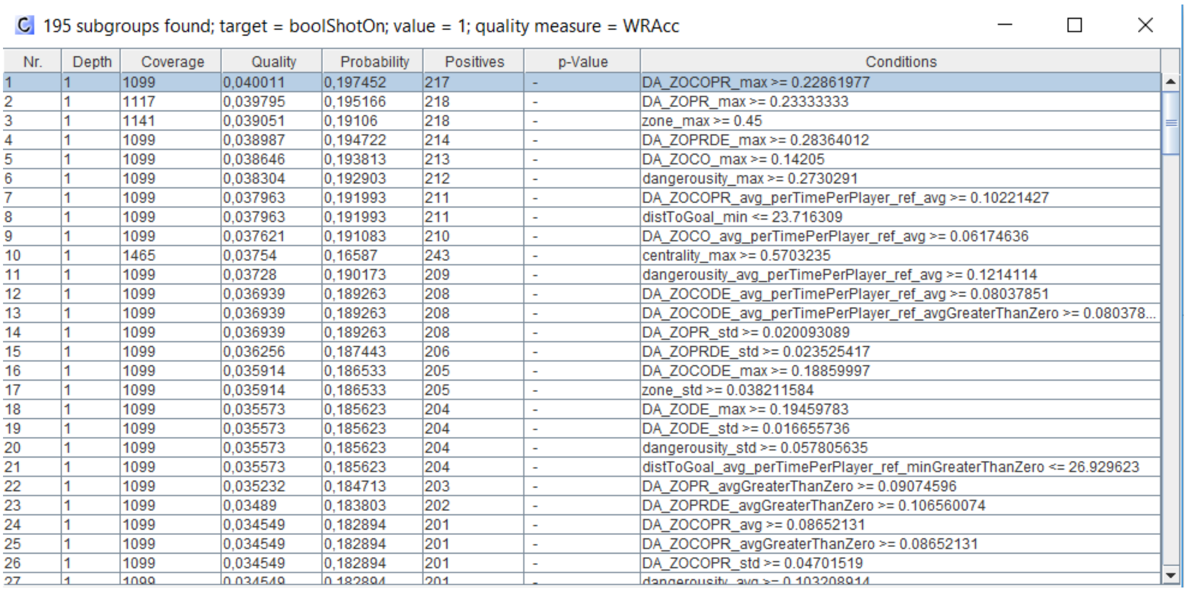
Figure C.1: Cortana output for the target “no shot“ at refinement depth 1.



264 subgroups found; target = boolShotOff; value = 1; quality measure = WRAcc

Nr.	Depth	Coverage	Quality	Probability	Positives	p-Value	Conditions
1	1	1099	0,04214	0,228389	251	-	distToGoal_min <= 23.716309
2	1	1465	0,041974	0,2	293	-	centrality_max >= 0.5703235
3	1	1099	0,041798	0,22748	250	-	DA_ZOCOPR_max >= 0.22861977
4	1	1117	0,041426	0,224709	251	-	DA_ZOPR_max >= 0.23333333
5	1	1099	0,040774	0,22475	247	-	DA_ZOPRDE_max >= 0.28364012
6	1	1099	0,03975	0,22202	244	-	distToGoal_avg_perTimePerPlayer_ref_avg <= 30.258686
7	1	1099	0,039408	0,22111	243	-	DA_ZOCO_max >= 0.14205
8	1	1099	0,039408	0,22111	243	-	dangerousity_max >= 0.2730291
9	1	1141	0,03911	0,216477	247	-	zone_max >= 0.45
10	1	1099	0,038384	0,21838	240	-	DA_ZOCOPR_avg_perTimePerPlayer_ref_avg >= 0.10221427
11	1	1099	0,03736	0,215651	237	-	DA_ZOCODE_max >= 0.18859997
12	1	1099	0,03736	0,215651	237	-	DA_ZODE_max >= 0.19459783
13	1	1465	0,036853	0,189761	278	-	DA_ZOPR_std >= 0.01317559
14	1	1465	0,036853	0,189761	278	-	centrality_avg_perTimePerPlayer_ref_avg >= 0.369549
15	1	1465	0,036853	0,189761	278	-	dangerousity_std >= 0.028110236
16	1	1099	0,036677	0,213831	235	-	DA_ZOCOPR_std >= 0.04701519
17	1	1465	0,036511	0,189078	277	-	DA_ZOCO_std >= 0.014778417
18	1	1465	0,036511	0,189078	277	-	DA_ZOPRDE_avgGreaterThanZero >= 0.07715965
19	1	1465	0,036511	0,189078	277	-	dangerousity_avg >= 0.075967506
20	1	1099	0,036336	0,212921	234	-	DA_ZOCO_avg_perTimePerPlayer_ref_avg >= 0.06174636
21	1	1099	0,035994	0,212011	233	-	distToGoal_avg_perTimePerPlayer_ref_minGreaterThanZero <= 26.929623
22	1	1465	0,035829	0,187713	275	-	DA_ZOCOPR_avg >= 0.061701693
23	1	1465	0,035829	0,187713	275	-	DA_ZOCOPR_avgGreaterThanZero >= 0.061701693
24	1	1465	0,035829	0,187713	275	-	DA_ZOPRDE_std >= 0.016087385
25	1	1465	0,035829	0,187713	275	-	DA_ZOPR_avgGreaterThanZero >= 0.0637853
26	1	1465	0,035829	0,187713	275	-	dangerousity_avg_perTimePerPlayer_ref_avg >= 0.07846523
27	1	1465	0,035829	0,187713	275	-	zone_std >= 0.024535866

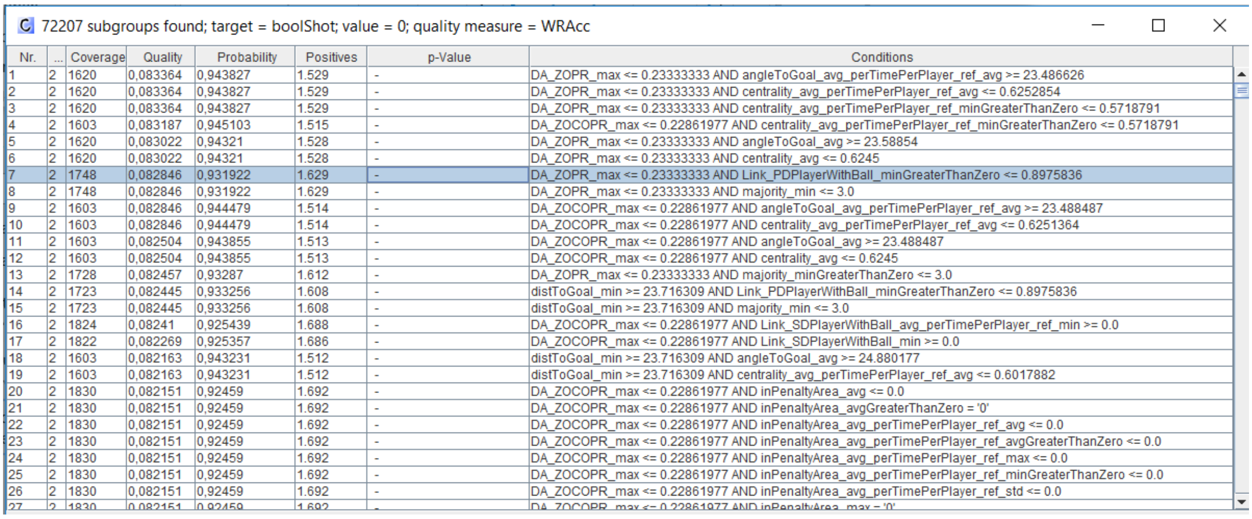
Figure C.2: Cortana output for the target "shot off target" at refinement depth 1.



195 subgroups found; target = boolShotOn; value = 1; quality measure = WRAcc

Nr.	Depth	Coverage	Quality	Probability	Positives	p-Value	Conditions
1	1	1099	0,040011	0,197452	217	-	DA_ZOCOPR_max >= 0.22861977
2	1	1117	0,039795	0,195166	218	-	DA_ZOPR_max >= 0.23333333
3	1	1141	0,039051	0,19106	218	-	zone_max >= 0.45
4	1	1099	0,038987	0,194722	214	-	DA_ZOPRDE_max >= 0.28364012
5	1	1099	0,038646	0,193813	213	-	DA_ZOCO_max >= 0.14205
6	1	1099	0,038304	0,192903	212	-	dangerousity_max >= 0.2730291
7	1	1099	0,037963	0,191993	211	-	DA_ZOCOPR_avg_perTimePerPlayer_ref_avg >= 0.10221427
8	1	1099	0,037963	0,191993	211	-	distToGoal_min <= 23.716309
9	1	1099	0,037621	0,191083	210	-	DA_ZOCO_avg_perTimePerPlayer_ref_avg >= 0.06174636
10	1	1465	0,03754	0,16587	243	-	centrality_max >= 0.5703235
11	1	1099	0,03728	0,190173	209	-	dangerousity_avg_perTimePerPlayer_ref_avg >= 0.1214114
12	1	1099	0,036939	0,189263	208	-	DA_ZOCODE_avg_perTimePerPlayer_ref_avg >= 0.08037851
13	1	1099	0,036939	0,189263	208	-	DA_ZOCODE_avg_perTimePerPlayer_ref_avgGreaterThanZero >= 0.080378...
14	1	1099	0,036939	0,189263	208	-	DA_ZOPR_std >= 0.020093089
15	1	1099	0,036256	0,187443	206	-	DA_ZOPRDE_std >= 0.023525417
16	1	1099	0,035914	0,186533	205	-	DA_ZOCODE_max >= 0.18859997
17	1	1099	0,035914	0,186533	205	-	zone_std >= 0.038211584
18	1	1099	0,035573	0,185623	204	-	DA_ZODE_max >= 0.19459783
19	1	1099	0,035573	0,185623	204	-	DA_ZODE_std >= 0.016655736
20	1	1099	0,035573	0,185623	204	-	dangerousity_std >= 0.057805635
21	1	1099	0,035573	0,185623	204	-	distToGoal_avg_perTimePerPlayer_ref_minGreaterThanZero <= 26.929623
22	1	1099	0,035232	0,184713	203	-	DA_ZOPR_avgGreaterThanZero >= 0.09074596
23	1	1099	0,03489	0,183803	202	-	DA_ZOPRDE_avgGreaterThanZero >= 0.106560074
24	1	1099	0,034549	0,182894	201	-	DA_ZOCOPR_avg >= 0.08652131
25	1	1099	0,034549	0,182894	201	-	DA_ZOCOPR_avgGreaterThanZero >= 0.08652131
26	1	1099	0,034549	0,182894	201	-	DA_ZOCOPR_std >= 0.04701519
27	1	1099	0,034549	0,182894	201	-	dangerousity_avg >= 0.103208914

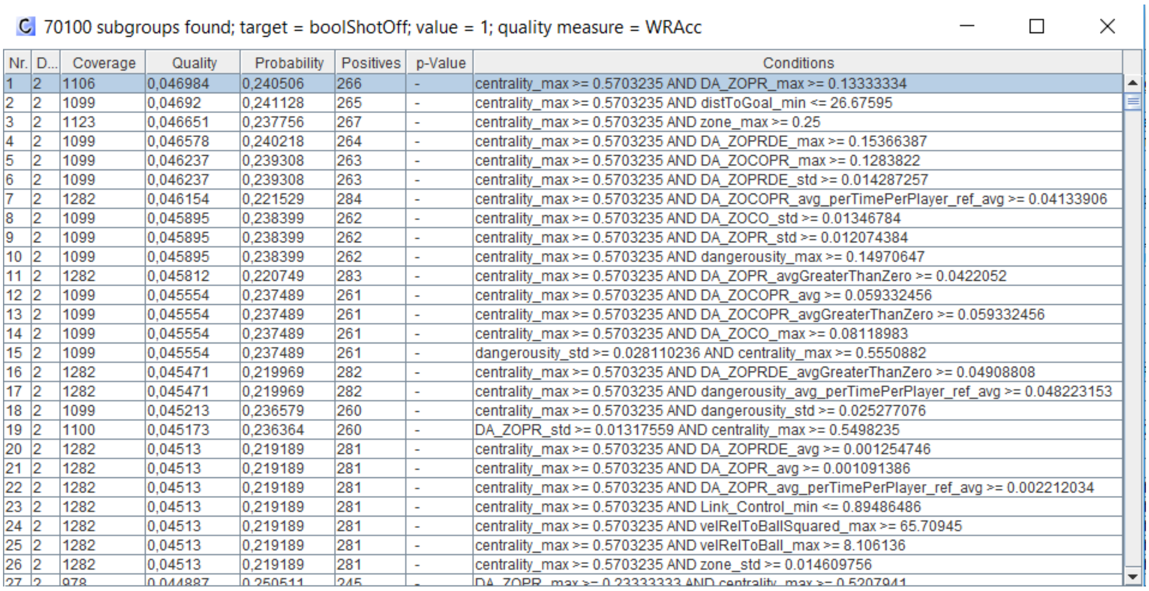
Figure C.3: Cortana output for the target "shot on target" at refinement depth 1.



72207 subgroups found; target = boolShot; value = 0; quality measure = WRAcc

Nr.	...	Coverage	Quality	Probability	Positives	p-Value	Conditions
1	2	1620	0,083364	0,943827	1.529	-	DA_ZOPR_max <= 0.23333333 AND angleToGoal_avg_perTimePerPlayer_ref_avg >= 23.486626
2	2	1620	0,083364	0,943827	1.529	-	DA_ZOPR_max <= 0.23333333 AND centrality_avg_perTimePerPlayer_ref_avg <= 0.6252854
3	2	1620	0,083364	0,943827	1.529	-	DA_ZOPR_max <= 0.23333333 AND centrality_avg_perTimePerPlayer_ref_minGreaterThanZero <= 0.5718791
4	2	1603	0,083187	0,945103	1.515	-	DA_ZOCOPR_max <= 0.22861977 AND centrality_avg_perTimePerPlayer_ref_minGreaterThanZero <= 0.5718791
5	2	1620	0,083022	0,94321	1.528	-	DA_ZOPR_max <= 0.23333333 AND angleToGoal_avg >= 23.58854
6	2	1620	0,083022	0,94321	1.528	-	DA_ZOPR_max <= 0.23333333 AND centrality_avg <= 0.6245
7	2	1748	0,082846	0,931922	1.629	-	DA_ZOPR_max <= 0.23333333 AND Link_PDPlayerWithBall_minGreaterThanZero <= 0.8975836
8	2	1748	0,082846	0,931922	1.629	-	DA_ZOPR_max <= 0.23333333 AND majority_min <= 3.0
9	2	1603	0,082846	0,944479	1.514	-	DA_ZOCOPR_max <= 0.22861977 AND angleToGoal_avg_perTimePerPlayer_ref_avg >= 23.488487
10	2	1603	0,082846	0,944479	1.514	-	DA_ZOCOPR_max <= 0.22861977 AND centrality_avg_perTimePerPlayer_ref_avg <= 0.6251364
11	2	1603	0,082504	0,943855	1.513	-	DA_ZOCOPR_max <= 0.22861977 AND angleToGoal_avg >= 23.488487
12	2	1603	0,082504	0,943855	1.513	-	DA_ZOCOPR_max <= 0.22861977 AND centrality_avg <= 0.6245
13	2	1728	0,082457	0,93287	1.612	-	DA_ZOPR_max <= 0.23333333 AND majority_minGreaterThanZero <= 3.0
14	2	1723	0,082445	0,933256	1.608	-	distToGoal_min >= 23.716309 AND Link_PDPlayerWithBall_minGreaterThanZero <= 0.8975836
15	2	1723	0,082445	0,933256	1.608	-	distToGoal_min >= 23.716309 AND majority_min <= 3.0
16	2	1824	0,08241	0,925439	1.688	-	DA_ZOCOPR_max <= 0.22861977 AND Link_SDPlayerWithBall_avg_perTimePerPlayer_ref_min >= 0.0
17	2	1822	0,082269	0,925357	1.686	-	DA_ZOCOPR_max <= 0.22861977 AND Link_SDPlayerWithBall_min >= 0.0
18	2	1603	0,082163	0,943231	1.512	-	distToGoal_min >= 23.716309 AND angleToGoal_avg >= 24.880177
19	2	1603	0,082163	0,943231	1.512	-	distToGoal_min >= 23.716309 AND centrality_avg_perTimePerPlayer_ref_avg <= 0.6017882
20	2	1830	0,082151	0,92459	1.692	-	DA_ZOCOPR_max <= 0.22861977 AND inPenaltyArea_avg <= 0.0
21	2	1830	0,082151	0,92459	1.692	-	DA_ZOCOPR_max <= 0.22861977 AND inPenaltyArea_avgGreaterThanZero = '0'
22	2	1830	0,082151	0,92459	1.692	-	DA_ZOCOPR_max <= 0.22861977 AND inPenaltyArea_avg_perTimePerPlayer_ref_avg <= 0.0
23	2	1830	0,082151	0,92459	1.692	-	DA_ZOCOPR_max <= 0.22861977 AND inPenaltyArea_avg_perTimePerPlayer_ref_avgGreaterThanZero <= 0.0
24	2	1830	0,082151	0,92459	1.692	-	DA_ZOCOPR_max <= 0.22861977 AND inPenaltyArea_avg_perTimePerPlayer_ref_max <= 0.0
25	2	1830	0,082151	0,92459	1.692	-	DA_ZOCOPR_max <= 0.22861977 AND inPenaltyArea_avg_perTimePerPlayer_ref_minGreaterThanZero <= 0.0
26	2	1830	0,082151	0,92459	1.692	-	DA_ZOCOPR_max <= 0.22861977 AND inPenaltyArea_avg_perTimePerPlayer_ref_std <= 0.0
27	2	1830	0,082151	0,92459	1.692	-	DA_ZOCOPR_max <= 0.22861977 AND inPenaltyArea_max = '0'

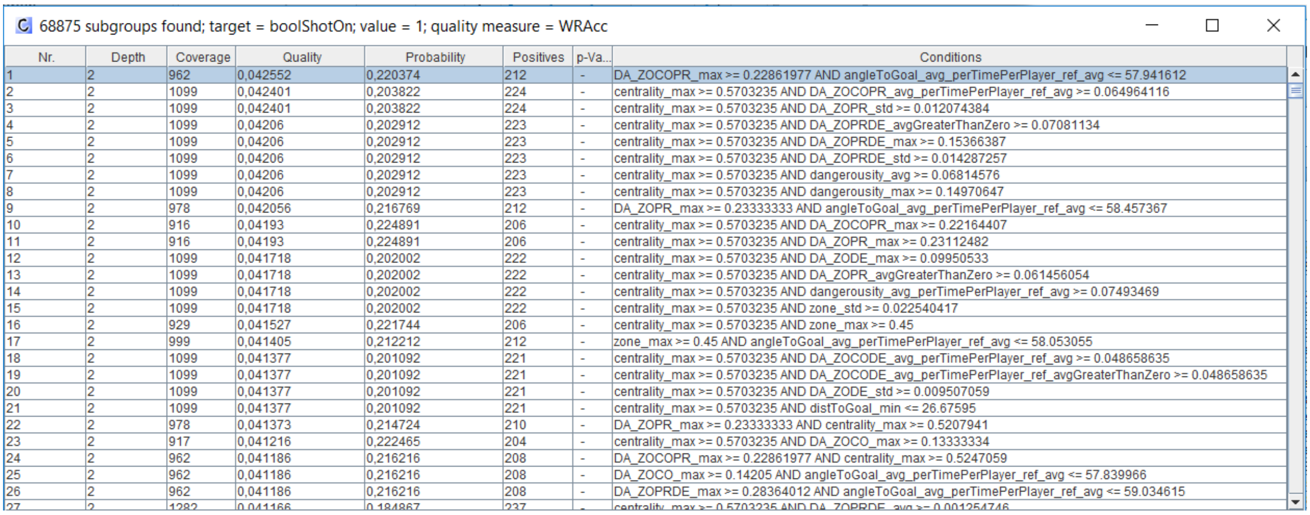
Figure C.4: Cortana output for the target “no shot“ at refinement depth 2.



70100 subgroups found; target = boolShotOff; value = 1; quality measure = WRAcc

Nr.	D...	Coverage	Quality	Probability	Positives	p-Value	Conditions
1	2	1106	0,046984	0,240506	266	-	centrality_max >= 0.5703235 AND DA_ZOPR_max >= 0.13333334
2	2	1099	0,04692	0,241128	265	-	centrality_max >= 0.5703235 AND distToGoal_min <= 26.67595
3	2	1123	0,046651	0,237756	267	-	centrality_max >= 0.5703235 AND zone_max >= 0.25
4	2	1099	0,046578	0,240218	264	-	centrality_max >= 0.5703235 AND DA_ZOPRDE_max >= 0.15366387
5	2	1099	0,046237	0,239308	263	-	centrality_max >= 0.5703235 AND DA_ZOCOPR_max >= 0.1283822
6	2	1099	0,046237	0,239308	263	-	centrality_max >= 0.5703235 AND DA_ZOPRDE_std >= 0.014287257
7	2	1282	0,046154	0,221529	284	-	centrality_max >= 0.5703235 AND DA_ZOCOPR_avg_perTimePerPlayer_ref_avg >= 0.04133906
8	2	1099	0,045895	0,238399	262	-	centrality_max >= 0.5703235 AND DA_ZOCO_std >= 0.01346784
9	2	1099	0,045895	0,238399	262	-	centrality_max >= 0.5703235 AND DA_ZOPR_std >= 0.012074384
10	2	1099	0,045895	0,238399	262	-	centrality_max >= 0.5703235 AND dangerousity_max >= 0.14970647
11	2	1282	0,045812	0,220749	283	-	centrality_max >= 0.5703235 AND DA_ZOPR_avgGreaterThanZero >= 0.0422052
12	2	1099	0,045554	0,237489	261	-	centrality_max >= 0.5703235 AND DA_ZOCOPR_avg >= 0.059332456
13	2	1099	0,045554	0,237489	261	-	centrality_max >= 0.5703235 AND DA_ZOCOPR_avgGreaterThanZero >= 0.059332456
14	2	1099	0,045554	0,237489	261	-	centrality_max >= 0.5703235 AND DA_ZOCO_max >= 0.08118983
15	2	1099	0,045554	0,237489	261	-	dangerousity_std >= 0.028110236 AND centrality_max >= 0.5550882
16	2	1282	0,045471	0,219969	282	-	centrality_max >= 0.5703235 AND DA_ZOPRDE_avgGreaterThanZero >= 0.04908808
17	2	1282	0,045471	0,219969	282	-	centrality_max >= 0.5703235 AND dangerousity_avg_perTimePerPlayer_ref_avg >= 0.048223153
18	2	1099	0,045213	0,236579	260	-	centrality_max >= 0.5703235 AND dangerousity_std >= 0.025277076
19	2	1100	0,045173	0,236364	260	-	DA_ZOPR_std >= 0.01317559 AND centrality_max >= 0.5498235
20	2	1282	0,04513	0,219189	281	-	centrality_max >= 0.5703235 AND DA_ZOPRDE_avg >= 0.001254746
21	2	1282	0,04513	0,219189	281	-	centrality_max >= 0.5703235 AND DA_ZOPR_avg >= 0.001091386
22	2	1282	0,04513	0,219189	281	-	centrality_max >= 0.5703235 AND DA_ZOPR_avg_perTimePerPlayer_ref_avg >= 0.002212034
23	2	1282	0,04513	0,219189	281	-	centrality_max >= 0.5703235 AND Link_Control_min <= 0.89486486
24	2	1282	0,04513	0,219189	281	-	centrality_max >= 0.5703235 AND velRelToBallSquared_max >= 65.70945
25	2	1282	0,04513	0,219189	281	-	centrality_max >= 0.5703235 AND velRelToBall_max >= 8.106136
26	2	1282	0,04513	0,219189	281	-	centrality_max >= 0.5703235 AND zone_std >= 0.014609756
27	2	978	0,044887	0,250511	245	-	DA_ZOPR_max >= 0.23333333 AND centrality_max >= 0.5207941

Figure C.5: Cortana output for the target “shot off target“ at refinement depth 2.


68875 subgroups found; target = boolShotOn; value = 1; quality measure = WRAcc

Nr.	Depth	Coverage	Quality	Probability	Positives	p-Va...	Conditions
1	2	962	0,042552	0,220374	212	-	DA_ZOCOPR_max >= 0.22861977 AND angleToGoal_avg_perTimePerPlayer_ref_avg <= 57.941612
2	2	1099	0,042401	0,203822	224	-	centrality_max >= 0.5703235 AND DA_ZOCOPR_avg_perTimePerPlayer_ref_avg >= 0.064964116
3	2	1099	0,042401	0,203822	224	-	centrality_max >= 0.5703235 AND DA_ZOPR_std >= 0.012074384
4	2	1099	0,04206	0,202912	223	-	centrality_max >= 0.5703235 AND DA_ZOPRDE_avgGreaterThanZero >= 0.07081134
5	2	1099	0,04206	0,202912	223	-	centrality_max >= 0.5703235 AND DA_ZOPRDE_max >= 0.15366387
6	2	1099	0,04206	0,202912	223	-	centrality_max >= 0.5703235 AND DA_ZOPRDE_std >= 0.014287257
7	2	1099	0,04206	0,202912	223	-	centrality_max >= 0.5703235 AND dangerousity_avg >= 0.06814576
8	2	1099	0,04206	0,202912	223	-	centrality_max >= 0.5703235 AND dangerousity_max >= 0.14970647
9	2	978	0,042056	0,216769	212	-	DA_ZOPR_max >= 0.23333333 AND angleToGoal_avg_perTimePerPlayer_ref_avg <= 58.457367
10	2	916	0,04193	0,224891	206	-	centrality_max >= 0.5703235 AND DA_ZOCOPR_max >= 0.22164407
11	2	916	0,04193	0,224891	206	-	centrality_max >= 0.5703235 AND DA_ZOPR_max >= 0.23112482
12	2	1099	0,041718	0,202002	222	-	centrality_max >= 0.5703235 AND DA_ZODE_max >= 0.09950533
13	2	1099	0,041718	0,202002	222	-	centrality_max >= 0.5703235 AND DA_ZOPR_avgGreaterThanZero >= 0.061456054
14	2	1099	0,041718	0,202002	222	-	centrality_max >= 0.5703235 AND dangerousity_avg_perTimePerPlayer_ref_avg >= 0.07493469
15	2	1099	0,041718	0,202002	222	-	centrality_max >= 0.5703235 AND zone_std >= 0.022540417
16	2	929	0,041527	0,221744	206	-	centrality_max >= 0.5703235 AND zone_max >= 0.45
17	2	999	0,041405	0,212212	212	-	zone_max >= 0.45 AND angleToGoal_avg_perTimePerPlayer_ref_avg <= 58.053055
18	2	1099	0,041377	0,201092	221	-	centrality_max >= 0.5703235 AND DA_ZOCODE_avg_perTimePerPlayer_ref_avg >= 0.048658635
19	2	1099	0,041377	0,201092	221	-	centrality_max >= 0.5703235 AND DA_ZOCODE_avg_perTimePerPlayer_ref_avgGreaterThanZero >= 0.048658635
20	2	1099	0,041377	0,201092	221	-	centrality_max >= 0.5703235 AND DA_ZODE_std >= 0.009507059
21	2	1099	0,041377	0,201092	221	-	centrality_max >= 0.5703235 AND distToGoal_min <= 26.67595
22	2	978	0,041373	0,214724	210	-	DA_ZOPR_max >= 0.23333333 AND centrality_max >= 0.5207941
23	2	917	0,041216	0,222465	204	-	centrality_max >= 0.5703235 AND DA_ZOCO_max >= 0.13333334
24	2	962	0,041186	0,216216	208	-	DA_ZOCOPR_max >= 0.22861977 AND centrality_max >= 0.5247059
25	2	962	0,041186	0,216216	208	-	DA_ZOCO_max >= 0.14205 AND angleToGoal_avg_perTimePerPlayer_ref_avg <= 57.839966
26	2	962	0,041186	0,216216	208	-	DA_ZOPRDE_max >= 0.28364012 AND angleToGoal_avg_perTimePerPlayer_ref_avg <= 59.034615
27	2	1282	0,041166	0,184867	237	-	centrality_max >= 0.5703235 AND DA_ZOPRDE_avg >= 0.001254746

Figure C.6: Cortana output for the target “shot on target“ at refinement depth 2.