UNIVERSITY OF TWENTE.

A data based scouting support system for Club Atlético San Lorenzo de Almagro

MSC thesis IEM UTwente
J.A. van Dongen – s1990624

Supervisors University of Twente:
Dr. G.W.J. Bruinsma
M. Koot MSc.

Supervisors San Lorenzo:
A. R. Zamora Patacchiola
A. E. González Demori

July 2022

Research information

Thesis title:	A data based scouting support system for Club Atlético San Lorenzo de Almagro
Author:	Joannes Antonius van Dongen
First supervisor:	Dr. G.W.J. Bruinsma
Second supervisor:	M. Koot MSc.
Company:	Club Atlético San Lorenzo de Almagro – Buenos Aires, Argentina
External supervisors:	A. R. Zamora Patacchiola
	A. E. González Demori
Date:	July 2022
Place:	University of Twente, Enschede, The Netherlands
Faculty:	Behavioural management and social sciences
Programme:	Industrial Engineering and Management – Master

This version includes pseudo-anonymization of players

Preface

This report is the result of the master thesis conducted to complete the master Industrial Engineering and Management at the University of Twente. The thesis is conducted at Club Atlético San Lorenzo de Almagro (in short: San Lorenzo), with supervision from both San Lorenzo and the University of Twente. To provide value to San Lorenzo, this thesis aims to provide a tool to support the scouting process at San Lorenzo.

First, I would like to thank my supervisors from San Lorenzo. The opportunity to conduct my thesis in South America was a dream come true and I have more than enjoyed this experience, going on an adventure and having some different kind of learning experiences as well! Because of my interests in, and the importance of data, together with a lifelong passion for football, it is interesting to work on a data related scouting project for a football club like San Lorenzo. I have had bi-weekly discussions with both supervisors, and many meetings with Andrés Zamora during my time in Buenos Aires. In these meetings I was always able to gather useful feedback and input to the thesis, while having friendly talks and being introduced to Buenos Aires and the (people at the) football club. It was an awesome experience which was made even better because of them.

Secondly, I want to thank my supervisors from the University of Twente, Guido Bruinsma and Martijn Koot. In our meetings, I have learned a lot about writing an academic thesis and being critical on my own work. As my first supervisor, Guido was also very helpful in structuring the plan for the thesis and was supportive in my plans to go abroad and experience an adventurous final part of my studies. The feedback from the meetings was always helpful to progress within the thesis and to guide me to the conclusion of the thesis. I have always enjoyed the positive spirit and friendly setting of the meetings with them.

Finally, I would like to thank my family and friends for their support throughout the execution of my thesis. I was always able to discuss my ideas and they supported me in my road towards the end of my studies.

Jan van Dongen

July 2022

Management Summary

Introduction and problem

San Lorenzo is an Argentinean first-division football club based in Buenos Aires. The scouting team is important in improving the squad and their performance, with a specific focus on scouting young talented players that they can sell on for a profit after a couple of years.

Improvements can be made within the scouting process, where a large part of the shortlisted players do not fit the requirements to replace a previous player. Through identification of similarity measures of players, the fit of shortlisted players is expected to be improved.

The problem described above is the result of multiple problems. The lack of data-driven analysis with technically founded results on the similarity between a to be scouted and a to be replaced player in the squad to help in the replacement of players at San Lorenzo, is the core problem identified and leads to the following main research question:

How can player data be used to derive similarity information for specific player positions?

Problem approach

The main research question is divided in different sections which all help to finally find an answer within the research. The CRISP-DM method is a six-phase process model, as shown in Figure M.1, with common approaches for data mining projects. As the core problem and research objective of the thesis are related with data mining, the CRISP-DM method is applied. The CRISP-DM approach is conducted to follow the different steps within data-mining research, related with Figure M.1, as summed up below.

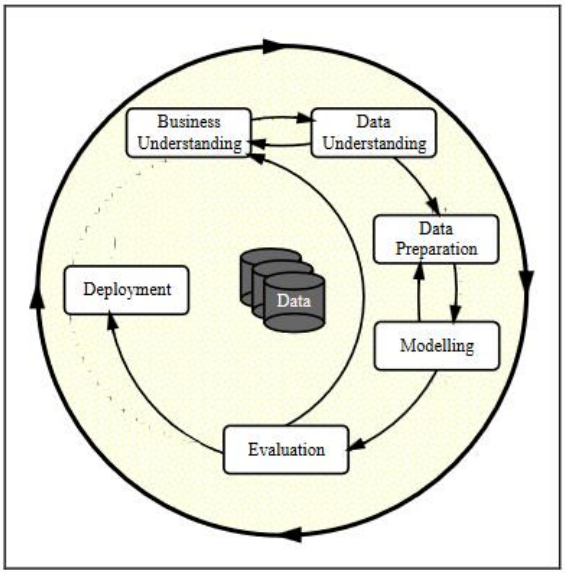

Figure M.1: Phases of the CRISP-DM method (Wirth & Hipp, 2000).

- A context analysis is done on San Lorenzo to get a general idea of the business and the problem to be solved. This is the first stage of the CRISP-DM method.
- Literature is studied to form a foundation for applying clustering and dissimilarity models in the to be designed method. These models use data to derive similarities between players and cluster players within groups.
- The CRISP-DM methodology is further specified and redesigned to include the relevant models found in literature. The redesigned method is specified for the objectives of the thesis and result in the development of a tool to shortlist potentially to be scouted players.

- The second main step in CRISP-DM is the understanding and preparation of data, which is the first applied phase of the redesigned method. An exploratory data analysis identifies the characteristics of the dataset and forms a foundation for the next steps in the redesigned method.
- After understanding of the data, the tool is designed to derive the similar players per position through the modelling of the similarity and clustering algorithms. The similarities between scouted and to be replaced players should help to solve the problem attacked in this thesis.
- The resulting tool is evaluated by the main scouting stakeholder, based on usability and value, to find results of the research. The usability and value provided by the tool is identified and gives a foundation to draw conclusions and recommendations from the research.

The designed method results

Based on the CRISP-DM approach and the clustering and dissimilarity methods identified, a new method was redesigned to solve the action problem.

The following steps are taken through applying this method:

- The football player dataset is collected and understood through an exploratory data analysis. This helps to identify relationships between football player attributes and different inputs to the clustering and dissimilarity models to be applied, such as the weights and importance of attributes for a specific player position group.
- To prepare the dataset for clustering and dissimilarity models, the data is pre-processed. This includes cleaning of the data, feature engineering, as well as transformation, standardisation, and selection of the data.
- The dissimilarity metric is designed based on the type of attributes in the final dataset. This can also be input to the clustering model, of which some require dissimilarity metrics to decide the clusters. The metric is based on the weighted Manhattan distance, which is chosen because it is one of the most widely used measures for numerical attributes, it is effective in high-dimensional data, and because of its integrated use within clustering algorithms in Python. Weighting of attributes is chosen because of the, on average, higher and thus better silhouette coefficient of -0.087 versus -0.204 in case no weights are used.
- The clustering method is chosen to decide the clusters in use of the designed tool. The most applicable and feasible methods are further analysed and the best fitting clustering model for the type of dataset is resulted. This is concluded as the GMC clustering model, based on the average and standard deviation of the silhouette coefficient of the clustering models as shown in Table M.1.
- With the use of the within-cluster sum of square, the quality of the applied clustering method can be evaluated for different numbers of clusters as input to the model. The optimal number of clusters is decided with the use of the elbow method and is saved as input to the clustering model.
- The clustering model is made with the final dataset and all the input resulted from the previous steps. The clustering models are evaluated with the silhouette coefficient and the best clustering method is decided.
- With the implemented clustering models, the tool can be completed. With the use of front-end and backend development packages, an interface is built to implement the clustering models and acquire the results. Results are acquired based on different examples. This resulted in the tool design as shown in Figure M.2. An example of the use of the tool is indicated with the possible buy of player 7 at the end of season 2018-2019, with his increasing value and thus good fit within the transfer philosophy of San Lorenzo.

- The tool is finally evaluated as a basis for possible future improvements, conclusions on the thesis as well as recommendations to the football club.

Table M.1: Summarization of silhouette scores per clustering model.

Clustering model	Cure	DBSCAN	GMC
Average silhouette score	-0.17686	-0.12362	-0.01467
Average standard deviation	0.1988	0.111653	0.049318

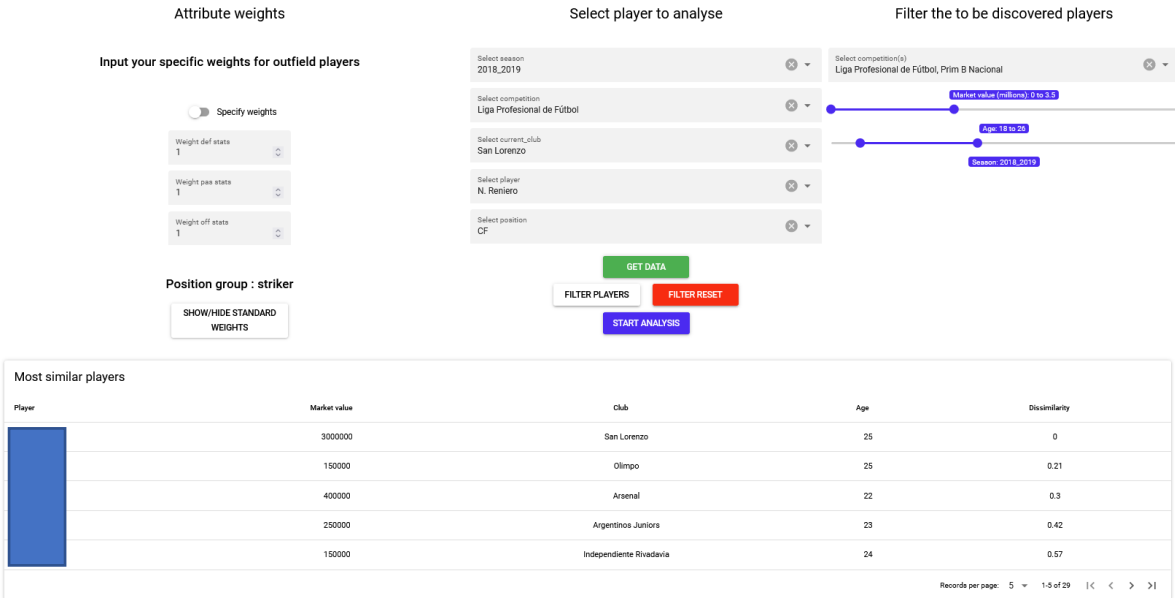

Figure M.2: Design of the developed tool including the clustering and dissimilarity algorithms.

Recommendations from applying the method

Through applying the method, recommendations can be made for the use of the tool by San Lorenzo to improve the probability of a to be scouted players to replace a player in the team.

The tool can be used by the scouting team to get a quick and easy overview of players that are similar to a to be replaced player as resulted from the stakeholder evaluation, because of the resulting figures as shown in Figure M.2. Next to the expected increase in probability that a scouted player will be able to replace the to be replaced player, the efficiency of the scouting process is expected to increase. This is expected as the results of the tool already provide a shortlist of potentially interesting players, which they can focus on within the scouting process instead of needing to go through all matches to find potentially interesting players.

The filter options available in the tool are recommended to be used to conform the transfer philosophy of San Lorenzo, which is based on mostly finding young talented players for cheap that can be sold for a profit. The use of this filter option combined with an analysis on the clustering results in increased knowledge on young talented players (see Figure M.3), which can also be identified through finding cheap players that are similar to more expensive players in terms of statistics.

The tool deployment, which is out of the scope of this thesis, is suggested as a next step for San Lorenzo to allow for quick and integrated use of the tool within the scouting process.

Currently, the weighting factors are rounded to halves, which can be further specified in future research to optimize the reflected influence of attributes on the classification of the players. With this, the theoretical background can be further improved, and the results will be more specific and focused.

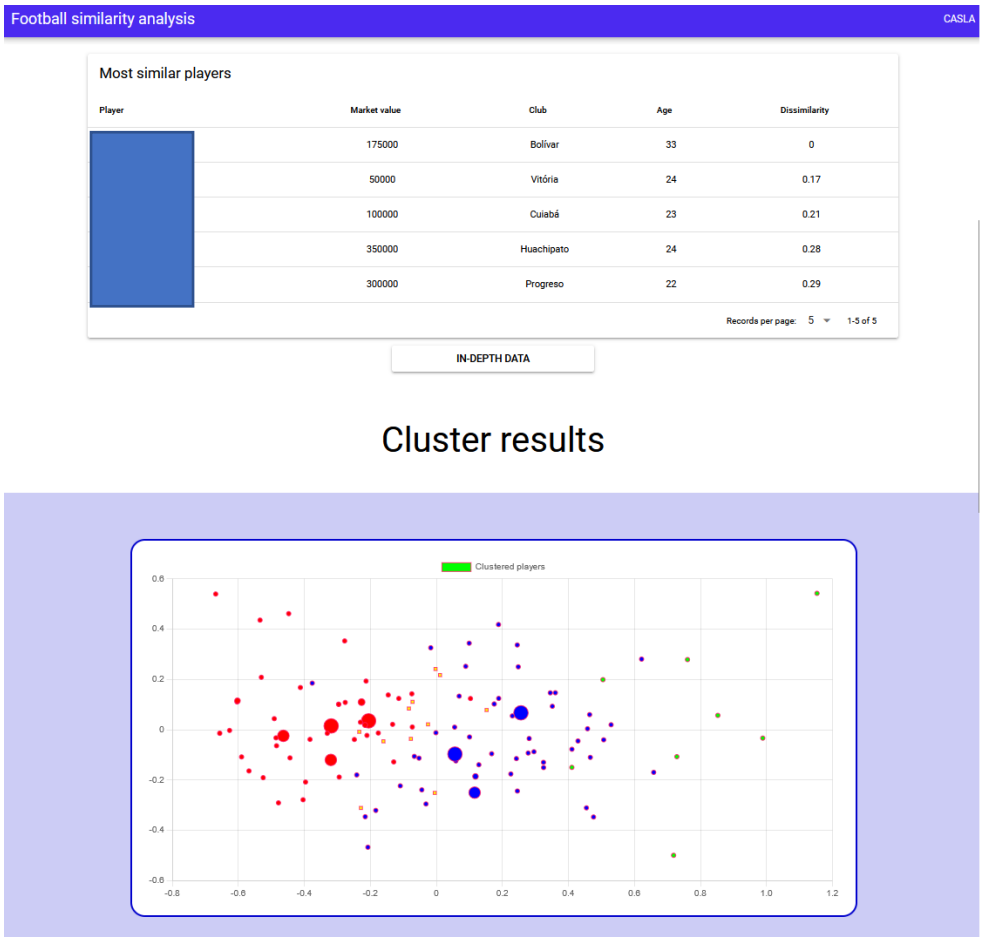

Figure M.3: Developed tool bottom part including the clustering results.

Reader's guide

To provide a clear overview of the structure in this thesis, a reader's guide is given. The content of the chapters apparent in the thesis is explained.

Chapter 1:
The first chapter introduces the reader to this thesis. A background to San Lorenzo is given through a context analysis, and the identification of the main problem is discussed. The problem approach is provided with the steps to reach the objectives of the thesis to solve the main problem.

Chapter 2:
The second chapter contains the literature study that is done to identify methods to find similarity information between to be replaced and to be scouted players. The relevancy of these methods is discussed within each section.

Chapter 3:
In the third chapter, the methods resulting from the literature study are combined with the general data mining research approach into a method designed for application for the scouting team at San Lorenzo. Requirements for designing an applicable method for San Lorenzo are outlined. Nine stages are passed before showing results of the tool.

Chapter 4:
The fourth chapter includes the application of the designed method in Chapter 3. The stages are followed, and the research process is described in detail.

Chapter 5:
This chapter includes the qualitative evaluation of the tool designed in Chapter 4. In cooperation with the stakeholders at San Lorenzo, the added value from the tool in solving the problem is evaluated.

Chapter 6:
Chapter 6 is the final chapter of this thesis. In this chapter, the conclusions and recommendations based on the applied method are given. The results and relevancy of the thesis is evaluated. Suggestions for future research are given, as well as a discussion on the shortcomings of the research.

The text contains references to certain sections. On a device, these references can be clicked to jump to this section.

Table of Contents

1 Introduction

This master thesis assignment is conducted at San Lorenzo de Almagro, in Buenos Aires, Argentina. The research project analyses the scouting systems currently in place at the football club, where the sub-optimal data driven approach limits the possibilities for the scouting department. Chapter 1.1 introduces the reader to the Football Club, Chapter 1.2 goes more in-depth on the context of the football club, Chapter 1.3 provides information about the problem resulting from an analysis of the football club, Chapter 1.4 gives an overview of the research, and Chapter 1.5 concludes the introduction of the thesis.

1.1 Background information

Club Atlético San Lorenzo de Almagro is a football club located in Buenos Aires, Argentina. The football club plays in the first division of the Argentine football leagues, called the Primera División. The 'big five' is the name of the five biggest football clubs in Argentina, including San Lorenzo. The football club was founded in 1908 and started playing in the first division of Argentine football in 1915. The first title arrived in 1923, which was the start of the glory of the football club. Many titles have been won since, including 15 Primera División titles and a Copa Libertadores. The club has seen many famous players, including Pablo Zabaleta, Ezequiel Lavezzi, and Iván Córdoba. (Meh, 2022)

At San Lorenzo, football players are scouted through traditional scouting as well as data driven scouting. For the data-driven scouting, tools such as 'Wyscout' are used. WyScout is a tool that offers datasets of players all over the world, including the South American competitions. The purpose of this research is to add value to the scouting department of San Lorenzo and help them improving their squad. Currently, troubles can arise when players need to be replaced by a new to be scouted player. Identification of the similarities between players can help to improve the assessment of players. With the current scouting process, the similarities between players are based on domain knowledge and watching players only. This is subjective. The lack of an objective similarity analysis can possibly be approached with data-driven distance measures or clustering methods as further explained in the thesis.

1.2 Context analysis

This section provides a context to the football club and gives a foundation for the thesis assignment. Altogether, this should give a good understanding of the current scouting process, the club, and the possible areas for improvement.

First, we dive deeper into the transfer philosophy and give a general background to the club's objectives and structure. Next, we analyse the scouting process at San Lorenzo. This includes traditional scouting, and data-driven scouting. The scouting results are then explained with possible improvements to be made through this research.

1.2.1 Transfer philosophy and background

1.2.1.1 The Argentinian league

Club Atlético San Lorenzo de Almagro plays in the Argentine first division called 'The Primera División'. The league works with two tournaments (opening and final stages) or parts of the season, which means that every year, two champions are crowned. Based on the average points collected within these two tournaments, the clubs are ranked to determine the relegated teams.

Even though Argentina is a big football country, the league is not the most competitive in the world. Generally, the best players in the Argentine league will transfer to the top competitions in Europe, such as the Premier League (England) or La Liga (Spain). Because of this, football clubs in South

America, including football clubs in Argentina, generally have a transfer philosophy of scouting young talents to later sell them to bigger leagues or clubs for a profit.

Argentina was always well represented within the top competitions in Europe. Throughout the last years however, less and less players from Argentina are sold to the top competitions. For example, 47 Argentine players played in the Serie A (Italy) in 2011-2012, which decreased to 24 in 2020-2021. One reason for the decreasing share of Argentine players in top competitions is the difficulty of scouting in the Argentinian league. Scouts have a more difficult time getting information about the players and getting access to games because the clubs are less open to scouts compared to for example Brazil. This does not explain the decreasing number of transfers of Argentine players however, as the accessibility to games did not change significantly. A noticeable difference, however, is the different in increasing quality of the youth academy compared to for example Brazil. Furthermore, scouting is more effective in leagues with a narrow quality spread between the teams. Also, the technology development within the years makes it easier to scout in lesser-known leagues such as the Chilean or Colombian league. (Smith, 2020)

Because of the politics in South American football, there is in general more uncertainty in the football clubs compared to for example European football. The culture of football and politics are related and politics can influence decisions or organization at football clubs which influences the consistency of the club (Hernández, 2018). The club's performance changes more radically because of this and the changes in the club's structure or organisation. For example, from around 2012, San Lorenzo had a couple successful years including winning the Copa Libertadores. Currently, they are closer to battling for relegation in the Primera División.

1.2.1.2 The transfer philosophy of San Lorenzo

San Lorenzo focuses on scouting young talent with the idea of helping them develop and sell them in 2-3 years for a profit. Apart from that, older and more experienced players are scouted to give an immediate quality boost to the squad. The club usually cycles through players and teams within 2-3 years, directly related with the scouting projects. Because of this, there is a quicker change in players compared to the changes in tactics of the club. There is usually insufficient money to keep hold of the players, which makes the club and the player part their ways. Because of this, a large part of the team is replaced within 2-3 years, and one-team players are less common.

Most players are scouted within other South American countries and leagues such as Uruguay, Colombia, Paraguay, and Chile. These leagues contain many players with relatively high quality and low transfer value. Other leagues such as the second division of Argentina, or the first division of Venezuela or Bolivia are less important for the scouting department. This is because the lower expected probability of finding potentially to be scouted players.

Within the last years, no signings were made primarily due to the pandemic and the resulting uncertainty. In general, however, San Lorenzo is a relatively big club in the Primera División, with more than average (financial) resources. San Lorenzo is a historically big club, but currently behind clubs such as River Plate and Boca Juniors in terms of financial resources.

1.2.2 The scouting process

San Lorenzo goes in the process of scouting players to try and bolster the quality in their squad, or to find the next talents that can achieve the quality to play for the team. Within the club, the usage of data was limited before 2020. Basic score-box methods were used to get an insight into the player's quality and performance. Score-box statistics gives an idea about the players performances through a visualisation of general player statistics. This includes for example goals, assists, clean sheets, fouls,

etc. The more advanced statistics, such as expected goals or expected assists, are not considered within these methods. These statistics are advanced as they are based on data science methods, and not necessarily easily interpretable. The five-step scouting process at San Lorenzo is explained below, where the organizational chart is given in Figure 1.1.

1. *Shortlisting players*
   First, a shortlist is created of players for each position that have or could have the quality to play for San Lorenzo. Players are shortlisted based on references from talking with other scouts, watching football matches, or analysing box-score statistics of players in interesting leagues. This is done by the scouting team (ST). Halfway through 2020, San Lorenzo started using analytics within the scouting process. The analytics are primarily used for shortlisting the players in the first step of the scouting process. Through a scouting recommender system, players are already identified and shortlisted. Data analytics rely on statistics and mathematical transformations to identify players that could provide value for the football club in terms of future market value and current team performances.
2. *Compile reports of players*
   The next step is to compile draft reports of the players included within the shortlist. The scouting team watches games of the players, possibly with the help of tools such as WyScout, and makes an analysis of these players. An overview of the positive and negative football characteristics is presented in this report, in terms of technical and mental abilities. This can for example be the good passing distribution (positive, technical) of the player, or the inability to keep their cool after being tackled heavily (negative, mental). Based on the club's criteria, certain players can already be dropped from the shortlist within this step. These are preliminary reports resulting from scouts watching videos of the players for several game weeks.
3. *Analysis of reports*
   The Chief Scout analyses the draft reports presented to him/her and decides together with the scouting team whether to continue monitoring the player in question. This is done through further analysis of the current requirements of the football club, and the potential value of the player in helping to achieve those requirements. This can, for example, be focused on performance or financial requirements.
4. *Compile detailed reports of final players*
   The previous step will have narrowed down the shortlist of players. For the remaining players that passed the criteria and fit to the requirements as denoted in steps 2 and 3, more detailed reports will be made. Within this step, a data-driven application is planned to be used in the future that quantifies these reports and rates the players to make a database of the scouted players. This would mean that the characteristics on which the players are analysed, would be identified based on the previously made reports and the positive or negative characteristics combined will form a rating for the players. This is expected to be done through adding up all positive and negative points of a player and comparing this to a standard to detect the effectiveness and quality of the player.
5. *Final analysis of decision maker*
   The detailed reports are presented to top-level personnel of the football club; the Technical Director (TD) and the Board of Directors (BD). These are the final decision makers, who will decide to physically go to a game of the player in question, to identify possible characteristics that are more difficult to see on the screen and decide whether to initiate contacts to try and sign the player.

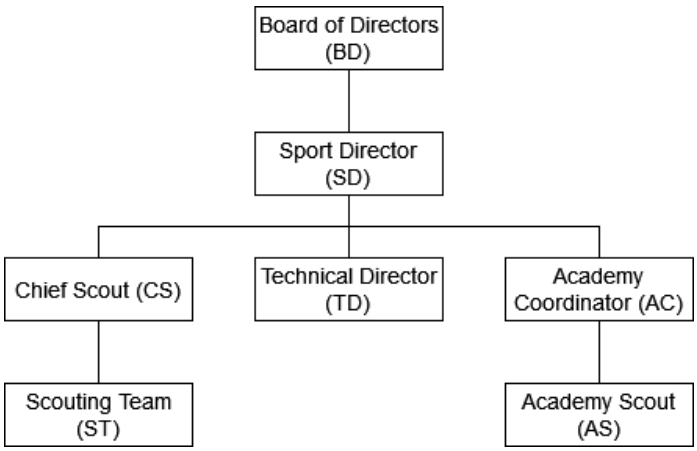

Figure 1.1: Organizational chart San Lorenzo

For the data analytics, data is used from players in the to be scouted leagues. General information is included in the data, such as their age, position, current club, etc. Also, football specific attributes are included, such as (expected) goals scored, assists, and percentage of defensive duels won. The dataset will be further explained in the method design in Chapter 3.

1.2.3 Scouting results

Following the steps described above in the scouting process, players can potentially be signed to the football club. Players are initially shortlisted through different methods, of which data analytics is one. With the data analytics methods, multiple players were identified. Mainly due to financial constraints, no signings were made recently. What can be seen however, is that these shortlisted players identified through data analytics were signed by Mexican and Spanish First Division football clubs. This could give an indication about the performance of the data analytics, as the identified players are signed to relatively high-quality football clubs. The potential of data analytics is seen, which is why a focus on further data-driven player analysis is expected to help in the scouting process. The current measurement of the scouting results and impact of scouted players is limited beyond subjective assessment, however. To measure the objectives for the scouting department and to indicate the performance of the data analytics and scouted players, currently only exploratory methods are used. Players can be continually assessed and monitored to measure the results. This, however, is expensive resource wise.

1.3 Problem identification

At San Lorenzo de Almagro, improvements can be made within the scouting department. This section outlines the action problem, gives a context to this problem, and identifies the core problem to be solved within the thesis.

1.3.1 Identification of the action problem

Player scouting is important for San Lorenzo. To improve the team and beat the competition, players are needed that have specific abilities for a specific position in the field. Players need to fit into a manager's approach. The player needs to have the ability to play a specific position within the squad, like the player that is to be replaced, considering their type of playstyle. This is important because of the quick cycles in players playing for the club. There is a need for identification of players that are similar to players that need to be replaced, to keep the possibility to play with similar tactics. In case tactics are to be changed, players that are deemed fitting for such a playstyle can again be analysed to see which similar (possibly more fitting and affordable) players are interesting for that playstyle. Also, because of the large number of players to be analysed or considered, the process is inefficient.

A standardized way of analysing players more effectively needs to be developed such that new players can repeatedly be discovered. Finally, the to be replaced player will have certain tasks on the pitch as demanded by the coaching staff, these need to be fulfilled as well.

Even though players are scouted partly with a data-driven approach at San Lorenzo, the similarity of these players with the current squad is unknown. Currently, players that are scouted by San Lorenzo generally will have the ability to play for the club but might find difficulties replacing the previous player and conform the squad roles for that position. Furthermore, the scouting process can be improved with a more accurate shortlisting of players. Because of the lack of knowledge about these players specific to the system of the San Lorenzo squad, there is a lower expected probability that a player will fit the system. This is an important issue, as the quick cycle through the players entails there are many changes in players while keeping similar tactics.

The action problem
A large part of the initial shortlist of players analysed by the scouting department of San Lorenzo are not deemed sufficient or have troubles replacing the previous player in a certain position.

The problem of scouted players not being able to replace a certain position, is a problem owned by the scouting department of San Lorenzo de Almagro. A solution needs to be found to achieve the norm of having a higher player scouting efficacy compared to the current situation.

1.3.2 Problem cluster and motivation of core problem

To come to the root of the problem at the scouting department of San Lorenzo de Almagro, the causes for the action problem given above, are identified, and related with each other in this section. The problems will be shortly described first, after which a problem cluster is made to visualize the problems leading to the action problem.

1. **Scouted players incapable of replacing the previous player**
   This is related to the action problem as given above. It is caused by the dissimilarities between the to be scouted and the to be replaced player (problem 3), and the lack of focus on the specific players that would be able to fit the team (problem 2).
2. **Insufficient focus on specific players that would fit the team**
   The scouting department needs to decide on which player would be able to fit the team and be the required improvement or replacement to the squad. However, because of the lack of information on similarities between players (problem 4), there is insufficient focus on similar players in this decision.
3. **Scouted players are not always sufficiently similar to replace a certain position**
   Scouted players generally will have the ability to play in a team, however, fitting in the squad and being similar to the to be replaced player could be more of an issue. This is caused by a lack of information about the similarities between the to be scouted and to be replaced player (problem 4).
4. **Lack of player similarity information**
   To determine whether a scouted player will be able to replace a certain position in the squad, information about the similarity of that player with the to be replaced player, is required. This is currently lacking, caused by insufficient well-grounded information coming from traditional scouting and a lack of data-driven similarity analysis (problems 5 and 6).
5. **Traditional scouting does not provide well-grounded similarity information**
   In traditional scouting, the ability of the player is deeply analysed. However, the information on the similarities between players is not sufficient. Scouts do compare players and identify

similar players; however, this is not well-grounded. This is caused by the bias of the scout towards the players similarity (problem 7).

6. **No data-driven player similarity analysis and technically founded results**
   Currently on the data side of player analysis, tools are used to analyse to be scouted players with a data driven approach. This approach does not focus on the similarities between the to be scouted and to be replaced player, however. This should be included in a non-biased and technically founded way to derive player similarity information.
7. **The similarity information in traditional scouting comes from an opinion and is biased.**
   In traditional scouting, players are watched, and an opinion is formed. Based on style of play, physique, character, on-the-ball and off-the-ball movement, the scout will have an opinion about which players would be similar to each other. This opinion, however, will be biased towards the scout's preferences and thus only provides one perspective. This is caused by the difficulty of quantifying similarity related statistics of the players (problem 8).
8. **Certain similarity relevant statistics are difficult to quantify with traditional scouting**
   A traditional scout has a lot of factors to include within his analysis of a player. When watching games, the performance of a player is judged. However, for the similarities between the to be scouted and to be replaced player, statistics can be relevant which cannot be easily quantifiable and thus include a bias. Because of this, it is difficult to extract similarity relevant statistics within traditional scouting.

Problem cluster

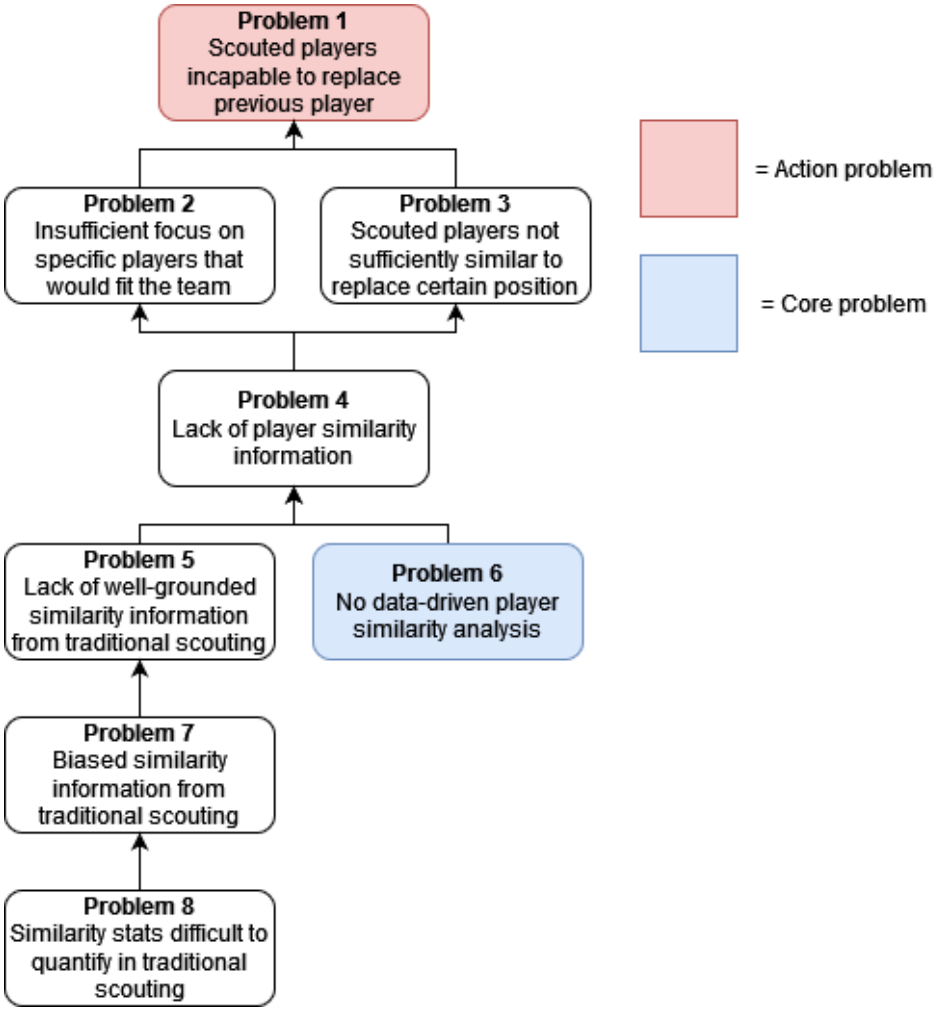

Figure 1.2: Problem cluster at the scouting department of San Lorenzo, arrows denoting the relation between the problems.

Core problem
There is no data-driven analysis with technically founded results on the similarity between a to be scouted and to be replaced player in the squad to help in the replacement of players at San Lorenzo.

Motivation and scope
The trilateral relationship between the management of the football club, the squad and the staff are important for the success of the club. The departments need to be aligned on their vision and values to achieve success. There are examples of football clubs such as Brentford and Watford that have changed their structure and aligned the organisation with shared visions and values, including the importance of data. Within these clubs, data is used to optimize the search for players that would fit the squad, but also to optimize training forms and improve the efficiency of processes within the club. This has led these clubs to achieve success. (Ernest, 2018)

The goal of a scouting department at a football club is to find players that have the quality to play for the club (soon). To determine whether a player will have the required quality based on the club's transfer philosophy, the scouting department analysed multiple aspects of the game. The similarities between the to be scouted and to be replaced player is one of these aspects. With the rise of the importance of data, data scouting in football is also becoming increasingly important ("The Growing Importance of Football Analytics," 2021). Through analysis of the data, it is possible to find similarities between players, and get insight into how similar a player will be to replace one of the current players in the squad. The similarities between the to be scouted and to be replaced player is currently not extracted from a data scouting perspective. Implementing a tool to derive this information and provide similar players for each position in the current squad, is expected to improve the efficacy of the scouting department at San Lorenzo. Because of the strategy of the club as explained in Chapter 1.2.1.2, the discovered similar players will help to find players that can fit the tactics, such that there is no need for changes in tactics. Similar players can be discovered through analysis of data and understanding of dissimilarities of players. Clustering algorithms and dissimilarity measures can be modelled for this objective, as this is a general approach for this goal (Saxena et al., 2017). The core problem is the result of finding an apparent problem without a direct cause. Furthermore, solving this problem should have effects towards solving the action problem (Heerkens & van Winden, 2017). To limit the scope of the project, the research is constrained to on-the-ball statistics and does not look at mental attributes such as adaptability or chemistry with the squad. With the timeframe available for this project (20 weeks), building the decision support system and making it applicable for the scouting department at San Lorenzo, seems feasible.

1.4 Research
Within this section, an outline is given for the research, including the research problem in Chapter 1.4.1, and the research design with research questions in Chapter 1.4.2.

1.4.1 Research problem
To goal of the research is to solve the core problem. There is an importance of data and the similarity between players, however, the correct approach of using this data needs to be explored. Based on the core problem, we can formulate the main problem statement in the form of a question:

How can player data be used to derive similarity information for specific player positions?

Through answering the above question, it will be possible to build a tool that will give player similarity information when replacing a certain position in the current squad at San Lorenzo. With this tool, the scouting department should then be able to focus their scouting on specific players that would fit the team as required, driven on data. The development of this tool is the objective of the research.

1.4.2 Research design and questions

The main problem statement formulated in Chapter 1.4.1 cannot be solved in one step. To come to a well-founded solution, several steps need to be taken. Answers need to be found to research questions, which will together form a solid foundation to answer the main problem. The research and sub-questions are derived from the main problem statement and will all have their specific share in arriving to a solution. The design is based on the CRISP-DM method, that goes through a data science life cycle such as the intended process in this research.

The CRISP-DM method is a six-phase process model, as shown in Figure 1.3, with common approaches for data mining projects. These phases can be performed in different orders, and backtracking within the phases will generally be necessary. The CRISP-DM method has a cyclic nature, which means that deployment does not mean the end of the project. All phases can have feedback that make it beneficial to reapproach another phase and improve the project altogether. These will be shortly described below. (Wirth & Hipp, 2000)

- **Business understanding**
  First, the projects requirements and goals need to be understood from a business perspective. With this knowledge, the problem can be defined in terms of data mining, with an initial plan to reach the goals.
- **Data understanding**
  Data mining projects require an understanding of the data to find value and answers to the described problems. The quality and problems with the data need to be identified such that it can be prepared for modelling in the next phase. Initial insights into the data can help to form hypotheses or a general idea of the possibilities with the data.
- **Data preparation**
  In this phase, the data is prepared to be ready for use within the modelling tools. Data preparation for example includes selection of data, cleaning of data, or constructing new attributes.
- **Modelling**
  To solve the data mining problem, modelling techniques need to be selected and applied. A model is made to achieve outputs and results that can help with solving the defined data mining problem and achieve the goals of the project.
- **Evaluation**
  The model needs to be evaluated to verify that it meets the business goals setup in the first phase. It is decided whether the results from the data mining models can and are to be used.
- **Deployment**
  The data mining results, and applied models, are usually only useful when the customer or user can use it. The project is to be deployed to make use out of the developed models.

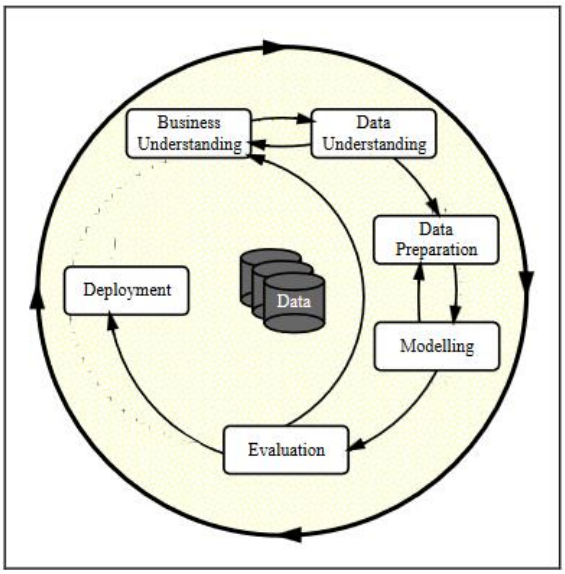

Figure 1.3: Phases of the CRISP-DM method (Wirth & Hipp, 2000).

Sub-question 1

With the context analysis done in this chapter, a general idea of the business and the problem is achieved, as is the objective of the first step of CRISP-DM (business and problem understanding). To complete this step and form a foundation for applying the methods in the to be designed method, a literature analysis is done.

The required models should use data to extract the characteristics of the data and derive how players are similar. As mentioned before, this can be achieved by clustering methods including dissimilarity measures. It is an effective method as data can be grouped to show which datapoints (players in the dataset) have similarities with other players and how they are different to other groups of players. Furthermore, the distance between the players is found which gives a strong general idea on the information we want to derive to give an interesting insight to the scouting team at San Lorenzo. The literature analysis is done in Chapter 2 to identify these methods including their relevancies for the objectives of this thesis.

1. **What similarity measurement and clustering models come forward in literature and can be useful for deriving similarity information for the football player dataset?**

Sub-question 2

The CRISP-DM method should be applied in a way that is specified for the objectives of this thesis. A more specific design of the to be conducted method is presented in Chapter 3. This method is developed with inclusion of the models and clustering approaches coming forward in Chapter 2. In this way, the data mining approach with the CRISP-DM method, is combined with clustering and similarity measurement approaches and models. The design of the decision support system will come forward in Chapter 3. To develop a tool that shortlists potential to be scouted players, the clustering methods need to be applied on the player dataset in an intuitive tool. This method is used to give more specific details on how certain sub-questions are answered and which steps are to be taken to extract and find the information required to reach the objective of this thesis. Chapter 4 and 5 are focused on the execution of this designed method.

2. **How can the CRISP-DM methodology be further specified and designed for application on the football scouting dataset and objectives?**

Sub-question 3

The second step in CRISP-DM, after business and problem understanding, is the understanding and preparation of data. The available dataset needs to be explored to form a foundation for relevant output through the to be applied models. Through an exploratory data analysis, the characteristics of the dataset can be identified, and data can be prepared for use in the tool. The analysis of this data is done in Chapter 4.1.1, where the data understanding is the first building block for further execution of the combined methodology. The preparation of the data in Chapter 4.1.3 is focused on pre-processing the data such that it is usable within application of the models identified based on the previous sub-question.

3. **What are the characteristics of the available dataset, and how can this be used and prepared for deriving similarity measures?**
    a. What are requirements for the tool from the stakeholders?

Sub-question 4

The main part of the research will be to design the tool to derive the similar players per position, to give insights into which players can replace other players in the squad. This includes the modelling and implementation of the clustering and dissimilarity algorithms to try and use to get the intended output from the dataset. To derive similar players in a data-driven manner, dissimilarity measures and clustering models can be combined. The methods derived from an analysis of the current state of literature, in Chapter 2, are used to get information of similarity per position and should be implemented to improve the tool based on the similarities between the to be scouted and to be replaced player. These are further designed and applied in Chapter 4.1.4 – 4.1.7.

Descriptive research is done to find the requirements from stakeholders, which need to be met by the final product of the tool.

4. **How can a tool be designed to derive similar players per position?**
    a. How do the relevant similarity measuring and clustering methods from literature fit in this tool?

Sub-question 5

When the tool is built, the tool can be evaluated to find results of the research. This is one of the final steps of the CRISP-DM process, evaluation of the models in Chapter 5. The tool needs to be evaluated to verify if it is useful for the scouting department at San Lorenzo. Through qualitative research, stakeholders will be interviewed on the usability and value provided by the tool.

5. **What is the potential added value of implementing the tool at San Lorenzo?**
    a. How would the tool be implemented for use by the scouting department at San Lorenzo?
    b. Are the results deemed valid and useful by the scouting department?

1.5 Conclusion

The financial situation and recent pandemic result in a lot of uncertainty in South American football clubs such as San Lorenzo. This impacts San Lorenzo's scouting department because of the increased financial constraints. In general, however, the idea in scouting remains the same. Young players with potential should be scouted to be sold in the future for a profit.

To scout young players with potential, a standardized scouting process is carried out. This process includes the shortlisting and monitoring of players, where increasingly detailed information will help in forming an opinion about the player. The final decision makers will take this opinion to decide whether to initiate contacts to sign this player. Data analytics methods support or improve efficacy in different steps within this process.

Similarity information between San Lorenzo's current players and to be scouted players should be identified to improve the scouting results. There is currently no data-driven analysis on the similarity between a to be scouted and a to be replaced player, which is the main problem. A tool should be designed and implemented to derive similarity information and provide similar players for each position in the current squad of San Lorenzo. This is expected to improve the efficacy of the scouting department. To achieve this, a literature study should be done on clustering players followed by the design and implementation of such a decision support system. The next chapters will include these steps, starting with the literature study in Chapter 2, answering the first knowledge question.

2 Literature study: data clustering and dissimilarity

Before we design the method to solve the problem as stated in Chapter 1, the current scientific state of the art data-driven clustering methods and dissimilarity metrics for player scouting need to be analysed.

Clustering football players and using dissimilarity metrics are required in this thesis to identify similar players to the current squad at San Lorenzo. With the use of these methods in the to be designed tool, information on the clusters and dissimilarity between players can be provided. The efficacy of the scouting department can be improved as the scouts are expected to have a more accurate and in-depth shortlist of to be scouted players.

Relevant clustering and dissimilarity concepts are found and studied, with the goal of developing an ensembled method through different similarity analyses in the next chapters. Two literature libraries, Scopus and Web of Science, are used to find the sources and information presented in this chapter. These libraries contain a wide variety of (types of) sources and with keywords focusing on the clustering or dissimilarity measures used in football or sports environments, relevant sources can be found. With keywords combinations of the research topic (clustering or dissimilarity) together with the methodology (method or modelling) and the sports topic (football or sports), relevant sources are found. With the different perspectives, a strong foundation is expected to be built in the design of the decision support system.

In Chapter 2.1, an outline is given for the theoretical framework. This includes the main constructs, the theoretical perspective and the research boundaries. The sub-question given below will be answered through 4 studies described and analysed in Chapter 2.2. Chapter 2.3 concludes this chapter.

> *What similarity measurement and clustering models come forward in literature and can be useful for deriving similarity information for the football player dataset?*

2.1 Theoretical framework

In this section, the theoretical framework for the literature study is set-up. The theoretical perspective of the research should be focused on the core problem identified in Chapter 1.2.2; the lack of data-driven analysis on the similarity between a to be scouted and to be replaced player in the squad.

Methods found in literature also consider player characteristics based on match event data (Bransen & Van Haaren, 2020; Decroos, Bransen, Van Haaren, & Davis, 2018; Maymin, Maymin, & Shen, 2011). For example, the chemistry, the effectiveness of players playing together, between players is assessed. While this is an interesting characteristic to determine in scouting, the current literature focuses on less widely available match event data. Match event data is data including key actions during the match, such as key passes or goals. These actions are then valued related to the final score of the match, and cooperation between players is assessed. The fit of a potentially to be scouted player in the team can then be assessed, or an optimal line-up in terms of chemistry can be determined. Because the data used in this thesis are only aggregate statistics per player, these methods cannot be immediately applied. This is useful for future research in case South American football data becomes more extensive.

Clustering football players considering suitable dissimilarity metrics should form the basis of solving the core problem at the scouting department of San Lorenzo. Dissimilarity metrics are used to measure the dissimilarity between data, where in our case the data will be the on-the-ball attributes of the to be scouted and to be replaced players at San Lorenzo. This can be used in the final decision

support system to aid the scouting department and providing an overview of similar, to be scouted, players for each position in the used formations. The player attribute dissimilarity metrics can be found in literature and will be elaborated on in the following subsection.

Because of the focus on the similarity and grouping of players, this core problem will be approached on the theoretical perspective of data clustering and dissimilarity metrics. In clustering, data is clustered within several groups to segment the data and classify new datapoints. The idea is that the datapoints within the clusters will have high intraclass similarities and low interclass similarities. This means that the groups are distant from each other, but the data within the groups are clustered closely. In short, cluster analysis is based on finding groups in data, mentioned by Kaufman and Rousseeuw (1990). Using this for the objective in this research, will make it possible to find players that are in the same cluster as the to be replaced player, considering the requirements. Furthermore, dissimilarity metrics are used in the clustering methods to identify which data points (players) are least dissimilar and thus could have similar impacts on San Lorenzo. An ensembled method can be developed to provide different perspectives on this issue. In the to be delivered tool, players can be suggested that are in the same clusters as the to be replaced players. However, the dissimilarity metrics can also be used on their own to give a suggested player list of players that are similar based on the metrics.

The above theoretical perspective is used in this research to find theories and models that are relevant for solving the core problem. For San Lorenzo, this will eventually be most relevant as the tool based on these models will help solving the action problem. Furthermore, the proposed methods can be relevant for further research in this domain.

2.2 Clustering and dissimilarity measurement methods

In this subsection, methods derived from the literature study are given. Through literature search on clustering and dissimilarity methods, preferably related with football player datasets, methods for clustering and applicable dissimilarity metrics are found. The four articles are elaborated upon in different subsections to segregate the differences in possible approaches. This makes it possible to look back into these articles and find the relevance and use from these approaches within the final design of the method in the next chapter. Within these methods, useful methodologies and functions will be given to form a theoretical foundation for building the scouting decision support system for San Lorenzo.

Multiple methods will be analysed in this section. These methods are all expected to provide different results. Within the article analyses, these methods will be analysed and the relevancy for the objectives of the thesis is identified. Based on the relevancy, the methods will be further discussed and prepared for use in the method design in Chapter 3.

Clustering is a technique to find groups in data. Numerical methods are used to group objects into partitions based on dissimilarity metrics as described by Kaufman and Rousseeuw (1990). There are many different methods that can be used for clustering data, all with their specific characteristics and use cases. To get the right clustering results for the objective of the thesis, a seven step guide can be followed from Milligan (1996). The steps are as follows:

1. Collecting data
2. Choosing variables used for clustering
3. Pre-processing data
4. Designing the dissimilarity metric
5. Choosing the clustering method

6. Determining the number of clusters for the method
7. Implementing the clustering method and acquiring results

The first three steps are mostly based on subject knowledge and will be carried out in Chapter 4. Within this literature study, we mostly focus on finding literature on steps 4 and 5, to be able to design the most suitable method for the objective of this thesis.

2.2.1 Dissimilarity metrics (Akhanli, 2019)

Part of the clustering steps is the (design of) the dissimilarity metric. Within this section, the resulting dissimilarity of values from different types of variables is discussed, as is required to be applied on the football dataset in the application of the clustering models explained in the next sections. These results are used to form a dissimilarity metric that includes all relevant attributes coming forward in the final dataset to be used within the clustering models.

Dissimilarity metrics specific for soccer players and their specific attributes are analysed by Akhanli (2019). Dissimilarity or distance metrics $d(x, y)$ indicate the distance between players x and y and is used for clustering players in the clustering method. In this way, the (statistical) difference of the players can be identified. This is an illustrative example based on 2D datapoints, where distance can also be measured in higher dimensions. The similarity between players can also be calculated with a similarity measure $s(x, y)$. The dissimilarity and similarity measures are complementary and defined given in Equation 2.1. The measures are now generally defined between 0 and 1.

$$d(x, y) = 1 - s(x, y) \tag{2.1}$$

As shown in Equation 2.2, a distance matrix can be constructed, including the distances between all datapoints (d_{xy} , where x and y are two datapoints).

$$M_d(D) = \begin{bmatrix} d_{11} = 0 & \cdots & d_{1n} \\ \vdots & \ddots & \vdots \\ d_{n1} & \cdots & d_{nn} = 0 \end{bmatrix} \tag{2.2}$$

Distance metrics are evaluated for different types of attributes. In football player datasets, categorical and numerical attributes are found. A categorical variable is a variable that takes one of a limited number of possible values. Based on qualitative assessment, they will fall in some category. For example, the preferred foot of the player. These categories are either ordinal or nominal. Ordinal means that there is a specific order in the categories, which is not the case for nominal variables.

Categorical variables can have two or more categories. In case of two categories, it is a binary variable. There are two possible outcomes for this variable, which can be equally valuable or not. Equally valuable binary variables are symmetric and require a different distance measure compared to not equally valuable asymmetric binary variables. As indicated and shown by Akhanli (2019), counts of binary outcomes between objects (see Table 3.1) are the starting point for finding the similarity of *nominal* binary variables. Possible similarity measures are given in Table 3.2. The first matching coefficient is related to the simple matching approach of Kaufman and Rousseeuw (1990), given in Equation 2.3. This formula can be used for multiple category variables as well, finding the similarity or dissimilarity between object on the base of matches (m) and number of nominal categorical variables (p). These metrics are useful in the thesis to define similarities between categorical variables of the players, such as their preferred foot.

$$s(x_i, x_j) = \frac{m}{p} \qquad d(x_i, x_j) = \frac{p-m}{p} \quad \text{where i and j are two players} \tag{2.3}$$

Table 3.1: Table with counts of binary outcomes between objects (Akhanli, 2019)

		Object j		
	Outcome	1	0	Total
Object i	1	a_{11}	a_{10}	$a_{11} + a_{10}$
	0	a_{01}	a_{00}	$a_{01} + a_{00}$
	Total	$a_{11} + a_{01}$	$a_{10} + a_{00}$	$p = a_{11} + a_{01} + a_{10} + a_{00}$

Table 3.2: Table with similarity measures for binary data (Akhanli, 2019)

Code	Coefficients	Formula ($s(\mathbf{x}, \mathbf{y})$)	Type
S1	Matching coefficient	$\frac{a_{11}+a_{00}}{p}$	Symmetric
S2	Jaccard coefficient	$\frac{a_{11}}{a_{11}+a_{01}+a_{10}}$	Asymmetric
S3	Kulczynski (1927b)	$\frac{a_{11}}{a_{01}+a_{10}}$	Asymmetric
S4	Rogers and Tanimoto (1960)	$\frac{a_{11}+a_{00}}{a_{11}+2(a_{01}+a_{10})+a_{00}}$	Symmetric
S5	Sneath et al. (1973)	$\frac{a_{11}}{a_{11}+2(a_{01}+a_{10})}$	Asymmetric
S6	Gower and Legendre (1986)	$\frac{a_{11}+a_{00}}{a_{11}+\frac{1}{2}(a_{01}+a_{10})+a_{00}}$	Symmetric
S7	Gower and Legendre (1986)	$\frac{a_{11}}{a_{11}+\frac{1}{2}(a_{01}+a_{10})}$	Asymmetric

Ordinal variables are specifically arranged. The distance or dissimilarity between ordinal variables can be calculated similar as continuous variables, such as with Euclidean distance (discussed below). The ranking of the variables can be used to give a value for all variables, possibly to be normalized to equal the weight of each variable. Normalization can be done through Equation 2.4. These measures can be useful in the thesis for evaluating variables such as the tier in which a player competes. The tier is the rank of the league in its country, such as the Premier League being the first tier, and the Championship being the second tier of English football.

$$z_{ik} = \frac{r_{ik}-1}{M_k-1} \quad r_{ik}\ is\ rank\ of\ object\ i\ and\ variable\ k, M_k\ is\ highest\ rank\ for\ k \tag{2.4}$$

Numerical variables are quantitative data and can be discrete or continuous. The numerical variables are scaled in one of two ways. Interval variables are scaled in a way that the difference between values is relevant, ratio variables are scaled in a way that the ratio between the two values is relevant. Most variables in a football player dataset will be numerical ratio variables, for example number of goals made by a player, or number of tackles won. These variables can either be *continuous* (from an uncountable set of values) or *count* (non-negative integer values) variables. For these variables, dissimilarity measures can be used as proposed in Table 3.3. The Euclidean and Manhattan distance measures are commonly used and use their specific formulas to calculate the distance between two

objects. These measures are useful for calculating the dissimilarity of player numerical variables within the thesis, such as the number of expected goals per 90 minutes for a specific player.

Table 3.3: Dissimilarity metrics for numerical variables (Akhanli, 2019)

Code	Measures	Formula ($d(\mathbf{x}_i, \mathbf{x}_j)$)
D1	Euclidean distance	$(\sum_{k=1}^{p}(x_{ik} - x_{jk})^2)^{1/2}$
D2	Manhattan (City Block) distance	$\sum_{k=1}^{p} \lvert x_{ik} - x_{jk} \rvert$
D3	Minkowski distance	$(\sum_{k=1}^{p}(x_{ik} - x_{jk})^q)^{1/q}$
D4	Canberra distance	$\sum_{k=1}^{p} \frac{\lvert x_{ik} - x_{jk}) \rvert}{(\lvert x_{ik} \rvert + \lvert x_{jk}) \rvert)}$
D5	Pearson correlation	$(1 - \rho(\mathbf{x}_i, \mathbf{x}_j))/2$
D6	(Squared) Mahalanobis distance	$(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{S}^{-1} (\mathbf{x}_i - \mathbf{x}_j)$

x_{ik} and x_{jk} are, respectively, the k^{th} variable value of the p-dimensional observations for objects i and j, $\rho(\mathbf{x}_i, \mathbf{x}_j)$ is the Pearson correlation coefficient, $\mathbf{S}$ is a scatter matrix such as the sample covariance matrix

Aggregation of mixed-type dissimilarity measures

The types of variables and distance metrics described above can be used to define the dissimilarity between players or objects in the data. However, more of the different types of variables described above are present in the player dataset. Because of this, these distance measures for the types of variables also need to be aggregated to one distance measure. This can be done by aggregating the individual distance matrices of all variables, or by aggregating the already aggregated variables per variable type.

For the first method, as proposed by Akhanli (2019), the Gower dissimilarity (Gower, 1971) can be used to calculate the aggregated dissimilarity. This can be seen in function 3.5, where the variable w_k is the weight and δ_{ijk} is 1 if x_{ik} or x_{jk} are present, and 0 otherwise.

$$d_G(\mathbf{x}_i, \mathbf{x}_j) = \frac{\sum_{k=1}^{p} w_k \delta_{ijk} d_k(x_{ik}, x_{jk})}{\sum_{k=1}^{p} w_k \delta_{ijk}}, \tag{2.5}$$

For the second method, the following function can be used, where l indicates the aggregated measure:

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sum_{l=1}^{h} w_l d_l(x_{il}, x_{jl}), \tag{2.6}$$

With this method, different dissimilarity measures can be used for different types of variables, as the aggregation is only afterwards.

The different distance metrics are thus combined into one distance metric suitable for complete player analysis and evaluation of similar players, perfectly suitable for the objective of the thesis.

Because of the different methods of the distance metrics, the results in distances between data will be geared towards that perspective. Different type of attributes also have different types of distance metrics within the study. Based on similarities between players, and the results from different distance metrics, the most accurate distance metrics can be identified per attribute type. The attributes are represented in one distance measure, called the performance distance. This is a combination of the different distance metrics per type of attribute. We will go through a similar process in the next chapter, where we will design the tool.

2.2.2 Clustering techniques review (Saxena et al., 2017)

In Saxena et al. (2017), a comprehensive study on clustering is presented. Different clustering methods are described and form an overview of all clustering techniques possible to be applied. Clustering methods use similarity between datapoints to divide the data into groups such that the datapoints within these groups are more similar each other compared to datapoints in other groups. These clustering techniques are reviewed, as the application of a clustering method is one of the main steps in the seven-step guide of clustering as explained in Chapter 2.2.

Hierarchical clustering methods iteratively divide patterns to form clusters. This can be either agglomerative (bottom-up) or divisive (top-down), respectively merging small clusters into larger clusters or dividing large clusters into smaller clusters. There are three types of linkages that are used for clustering the data. In single-linkage clustering, datapoints are grouped based on the two closest datapoints in two different clusters. The two clusters that have the minimum distance between the clusters are grouped together. In complete-linkage clustering, the minimum distance between the two most distant datapoints in two different clusters decides which groups of datapoints are clustered. In average-linkage clustering, the minimum distance of the average of the distances between clusters is the deciding measure. These methods prevent the possibility of points switching from cluster, which is possible in certain enhanced hierarchical clustering methods, such as BIRCH, CURE, ROCK, or CHAMELEON.

Partition clustering methods cluster datapoints in *k*-clusters based on optimizing a criterion function. This is usually the dissimilarity metric, for example the Euclidean distance, where the minimum distance between datapoints and clusters are derived. Examples of partition clustering methods are k-means clustering and fuzzy c-means clustering. In k-means clustering, the dataset is classified in *k* clusters. The minimum distances between the datapoints and the cluster centres decide the classification of a datapoint to the cluster. In fuzzy c-means clustering, one datapoint can belong to multiple clusters. The fuzziness of the boundaries between clusters here decides the datapoints that are included in multiple clusters. CLARANS is an example of a partition clustering method.

Graph clustering represents clusters in graphs. Nodes are divided into clusters considering the edge structure of the graph, with higher edge densities within clusters as opposed to between clusters, representing the intraclass and interclass similarities, respectively. For example, the CLICK algorithm identifies tight clusters through graph-theoretic and statistical techniques.

Model based clustering methods cluster datapoints based on mathematical models. Feature details are detected on each cluster, representing a specific class. Decision trees and neural networks are the most frequent used model-based clustering methods. In the decision trees algorithm, a hierarchical tree represents leaves with a probabilistic description of that class. Neural networks are based on neurons, with weighted inputs and outputs, connected with each other.

Mixture density-based clustering uses probability to cluster datapoints in a specific cluster. The clusters and their distribution are identified, with the datapoints derived from multiple possible

density functions. Examples are the DBSCAN algorithm and the expectation-maximization (EM) algorithm. DBSCAN finds arbitrary shaped clusters and clusters with noise based on minimum points to define a cluster and the distance between neighbourhoods. EM is an iterative method based on iterative derivation of log-likelihood values of the estimate of the parameters and maximization of the log-likelihood through computation of new parameters.

Grid-based clustering forms a grid through partitioning of the space. This grid is than filled with the appropriate datapoints or objects, and the density of the cells determines whether they are eliminated. The grouping of these dense cells than forms clusters. For example, CLIQUE automatically lowers the dimensionality of a high dimensional dataspace making clustering more effective.

The above analysed clustering techniques all have their specific positives and negatives, which need to be related with the objectives in this thesis to decide on which cluster to choose. Some of these methods are compared in terms of characteristics which should be checked based on the application of interest. This includes for example the scalability and time complexity.

Table 3.3: Clustering algorithms compared based on characteristics. (Xu & Tian, 2015)

Category of clustering	Algorithm name	Time complexity	Scalability	Suitable for large scale data	Suitable for high dimensional data	Sensitive of noise/outlier
Partition	*k*-means	Low $O(knt)$	Middle	Yes	No	High
	PAM	High $O(k(n-k)^2))$	Low	No	No	little
	CLARA	Middle $O(ks^2+k(n-k))$	High	Yes	No	Little
	CLARANS	High $O(n^2)$	Middle	Yes	No	Little
Hierarchy	BIRCH	Low $O(n)$	High	Yes	No	Little
	CURE	Low $O(s^2*logs)$	High	Yes	Yes	Little
	ROCK	High $O(n^2*logn)$	Middle	No	Yes	Little
	Chameleon	High $O(n^2)$	High	No	No	Little
Fuzzy based	FCM	Low $O(n)$	Middle	No	No	High
Density based	DBSCAN	Middle $O(n*logn)$	Middle	Yes	No	Little
Graph theory	CLICK	Low $O(k*f(v,e))$	High	Yes	No	High
Grid based	CLIQUE	Low $O(n+k^2)$	High	No	Yes	Moderate

2.2.3 Cluster of variables around Latent Variables (Carpita, Ciavolino, & Pasca, 2021)

In Carpita et al. (2021), an approach for clustering of football player attributes is proposed. Cluster of variables around Latent Variables (CLV) is a clustering approach based around latent variables. Attributes can be combined into a composite attribute for easier player role and position identification. The used attributes are taken from EA Sports, where experts identified performance composite indicators. Some of these variables distributed in composite indicators are shown in table 3.4. These indicators are similar to the data as present in the database made available for this project. Because of this, the described clustering techniques should be applicable for this research's objective.

Table 3.4: Overview of attributes combined into composite indicators. (Carpita et al., 2021)

Table 1 The 33 performance attributes (variables) of *sofifa*, *fifa_04* and *fifa_07* with experts' dimension classifications

Attributes (variables)	Long names	Dimension classifications of:		
		sofifa (LABEL)	*fifa_04*	*fifa_07*
y_1	Shot power	Power (POW1)	Technical	Shooting
y_2	Jumping	Power (POW2)	Physical	Physical
y_3	Stamina	Power (POW3)	physical	Physical
y_4	Strength	Power (POW4)	Physical	Physical
y_5	Long shots	Power (POW5)	Technical	Shooting
y_6	Aggression	Mentality (MEN1)	Mental	Physical
y_7	Interceptions	Mentality (MEN2)	Mental	Defending
y_8	Positioning	Mentality (MEN3)	Mental	Shooting
y_9	Vision	Mentality (MEN4)	Mental	Passing
y_{10}	Penalties	Mentality (MEN5)	Technical	Shooting
y_{11}	Dribbling	Skill (SKI1)	Technical	Dribling
y_{12}	Curve	Skill (SKI2)	Technical	Passing

The clustering method is focused on clustering the variables into composite indicators, which could give a more narrow and clear indication for specific player positions and roles. The CLV procedure identifies *K* clusters of player variables and *K* latent components, with high correlation between variables in the cluster and the corresponding latent component (Evelyne Vigneau & Chen, 2016; E. Vigneau & Qannari, 2003). The covariance between variables in a cluster is maximized through hierarchical clustering and partitioning algorithm through Equation 2.7. In this equation, the covariance between x_j and c_k latent components come forward. Furthermore, δ_{kj} is 1 if variable *j* is in cluster G_k and 0 otherwise.

$$T = n \sum_{k=1}^{K} \sum_{j=1}^{P} \delta_{kj} Cov^2(x_j, c_k) \tag{2.7}$$

In the algorithm, variables move between the clusters to try and increase value *T*. In an iterative process, the latent components related with a specific cluster have a higher covariance with the variables in the cluster compared to the other latent components. With the variation of *T* within these iterations, can be analysed to choose the number of clusters. Similar as to in Principal Component Analysis (PCA), the number of clusters can be chosen where the variation (delta) of the *T* levels off.

Because of the objective of identifying similar players for a specific position and role, clustering statistics from the available dataset at San Lorenzo could thus be useful. This is, however, not a complete method to cluster the players themselves. For that, clustering of players based on these (composite) indicators is required.

2.2.4 Gaussian Mixture Clustering (Soto-Valero, 2017)

Soto-Valero (2017) presents a way to characterize football players. The Gaussian Mixture Clustering models (GMC) can be used in combination with principal component analysis to characterize professional football players. Attributes are taken from EA Sports' FIFA video game series and applying the clustering method resulted in different team roles.

Principal component analysis (PCA) is first used to reduce the dimensionality as preparation for the clustering algorithm. The GMC clustering model then identifies groups of datapoints, or observations based on their similarities through distance metrics.

PCA is a technique to reduce the dimensionality of multivariate data. The dimensionality is reduced by replacing many correlated variables with a few uncorrelated variables that retain most information. Equation 2.8 is the first principal component and is a weighted linear combination between *k* variables.

$$PC_1 = a_1X_1 + a_2X_2 + \cdots + a_kX_k \tag{2.8}$$

The first principal component will account for the most variance, and every next principal component will account for less variance while avoiding correlation (being orthogonal) with the first principal components (Jolliffe, 2011). Through applying the PCA algorithm, the reduced dimensionality can allow for easier visualisation.

GMC is a probability-based clustering model, clustering datapoints with high intraclass similarities and low interclass similarities. This means that the datapoints are closely related within a cluster but separated between clusters. In finite mixtures models (McLachlan & Peel, 2004), the clusters are related with the component probabilities. Independent and identically distributed data can be found from the mixture density as given in Equation 2.9. In this equation, f_k is the probability density function for cluster k (from a total G clusters or components). τ_k gives the probability that a datapoint belongs to mixture component k.

$$f(x) = \sum_{k=1}^{G} \tau_k f_k(x) \tag{2.9}$$

The probability density functions for the components is given in Equation 2.10, and is based on the mean μ_k and covariance matrix $\sum_k$. These parameters, as well as τ_k can be estimated via the expectation-maximization algorithm. Which is an iterative method for finding maximum likelihood estimates. (Dempster, Laird, & Rubin, 1977)

$$\phi\left(x_i;\mu_k,\ \sum_k\right) = \frac{\exp\{-1/2(x_i-\mu_k)^T\ 1/\sum_k\ (x_i-\mu_k)\}}{\sqrt{\det(2\pi\sum_k)}} \tag{2.10}$$

The results of the GMC make it possible to find players with a specific role and position in the team. The clustering model can cluster the players based on principal components or an aggregated set of variables. This is very useful for the objectives of this research, as we want to find players that can replace another player that will have had a specific role in the current squad as well.

2.3 Conclusion

In this chapter, different clustering and dissimilarity methods are analysed. These methods can be used to build a scouting decision support system and improve the efficacy of the scouting department at San Lorenzo. The methods can be analysed and evaluated to select most applicable methods for further use in the thesis, to limit the scope and remain focus. The clustering method considers different steps of which three are primarily discussed in this chapter.

To apply a clustering technique, dissimilarity metrics are required. As found in literature, different measures can be used for different types of variables, useful for application within the thesis, as for the different types of variables present in the dataset. These measures can be aggregated in different methods as well, based on for example clustered variables around latent variables (CLV) or Principal Component Analysis. Aggregation of variables can be useful for visualisation, important in the

intended deliverable of this research. With the aggregated variables, the clustering method itself can be executed. The mixture components can be used in Gaussian Mixture Models as base for probability-based clusters.

The clustering techniques and dissimilarity metrics form a method to be used for identifying player similarities. The application of these algorithms is specific to the type of dataset and type of results to be discovered. Within the design and execution of the method, this will be further analysed.

With this information, an ensembled method can be further developed in the design of the method in Chapter 3. Based on the requirements for the intended deliverable, the right methods can be combined. The combined method is served as the foundation of the results found in the decision support system.

3 Design of method

Chapter 2 gives a foundation on applying clustering models on football player datasets, as part of the objective in this thesis. Within this chapter, the methods found in literature are combined and redesigned to be applicable for San Lorenzo, based on the requirements and philosophy of the football club. In addition to the clustering methods discussed in Chapter 2, the design of the final tool is discussed in this chapter. The clustering methods are applied on the dataset and visualised in an intuitive tool. The following research question is answered:

How can the CRISP-DM methodology be further specified and designed for application on the football scouting dataset and objectives?

To answer this research question, the tool requirements based on the stakeholders are first set-up in Chapter 3.1. The to be applied method will then be constructed in Chapter 3.2, including all relevant steps from dataset to decision support system. The chapter is concluded with an answer to the research question in Chapter 3.3.

3.1 Requirements method and tool

To design a method that will result in an effective decision support system, a list of requirements needs to be set up and followed. The methods and techniques found in the literature will have their specific applicability within the development of the tool. This section provides an overview of the requirements within the final design. These requirements are setup with input from the stakeholders in this process, the scouts at San Lorenzo, and the context analysis done on the club in Chapter 1.2. The necessity in these requirements come from the applicability of the to be designed tool in finding results and making a positive impact in the efficiency of the scouting process. These requirements are followed within the design of the tool, as described below, and are to be evaluated and measured in evaluation of the tool in Chapter 5.

- **Philosophy San Lorenzo**
  The scouting decision support system should be designed such that it is conform the scouting and transfer philosophy of San Lorenzo. As analysed in Chapter 1.2 for example, San Lorenzo is a football club that scouts primarily in South America, for cheap players that can later be sold for a profit. This should indicate for the decision support system that the focus should be on suggesting cheap South American players with a high potential value.
- **Intuitive and future-proof usage**
  The decision support system will be developed with the intention for it to be used by the scouting department of San Lorenzo. To achieve this, the tool should be usable by the scouts, and provide intuitive results. An interface is required which includes a visualisation of the results. Also, because of the quick changes in football, the tool should be designed such that it can be used with updated datasets, with different players and statistics for example.
- **Results and evaluation**
  The decision support system should provide a list of suggested players per position, to give the scouting department a shortlist of players potentially to be scouted. The tool should give an indication of the similarity with the current player on that position, through the means of dissimilarity measures, clusters, and an overview of the general attributes of the compared players. General information is provided to allow for easy navigation to that player by the scout, including information such as market value indicating the quality and reachability of the player. This is required to give a well based recommendation for the scouting department, and for the data to be represented in a way that is understandable.
- **Perspectives of results for completeness**

Clustering techniques and dissimilarity measures are all different. No clustering technique will result in the same outcomes, which gives them their specific perspective. To give the user a complete indication of the player and similarity with other players, different perspectives should be present in the results of the tool. This can be done through both indicating results from clustering techniques and providing the dissimilarity measures. In this way, the similarity with other players is represented from multiple perspectives.

- **Customizability**
  To give a good user experience to the scouting department at San Lorenzo, customizability is required. Where the suggested players per position in the current squad of San Lorenzo can be a rather long list, filtering should be applicable to narrow down the suggested players to the scouts' specific needs. This should for example be the age of the player, and the general market value.
- **Theoretical foundation**
  The suggested players resulting from the use of the tool should be theoretically founded to give valid input to the scouting department. Dissimilarity measures and clusters resulting from the use of the tool should indicate similarities between players based on scientific methods.
- **Ethical consideration**
  The decision support system will be largely based on player data. The tool can be used for scouting objectives and could lead to positive or negative information to be interpreted by the scouting department of San Lorenzo. Decisions however will not be made through the support system, as it is only meant as a support tool. Sheridan and Verplank (1978) postulates that there can be a level of automation, instead of either entirely automating a task or not. A scale of levels is proposed in the study, as can be seen in Table 4.1. To relate the intended decision support system to this scale, we can conclude a low automation level of 2 or 3. Similar players will be presented by the tool, as a set of 'decision alternatives' and narrows it down based on similarity with the current squad. The interpretation and scouting decisions will be made by the scouting department employees at San Lorenzo, however. Furthermore, the decision support system will be transparent, to keep integrity. For players, this implicates that their first assessment can follow from data methods. These data methods are not a complete representation of the player quality, and players are thus never completely chosen because of data algorithms. To prevent manipulation from players, to for example player differently to fit a category of players more, the specifics in the data algorithms and where the scouting department focuses their analysis on, is not to be shared.

Table 4.1: Levels of automation. (Sheridan & Verplank, 1978)

Low	1	The computer offers no assistance, human must take all decisions and actions
	2	The computer offers a complete set of decision/action alternatives, or
	3	Narrows the selection down to a few, or
	4	Suggests one alternative, and
	5	Executes that suggestion if the human approves, or
	6	Allows the human a restricted veto time before automatic execution
	7	Executes automatically, then necessarily informs the human, and
	8	Informs the human only if asked, or
	9	Informs the human only if it, the computer, decides to
High	10	The computer decides everything, acts autonomously, ignores the human

3.2 Redesigned method

The different stages within the method are presented in this section. These stages are the result from analysis of the literature, and applicability to the intended results and objectives for San Lorenzo. As found in Chapter 2.1, the clustering process is based on seven stages (Milligan, 1996). These stages are used as a basis for the redesigned method, approached in an iterative manner. The iterative characteristic of the redesigned method entails that even though the stages are represented in one 'path', these are not applied in one clear structure. A standard process model for data mining, CRISP-DM, is integrated with the seven steps, to account for a complete data science life cycle. This results in a, for the objectives of this thesis, more complete guide with the nine-step process summed up below. The deployment of the tool is not considered as it is out of the scope of this thesis. CRISP-DM is a process model that relates to the data science life cycle. The first step is getting an understanding of the business and the requirements, this is already done through the context analysis and requirements setup. Data science projects follow this cycle to help plan, organise and implement the project (Wirth & Hipp, 2000). The cycle follows different stages as described below without maintaining one specific structure, but with possible iterations and thus coming back to previous stages. To achieve the intended decision support system, the development of the tool and interface itself should be included within this list, resulting in the nine stages as given below:

1. Collecting and understanding data
2. Choosing variables used for clustering
3. Pre-processing data
4. Designing the dissimilarity metric
5. Choosing the clustering method
6. Determining the number of clusters for the method
7. Implementing the clustering method, acquiring results and evaluation of the models used
8. Developing the tool including the results
9. Evaluation of the tool

These steps will be further analysed and explained within the following subsections.

3.2.1 Stage 1 – Collecting and understanding data

The first stage continues on understanding the dataset as already shortly described in Chapter 1.2.2. The dataset required to accomplish the objective of the thesis is first collected. Next, the data is examined to identify the format, number of records, and types of attributes for example. The data is supplied by the football club, and includes a total of 31691 players, and 122 player attributes. After examining the dataset, the data is explored more intensively. To explore the dataset and find valuable insights, objectives are set up. These objectives relate with the applied methods, which require certain input or general understanding of characteristics in the dataset (rows, columns, values, etc.). Through visualisations and exploration in terms of correlation and distribution plots, relationships and characteristics of the data are identified. Correlation plots are made to identify relations between attributes and important attributes to be weighted in the clustering algorithm. This can for example be a relation between the number of goals scored by a player, and the market value of this player. Distribution plots are made to identify the shape of the data and allow for data cleaning in preparation for the clustering and dissimilarity algorithms. In this step, the quality of the data is thus assessed and improved for further use in the method.

The output of the first stage is an overview of the dataset together with its characteristics.

3.2.2 Stage 2 – Choosing variables for clustering

With a deeper understanding of the dataset resulting from stage 1, the useful data are selected for the objectives of the thesis. Based on requirements and relevancy for San Lorenzo, certain parts of

the data are argued to be used in the final design of the decision support system. The 122 attributes present in the dataset are assessed based on relevancy. Valuable attributes to explain the characteristics of the football player are chosen. These attributes are standardized attributes which are comparable within the dataset and give additional information on the player. Choosing these variables is further focused on in the next stage.

The output of the second stage is an overview of the importance of certain attributes to be further used in the next stage.

3.2.3 Stage 3 – Pre-processing data

The final dataset used for modelling is prepared. This stage ensures that the quality of the data is sufficient, and prevent the garbage-in, garbage-out principle for example. Data needs to be prepared such that the resulting dissimilarity metric in the next stage represents how it is interpreted in the to be built tool (Hennig & Hausdorf, 2006). This stage consists of the following five steps:

1. **Cleaning data**
   A dataset includes values that do not make sense or are erroneous. These values are removed or corrected to make sense of the dataset and reconstruct it for further use. As a result of the first stage, the dataset is described visually and statistically. The exploration of the dataset gives insights into null values and outliers that make data invaluable. Null values indicate missing data, which leads to removal of all data from the player. Outliers are individually assessed as it is common to have outliers that do make an impact on the results but are valid. This can for example be the top goal scorer of the league, which can be an outlier value, but still is valid. The cleaning of the data results in a dataset which gives accurate information on the players. Furthermore, cell values that heavily impact distributions of attributes while these specific cell values are not valuable, are removed to return more accurate distributions.
2. **Feature engineering**
   Certain attributes have a more interesting or effective meaning when combined. In this step, new attributes are defined that are helpful in the rest of the method. The relevant information from the variables is extracted through for example summarisation of data. For our use case, this means that certain variables will be represented relative to the time played by the player. For example, the total number of goals scored by a player is not as relevant, as the number of goals scored per 90 minutes. The first type of the attribute will give large variety in possible values because of the differences in number of games played by the players. Because of this, the variables are standardized and can be compared between players.
3. **Integrating data**
   If there are multiple datasets or data origins, these are integrated into one dataset. This step includes standardization and adaptation in case of different types of attributes or data types. Even though it is one of the usual steps in data-preparation, this project only uses one dataset, and this step is thus skipped.
4. **Data transformation and standardisation**
   Data transformations are applied to improve the usability of data in the modelling algorithms. For example, power transformations affect the impact of outliers on the results of the cluster analysis. The variables are transformed in such a way that the distance measures resulting from applying dissimilarity metrics, relates to the differences as can be interpreted in the application of interest, our case the decision support system (Hennig & Liao, 2013). In the dataset, attribute distributions following from the exploratory data analysis, are skewed to the right or the left. This skewness is reduced with power transformations to make small differences in lower data values be represented similarly as higher differences in higher data

values (squared transformation) or the other way around (square root transformation). Data standardisation is applied to remove the impact of scaling on the dissimilarity metrics, and make data comparable, as will be used in the next stage. Data is reformatted if necessary. Similar as in integration of the data, certain data values have datatypes that cannot be easily used in the next stages.

5. **Variable selection and dimension reduction**
   Variable selection and dimension reduction is applied for better representation in the cluster analysis and reducing computational costs for high dimensional data. Variable selection entails ignoring certain variables in the modelling phase, where dimension reduction combines variables into a lower dimension. For example, variables are selected for removal if they have no added value in describing or characterizing the player. Also, characteristics of the football club make a part of the dataset useless, for example in budget considerations. Principal component analysis (PCA) is one of the dimension reduction techniques, which will be further explained in the next stage.

The output of the third stage is a cleaned and ready-to-use dataset.

3.2.4 Stage 4 – Designing dissimilarity metric

At this moment, the data is prepared and ready for input to the model. Clustering is based on grouping similar players together, which is determined based on the distances between these players. The distance or dissimilarity is measured through the different player attributes. In this stage, we apply a combination of the methods as found in literature, in Chapter 2. The dissimilarity measure follows from applying the method as explained in Chapter 2.2.1, where mixed-type attributes are measured and aggregated. The type of variables is assessed to decide on the correct way of measuring dissimilarity. With the scaling and standardization completed in the previous stage, the attributes have the same impact on the dissimilarity measure. From this point onwards, the different attributes can be weighted on importance for that specific type of player as resulting from the exploratory data analysis and domain knowledge, where for example for a defensive player, the amount of successful defensive actions will be more important compared to an offensive player. The PCA algorithm as explained in Chapter 2.2.4 is applied to further reduce the dimensionality allowing for visualisation and faster computing time. A minimum of 60% of variance is found in these principal components to still have an accurate representation of the dataset (Sarstedt, 2019). This is taken as a minimum after which an analysis results in a final definite percentage in Chapter 4.1.7. After applying these methods, the metric is ready for input to the clustering model in the next stage.

The output of the fourth stage is the dissimilarity metrics to be used in the clustering method, to derive the dissimilarities between players in the dataset.

3.2.5 Stage 5 – Choosing clustering method

For choosing the clustering method, different steps need to be taken. First, a selection is made of clustering methods to be tried. Based on the literature study in Chapter 2.2, this selection is evaluated and the techniques most relevant for the objectives of this thesis come forward. The models need to be assessed based on domain knowledge, relevancy, and method requirements. As mentioned in Saxena et al. (2017), each clustering algorithm has its specifics strengths and weaknesses, for example on the size of the dataset or the quickness of the algorithm. These are more or less important based on the prepared dataset, as for example the PCA algorithm drastically reduced the dimensionality of the to be clustered dataset. The strengths and weaknesses are evaluated with the characteristics of the dataset and the objectives of this research, to make the decision. To limit the scope, only the most

applicable methods are further analysed and used at first. The overview of clustering methods and their specific characteristics is presented in Chapter 2.2.2.

The output of the fifth stage is an examination and decision of the clustering method to be applied.

3.2.6 Stage 6 – Determining number of clusters

The quality of applying the clustering method indicates what number of clusters is optimal for that specific clustering methodology. This quality is related with some clustering validation index, for example the average intraclass dissimilarities (Caliński & Harabasz, 1974) or more specifically the WCSS (Within-Cluster Sum of Square) as to be used (Saji, 2021). This type of objective criteria is used to estimate the number of clusters that is optimal for the dataset for applying the clustering method. With the optimality, the general separability of the clusters based on the WCSS value is meant. This relates to how a group of players will be separated with another group of players, based on certain attribute values. This estimation is evaluated based on application specific characteristics, where a specific number of clusters makes the most sense.

The WCSS is the summed squared distances between the centroid and the individual datapoints of each cluster. By plotting these values against the number of clusters, a visual elbow is encountered, which represents the optimal number of clusters as in Figure 3.1. The elbow is found where the graph changes quickly. This is the point where the proportion of explained variance decreases, and the additional 'won' WCSS value becomes insignificant compared to the costs of adding more clusters.

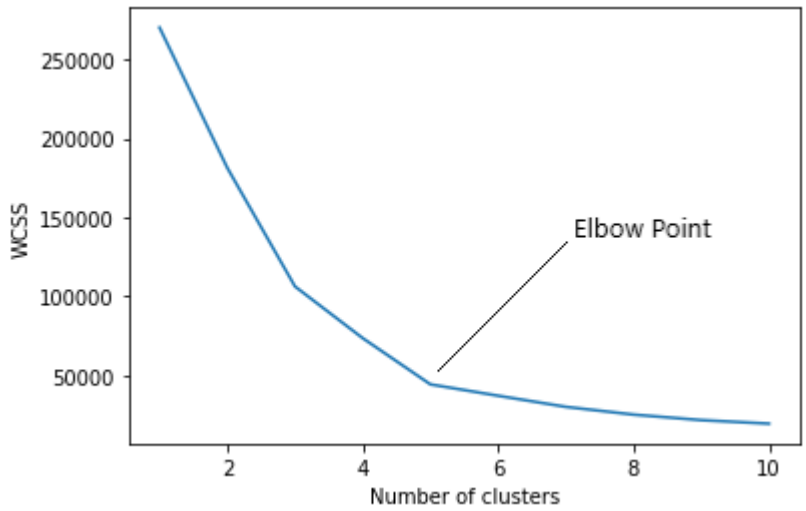

Figure 3.1: Elbow method used for determining the optimal number of clusters. (Saji, 2021)

Manual visual inspection can be used to determine the number of clusters, to be done before further use of the clustering algorithm. To automate this process, the so-called elbow strength can be calculated to find this elbow point. By analysing the differences in the criterion (in this case the WCSS) and finding out where the difference between the change in value and second-order difference (the elbow strength) is the greatest, the optimal number of clusters is automatically extracted. (Granville, 2019)

The output of the sixth stage is the number of clusters as input to the clustering model in the next stage.

3.2.7 Stage 7 – Implementing clustering method, acquiring results and model evaluation

In this stage, the clustering model is built. The code is made in Python to analyse the dataset according to the chosen clustering method, dissimilarity metrics, and number of clusters as derived in the

previous steps. Programming packages for the above-described methodologies are found and implemented. The dataset used is of course the acquired and pre-processed dataset from the first three stages of the method. Preliminary results are shown to give an indication of the results to be acquired in the development of the tool.

To acquire the best possible results, the clustering methods resulting from stage 5 are implemented for evaluation. With the use of evaluation metrics, the most applicable clustering method is found. With this knowledge, the standard to be used clustering method is appointed. Examples of evaluation metrics are Silhouette coefficient and Dunn's index. These are seen as the most widely used and most accurate evaluation metrics, which is why they are chosen for further evaluation and application within this thesis (Legány, Juhász, & Babos, 2006). These evaluation metrics will be shortly described and used for evaluation of the four clustering methods in the execution of the method. The inter-cluster distance is the average distance between all clusters, and the intra-cluster distance is the distance between the points within a cluster.

- **Silhouette coefficient**
  The silhouette coefficient is a measurement that interprets and evaluates the consistency of datapoints within clusters. The higher the silhouette coefficient, the higher the similarity or cohesion of datapoints within a cluster. The silhouette coefficient is calculated with the average distances between datapoints within a cluster, where the smaller value entails a better cohesion. This value is compared with the average distances between the datapoint and the datapoints of the closest (so called neighbouring) cluster. Based on these two values, the datapoints can be evaluated based on how well they match their cluster, and how well they are separated from other clusters. (Rousseeuw, 1987)
  To give an example of the Silhouette coefficient, Figure 3.2 shows different silhouette coefficients for two or three number of clusters in K-means clustering. Within these clusters, the separability of the clusters leads to silhouette coefficient values which conclude that two clusters lead to most separable clusters in this case, with a silhouette coefficient of 0.705, compared to a value of 0.588 for three clusters. The formula for the silhouette coefficient is given in Equation 3.1. (Pedregosa, 2011)

$$\frac{(b-a)}{\max(a,b)} \quad \text{where a: avg intra} - \text{cluster distance and b: avg inter} - \text{cluster distance} \qquad (3.1)$$

- **Dunn's index**
  The Dunn's index is an index that calculates the variance within and between clusters, with the aim of finding a set of compact clusters that are separated from each other. The interclass similarities are compared with the intraclass similarities and a higher rate (relatively higher distance between clusters opposed to within clusters) is desirable, entailing that a higher Dunn's index is better. The computational costs increase heavily however, with increased size of data. The formula for the Dunn Index is given in Equation 3.2. (Dunn†, 1974)

$$DI_m = \frac{\min\limits_{1 \leqslant i < j \leqslant m} \delta(C_i, C_j)}{\max\limits_{1 \leqslant k \leqslant m} \Delta_k} \quad \text{with } \delta(\text{Ci}, \text{Cj}) \text{ as the inter-cluster distance metric} \qquad (3.2)$$

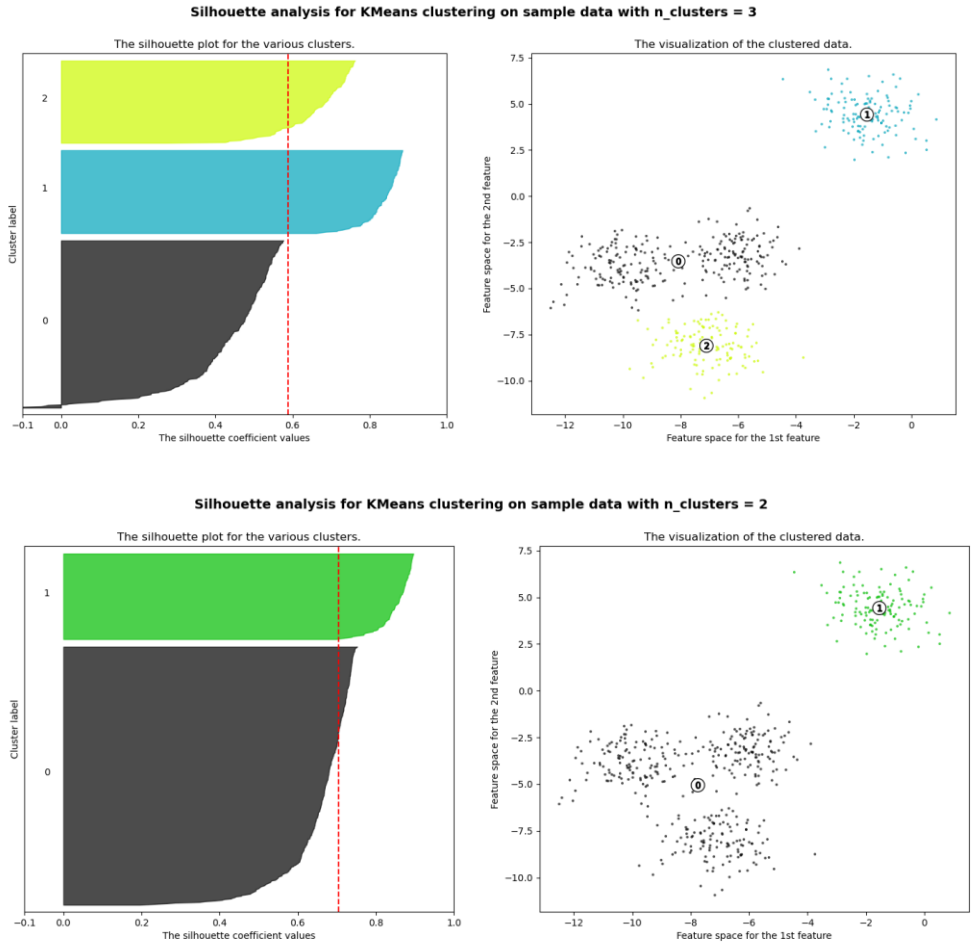

Figure 3.2: Difference in silhouette coefficients with KMeans clustering for respectively 3 and 2 clusters. (Pedregosa, 2011)

Because of the disadvantage of Dunn's index having high computational costs on larger datasets, Silhouette coefficient is chosen as evaluation metric with the objectives of this thesis. For evaluation, three position groups are analysed and the Silhouette coefficient is returned. Based on the best average Silhouette coefficient, the standard clustering method for the tool is chosen. The average Silhouette coefficient is taken over tests for each position group and is discussed below. Furthermore, the computation time is derived, as relatively quick results are important in the future application of the tool.

Also, the difference in model results including or excluding weighting per position variables is evaluated. The hypothesis was that weighting the attributes that are more important for that specific position group will enhance the possibility to distinguish the players from each other. This is evaluated per clustering type as well.

Finally, the threshold for the variance to be represented by the principal components is evaluated and optimized. In the analysis of the model, different thresholds are tried to evaluate which percentage of variance will entail a high represented variance by a small number of principal components.

The output of the seventh stage is the resulting suggested similar players per player, considering the implemented clustering method.

3.2.8 Stage 8 – Develop tool including the results

Within this stage, the results from clustering are transferred to a tool for the scouting department at San Lorenzo. A tool is created that applies the clustering methods on an inputted dataset, and gives

an intuitive visualisation of the results, in terms of suggested players and their comparison to the current squad at San Lorenzo.

The clustering results following from the previous steps are incorporated in a tool to allow for analysis by the scouting department of San Lorenzo. With the use of dashboarding and web development tools, the tool is built on a local host at first. To integrate the Python code used for building the clustering algorithm and the dissimilarity measures, together with visualisation of the resulting similar players and statistics, several tools are used together. The Python code is integrated in a Python web framework called Django. Django allows for development of websites, focusing on the backend of the tool, which is the data access layer of the software. Django communicates with the front-end developed in Vue.js. Vue.js is a frontend JavaScript framework used for building applications, building on top of the standard HTML, CSS, and JavaScript languages. Vue.js focuses on the frontend of the tool, the presentation layer of the software. Using these tools, the data can be loaded into the back-end server, and with the user input of the player to be clustered and compared, the back-end server can be run to give these results back to the front-end and allow for visualisation in the locally hosted site.

The output of this stage is an easy-to-use tool and interface for the scouting department at San Lorenzo to use.

3.2.9 Stage 9 – Evaluate tool

Before this stage, the tool is created and gives results to be assessed by the scouting department at San Lorenzo. In this stage, the tool needs to be assessed and evaluated based on the requirements from San Lorenzo. These requirements are mostly related to the requirements setup in Chapter 3.1, and based on the problem and context stated in Chapter 1. However, they are revisited for pure evaluation of the tool, the usability, and the results itself within the potential use within the scouting process at San Lorenzo. Below, they are summed up including the way of evaluation and decision on sufficiency.

1. **Transfer philosophy**
   The tool should be applicable in search of players that conform the transfer philosophy of San Lorenzo. The tool should be customizable, including filter types to search for players that are conform the transfer budget of San Lorenzo and fit the type of player the scouting department is looking for at that moment (in terms of age, type of position, market value, competition). Furthermore, the weighting of attributes should be customizable to fit the scouts' specific preferences in the scouting process. Because of the usual quick cycles within the squad at San Lorenzo, this should help with the efficiency of finding to be further scouted players. Whether the tool conforms this requirement can be assessed by the scouting department of San Lorenzo through subjective evaluation. The tool is sufficient in case the scouting department can search for players that conform their transfer philosophy and current vision of characteristics of a replacement player (see Chapter 1.2.1.2).
2. **Intuitive and repetitive use**
   The tool should be usable by the scouting department, and not give problems in use in terms of difficulty of finding the options the tool features. With updated datasets which will become available in the next years, there should not be any difficulty in updating the dataset within the tool either. An effective and repeatable method to analyse and derive new potential to be scouted players is to be designed which requires the reusability and ease of use of the tool. To confirm whether the tool conforms this requirement, the tool can be subjectively evaluated by the scouting department at San Lorenzo. The easiness of using the tool will be measured

by the scouting department and measured with either a sufficient or insufficient usability, as well as possible use on other datasets, within the tool. This decides whether the tool is sufficient based on this requirement.

3. **Results**
   The tool should give results based on the applied clustering models, in the form of players similar to the compared and inputted player, to solve the core problem stated in Chapter 1.3.2. The resulting players should be visualised including general information of the player, as well as technically founded dissimilarity measures and clustering results. This is the main idea of the tool, solving the core problem approached in this thesis. To confirm whether the tool conforms this requirement, subjective evaluation by the scouting department is done on the completeness of results and whether different perspectives are provided. Furthermore, tests can be done to check if certain players come forward that have come forward in the scouting process before, which could indicate only an increase in speed of the process, or that mostly new unknown players are identified which could be interesting and provide added value to the scouting process. The shortlist resulting from the tool should aid in the efficacy of the scouting process, where the shortlist gives an initial overview of the players and aids the scouting team in which players should and should not be scouted more in-depth. Finally, because of the transfer philosophy of the club wanting to get young potentially good players and selling them on for a profit, there should somehow be an indication of the potential value of a player.

This gives the following checklist to be evaluated by the scouting department, which will be elaborated on in assessment, to allow for further improvements in future research. A tool is deemed sufficient if the following points are checked off.

- ☐ The scouting department can use filter options within the tool to search for players conforming the transfer philosophy and characteristics of a to be discovered player.
- ☐ The tool is sufficiently usable in terms of application of the included features.
- ☐ The tool improves the efficacy of the scouting process.
- ☐ The tool gives technically founded results based on the clustering models, including general information (such as age, club, market value) as well as statistical measures for each similar player.
- ☐ The tool gives different perspectives on the similarity and clustering results of the analysed model.
- ☐ The tool helps to solve the problem about missing similarity information.

The assessment and evaluation are done through subjective evaluation. Qualitative research can be done with the stakeholders to evaluate the tool subjectively, based on usability and relevancy of the tool. The process needs to be reviewed to derive any problems with the method and give options for future improvement. This makes it possible to conclude whether to go to the next step of deployment or iterate in the development to further improve the tool.

The output of this stage should be an evaluation of the tool, with possible improvements or adjustments to be made.

3.3 Conclusion

In this chapter, literature is combined into one method applicable for the objectives of this thesis. The method designed in Chapter 3.2 can be applied based on requirements in Chapter 3.1 to achieve clustered groups in the football player dataset available for the intended deliverables of this thesis. The method is standardized and applicable for football player datasets, for example the dataset as

available within this thesis. The method is designed to achieve the main goal, improving the efficacy of the scouting department at San Lorenzo, considering the requirements and criteria as assessed. Applying the method will result in a tool that can be used to find an overview of similar players per position in the current squad at San Lorenzo, based on the dissimilarity metrics and clustering methods applied.

The designed method is implemented in the next chapters to build the intended deliverable and provide practical solutions and suggestions to San Lorenzo. Steps 1 to 8 are presented in Chapter 4, and steps 9 and 10 are presented in Chapter 5.

4 Results

4.1 Implementing the method

In this section, the stages of the designed method in Chapter 3 will be followed in an iterative manner.

4.1.1 Collecting and understanding data

Before applying clustering methods and going in depth on the theoretical models, the dataset is explored and understood. Exploratory data analysis is performed to get an overview of the data, possible redundant attributes or errors in the dataset that need to be processed in the next steps.

To have a focus in the exploration of the dataset, objectives for the exploration are setup. This focus will make sure that characteristics, that are interesting for the to be applied methods, are identified.

- A general overview of the types of attributes, average data values per position group and size of the dataset is to be identified.
- Groups of players should be identified that can be compared based on their attributes. Players in different positions are expected to have different important attributes to evaluate the quality of players and compare them between each other.
- Following on the previous point, the importance of the attributes should be understood and evaluated. Which of the attributes for example have the most impact on the market value of players in a certain position? This will be useful for focusing on the right attributes for a player position. Intuitively, for attackers, the amount of defensive sliding tackles will for example be less important compared to defenders. Correlation plots can be used for this purpose, identifying which attributes have the most influence on the market value of a player.
- In the iterative exploration process, data can also be pre-processed partly before the third stage. Obvious outliers or null values can be handled for example.

Data is released by the football club for purposes of this research and is originated from 'WyScout', which is a third party that collects all this necessary data, and San Lorenzo has access to this data. The dataset is called 'conmebol_leagues_2018_2021' and contains player information from the Conmebol competitions between 2018 and 2021. The rows are specified per season, so the dataset contains information for players per season, instead of aggregated over the multiple years. To further explain, the identifier of the rows is the player's name, club, and season together. Conmebol is the South American football association, meaning that the dataset contains data from competitions in South American countries (Argentina, Bolivia, Brazil, Chile, Colombia, Ecuador, Paraguay, Peru, Uruguay, Venezuela). The dataset contains players that played at least one game during these three years. The dataset has 31691 rows and 122 columns. The rows represent the player data for a specific season, and the columns represent the attribute or characteristics of the player.

This dataset contains a lot of relevant attributes that altogether give a complete view on the player. Because of this, it is useful and complies with the requirements setup in Chapter 3. The models to be implemented on the dataset use these attributes to characterise the players and with a valid scientific background, usable results will be achieved from implementation. The attributes are discussed in detail as these will be important in the dissimilarity measures as well as input to the clustering models in the final designed tool, which are explained in-depth in Chapter 2.2.

The dataset contains different types of variables, described in the next part. The variables that are useless or obvious, such as the continent and region with the same values, are not considered in the description. Finally, while exploring the dataset, the dataset can be partly cleaned in an iterative exploring process as well. Exploring the dataset will lead to the discovery of for example null values or

outliers, which can be cleaned. The cleaning process of the data is then further finalised in the third stage.

General attributes

General information on the players is explained in multiple attributes as can be seen in Table 4.1. The attribute names and their description and relevancy are explained. In the next section, the attribute distributions will be further explained and visualised in Figures 4.3 and 4.4.

Table 4.1: Explanation of general information attributes in the dataset.

Attribute	**Explanation**
player	This variable contains the player's name. In the thesis, this variable will be used to identify which datapoint from the analysis relates to which player, to give useful suggestions to the football club.
club	This variable relates to the club the player plays at, at this current moment.
current_club	This variable relates to the club the player played at, at the time of measuring the statistics, so in that season.
ws_player_general_age	This variable is the current age of the player.
ws_player_general_market_value	This variable represents the market value
ws_player_general_contract_expires	This variable represents the expiry date of the contract of the player with his current club.
ws_player_general_birth_country	The country in which the player was born.
ws_player_general_passport_country	The countries of which the player has a passport.
ws_player_general_foot	The preferred foot of the player.
ws_player_general_height	The height of the player in centimetres.
ws_player_general_weight	The weight of the player in kilograms.
ws_player_general_on_loan	This variable represents if the player is on loan or not.
season	This variable is important to represent in the results of the decision support system. Players are suggested, but players that have played similar in the last year to a to be replaced player are more recent and thus more relevant. The same player can thus be suggested from multiple seasons, with the last season as most relevant.
competition	The competition in which the player played that season.
tier	The tier of the division in which the player played in that country.
nation	The country in which the player played that season.

The competition and tier a player play in, is an interesting attribute for the player. The different leagues in the Conmebol federation have different qualities. Because of this, the attributes do not have the same meaning for measuring the quality of a player when looking at different competitions. For example, a player might be able to score more goals in a competition where there is less resistance of quality defenders. The KA leagues ranking ("4Q December 2021 CONMEBOL Ranking," 2021) is

taken as a measure for the quality of the leagues and is added to the database. The second-tier divisions are expected to have lower qualities than the first-tier divisions and are adapted based on this assumption.

For the general attributes, all players combined are first considered before discussing some differences between the position groups. The expected most interesting numerical variables of this category are visually described in Figure 4.1. Density histograms of the numerical variables are shown, giving an insight into some general characteristics of the player base. In iterative process, for example the null values of the height and weight are cleaned, as well as deletion of players with a market value above the means of San Lorenzo.

Figures 4.1 and 4.2 show the density histograms of general performance attributes and the attacking performance attributes, respectively. All these attributes are explained in Table 4.3, where the attributes are chosen as the attributes that give a general overview of the players in the dataset. As will be explained below, players that played only a limited number of games are deleted from the dataset. Something interesting that can be interpreted from this figure is that most players have around 20 duels per 90 minutes played, but also quite a large group only have between 0 and 2 duels per 90 minutes played. When we further investigate this data, we identify that this is the group of goalkeepers in the dataset, which does make sense, as the goalkeepers are not probable to get into duels often. This does however indicate that it might be important to divide the dataset into goalkeepers and outfield players.

In Figure 4.2, some of the variables, such as goals or shots, are skewed to the right. This results from a lot of players (defenders and goalkeepers mostly) expected to not take any shots per game. Furthermore, for example the goals distribution is filled with certain players that scored >40 goals in that season, while most players scored only a couple at most. Because the players that scored a lot of goals, often also played a lot of games, this distribution and attribute does not give a lot of information on its own. This entails that it will be more realistic to only analyse percentage and per 90 minute statistics. Also, as will be further discussed in Chapter 4.1.3, the data distributions will be transformed to be less skewed for input to the clustering models as this improves the accuracy of the clustering models. Furthermore, the dataset should be split into players per position group, to analyse the difference in players based on important statistics for that position group. This will make the dissimilarities between players in the results of the tool, the objective of the thesis, more accurate as they will be based on important statistitcs for that position. The already known differences in players in terms of positions and general differences in statistics between these positions will thus not interfere with our objective of finding similarity information on specific players and their respective positions. To give an example of differences in attribute importance; for strikers, the amount of goals per 90 minutes is expected to be more important compared to the percentage of tackles completed.

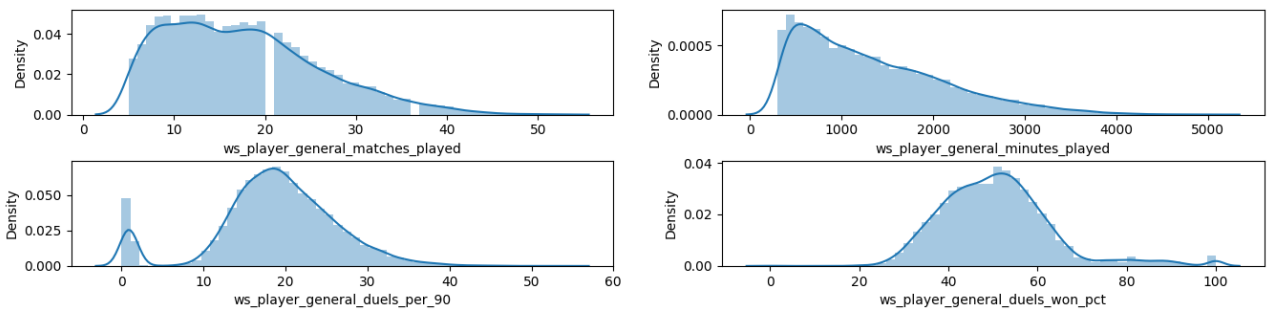

Figure 4.1: Distributions of general performance attributes in the dataset (see Table 4.3)

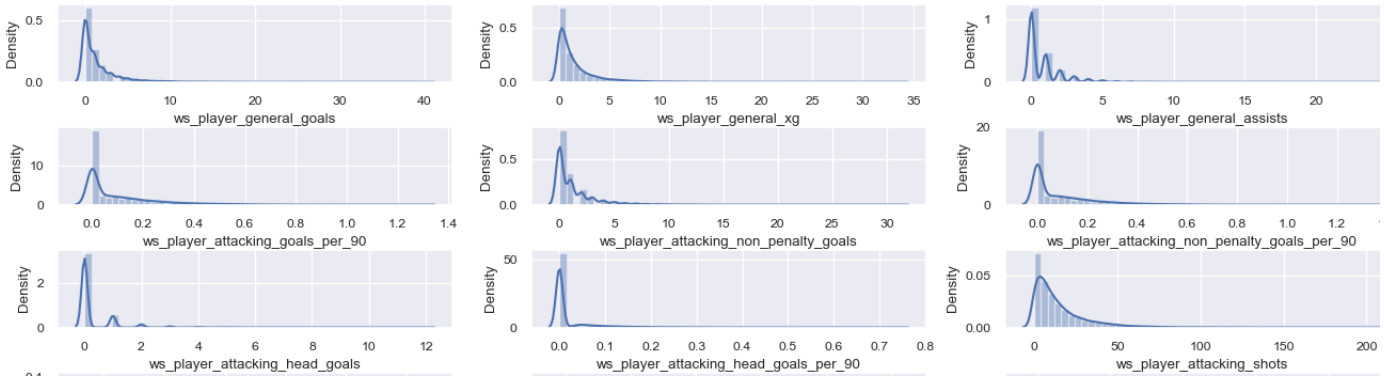

Figure 4.2: Distributions of attacking performance attributes in the dataset (see Table 4.3)

Position attribute

As previously explained, it is important to split the dataset per position group, to analyse players based on important statistics for that specific position group. With this, the to be applied models focus more on the attributes that are more relevant for a particular position group. This is based on the correlations found in attributes with the market value which is the closest direct indicator of the performance of the player, where a higher correlation suggests that that attribute is more important for that position, as the market value is generally more impacted by that statistic. The correlations between attributes and the market value are displayed in Figure 4.6 and Figures A1-3 in the Appendix. With the analysis specific per position group, the difference in characteristics of players can be discovered more specifically. In an analysis of all players combined, the main results are expected to show the different position groups, which is what we already know and what does not give the extra required information that will add value to the scouting department. In the dataset, the position of the player is represented. 'ws_player_general_position' gives a list of positions in which the player has played. The possible player positions are explained in Table 4.1. Here, the position letter means the following:

- A = attacking
- B = back (defender)
- C = centre
- D = defensive
- F = forward
- GK = goalkeeper
- L = left
- MF = midfielder
- R = right
- W = wing

To transform the player positions into usable data, a binary variable will be used for all possible player position groups. The player position groups are the groups of positions that can be compared with each other. This is important for weighting certain attributes differently for different player roles and positions. Players can have multiple positions being assigned to them, meaning that they can also be appointed to multiple roles. Based on domain knowledge, an initial grouping of player positions is shown in Table 4.2. This is subject to change in further exploration of the dataset.

Table 4.2: Player roles, positions including the characteristics of that position.

Position group	Positions	Explanation
Goalkeeper	GK	Individual position, no comparison with outfield players.
(Wing)backs	RWB RB LB LWB	Compared to the centre backs, the (wing)backs can have more duties of utilising the space on the wings to help creating chances in the attack.
Centre backs	RCB CB LCB	Different from the wingbacks, the centre backs are even more focused on preventing the opponent from scoring. Besides building up from the back, there is no expected importance in crosses or through passes. Central defenders mainly have the duty to prevent the possibility of attackers of the opponent to create chances to shoot on the goal and possibly score.
Defensive midfielders	RDMF DMF LDMF	Defensive midfielders mainly have the duty to protect the backline and keep the spaces small.
Central midfielders	RCMF CMF LCMF	Central midfielders are more controlling and are the link between defence and attack.
Attacking midfielders	RAMF AMF LAMF	Attacking midfielders help create chances and score goals.
Wingers	RW RWF LWF LW	Wingers are expected to be important in creating chances for the striker. Crosses from the wing or dribbles beating opposition defenders can be important attributes.
Strikers	CF	Strikers mainly have the duty to score goals. This is also the line where the pressure on the opponent begins.

Throughout the exploration of the dataset, the position groups defined in Table 4.1 will be used to compare attributes and correlations between attributes to find importance and characteristics of variables per position group. The market value will be considered in all attribute groups, to identify which attributes have a positive or negative influence on the market value, which can be intuitively directly related to the quality of a player. This will make it possible to weight attributes for different position groups.

Figures 4.3 and 4.4 show plots of general attributes of the full dataset (including all player positions) and are compared to plots filtered on a specific position group. This is done to get an overall idea of the dataset, see what an average player will look like based on attributes, to get an understanding and background for specifying the differences in players further in the thesis. This is still related to the data understanding part of the designed method. The preferred foot, the height and the weight variables given in the dataset are not completely accurate but do give a good general indication of differences in position groups. For the position groups, some differences are found compared to the complete set of players. For example, the height, age, and weight distributions of the goalkeepers are on average higher. This confirms the intuitive hypothesis that goalkeepers are the tallest players in the squad and can usually continue playing until a later age because of the smaller physical efforts. Apart from that, the favourite foot of wingbacks is more evenly distributed compared to centrally positioned players, suggesting that most players are right footed, but relatively most left-footed players play on the left-back position. For wingers, the possibility to cut inside makes it intuitively less important to have their preferred foot on that side of the pitch. Attacking midfielders and wingers are on average the smallest

players compared to other position groups. Apart from a positive correlation between height and weight of the players, no further to be discussed correlations are found in the general attributes for the different position groups.

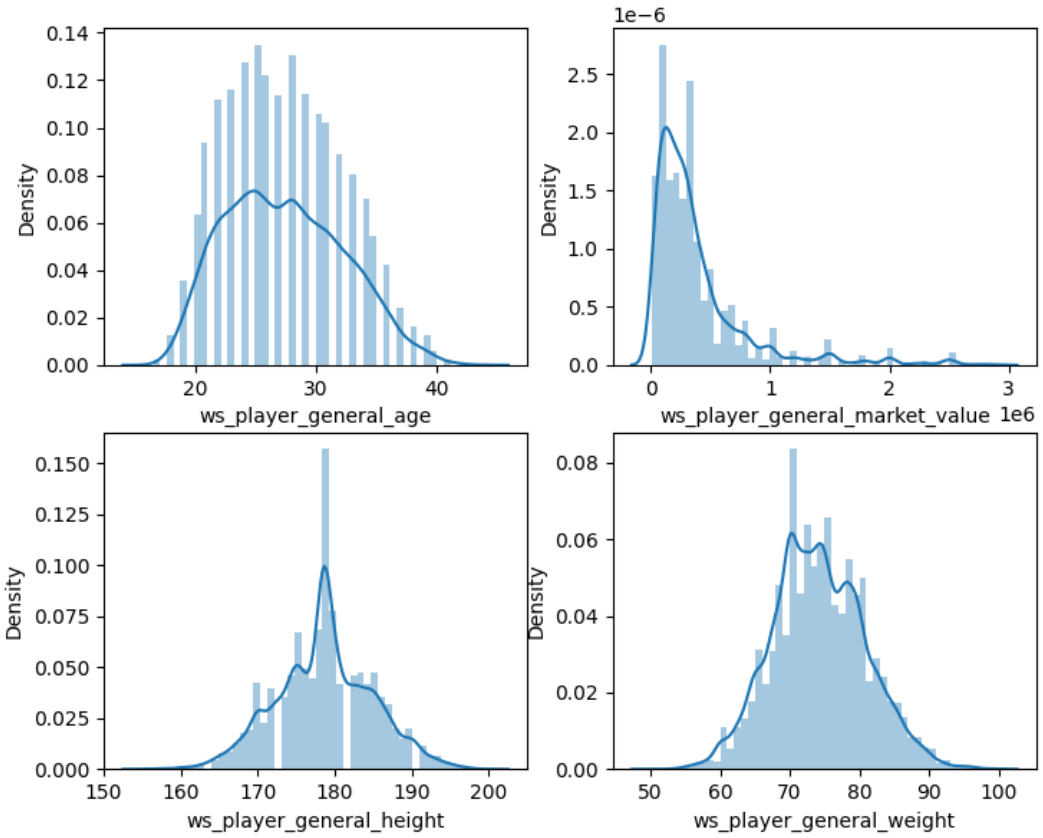

Figure 4.3: Distributions of general numerical variables in the dataset.

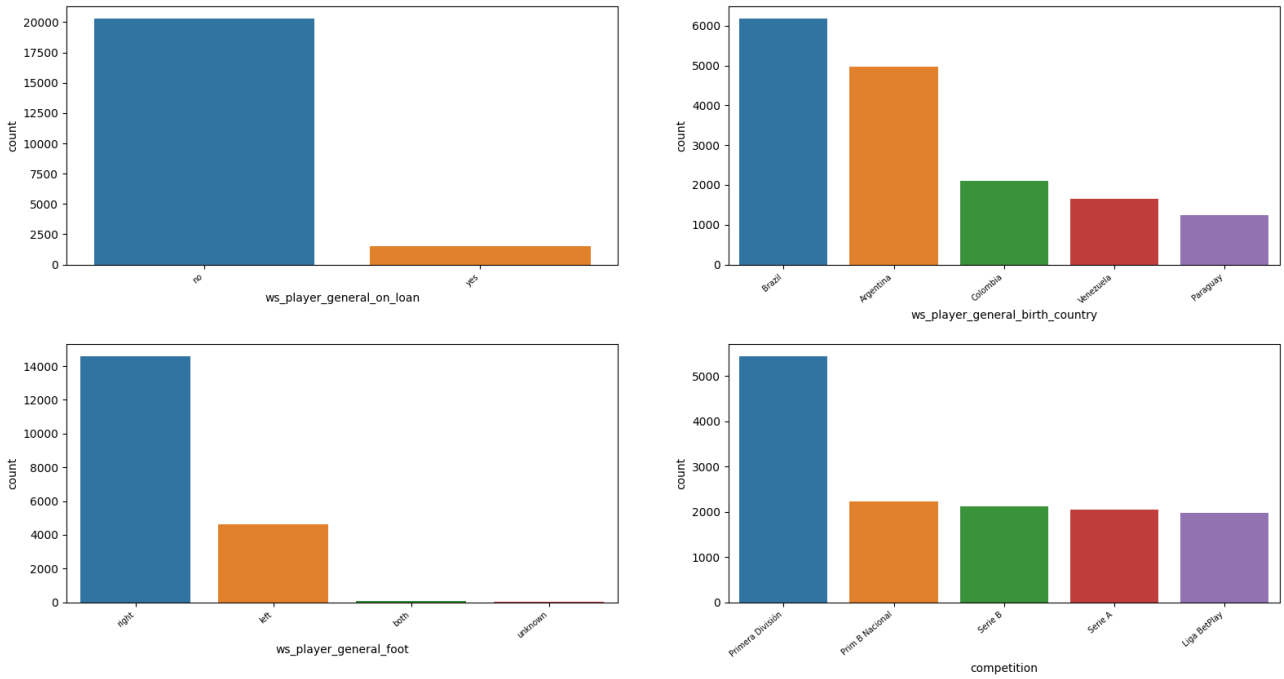

Figure 4.4: Count plots of the categorical general variables in the dataset.

Differences in characteristics of position groups can be seen in Figure 4.5. Spider charts are generated which show the average measure for certain attributes that make it easier to understand what type of player a player in a specific position group is. Five attributes that are comparable as well as interesting in terms of divisibility of player characteristics are compared between the different position groups. For example, the average number of forward passes decreases when moving towards more attacking position groups, where the average number of duels increases.

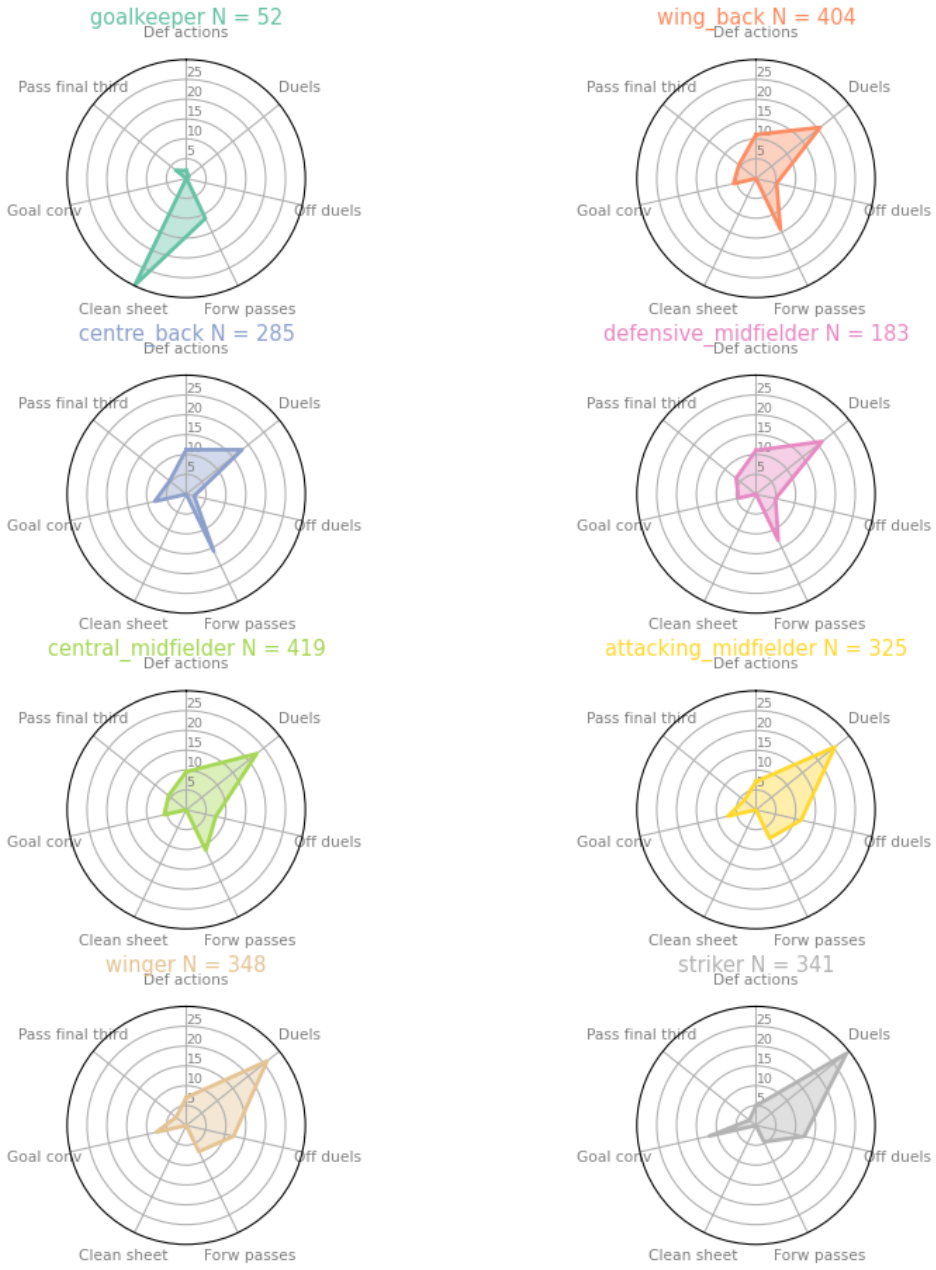

Figure 4.5: Spider charts with average measures for five attributes dividing the players in groups. (N is sample size)

Performance attributes

Apart from the more general information on the player, performance attributes are represented in the dataset as well. These can be divided into 5 categories explained in Table 4.3. This also includes the main concepts of the attributes and their explanation.

Table 4.3: Explanation of performance attributes in the dataset.

General	Matches played	Number of games played during that specific season
	Minutes played	Number of minutes played during the season
	Duels	Duel for the ball between the player and an opponent
	Free kicks	Standard situation where a player can freely kick the ball from a certain position
	Direct free kicks	A free kick that is allowed to go directly to the opponent's goal
	Corners	Standard situation where the ball can be played into the box from the corner of the field
	Penalties	Standard situation where the player can kick the ball at the goal from the penalty spot
Attacking	Goals	The number of goals made in that season
	XG	The expected number of goals made
	Successful actions	The number of successful attacking actions
	Non penalty goals	A goal not coming from a penalty kick
	Head goals	A goal scored with the head
	Shots	The number of shots taken
	Shots on target	A shot taken at the opponent's goal
	Goal conversion	A shot being converted into a goal
	Assists	The final pass to the player that scored a goal
	Crosses	A (high) pass towards another player
	Crosses from left	Cross from the left side of the pitch
	Crosses from right	Cross from the right side of the pitch
	Crosses to goalie box	Cross that arrived in the opponent's penalty area
	Dribbles	A run with the ball
	Offensive duels	Duel in the attack
	Touches in box	Touch of the ball in the opponent's penalty area
	Progressive runs	A run towards the opponent's goal
	Accelerations	Speed up of the player with the ball
	Received passes	Passes received from another player
	Received long passes	Long pass received from another player
	Fouls suffered	An offence from an opponent player
Defensive	Duels	Duel in the defence
	Actions	Defensive action to try and win the ball
	Aerial duels	Two players competing for the ball in the air
	Sliding tackles	Slide on the ground to dispossess the opponent
	Shots blocked	Blocking a shot of an opponent
	Interceptions	Active interception of a pass or shot of the opponent
	Fouls	A committed offence on an opponent player
	Yellow cards	Disciplinary action, as a received warning
	Red cards	Disciplinary action, being sent off the field
Passing	XA	Expected assists
	Passes	Pass to a teammate
	Forward passes	Pass to a teammate towards the opponent goal
	Back passes	Pass to a teammate backwards
	Lateral passes	Pass in the lateral direction
	Short-medium passes	Pass in a short to medium distance
	Long passes	Pass in a long distance
	Average pass length	Average length of passes taken

	Average long pass length	Average length of long passes taken
	Shot assists	The final pass of a ball before a shot being taken
	Second assists	Last pass before an assist
	Third assists	Last pass before a second assist
	Smart passes	Creative and penetrative pass to break the defence
	Key passes	Pass that creates a goal scoring opportunity
	Passes to final third	Pass towards the final third of the pitch
	Passes to penalty area	Pass to the penalty area of the opponent
	Through passes	Pass played into the space behind the defensive line
	Deep completions	A non-cross pass to the opponent's penalty area
	Progressive passes	A pass to advance the team closer to the goal
Goalkeeping	Conceded goals	Goal scored by the opponent
	Shots against	Shot taken by the opponent
	Clean sheets	Zero goals being scored against the team in a game
	Save rate	Shots saved as percentage of total shots taken
	XG against	Expected conceded goals
	Prevented goals	Goal being prevented by a save or interception
	Back passes received	Pass from a teammate received
	Exits	Exiting the ball away from the goal
	Aerial duels	Competing to win the ball with an opponent player

In the table above, a general overview of the attributes and their concepts are given. In the dataset, the clean sheet percentage variable is engineered from the total number of games played and the number of clean sheets as original attribute value. Additional attributes are represented with either 'per_90', 'padj', or 'pct'. 'per_90' transforms the variable into a measure per 90 minutes (one full football game). For example, 'ws_player_attacking_goals_per_90' gives the average amount of goals per 90 minutes played. 'padj' gives the possession adjusted value for defensive actions mostly. Possession adjusted attributes are generalized based on the effective gametime and possession of the opponent ("Possession-adjusted," 2022). Not all teams have the same amount of possession, which has an impact on defensive count statistics. For example, 'ws_player_defensive_padj_sliding_tackles' gives the average sliding tackles per player, based on 50%/50% possession between the two teams. For a player that plays in a more defensive team, the statistic will thus be lowered, as he had more opportunities to tackle his opponents as they were attacking most of the time. 'pct' transforms the variable into a percentage completion variable. For example, 'ws_player_attacking_shots_on_target_pct' gives the percentage of shots taken by the player that were on target (shots on target / shots taken).

Relations between attributes

Per category, the performance attributes for the most insightful position group will be visually described below. Density histograms of the numerical variables are not shown but are used to pre-process the data for better use in the clustering models and better visualisation as explained in Chapter 4.1.3. Focusing on a specific position group, correlations with the market value can help to indicate important attributes for players playing in a specific position (group) and will be shown below. For strikers for example, the relatively high correlation between goals per 90 minutes and market value indicates that a high scoring striker is valued higher than low scoring strikers, which is intuitive. These more important variables can then be weighted more highly in the design of the dissimilarity metric, which will show the similarity information between players as part of the goal in this thesis. A general rounding approach is considered where the important attributes get the high weight, and the unimportant attributes get a low weight. Depending on the domain knowledge, the attribute group

will be weighted as highly important (weight of 1), normal importance (weight of 0.5) or unimportant (weight of 0).

A correlation analysis on goalkeeping attributes is applied for goalkeepers, as seen in Figure 4.6. The values that are presented in this figure show the correlation between the attributes of the row and the column. A high value, such as seen for conceded goals per 90 minutes, and shots against per 90 minutes, thus indicates a high positive correlation. The correlation between attributes is presented between -1 and 1, where -1 shows a high negative correlation, 0 shows no correlation, and 1 represents a high positive correlation. The legend on the right in the figure shows the colour codes of the correlations, which are also presented in the figure cells. With this information, we can thus identify which attributes correlate to each other, and which attributes might have a positive or negative impact on other types of attributes.

Intuitive correlations are found in this plot, with positive correlations with the market value for preventing goals or having clean sheets. Slightly less intuitive is the slight positive correlation between market value and received passes, which might indicate ball playing goalkeepers to be more valuable. Finally, this could be further argued because of the slight negative correlation between market value and the number of exits per 90 minutes. Compared to the other attributes however, these absolute correlations are lower, and thus valued less in analysis of the goalkeepers.

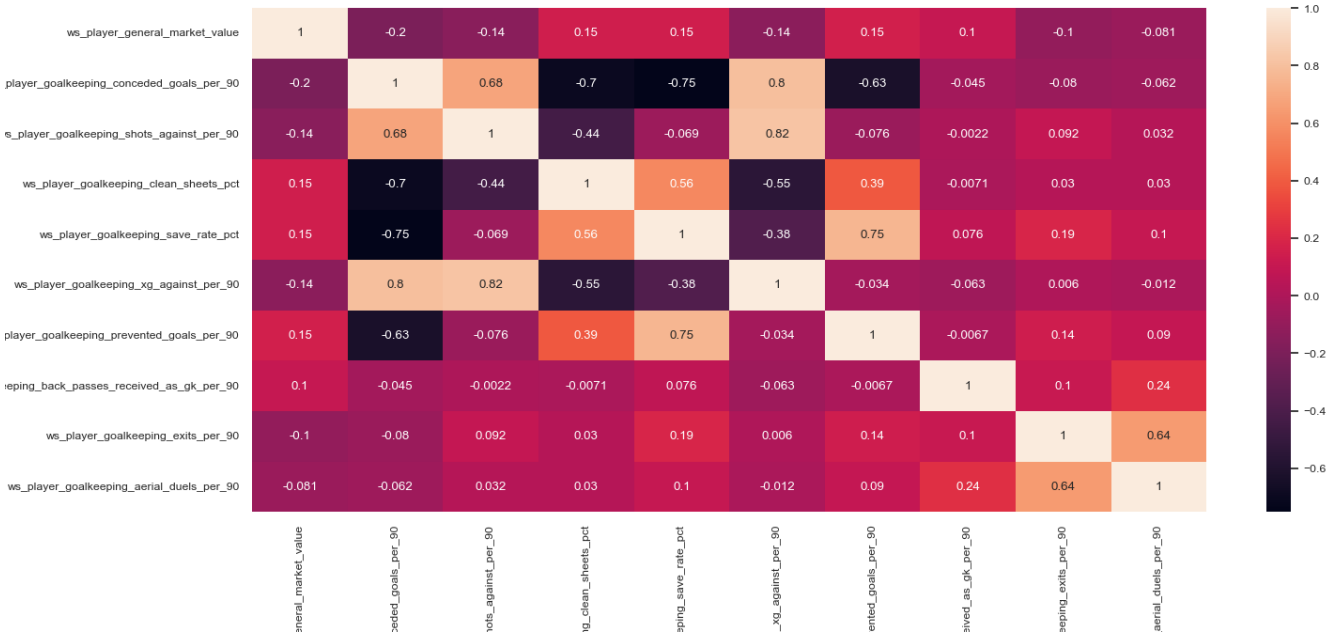

Figure 4.6: Correlation between goalkeeping attributes of goalkeepers.

A correlation analysis on attacking performance attributes is applied for strikers, as done for the goalkeeping attributes for goalkeepers. The correlation plot is explained below and shown in Appendix Figure A.1. A high correlation is for example seen for attacking goals per 90 minutes, and penalty goals per 90 minutes, thus indicates a high positive correlation. This suggests that players that have a lot of penalty goals per 90 minutes, will also have a lot of goals per 90 minutes (as expected).

Like other attacking position groups (wingers), a slight negative correlation between headed goals and the market value is found. Even though the dataset is split up per position group, it might also be interesting to focus on different types of players per position group. In this example, a target man that is more focused on headed goals compared to other types of goals, is expected to have more importance on that attribute as well. Furthermore, interesting to see is the positive correlations

between progressive share in possession, such as successful attacking actions and progressive runs, and the market value. Bringing the ball forward to more valuable positions seems to be an important factor in player attributes.

What also can be extracted from Figure A.1 is the high correlation between certain attributes. For example, correlations are found between attacking goals, non-penalty goals, penalty goals and headed goals. This confirms the data characteristics where certain attributes are split in more specific attributes. The overarching attribute (in this case goals) can be discarded as it is already represented in the more specific attributes. This is the same for crosses and crosses from the left or right flank or crosses towards the goalie box. High correlation between statistics also suggest that these attributes together do not divide players in the dataset much more than one of these attributes will do. This suggests what comes forward in the principal component analysis, where the components are combined which together give a higher explained variance of the points in the dataset. The attributes with high correlation are expected to be found in the components that do not give a high explained variance on the players in the dataset.

Other interesting correlations are found by analysing the other position groups. For players that are more focused on the strikers, higher positive correlations with the market value are found in attributes such as received passes or crosses compared to goal scoring attributes, indicating that it can be more important for these players to create chances rather than finishing them.

A correlation analysis on defensive performance attributes is applied for centre-backs, as seen in Figure A.2. This is done to identify which attributes are important for defenders, which attributes correlate with each other, all in preparation of how to approach these attributes in handling of the data and method application. What is interesting to see is that the number of defensive actions such as duels, tackles, or fouls are slightly negatively correlated to the market value, where the percentage won of these attributes have a slight positive correlation. This indicates that it might be most important for defenders to not be too aggressive in duels and be calmer and more collected assuring that a duel will be won. Obvious negative correlations are found between fouls or cards and market value.

Looking at other position groups, the correlations between defensive attributes and the market value decrease and are generally negative. Besides the intuitive idea that defensive attributes are less important for attacking oriented players, more defensive active players might be less valuable focusing on attacking positioned players.

A correlation analysis on passing attributes is applied for central midfielders, as seen in Figure A.3. Positive correlations are found mostly on shorter distance passes that progress the play compared to long passes or crosses. This might be the result of these passes being too risky in keeping possession of the ball and thus not desirable.

Looking at more offensively oriented players, the correlations between passing attributes and market value is lower. For defenders however, these correlations are similar or even higher. Passing attributes are one of the main attributes for centrally positioned players, and less so for attackers, as further argued by these correlation values. Also, what is interesting to see, there is a negative correlation between market value and third assists

4.1.2 Choosing variables for clustering

Based on the previous section, combined with the requirements set up in Chapter 3.1, we can narrow down the variables to use for the rest of the method. This is an important step in improving the focus and the speed in the cluster modelling.

The attributes present in the dataset are extensive, with a total number of 122 attributes or columns. Based on the exploration of the dataset in the first stage, combined with the requirements, certain variables can be removed.

The 'region', 'continent', and 'confederation' columns do not give additional information or value to the players in the dataset, as the values for all players are the same. The players play in South America, in the Conmebol confederation, which is a given for the further analysis of the dataset.

Because of the difference in matches and minutes played by players in the dataset, count variables do not give sufficient information on their own. Attributes such as the total goals scored, do not give an insight into the performance of the player, if the number of games played is not considered. Because of this reason, the count attributes are discarded from the dataset, and we will further focus on statistics standardized to per 90 minutes, or percentages.

The dataset is reduced in dimensionality, with the non-performance related attributes being discarded. The non-performance related attributes are not considered within the modelling of the clustering algorithms as they do not give information about the performance of the players and thus do not aid in the goal of scouting players that are more likely to replace the to be replaced player in the current squad. These attributes are however still important to get an understanding of the dataset and support the decisions in grouping the players in certain position groups for example. The position changes of players are incorporated within the model, as all played positions by that player are included within the dataset. For example, a player that played the main part of the season as a midfielder, and three matches as a winger, will be included in the analysis of midfielders as well as of wingers. Because of this, the different possible to be played positions of a player are incorporated within a model. The information of the player for the season will decide which positions are played, and thus are incorporated within the model. With further dimensionality reduction analysis in the next two stages, the selection of attributes used for clustering will be finalized.

4.1.3 Pre-processing data

Cleaning data

In Chapter 4.1.1, statistical descriptions of certain variables are visually presented. In data pre-processing it is important to analyse the distribution of attributes to prepare the data for statistical analysis. Outliers can for example endanger the results of clustering in this method. Outliers resulting from the statistical description of the attributes are further analysed. Because of the variate nature of a lot of the count and percentage attributes, outliers are visually checked and evaluated. Because players are likely to have percentage attributes valued from 0 to 100, in case a player only had a couple occurrences of that specific event, visual inspection of the attribute distribution is required. These steps are all taken throughout the exploration of the data in the first stage.

Null values and outliers

For rows (players) that contain attributes with null values or values that are outliers within the dataset, multiple approaches can be followed. The rows of data can be discarded in case the data is not deemed valuable anymore because of the null values or outliers. Apart from discarding the rows, the null values or outliers can also be replaced with a particular value. This value can be for example the mean, median, or mode, a completely new category (in case of categorical data), or a value predicted based on other attributes. To keep a pure dataset, rows of data that include null values in relevant columns (for relevant attributes), will be discarded. Null values are thus decided not to be replaced with other values.

Because of the high dimensionality of the data (122 attributes per player), there are missing values in some of these attributes for specific players. For numerical variables, null values will lead to removal of the row generally (because of inaccurate information). Outlier values are visually inspected, as it is probable that these values are still valuable for the algorithm, for example where the top goal scorer of the competition might be an outlier, but still with valid values. Certain variables can, based on domain knowledge, be predicted by other variables. For example, the weight of the player is correlated with the height of the player (if the player is taller, he is more likely to be heavier as well). Furthermore, the number of successful defensive actions of a player will most likely have a correlation with the number of defensive duels of that player.

For the height and weight of the players, the one attribute is correlated with the other attribute, which makes it possible to predict the missing values. If both are missing, the average of the dataset can be taken. As opposed to what is said before, these missing values are not used in analysis of the player performance, but only used as a general overview of the player once they are returned from the analysis for example. There is thus no impact from this method on the clustering and dissimilarity methods itself.

Finally, the skewness and chaos in some of the attribute distributions are the result of zero values in the dataset. For example, defenders are not likely to score headed goals, so there are a lot of datapoints for the variable 'ws_player_attacking_headed_goals_per_90' with the value of '0'. This variable, however, is mostly important for attacking players, which is why the influence of the defenders on this variable needs to be reduced to also reduce the skewness of the distribution. This will be further explained in the next sections, where attributes are weighted per position group, such that the distributions of the attributes are less skewed and avoid the above problem. This will be further explained in the data transformation and standardisation section.

Feature engineering

Because of the deletion of count variables, certain types of statistics might not be represented in the dataset anymore. Most variables are already represented in the dataset standardized to per 90 minutes (standardized to the standard game time of a match), or in percentage accurate. The only attribute that is not represented in a standardized way is the clean sheets of the goalkeepers. To construct this variable, the number of total clean sheets is divided by the total number of games played, to get a percentage of clean sheets per game (the probability of a clean sheet for that goalkeeper). This is represented in the variable called 'ws_player_goalkeeping_clean_sheets_pct'. The other attributes represented in the dataset give an in-depth view on the performance and characteristics of the players.

Integrating data

Data comes from one dataset and does not need to be integrated with other datasets.

Data transformation and standardisation

The first data transformation applied on the dataset is the 'per_90' and 'pct' transformations. All count variables are transformed to 'per_90' attributes with the help of the total minutes played by that specific player, as previously explained.

Distances between players are the result of total distances in attributes combined in a distance measure, as will be further analysed in stage 4. The distances between players on a particular attribute should represent how different players are in real life. As seen in the exploratory data analysis and plots of the distributions of the attributes, a lot of attribute distributions are skewed (to the right). This is the result of a lot of players not 'participating' in a certain attribute as much as a smaller other

group of players. For example, the smaller group of defenders in the dataset will represent most of the high values for defensive attributes such as defensive sliding tackles. A larger group of attacking midfielders or attackers will have very low values for these attributes, which makes the distributions skewed to the right. Reducing this skewness of attribute distributions is vital, as small differences in low values for these types of attributes should represent the same difference as higher differences in high values. Non-linear transformations are required to make differences between for example 0 and 0.5 defensive sliding tackles per 90 minutes similar to the difference between 2 and 3 defensive sliding tackles. Non-linear transformations change linear relationships. For example, taking the square root of an attribute will make the differences in originally higher values, lower relative to the differences in lower values. For this purpose, the square root is applied as a non-linear transformation to highly right skewed attributes following from the exploratory data analysis in the first stage. Similarly, the square transformation is applied to highly left skewed attributes.

Standardizing the variables can be used to reduce the impact of scaling on dissimilarity metrics and make the variables comparable. We scale all important variables that are already transformed into variables between 0 and 1 (pct to 0-1 and count per 90 to 0-1).

Different attributes can have different importance in assessing the performance of the player. This is dependent on the position a player plays in, the type of football a team plays, what kind of player is required, etc. Because of this, the final tool will include the possibility to adapt weights of variables to customize the results for the scouts' preferences.

Variable selection and dimension reduction

Attributes that are deemed useless through domain knowledge and visual exploration of the data, can be deleted to reduce dimensionality, and improve speed in the algorithm modelling. For example, the 'continent' and 'region' attributes both give the same value for all players in the dataset.

Players with a market value higher than 3 million euros are filtered out of the results of the algorithms, to only analyse players within the transfer budget of San Lorenzo. This market value is considered as a representative of the real value of the player in the case a transfer would occur. This real value depends on multiple additional factors such as the players' desire to leave or the importance of the player within the squad. Also, in case of a player having played only a small number of games, the information is unreliable. For the last case, we make a threshold of players needing to have played at least 300 minutes, and at least 5 games. Data is unreliable because certain attributes are denoted 'per_90', meaning it is standardized to per 90 minutes. In case a player would have only played 10 minutes and scored one goal, the goals per 90 minutes of this player would be 9, which is an improbable statistic based on domain knowledge as decided together with the stakeholders. The group of players with limited gametime are thus not considered as they can spoil the analysis through unrealistic attribute values. For players with more gametime, the statistics are standardized and comparable. Generally, the dataset will include players that have played a lot of matches, but relatively less minutes compared to other players. These will be players that for example often get substituted within the field for the last 30 minutes of the game. Expected is that there is a difference in how a player will perform getting subbed on 30 minutes for 3 games, compared to a player playing one game for 90 minutes. This could give slight biases within the analysis, however, because of the quantity of the data and the rarity of these players, the impact is expected to be limited. The players having played at least 300 minutes on their positions, will at least give an indication of the player's performance on a specific position within the field. This is further discussed in Chapter 6.

As already explained in the previous stage, count variables are disregarded from the dataset because of the lack of value of these individual attributes. The 'per_90' or 'pct' attributes are considered

because of the standardized nature of these attributes. Count variables that do not have corresponding 'per_90' attributes, are added into the dataset. Variables such as height, weight, and age are not considered in the clustering and dissimilarity analysis because these are not performance metrics but can be used to filter a type of player if wanted. This will be available in the tool to check up on a specific type of player.

Within the following stage, further dimensionality reduction is applied to design a dissimilarity metric.

4.1.4 Designing dissimilarity metric

Dissimilarity metrics will be used for representation of differences between players, as well as for input to the clustering methods. This will thus lead to dissimilarity scores as part of the objective of this thesis. All to be considered attributes are numerical variables. These variables are numerical ratio variables, as the ratio between two possible values is more relevant compared to the difference between the two values. In pre-processing of the data, the variables are normalized and scaled such that they are transformed into continuous variables between 0 and 1. The distance between players considering these attributes can be measured with distance measures as shown in Table 3.3. For the different position groups, the attributes will be weighted differently based on the domain knowledge and correlations found in the exploratory data analysis in the first stage. The weights are the result of visual analysis of the correlation plots of the specific position group attributes with the market value. High correlations between the attribute and the market value suggests a higher importance for that variable for that position group compared to an attribute that has little effect on the market value of a player. Because of these reasons, the different attributes do not have the same weight or value in measuring similarity between players. With the weighted Manhattan distance measure, a slight variation on the Manhattan distance is used for measuring the dissimilarity between players considering the weight of different attributes. The Manhattan distance is one of the dissimilarity metrics for numerical variables as explained in Chapter 2.2.1, which is useful as the to be considered attributes are numerical. The Manhattan distance is chosen because it is one of the most widely used measures, it is effective in high-dimensional data, and because of its integrated use within clustering algorithms in Python. (Akhanli, 2019)

Weighting of attributes can be used to focus the analysis on more important statistics for the player position group that is analysed. Within the evaluation of the clustering model and dissimilarity metrics, the analysis including weighting is compared with the analysis excluding weighting. The expected difference will be that the differences between players in specific position groups will be easier to evaluate if weighting is applied, as the characteristics that form the player performance for that position are focused on. The weighting of the different attributes depends on the position group that is analysed. Based on the exploratory data analysis and domain knowledge, the factors for (groups of) attributes are decided and shown in Table 4.4. The weight factor is the number with which the distance will be multiplied to appoint either more or less weight and importance to an attribute. Groups of attributes and individual attributes are weighted, as many of the grouped attributes will have similar weights, with certain exceptions individually noted. For attacking minded players, attacking attributes will intuitively be more important in evaluation, while for defensive minded players, defending attributes are intuitively more important. Furthermore, certain higher positive or negative correlations between attributes and the market value are found for specific position groups which will lead to specific differences in weighting. To deal with players that can play in different positions, such as a winger with striker potential, the player will be incorporated within the analysis if they have played even one game within that position. A winger with striker potential will probably have played within the striker position and will thus come forward in the analysis if they are similar to the to be

compared player. The weighting of these attributes in the application of the tool will be subject for change based on the user's input and will thus be dynamic.

Table 4.4: Weighting of (groups of) attributes per position group.

Position group	(Type of) attribute	Weight factor
Goalkeeper	Performance – goalkeeping (group of 6 attributes)	1
	back_passes_received_as_gk_per_90 (individual attribute)	0.5
	exits_per_90 (individual attribute)	0.5
	aerial_duels_per_90 (individual attribute)	0.5
	Performance – others (group of 63 attributes)	0
(Wing)backs	Performance – defensive (group of 11 attributes)	1
	Performance – passing (group of 24 attributes)	1
	long_passes_per_90 (individual attribute)	0.5
	average_pass_length_m (individual attribute)	0.5
	average_long_pass_length_m (individual attribute)	0.5
	deep_completions_per_90 (individual attribute)	0.5
	deep_completed_crosses_per_90 (individual attribute)	0.5
	Performance – attacking (group of 23 attributes)	0.5
	Performance – goalkeeping (group of 9 attributes)	0
Centre backs	Performance – defensive (group of 11 attributes)	1
	Performance – passing (group of 24 attributes)	1
	long_passes_per_90 (individual attribute)	0.5
	average_pass_length_m (individual attribute)	0.5
	average_long_pass_length_m (individual attribute)	0.5
	deep_completions_per_90 (individual attribute)	0.5
	deep_completed_crosses_per_90 (individual attribute)	0.5
	Performance – attacking (group of 23 attributes)	0
	Performance – goalkeeping (group of 9 attributes)	0
Defensive midfielders	Performance – defensive (group of 11 attributes)	1
	Performance – passing (group of 24 attributes)	0.5
	long_passes_per_90 (individual attribute)	0
	average_pass_length_m (individual attribute)	0
	average_long_pass_length_m (individual attribute)	0
	deep_completions_per_90 (individual attribute)	0
	deep_completed_crosses_per_90 (individual attribute)	0
	Performance – attacking (group of 23 attributes)	0.5
	Performance – goalkeeping (group of 9 attributes)	0
Central midfielders	Performance – defensive (group of 11 attributes)	0.5
	Performance – passing (group of 24 attributes)	0.5
	long_passes_per_90 (individual attribute)	0
	average_pass_length_m (individual attribute)	0
	average_long_pass_length_m (individual attribute)	0
	deep_completions_per_90 (individual attribute)	0
	deep_completed_crosses_per_90 (individual attribute)	0
	Performance – attacking (group of 23 attributes)	0.5
	Performance – goalkeeping (group of 9 attributes)	0
Attacking midfielders	Performance – defensive (group of 11 attributes)	0.5
	Performance – passing (group of 24 attributes)	1
	long_passes_per_90 (individual attribute)	0.5
	average_pass_length_m (individual attribute)	0.5
	average_long_pass_length_m (individual attribute)	0.5
	deep_completions_per_90 (individual attribute)	0.5

	deep_completed_crosses_per_90 (individual attribute) Performance – attacking (group of 23 attributes) Performance – goalkeeping (group of 9 attributes)	0.5 1 0
Wingers	Performance – defensive (group of 11 attributes) Performance – passing (group of 24 attributes) long_passes_per_90 (individual attribute) average_pass_length_m (individual attribute) average_long_pass_length_m (individual attribute) deep_completions_per_90 (individual attribute) deep_completed_crosses_per_90 (individual attribute) Performance – attacking (group of 23 attributes) Performance – goalkeeping (group of 9 attributes)	0 1 0.5 0.5 0.5 0.5 0.5 1 0
Strikers	Performance – defensive (group of 11 attributes) Performance – passing (group of 24 attributes) long_passes_per_90 (individual attribute) average_pass_length_m (individual attribute) average_long_pass_length_m (individual attribute) deep_completions_per_90 (individual attribute) deep_completed_crosses_per_90 (individual attribute) Performance – attacking (group of 17 variables) crosses_from_left_flank_per_90 (individual attribute) crosses_from_left_flank_pct (individual attribute) crosses_from_right_flank_per_90 (individual attribute) crosses_from_right_flank_pct (individual attribute) crosses_to_goalie_box_per_90 (individual attribute) received_long_passes_per_90 (individual attribute) Performance – goalkeeping (group of 9 attributes)	0 0.5 0 0 0 0 0 1 0.5 0.5 0.5 0.5 0.5 0.5 0

For purposes of further dimensionality reduction, faster computation, and visualisation of clusters in analysis, Principal Component Analysis is applied. The correlated variables in the final selection of attributes for the specific position group are replaced by fewer principal components that have lower correlation. In the application of the model, a minimum percentage of the variance will be represented by the resulting principal components. As an initial number, 60% will be used. However, different thresholds will be analysed to see where the trade-off between variance and number of features is generally optimal. This should be the percentage where a lot of the variance is represented in a small number of components. In the evaluation of the model, this is assessed and further optimized. The dimensionality reduction due to the Principal Component Analysis, thus allows for faster computation in the application of the clustering algorithms.

To return one value, the distances between players based on the variables are aggregated. Based on the weighting of the attributes, the dissimilarities are aggregated as shown in Equation 2.5.

4.1.5 Choosing clustering method

A fitting clustering method must be chosen for application in the modelling of the tool to achieve accurate results. The characteristics of different clustering algorithms are compared in Table 3.3. Through extracting the characteristics from the to be used dataset and considering the requirements of San Lorenzo, a suitable clustering algorithm can be chosen when evaluating these characteristics with the clustering techniques in Table 3.3.

The clustering analysis will focus on specific position groups, which lowers the initial size of the dataset. However, the many to be considered players and attributes still entail a large-scale dataset.

The dimensionality of the dataset is also large, as many attributes are considered in the dissimilarity metric. However, with application of the PCA algorithm as explained in the previous section, this dimensionality can be reduced. Time complexity is preferred to be low, as the tool should be able to be used quickly and effectively by the scouting team of San Lorenzo. The choice on these algorithms come from literature study in Chapter 2.2.2 where the suitability of these clustering algorithm based on the above characteristics is evaluated in Table 3.3. The clustering algorithms should be relatively scalable, work on large and high-dimensional datasets as is a characteristic of the football player dataset (many rows and columns), and not be sensitive to possible outliers. Based on the above characteristics and usability in programming tools, the following clustering algorithms are further analysed and evaluated in application on the dataset (next stages).

- **CLARANS**
  CLARANS is a partition clustering method. The datapoints are clustered based on optimizing the dissimilarity metric. It is a partition around medoids method, but more time-efficient through use of subsets. Iteratively, points are chosen as medoids and one of the medoids and a random other point are chosen for evaluation in terms of distance to the clusters. By choosing the medoids with the closest total distances, local minima can be found. In iterations, multiple local minima are found, and the best local minima is returned at the end of the algorithm. The use of medoids reduces the sensitivity to outliers. A balance is kept between computational costs and the effects of sampling data on the formation of the clusters. (Ng & Han, 2002)
- **CURE**
  CURE is a hierarchy clustering method. In this method, a balance is kept between centroids and extremes of the points to avoid problems with non-uniform clusters. Groups of scattered points are shrunk and represent the 'old' group of points. The closest pair of these representatives are then merged in steps. This makes it more robust to outliers. The algorithm is efficient for large databases, as the dataset to be analysed for the objective of this thesis. (Guha, Rastogi, & Shim, 1998)
- **DBSCAN**
  DBSCAN is a density-based clustering method. Because it is based on density of points, it performs well on detecting and clustering outliers. This, however, is already dealt with in pre-processing of the data in the used method. Furthermore, it is not required to input a number of clusters within the algorithm as opposed to other algorithms. The algorithm works with the idea that a cluster of datapoints have a high density. Points are density reachable if they are within a certain distance (eps, input to the algorithm) from each other. Points are connected if they are reachable through other points. Iteratively, points are picked in the dataset, and if there are a minimum number (minPoint) of points density reachable, they are considered in the same cluster. Continuing this algorithm provides clusters where points are connected. (Ester, Kriegel, Sander, & Xu, 1996)
- **GMC**
  GMC is a probability-based clustering method. The resulting clusters can be interpreted as more intuitive in terms of shapes, based on the used dataset. See Chapter 2.2.4 for a more detailed explanation.

The above-mentioned clustering methods will be used in the following stages and simultaneously evaluated for relevancy in this thesis. This is done because there is no clear distinction in suitability of the clustering methods with the objectives of this thesis. The evaluation of the four clustering methods

will support the final decision on which clustering algorithm is most suitable for the objectives of this thesis.

4.1.6 Determining number of clusters

Most of the clustering algorithms require user input for the number of clusters to be returned by the algorithm. The in the method described Elbow method based on the WCSS measures is used for each cluster requiring user input for the number of clusters. This algorithm is implemented in the code to automatically extract the optimal number of clusters in the clustering algorithm.

The code loops through multiple options for the number of clusters, from 1 to 20. For every number of clusters, the method is executed. This results in the cluster of players with the characteristics of this cluster model, including the positions of the players (datapoints) and which cluster they are appointed to. The method described in Chapter 3.2.6 is used to calculate the relative strength. The place with the highest relative strength then determines the optimal number of clusters for this analysis method.

4.1.7 Implementing clustering method, acquiring results and model evaluation

As a result of the previous stages, enough information is acquired for implementation of the clustering algorithms. These algorithms are implemented in Python and the clusters of players as well as dissimilarity metrics are returned.

Results clustering position groups

Examples of results are shown below. For example, if a specific position group is analysed with the clustering algorithms, results as shown in Figures A.4 and 4.7 are shown. These are presented to give an idea of how results follow from the dataset, and to get an idea which attributes and characteristics to focus on in analysis of players. Principal components are returned based on the most important attributes, the weighted Manhattan distances is calculated, and clusters of players are returned. Furthermore, if a specific player is given as input, the dissimilarity to other players is returned. Figure A.4 shows a selected number of goalkeepers from the dataset with the information on selected attributes and the resulting principal components. In Figure 4.7, this dataset is shown in graphs where the importance of the first principal component can be easily seen, as clustering is mostly based on that variable. The clustering based on the second and third components is less clear because of the lower variance coming from these components. The above gives sufficient information for use in development of the tool, where the user will be able to input a certain player to find similar players in the same cluster, as well as dissimilarity measures related to those players. Finally, Figure 4.8 shows the clustering graphs for central midfielders, done to give an idea of differences between position groups and the considered attributes and consequences thereof. Because there are many more attributes involved in the characteristics of a central midfielder compared to a goalkeeper, there are more principal components derived to have a high explained variance. Because of this, the clustering seems less obvious, as the importance of the components (the most important three analysed) is less compared to the case of the goalkeepers.

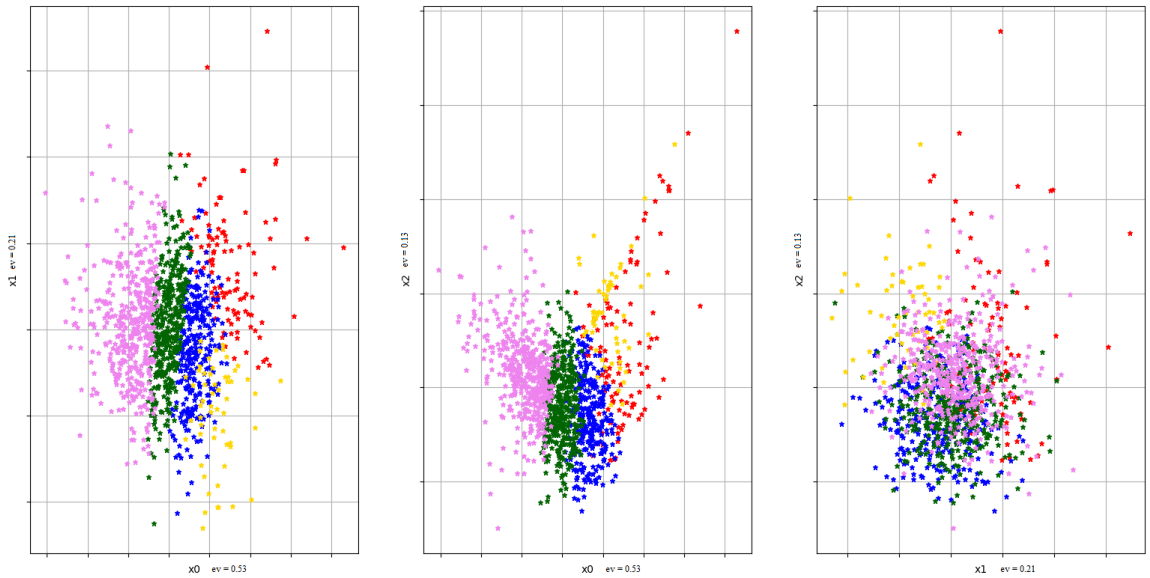

Figure 4.7: Clustering graphs for goalkeepers. 'x0' shows the value for the first principal component (with the highest included variance), 'x1' and 'x2' show the second and third principal components.

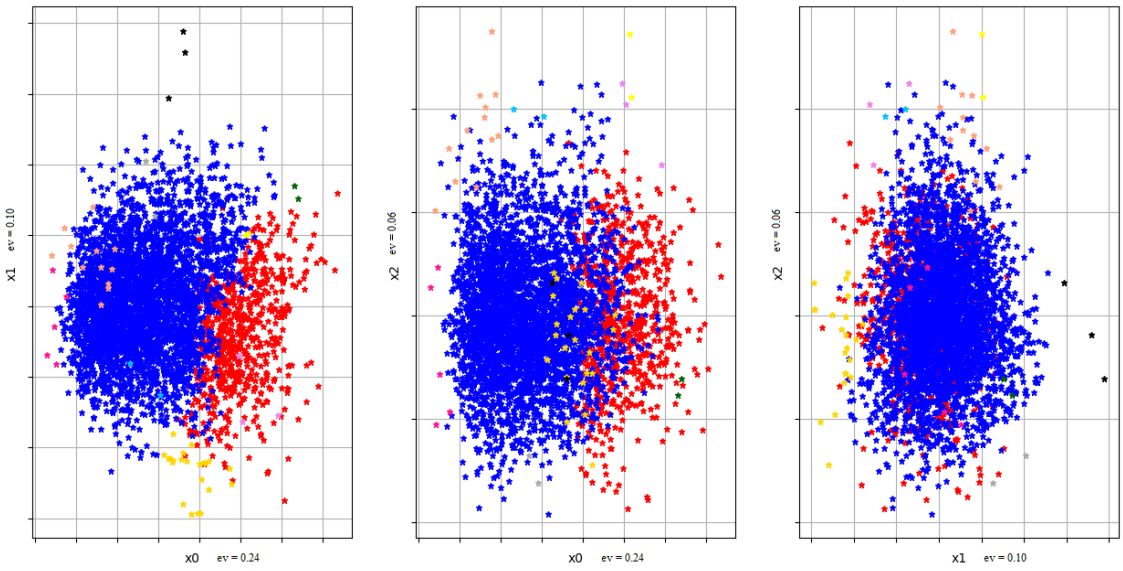

Figure 4.8: Clustering graphs for central midfielders. 'x0' shows the value for the first principal component (with the highest included variance), 'x1' and 'x2' show the second and third principal components.

The principal components are a reduced number of components from the original number of attributes characterizing the players. Figure 4.8 shows the clustering graphs for the central midfielders in the player database, and the explained variance of the first three principal components is shown. This thus gives an idea on how the applied PCA reduces the number of attributes to principal components and how they explain the players (data-points) with compositions of indicators. This is further explained below.

Principal component analysis

To explain a useful amount of variance from the principal components and cluster the players however, more principal components are used. In Figure 4.9, the explained variance ratio is presented as a function of the number of principal components, for the central midfielders. This shows that a trade-off needs to be made in reduction of the number of components, and the explained variance

(which is directly related to the clustering performance). The principal components that explain sufficient variance are finally chosen in the applied clustering models to improve the efficiency while keeping well-founded results. For the central midfielders, at a cumulative explained variance of around 0.8 and 20 principal components, an increasingly higher addition of principal components is required to improve the explained variance. For this reason, an explained variance of 0.8 can be concluded to be the optimal trade-off. Figure 4.10 shows the explained variance for the first 12 principal components, which gives an idea on how the added explained variance reduces as more principal components are added. Because of the high initial number of attributes (63 in this case), the number of principal components required is intuitively higher, as a similar percentage of attributes is required when compared to starting with fewer variables. In further evaluation of all position groups, these characteristics of the PCA analysis will be included in concluding what explained variance should be used as a threshold for input in the generally applicable tool.

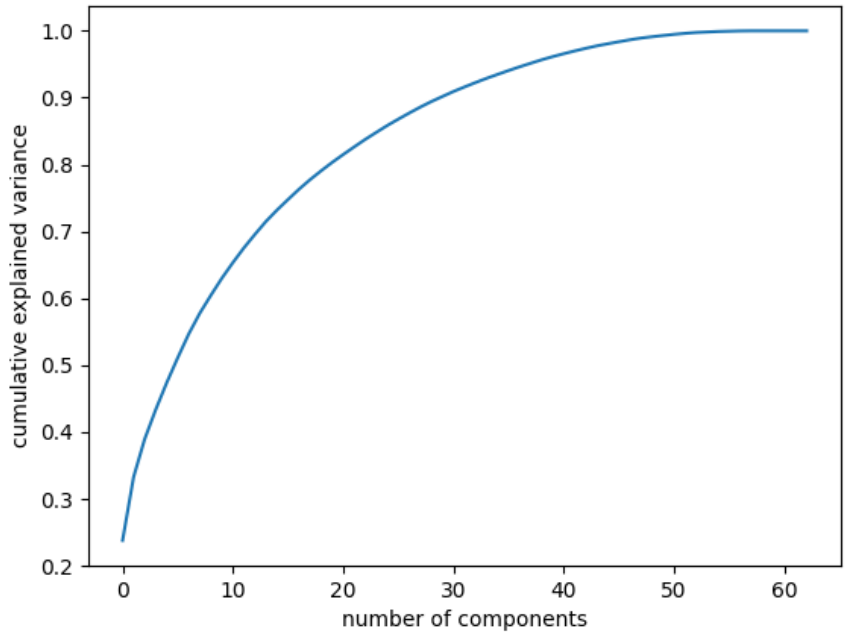

Figure 4.9: Cumulative explained variance for the central midfielder's position group as a function of the number of principal components.

```
[2.37985053e-01 9.48638833e-02 5.65095110e-02 4.40715621e-02
 4.03819711e-02 3.73813519e-02 3.56455712e-02 3.09570756e-02
 2.66744109e-02 2.56548819e-02 2.33691196e-02 2.20564494e-02
```

Figure 4.10: Explained variance per principal component for the central midfielder's position group.

Aside from the explained variance analysis in the principal components, the principal components are made up out of the attributes by identification of attributes that explain the variance of the dataset in themselves. These attributes are weighted in terms of how much they influence and contribute to the principal components. The groups of attributes contributing to the principal component are highly correlated and thus in itself form clusters which represent the principal component. In the execution of the principal component analysis, the eigenvector of the covariance matrix represents the coefficients of the linear combination of the original variables (the loadings) to construct the principal components. (Jolliffe, 2011)

Figures 4.11 and 4.12 show the sorted values of the weighting of attributes in the first two principal components in analysis of the central midfielder's position group, in order to give an idea of the most important and explaining attributes for a specific position group. The values represent the loadings or the weighting of the attributes in construction of the principal component. What we see from this figure is that the principal components are expressed by different attributes, where the first principal component is mostly described by several attacking attributes, the second principal component is mostly described by passing attributes. When looking back into the clusters shown in Figure 4.8, players distant on the first principal component axis are thus expected to have higher differences in attacking influence in the game.

To test this and show the working of the principal component analysis, the players with the highest and lowest values for the first principal components are compared. The player with the highest value is named 'X', and the player with the lowest value is named 'Y'. Expected will be that player 'X' and 'Y' differ a lot in the attributes that contribute heavily to the first principal component. The original dataset shows that player 'X' is offensive minded with many goals and assists (14 and 12 respectively in 29 games), 3.88 successful offensive actions and 12.11 key passes to the final third, but only 2.19 successful defensive actions (all per 90 minutes). Player 'Y' on the other hand is defensive minded with a lot of defensive actions and a limited number of goals (1 goal in 35 games), 8.67 successful defensive actions but only 0.17 successful attacking actions and 0.4 passes to the penalty area (per 90 minutes). This suggests that the principal components account for a distinction in types of players, where the similar players to player 'X' will have similarly high attributes for the given more attacking-minded attributes.

	PC1	PC2
ws_player_key_passing_passes_to_penalty_are...	0.27079	0.10675
ws_player_attacking_successful_attacking_...	0.26698	0.00098
ws_player_key_passing_xa_per_90	0.25583	0.08314
ws_player_attacking_offensive_duels_per_90	0.23766	-0.05774
ws_player_key_passing_shot_assists_per_90	0.23611	0.12566
ws_player_attacking_dribbles_per_90	0.23462	-0.07791
ws_player_attacking_touches_in_box_per_90	0.22186	-0.12803

Figure 4.11: Eigenvalue representation of contribution of attributes in the principal components, sorted by contribution to principal component 1 (top 7 attributes shown).

	PC1	PC2
ws_player_key_passing_passes_to_final_third_p...	0.01155	0.31398
ws_player_key_passing_progressive_passes_pe...	0.01829	0.29721
ws_player_passing_lateral_passes_per_90	-0.04213	0.29223
ws_player_passing_forward_passes_per_90	-0.02287	0.27607
ws_player_attacking_received_passes_per_90	0.02412	0.27533
ws_player_passing_accurate_short_medium_pa...	-0.10067	0.27072
ws_player_passing_short_medium_passes_per_...	-0.01398	0.26828

Figure 4.12: Eigenvalue representation of contribution of attributes in the principal components, sorted by contribution to principal component 2 (top 7 attributes shown).

Evaluation clustering methods

As explained in the method design, it is important to evaluate the models and find the most applicable clustering methods as well as variables required as input to the designed models. In this section, each clustering method is shortly analysed based on the silhouette values, number of principal components used and computational time, for different input of variables. This results in a general idea and standard implementation of the models for the design of the tool in the next stage. For all four clustering algorithms, silhouette values and computational times are denoted for different configurations, as shown in Tables 4.5-4.8. The silhouette values are given in the cells with, in brackets,

the number of features considered. The higher the silhouette coefficient, the higher the similarity or cohesion of datapoints within a cluster, which is the objective of clustering. Positive values close to 1 are good values for the silhouette coefficient, which denote good separation between clusters. Values closer to 0 denote clusters that have more overlaps. In this analysis, because the players are already grouped in their position groups and it is harder to distinguish between players, a lower silhouette value is expected as compared to when the dataset used for clustering would contain all different players. Because of this, all positive silhouette values are seen as good values. Finally, the computation time in minutes is given per position group per clustering algorithm. Three different types of position groups are included, who all have a distinct importance in included attributes. Because the position groups are entirely different, the importance of all attributes is expected to be included within the complete analysis. The EV value denoted in the tables refers to the explained variance threshold that is considered within the identification of the amount of considered features (principal components). A higher explained variance would thus include more principal components, as more principal components lead to a better representation (explained variance) of all included variables.

The results shown in Table 4.5 are presented because it is an initially interesting clustering algorithm. As shown under the central midfielder and winger position groups however, the clustering algorithm takes around one hour to complete, which makes it too slow in application within the scouting process. Because of this, the silhouette values are not evaluated anymore for the central midfielder and winger position groups, and the option to choose this clustering algorithm as the final algorithm is discarded.

For an explained variance of 0.8, differences in the analysis are considered between the attributes being weighted, versus the attributes not being weighted. The difference between these analyses, is the application of the weights on the attributes, denoted in Table 4.4. The analysis without weighted attributes thus does not consider the weights from Table 4.4 and includes all position group attributes with the same weighting. The difference is to be analysed to see the importance and effect of weighting on the results in clustering.

Table 4.5: Model evaluation for CLARANS clustering algorithm.

CLARANS clustering algorithm				
EV (variance)	Goalkeeper (60 min)	Central midfielder (60 min)	Winger (60 min)	Weighted
0.9	0.278 (4)	Clustering algorithm found to be too time consuming, no further analysis.		✓
0.85	0.220 (3)			✓
0.8	0.220 (3)			✓
0.8	0.060 (3)			X
0.75	0.220 (3)			✓
0.7	0.319 (2)			✓
0.6	0.319 (2)			✓

Table 4.6: Model evaluation for CURE clustering algorithm.

CURE clustering algorithm				
EV (variance)	Goalkeeper (0.5 min)	Central midfielder (1.5 min)	Winger (1 min)	Weighted
0.9	-0.251 (4)	-0.282 (30)	-0.338 (25)	✓
0.85	0.057 (3)	0.342 (25)	-0.338 (21)	✓
0.8	0.057 (3)	-0.329 (20)	-0.301 (17)	✓
0.8	-0.386 (3)	-0.314 (20)	0.142 (22)	X
0.75	0.057 (3)	-0.329 (17)	-0.338 (14)	✓
0.7	0.102 (2)	-0.512 (14)	-0.301 (12)	✓
0.6	0.102 (2)	-0.216 (9)	-0.338 (8)	✓

Table 4.7: Model evaluation for DBSCAN clustering algorithm.

DBSCAN clustering algorithm				
EV (variance)	Goalkeeper (0.2 min)	Central midfielder (0.5 min)	Winger (0.4 min)	Weighted
0.9	0.278 (4)	0.132 (30)	-0.636 (25)	✓
0.85	0.220 (3)	0.132 (25)	-0.636 (21)	✓
0.8	0.220 (3)	0.132 (20)	-0.636 (17)	✓
0.8	0.149 (3)	-0.601(21)	-0.696 (22)	X
0.75	0.220 (3)	0.132 (17)	-0.636 (14)	✓
0.7	0.319 (2)	0.132 (14)	-0.636 (12)	✓
0.6	0.319 (2)	0.132 (9)	-0.636 (8)	✓

Table 4.8: Model evaluation for GMC clustering algorithm.

GMC clustering algorithm				
EV (variance)	Goalkeeper (0.5 min)	Central midfielder (3 min)	Winger (3 min)	Weighted
0.9	0.125 (4)	-0.097 (30)	0.061 (25)	✓
0.85	0.126 (3)	-0.105 (25)	0.061 (21)	✓
0.8	0.126 (3)	-0.117 (20)	0.061 (17)	✓
0.8	0.060 (3)	-0.056 (20)	-0.136 (22)	X
0.75	0.109 (3)	-0.208 (17)	-0.028 (14)	✓
0.7	0.069 (2)	-0.175 (14)	-0.060 (12)	✓
0.6	0.063 (2)	-0.126 (9)	-0.061 (8)	✓

The results show that in general, a lot more features are required to explain the variance for position groups other than goalkeepers. This is because these position groups consider many more types of statistics.

Before the clustering is applied, the players are grouped based on position group, which makes the dataset used for clustering narrow in terms of differences between players. Because of this, the clustering graphs do not seem as grouped, as for some configurations, there is not a lot of distinction made between the clustered groups of players (looking at the silhouette value). To motivate this idea, the clustering model is evaluated for all outfield positions (so all players excluding goalkeepers) together, with the results shown in Figure 4.13. Here you can see that the players are grouped more distinctively, where the groups mostly represent different position groups. For the analysis per position group, it is possible to see the detailed differences between players within a certain group. This will come forward in the designed tool and gives the user information on different types of players within a position group.

Figures 4.13 and 4.15 show summary statistics of the different clusters as presented in Figure 4.20. An average configuration is used for these results with an explained variance threshold of 0.8 and inclusion of weighting of attributes. This gives an idea of what kind of players appear in each cluster, to show how the clustering makes sense to distinguish types of players in the to be delivered tool and be valuable for the scouting department to find specific types of players. The different clusters generally represent position groups, or the type of position in which a player plays. In Figure 4.13, the first cluster (first columns in the histogram) contains the goalkeepers and defensive minded players (centre backs and defensive midfielders). The second cluster contains mostly attacking minded players; attacking midfielders, wingers, and strikers. The third cluster mostly contains position groups where defensive and passing attributes are intuitively important (wing backs, centre backs, defensive midfielders, and central midfielders). This shows that the clustering method was able to group the

players based on their attributes which were different based on their played position. Attacking minded players intuitively have higher passing and chance creation statistics compared to defenders, which is why these are clustered together. Defensive minded players, such as in the first cluster, are clustered based on their defensive attributes, such as the successful defensive actions as represented in Figure 4.15. This figure shows that the clusters are thus partly divided based on defensive attributes, thus clustering defensive minded players and more attacking minded players separately. The distinction between the second and third clusters (cluster 1 and 2) is also represented in spider charts. These charts show how the clusters differ in terms of average attributes for some characterising attributes. Cluster 1 shows attacking minded players with many offensive actions, where this is lower for the more defensive minded players in cluster 2.

This is further supported by the contribution of attributes on the two principal components as described in Figures 4.16 and 4.17. This shows that the first principal component is mostly based on attacking attributes (such as offensive duels and received long passes), and the second principal component is mostly based on key passing attributes (such as progressive passes and passes to the penalty area). Because of this, the clustering algorithm divides offensive and defensive minded players, as well as creative players. Players on the right of the clustering graph are more focused on attacking attributes (compared to players on the left), and players in the top part of the clustering graph are more focused on the passing attributes (compared to players on the bottom part of the clustering graph).

Useful information here is that it is already a helpful first step to group the players based on positions before clustering, as this can make for a more specific analysis of the groups of players. The players in the positions are thus usually similar in types of attributes, and looking specifically within those position groups, it should be possible to find more detailed differences in the players.

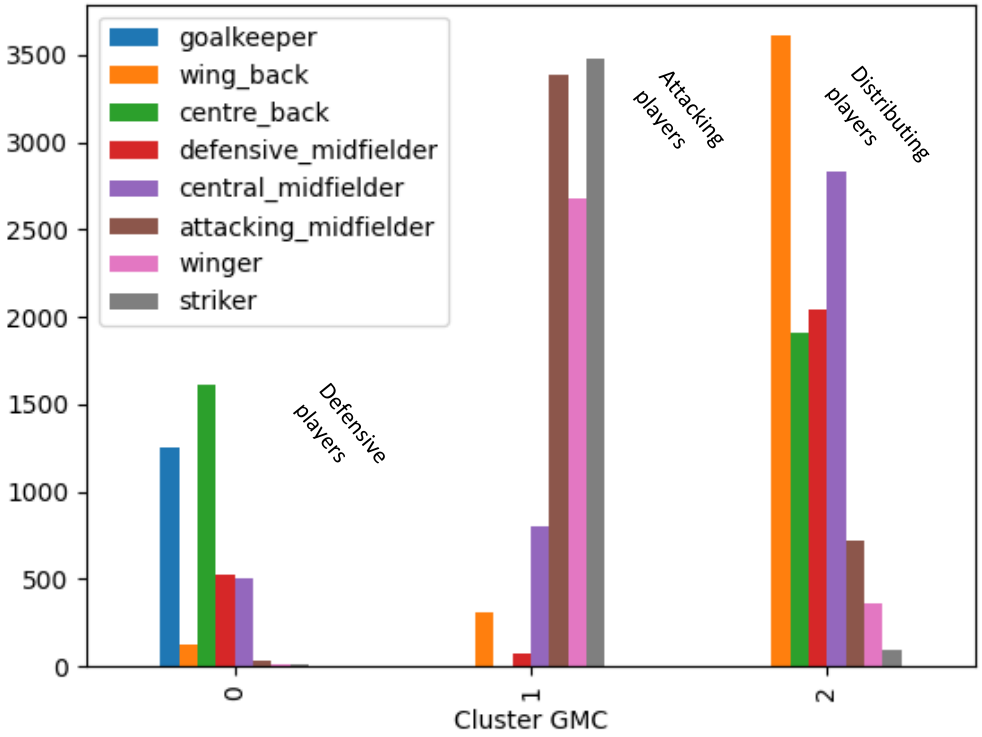

Figure 4.13: Representation of position groups in clusters in analysis of all players combined.

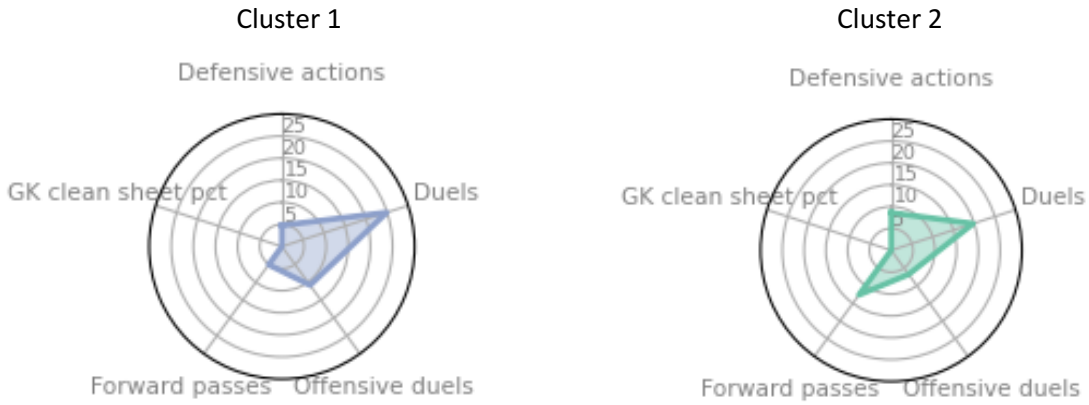

Figure 4.14: Spider charts for cluster 1 and 2 represented in Figure 4.13, in analysis of all players combined.

	ws_player_general_age	ws_player_general_market_value	ws_player_defensive_successful_defensive_actions_per_90	ws_player_attacking_goals_per_90
0	27.93964	377724.75295	10.12237	0.10266
1	27.79460	479825.40686	7.70082	0.15912
2	29.07837	380987.24268	6.91136	0.10705
3	27.43614	415667.95866	4.57469	0.32641

Figure 4.15: Summarization of average attributes in clusters of all players combined.

	PC1	PC2
ws_player_attacking_offensive_duels_per_90	0.26570	0.08087
ws_player_attacking_received_long_passes_per...	0.23786	0.04974
ws_player_attacking_touches_in_box_per_90	0.23214	-0.15520
ws_player_attacking_successful_attacking_acti...	0.22457	0.14592
ws_player_attacking_shots_per_90	0.21610	-0.06003
ws_player_attacking_dribbles_per_90	0.19985	0.11130
ws_player_attacking_xg_per_90	0.18060	-0.17474

Figure 4.16: Eigenvalue representation of contribution of attributes in the principal components in clustering on all players, sorted by contribution to principal component 1 (showing top 7 contributing attributes).

	PC1	PC2
ws_player_key_passing_progressive_passes_pe...	-0.14228	0.30254
ws_player_key_passing_passes_to_penalty_are...	0.17981	0.29712
ws_player_key_passing_passes_to_final_third_p...	-0.09342	0.28842
ws_player_passing_forward_passes_per_90	-0.18464	0.25427
ws_player_passing_lateral_passes_per_90	-0.07586	0.20663
ws_player_attacking_received_passes_per_90	-0.02510	0.20529
ws_player_defensive_successful_defensive_acti...	-0.13821	0.19885

Figure 4.17: Eigenvalue representation of contribution of attributes in the principal components in clustering of all players, sorted by contribution to principal component 2 (showing top 7 contributing attributes).

Figure 4.18 shows a summarization of average attributes for the clusters of central midfielders as presented in Figure 4.8. This figure is presented to give an idea of the differences in separability between players when looking at all positions versus one specific position group. Because the clustering is done specifically on a position group, it is intuitively more difficult to distinguish the differences between the players, and thus separate them in clusters. The performance of the clustering algorithm in terms of separability within the dataset is thus worse compared to the analysis on all players combined. As explained in the principal component analysis however, when two different players with high differences on the first principal component are compared as an example, the specific differences do come forward. Because of this, the dissimilarity metrics calculated from differences in the attributes of the players is more valuable, to be represented in the to be designed tool.

	ws_player_general_age_x	ws_player_general_market_value_x	ws_player_defensive_successful_defensive_actions_per_90_x	ws_player_attacking_goals_per_90
1.0	27.89809	445854.43038	8.52449	0.12988
2.0	28.00396	395573.87863	8.54482	0.15077
3.0	28.14286	250000.00000	8.92000	0.16667
4.0	27.33333	783333.33333	5.55667	0.21000
6.0	19.00000	50000.00000	8.95000	nan
8.0	23.00000	650000.00000	7.34000	0.03000
11.0	30.00000	262500.00000	10.02500	0.14500
12.0	25.00000	1100000.00000	5.98000	0.11000
13.0	34.00000	400000.00000	10.25000	nan

Figure 4.18: Summarization of average attributes in clusters of central midfielders.

A consensus found in the above experiments suggests the use of GMC clustering model because of the, on average, more consistent and higher silhouette scores. A higher silhouette score entails a high similarity within a cluster and lower similarity with other clusters, thus positive. GMC as the most effective clustering model for the objectives of this thesis can be concluded with an analysis on the average and standard deviation of the silhouette scores, as found in Table 4.9. GMC comes forward as the clustering model with the highest average silhouette score, as well as the lowest average standard deviation. This entails that the separability of the datapoints between the GMC clusters is the highest relative to the other clustering models. Furthermore, the lowest average standard deviation shows that there are smaller differences between the silhouette scores between the different configurations, which makes the consistency of that clustering model the highest compared to the other clustering models.

Table 4.9: Summarization of silhouette scores per clustering model.

Clustering model	**Cure**	**DBSCAN**	**GMC**
Average silhouette score	-0.17686	-0.12362	-0.01467
Average standard deviation	0.1988	0.111653	0.049318

Through summarizing the differences in silhouette scores when adding weighting on the attributes compared to no weights, a conclusion can be made on whether to weight the attributes in the final model as well. Weighting the different attributes gives on general better silhouette scores, with an average silhouette score of -0.087 versus -0.204 respectively between weighted and non-weighted attributes. Because of this, the weighting of attributes is implemented in the tool as a standard.

Finally, an explained variance (EV) of 0.8 is chosen as it generally gives high silhouette scores with a limited number of principal components representing the attributes. For this conclusion, the cumulative explained variance and number of required principal components are analysed to make a trade-off. As explained above, the point is discovered where an increase in explained variance requires a significant higher number of principal components. At this point, the added value of the explained variance is not worth it when compared to the required principal components. In Figure 4.19, this is visualised with the average silhouette values for different configurations of explained variance thresholds. It can be easily seen that for an explained variance (EV) of 0.8, there is a big increase in average silhouette value with a relatively small increase of explained variance. This is identified in a similar way as the elbow method discussed in Chapter 3.2.6.

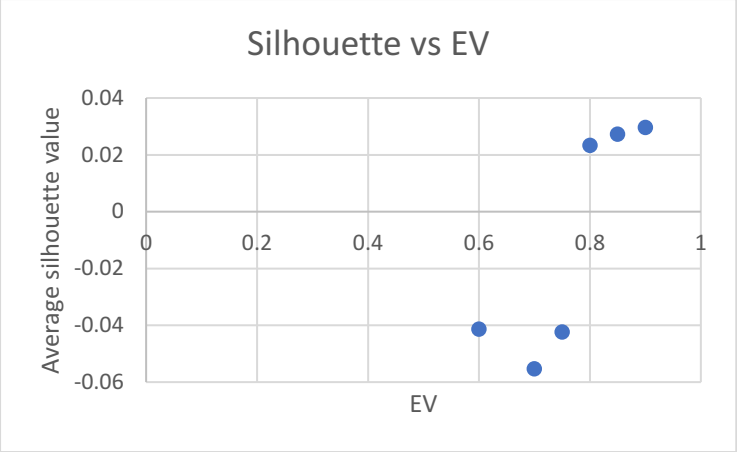

Figure 4.19: Average silhouette values for different configurations of the explained variance (EV) threshold.

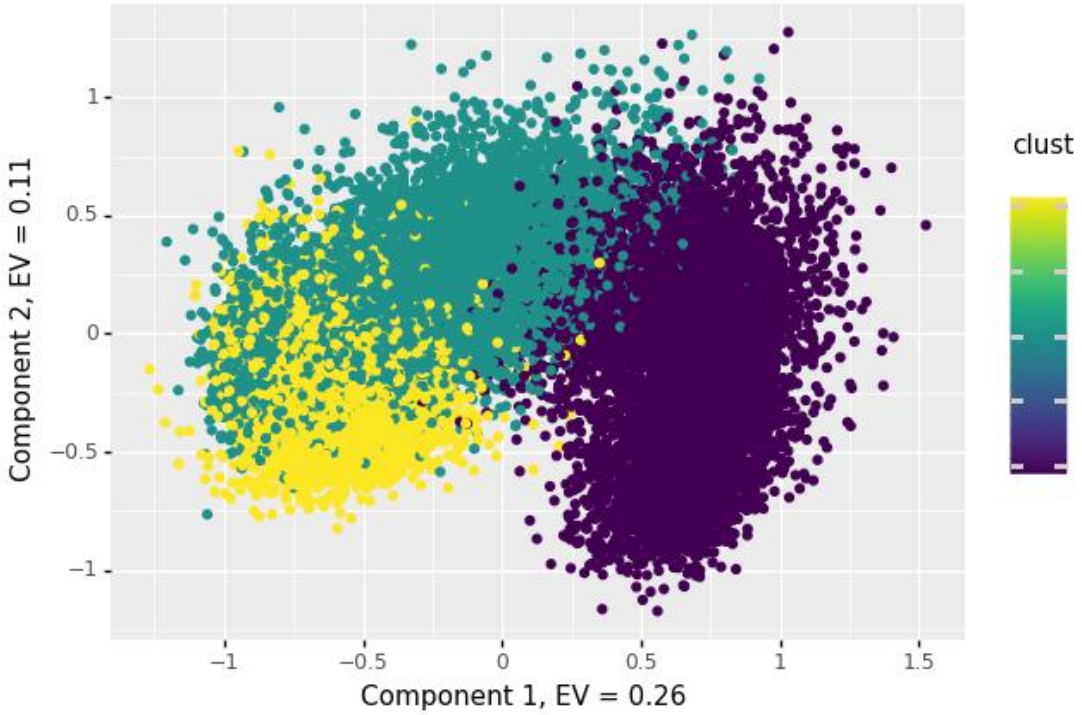

Figure 4.20: Clustering graphs for all outfield positions including the explained variance for the first two components.

4.1.8 Develop tool including the results

Based on the previously stated research requirements of the scouting department at San Lorenzo (Chapter 3.1), as well as the characteristics of the dataset and configuration of variables as input to the clustering models, a tool can be designed. This can be achieved using back-end development tool Django, including incorporation of the Python code for modelling, and front-end development tool Vue.js as explained in Chapter 3.2.8.

Generally, the designed models require input from the user to achieve results that are useful for the user. The user wants to discover which players might be similar or in the same cluster as a specific type of player, for example one that must be replaced in the current squad. The to be analysed dataset is inputted. The model uses this information to discover the players that are most similar and related

in terms of clusters to the queried player. Based on the position of the player, attributes are weighted differently considering importance in that specific position group. Afterwards, a clustering algorithm and dissimilarity metrics are applied to discover the required information. This information is then to be displayed to the user, such that they can use this information to improve their analysis. Altogether, this shortly described the functionality of the to be designed tool. This is structured below, including which features the to be designed tool should or could have to allow for results that are relevant for the issues approached within this thesis. This also considered the requirements of the tool as setup in Chapter 3.2.9. These requirements are numbered from one to three and are referred to within this section. As mentioned in Chapter 3.2.9 itself, these requirements also follow the initial requirement setup in Chapter 3.1. Explanations of why certain features are important for the goals of this thesis are formulated following by the listed feature.

The reliability of the scouting tool should be considered in relation to the importance and weight in which the suggestions or shortlisted players are taken within the scouting process. The shortlisted players are based on the modelled clustering and dissimilarity algorithms and thus have a technical foundation. However, with the inclusion of the threshold of the explained variance and the included weights, it is important that the shortlisted players are only taken as suggestions and should be scouted further and in detail within the scouting process to form a complete analysis on these players.

The main goal within this thesis is to solve the lack of similarity information on players. The tool should give information on the similarity between a specific player and a to be scouted player. To search for similar players compared to a specific current player, the current player needs to be filtered out of the dataset. Furthermore, San Lorenzo has a specific transfer philosophy which means that players that are for example too expensive, should be filtered out of the possible results as well. This relates to the first requirement in Chapter 3.2.9, which includes the possibility to search for players conforming the transfer philosophy at San Lorenzo.

1. User input: which player to analyse and what dataset to use
    a. Filtering and selection buttons to navigate to the to be analysed player:
    b. Filtering of players that could possibly result from analysis:

Scouts maintain their specific characteristics and ideas, which is why manual adaptation of attribute importance should be possible. The tool is not meant to take over the role of the scout, but just as a supporting tool within the scouting process making a more specific shortlist of players for the scouts to further analyse. Because of this, the tool should be customizable by the user of the tool and give their own weights and importance to attribute groups when analysing a certain player. This also relates to the first requirement from Chapter 3.2.9, where in customizability, the inclusion of weighting and the specification of the scouts' ideas is included.

2. Weighting of attributes
    a. Allow for manual adaptation of weights on attributes, for customizability

The required similarity information between players, as is the goal of the thesis, is the result of clustering algorithms and dissimilarity metrics. To display the similarity information, these algorithms should be executed within the tool. This is related to the third requirement from Chapter 3.2.9, which focuses on the results formulated within the tool, based on the applied clustering and dissimilarity algorithms.

3. Execution of clustering algorithm and dissimilarity metrics

The similarity information that should result from the execution of the algorithms needs to be displayed to be useful. A shortlist of most similar players should be provided to support the scout within the scouting process. With the list of most similar players, the scout can continue the scouting process with a focus on a smaller subset of players compared to when they start the process with limited information. This makes it a requirement to give an overview of the most similar players compared to the analysed player, including in-depth statistics to allow for complete analysis. The clustering graph should be returned to give a different perspective and make it possible for the scout to find hidden talent, which gives them an extra opportunity to find player that could fit their transfer philosophy. This relates to the third requirement formulated in Chapter 3.2.9, including the formulation and visualisation of the initial shortlist and overview of resulting analysed players with the statistically founded results.

4. Visualisation of results
    a. Basic overview of to be analysed player, and 5 most similar players
    b. Statistical results from dissimilarity metrics and clustering (different perspectives) displayed
    c. More extensive statistics to be compared between the analysed and resulting players
    d. Return of the clustering graph, to see what type of players are close to the analysed player, as well as to identify players with high potential market value

The above-described structure can be further visualised in an (initial) example design of the tool as shown in Figure 4.21. The above features are included in this design, and some dummy data is used to represent the working of the tool. The scout should start in the 'Select player to analyse' part of the tool, to search for the player they want to compare. To the left and to the right of this section, respectively the weights of the attributes and the filtering on the to be returned and analysed players can be specified. When clicking on the blue button 'Start analysis', the scout should wait for the results presented in the lower section of the tool, which can be interpreted directly after.

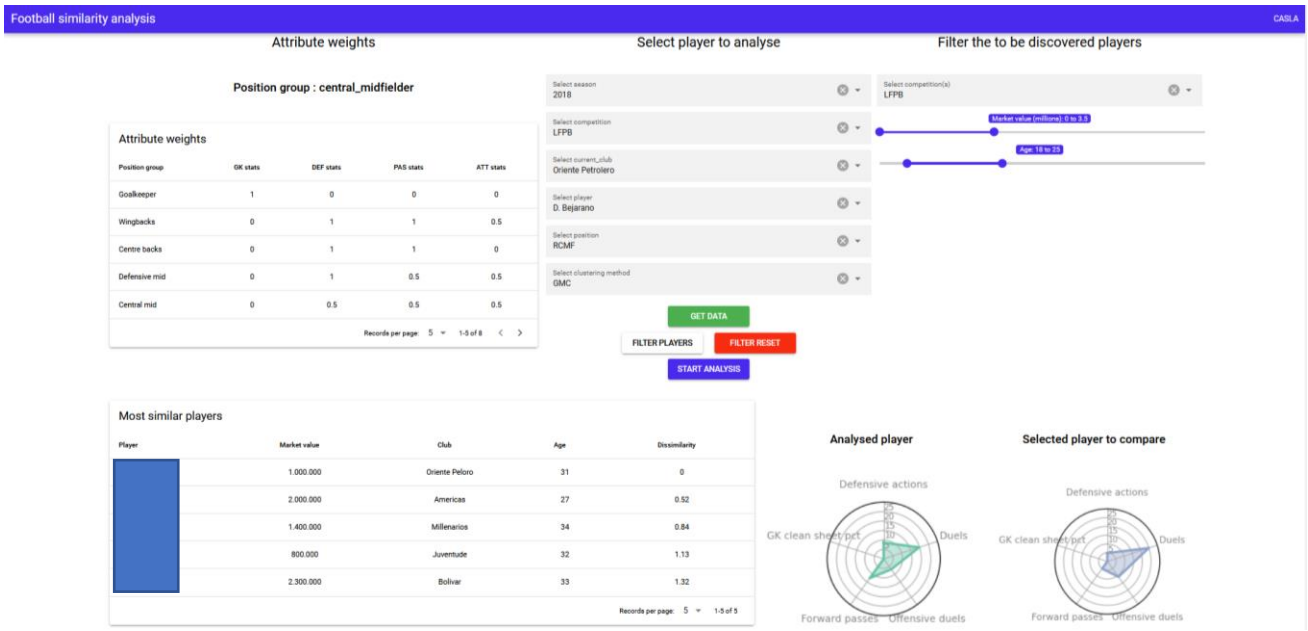

Figure 4.21: Initial design for to be programmed tool.

The tool is designed in the programs described, which allow for a responsive and intuitive design and execution of the models.

Within the development some small adaptations are made based on the plan described above. The developed tool is shown in Figures 4.22 and 4.23, showing certain differences as opposed to the initial design in Figure 4.21. Figure 4.22 shows the top part of the tool, including the input to the to be run model. Figure 4.23 shows the results from the models.

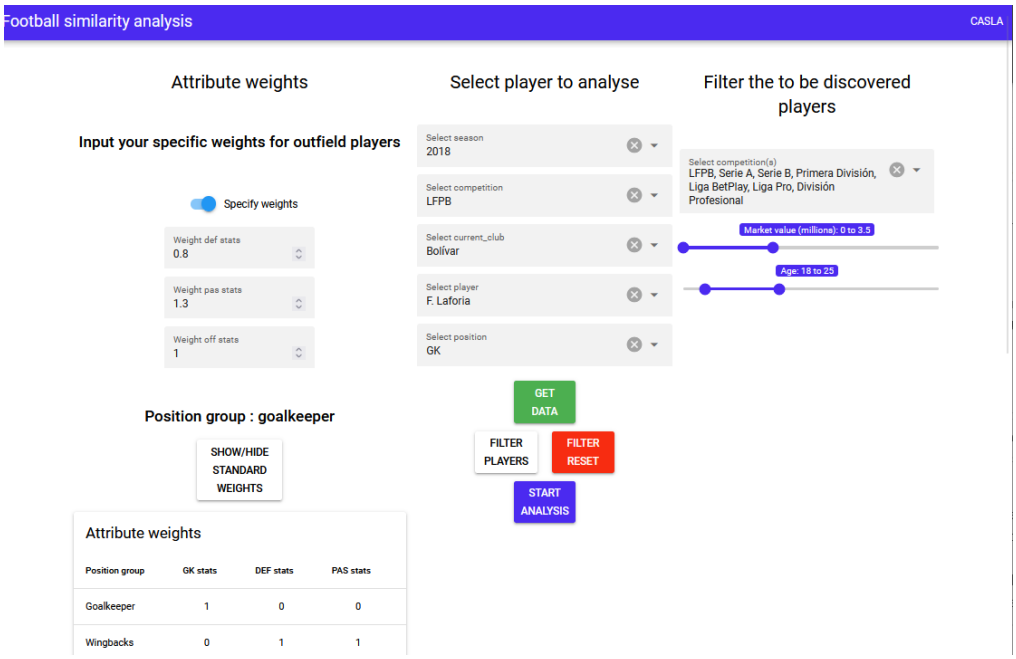

Figure 4.22: Developed tool top part including the input to the model.

First, the specification of the weights as a required input to the model is realised through certain simple input elements which can be changed by the user. The standard weighting of the attributes is given as an option to show to inform the user what the standard input to the weighting of the attributes will be.

Second, the option to choose the clustering method is disregarded in the process as the evaluated clustering algorithms lead to an optimal one, on which the tool is developed and modelled. The GMC clustering method is thus used as a standard. Together with the filters shown on the top right, this is the input into the model before the results are shown in the lower part of the tool. The data on which the analysis is performed is thus filtered and the to be analysed player is included within the clustering and dissimilarity algorithms. Once the model is run, the output is shown in the lower part of the tool.

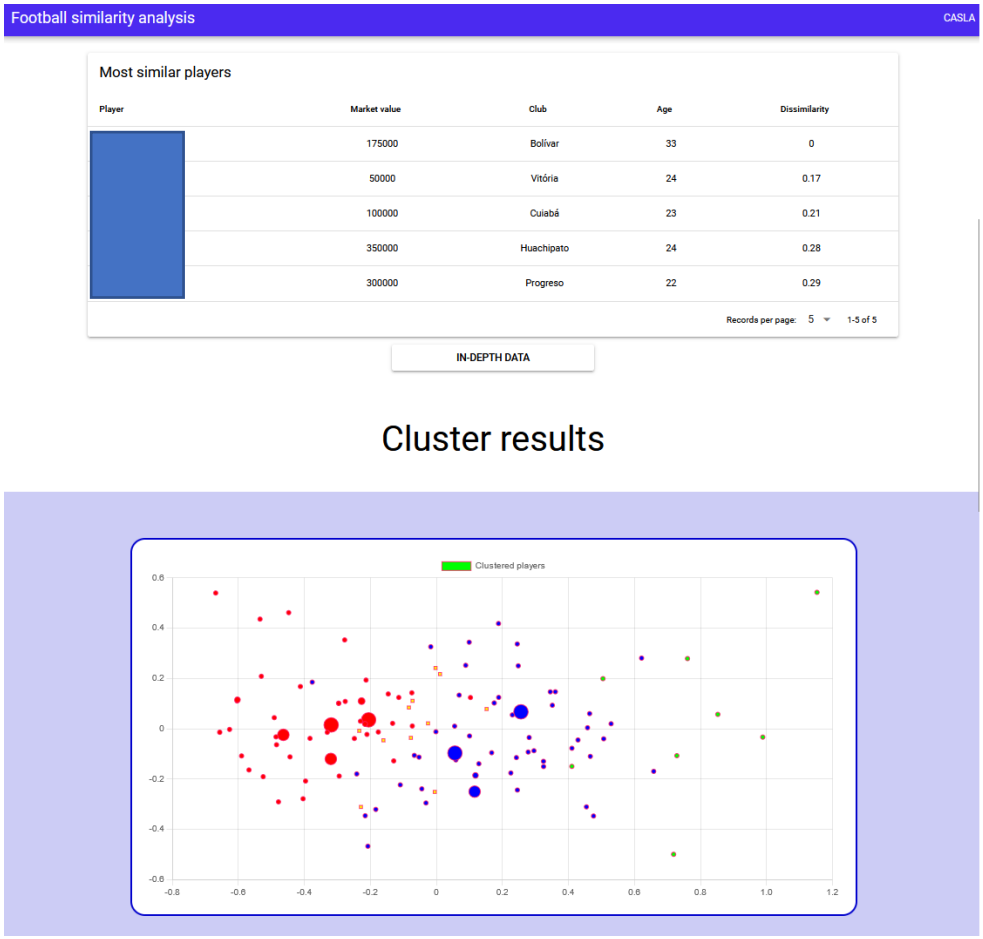

Figure 4.23: Developed tool bottom part including the results.

In the results of the tool, a general overview is given of the analysed player with its most similar players from the filtered dataset. Furthermore, an option is given to observe in-depth data on these players, which provides all statistics of the players in a table. Finally, the clustering results are returned within a graph including differences in sizes of the players (data points in the graph) based on their market value. This provides the possibility to identify lower valued players that are close to higher valued players in terms of average statistics. The analysed player and its similar players are returned with square data points to identify them quickly within the graph.

Tool results

Using the designed tool will present similar and potentially to be scouted players for the scouting department of San Lorenzo. In this section, some examples are provided with analysis of players that played for San Lorenzo in 2021, the most recent season considered within the available dataset, as well as transfers made at the end of season 2018-2019.

Player 1

Player 1 was an important attacking minded player within the squad at San Lorenzo in the 2021 season. As this player is increasing in market value, it is probable that he will be sold in the near future. This means that this player should be replaced, ideally by a young relatively cheap talent as fitting the transfer philosophy. For demonstration purposes, this player is analysed where the search for a replacement player is based on a young player (18 to 25 years old) that is worth maximally 3.5 million and is playing in the same league as San Lorenzo.

The results of the above search are presented in Figures 4.24 and 4.25. Player 2 of Vélez Sarsfield is considered the most similar player based on the filters. With a market value of 1.8 million, this could

be an interesting player to consider in the scouting process and analyse more in-depth. Figure 4.25 shows the clustering results, where Player 1 and Player 2 are shown in the yellow rectangles within the small, annotated box (names to be seen within the use of the tool itself when hovering over the datapoints). The size of the datapoints gives an indication of the market value of the players, where we can see that these players are also relatively close to a cluster of higher valued players (the blue circles). This entails that the players have relatively similar statistics to highly valued players, which could indicate that Player 2 is a player with potential to grow in terms of market value and intuitively also performance and quality.

The data from Player 2 was based on season 2018-2019, which was the season before he was sold for 10.5 million euros ("Market values," 2022). This case would be interesting for San Lorenzo, as this could have potentially been a transfer with a profit.

Most similar players

Player	Market value	Club	Age	Dissimilarity
	6000000	San Lorenzo	29	0
	1800000	Vélez Sarsfield	24	0.29
	450000	Godoy Cruz	22	0.7
	450000	Godoy Cruz	22	0.7
	3000000	Gimnasia La Plata	22	0.84

Records per page: 5 1-5 of 6

Figure 4.24: Results of similar players on analysis of Player 1.

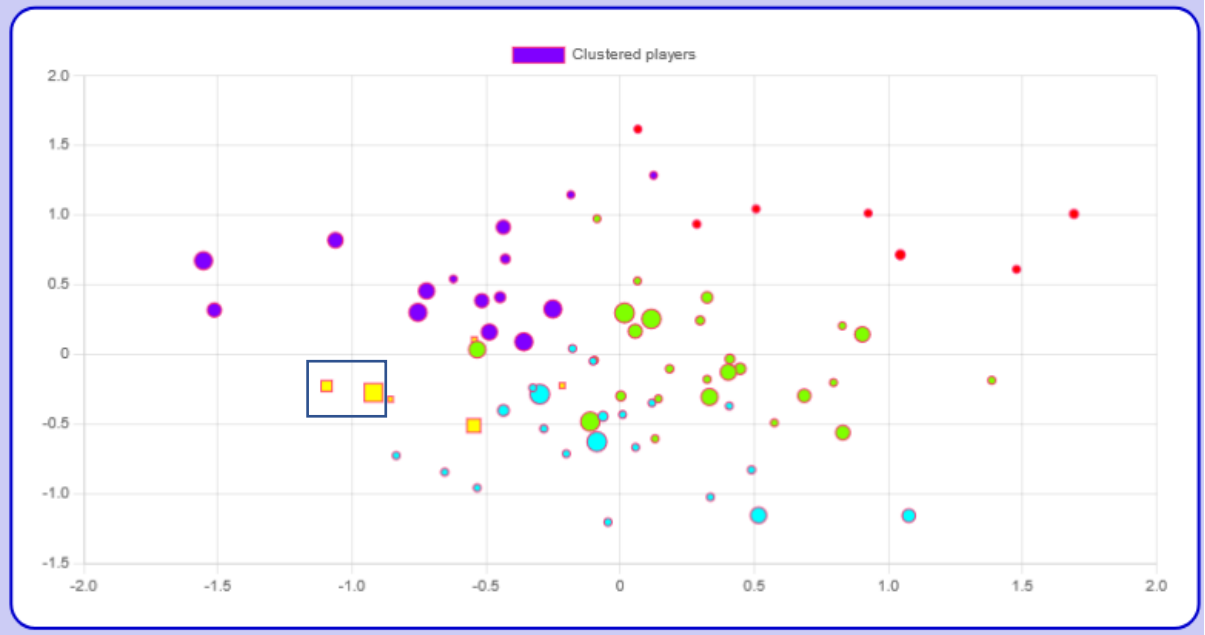

Figure 4.25: Clustering results on analysis of Player 1.

Player 3

Player 3 was an important defensive minded player within the squad at San Lorenzo in the 2021 season. Because of the increasing age of this player, it is probable that he will be needed to be replaced soon. This means that this player should be replaced, ideally by a young relatively cheap

talent as fitting the transfer philosophy. For demonstration purposes, this player is analysed where the search for a replacement player is based on a young player (18 to 25 years old) that is worth maximally 3.5 million and is playing in the same league as San Lorenzo.

The results of the above search are presented in Figures 4.26 and 4.27. Player 4 of Rosario Central is considered the most similar player based on the filters. With a market value of 350.000, this could be an interesting player to consider in the scouting process and analyse more in-depth. Figure 4.27 shows the clustering results, where Player 3 and Player 4 are shown in the yellow rectangles within the small, annotated box. These players are also relatively close to many higher valued players (blue circles and yellow rectangles). This entails that the players have relatively similar statistics to highly valued players, which could indicate that Player 4 is a player with potential to grow in terms of market value and intuitively also performance and quality.

Most similar players

Player	Market value	Club	Age	Dissimilarity
[redacted]	150000	San Lorenzo	36	0
	350000	Rosario Central	21	0.13
	3100000	Vélez Sarsfield	24	0.32
	300000	Rosario Central	23	0.44
	1800000	Independiente	21	0.51

Records per page: 5 1-5 of 18

Figure 4.26: Results of similar players on analysis of Player 3.

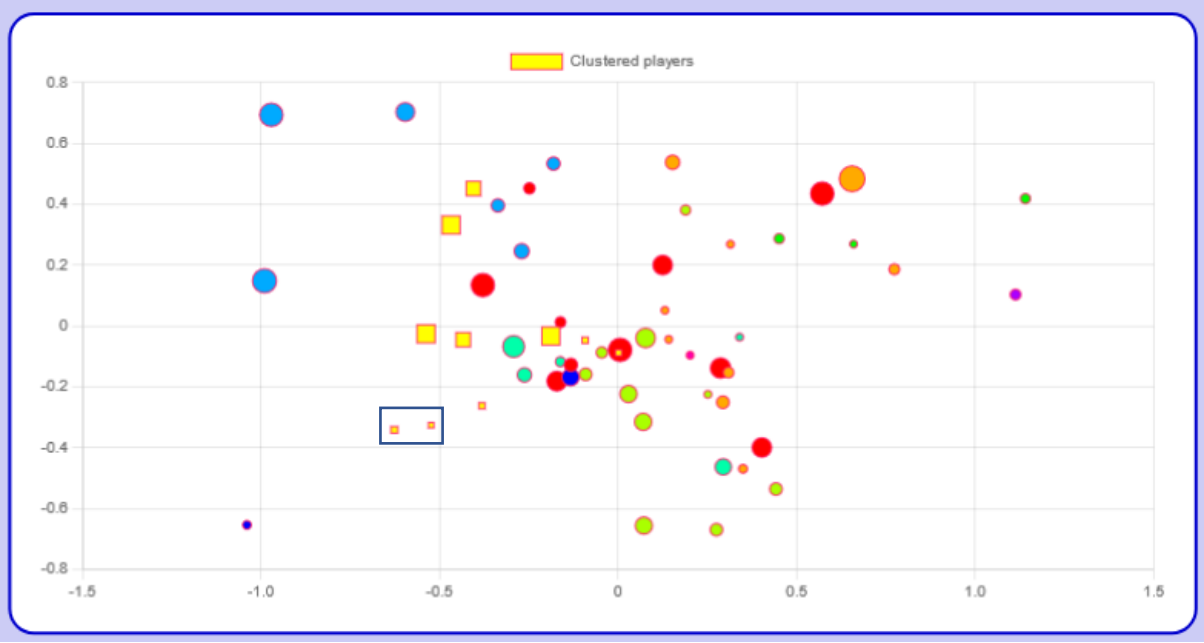

Figure 4.27: Clustering results on analysis of Player 3.

Potential replacement of Player 5

In the end of season 2018-2019, striker Player 5 was sold to Racing Club for €3.25m. One of the players that was signed at this time, was Player 6 from Defensa for €1.70m. If the tool would have

been used at this point, and an analysis would have been done on Player 5, Player 6 would not have come out as one of the most similar players. This was a relatively expensive player at that moment in time, and the market value of this player has dropped over the last years. The tool, on the other hand, would have had the results as shown in Figure 4.28. With the transfer philosophy of contracting young players with potential, and selling them on for a profit, the young player Player 7 would have been an interesting player to further scout. With a significantly lower price, if the player would have been deemed interesting throughout the entire scouting process, the player could have been signed at a significantly lower price. Currently, this player is showing his potential, with a current market value of €1.5m. ("Market values," 2022)

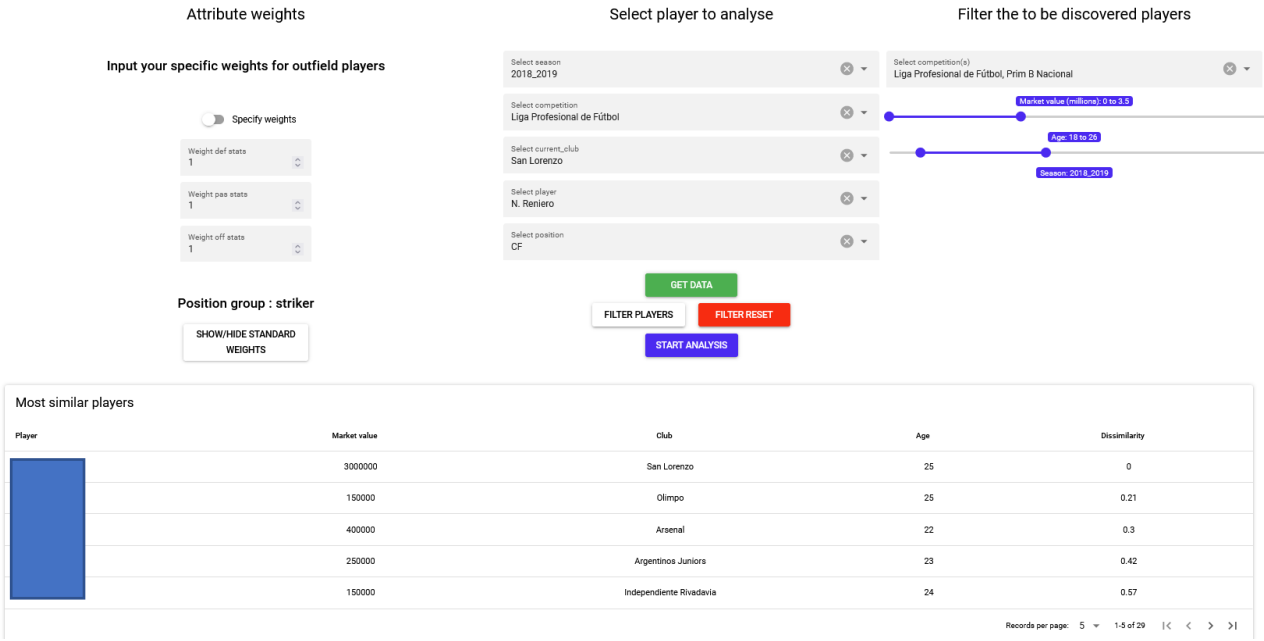

Figure 4.28: Potential replacements for Player 5 at the end of season 2018-2019.

Potential replacement of Player 8 and Player 9

At the end of season 2018-2019, Player 8 and Player 9 left San Lorenzo. These two experienced central or attacking minded midfielders needed to be replaced. San Lorenzo bought a central midfielder named Player 10 for €2.00m. To find a relatively experienced player, with room to grow, a midfielder with experience was searched for in the tool. The results in Figures 4.29 and 4.30 show the dissimilarity of midfielders compared to the two to be replaced players. For both players, Player 10 was seen as a relatively similar player, and attention for this player would thus also have been caught with the use of this tool. With the current increasing market value of the player, and the transfer to Boca Juniors for a profit in the end of season 2020-2021, this can be seen as a successful transfer. The tool in this case could have had a supporting role in the decision to buy the player, as it was seen as a fitting player compared to the to be replaced players. ("Market values," 2022)

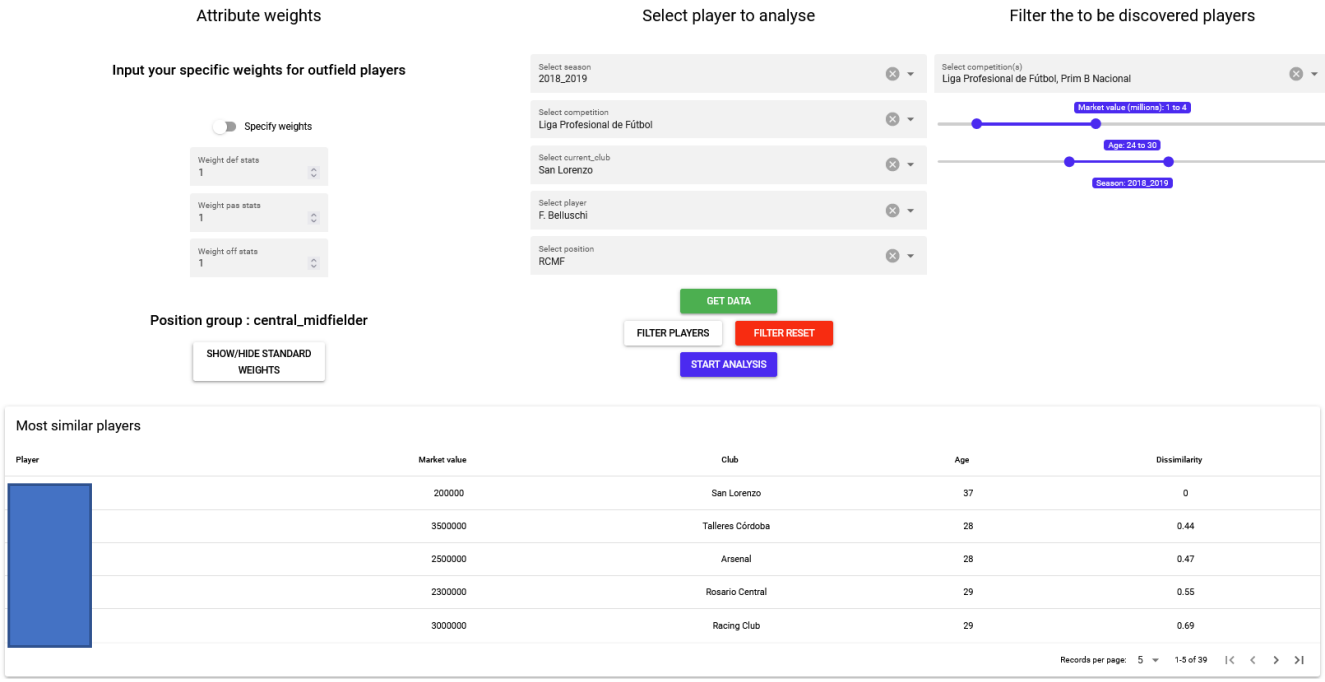

Figure 4.29: Potential replacements for F. Belluschi at the end of season 2018-2019.

Figure 4.30: Potential replacements for G. Castellani at the end of season 2018-2019.

4.2 Conclusion

In this chapter, the designed method from Chapter 3 is followed and results in a functional scouting tool. The tool can be used to find shortlisted and clustered players that are similar to a to be compared player, which is input to the tool. The efficacy of the scouting process can be improved using this tool, where the initial shortlist of potentially to be scouted players can be easily and quickly determined from one quick search within the tool. Practical suggestions and historical events are analysed to evaluate the tool and observe where the tool would have been practical. Results are found that show for example initially shortlisted players that include successfully bought and sold (for a profit) players from the last couple of years at San Lorenzo.

The clustering algorithms and dissimilarity measures are concluded through analysis of the clustering performance. This is done with the evaluation of silhouette scores, concluding the most effective clustering model as the GMC model. The weighted Manhattan distance measure is used to conclude dissimilarities between players.

The next chapter will include a subjective evaluation of the tool, through an analysis of possible application together with the stakeholders, and more specific analysis with the main scout at San Lorenzo.

5 Tool evaluation

In this chapter, the in Chapter 4 created tool will be evaluated for use at San Lorenzo.

5.1 Evaluate tool

Models are evaluated as shown in the previous chapter, but here we evaluate the tool and use of the tool subjectively. Based on the requirements setup in Chapter 3.2.9, the tool can be assessed and deemed sufficient or not. In this section, all requirements in the checklist are evaluated in cooperation with the scouting department at San Lorenzo, including evaluation of sufficiency and thus whether the requirements can be checked off.

The evaluation in this section will be done through qualitative research with cooperation of the two supervisors for this thesis, including the main scout. The points of evaluation will be addressed, and the responses are structured per evaluation topic.

Based on the qualitative research, three main evaluation topics can be concluded as well as spider charts for overall effectiveness for the scouting process before and after (imaginary) usage of the tool. The evaluation topics will be further discussed referring to the requirements stated below. For example, the scouting process is mentioned within this section, which is explained in Chapter 1.2.2.

- Features

The features that are present within the tool are seen as complete and sufficient in general. For example, filtering the players is easy and functional and should remain within the tool. The filters for the to be discovered players make it possible to search for players that are within their transfer philosophy. In general, this means that requirement 1 can be checked off. For possible improvements however, it can be useful to add certain filters. To search for players that have played within a specific range of minutes, to discover for example very game-fit players, a filter can be added to specify the range of minutes a player should have played to be part of the results. Furthermore, the preferred foot of the player can also be added to make it possible to for example find a left-footed centre back, which can generally be harder to find because most players are right-footed, but which can have additional value as a left sided centre back. Finally, the possibility to change the number of players to be resulting from the analysis would be a good addition to give additional options to the scouts when analysing a player.

The resulting tables and clustering graph are also seen as good inclusions within the tool. The table with the general overview of the resulting players with the dissimilarity information gives a quick, easy, and sufficient idea of the players. This is very useful for the scouting department at San Lorenzo, as they will already have a general idea of the players and can quickly see what type of players are resulting from the tool. The more in-depth data that can be checked is a feature which is in general relevant for the experienced scouts, as this will give them a more in-depth view of the players. This is also seen as a good feature for scouts to get a better idea of players they might not know as well. This can for example be useful when scouting in new areas that are less explored. The graph with the clustering results is also valuable for the scouts and should remain within the tool.

Another interesting addition would be a schematic football field visualisation in the tool where the scouting team can find similar players for all positions compared to the current squad at San Lorenzo. The scouting team can look for similar players based on the filters and find a completely new team of potential replacements in a quick effective way.

- Usage and results

The tool is seen as a usable tool within the scouting process at San Lorenzo. The tool can be used by the scouts to get an idea of similar players when analysing a certain position. When the scouting team is looking for a specific position, they can search for the current player at San Lorenzo at that position to see which players are similar to replace them. However, what is seen as another option, is that the scouting team can have a specific player in mind from which they like the playstyle, but who can be out of their budget for example. Because of this, they cannot approach them in the possibility of signing them, which makes them want to look at players that are similar to him, which can result from using the tool. This resulting shortlist for similar players to a to be replaced player in the current squad, or an interesting player not in the club, is seen as added value within the scouting process. For example, when a new coach would arrive to San Lorenzo, who likes a specific playstyle and is looking for a midfielder with a specific profile, the scouting team can find a player they know with that profile and look for players that are similar to him in terms of statistics. This will then give a shortlist of players they can further analyse within the scouting process. The similarity information is seen as added value to the scouting team and give more information on the probability that a player will be able to replace a certain player in the current squad, or to have a similar playstyle to a highly rated player, which checks off requirement 6.

The resulting table and clustering graph give a good feeling of the similar players for the scouting department. This can for example be a good option to send to the sporting director when they want to show an overview of the players that they are looking at in the scouting process. The in-depth information in the table is mostly relevant for the scouts themselves, as this is too broad to send within a scouting report to the sporting director for example. The above evaluation makes the tool sufficiently usable in terms of application of the features, checking off requirement 2. The results from the tool give sufficient information about the players from different perspectives and are interpretable by the scouting team, checking off requirement 4 and 5.

The tool is also an interesting addition in terms of efficacy of the scouting process. Currently, depending on how known the player is to the scouting department, a player can be analysed by watching their games in around one to two hours. If this results in the conclusion that a player is not deemed fitting to the required type of player, this time is used less effectively. With the use of the tool, a not fitting player is expected to not come forward within the shortlist, which prevents the scouts from having to watch a lot of games of this specific player. It is difficult to say how much time this saves exactly, but it is expected to be an improvement, checking off requirement 3.

- Implementation

To be discussed in Chapter 5.2.

- 1. The scouting department can use filter options within the tool to search for players conforming the transfer philosophy and characteristics of a to be discovered player.
- 2. The tool is sufficiently usable in terms of application of the included features.
- 3. The tool improves the efficacy of the scouting process.
- 4. The tool gives technically founded results based on the clustering models, including general information (such as age, club, market value) as well as statistical measures for each similar player.
- 5. The tool gives different perspectives on the similarity and clustering results of the analysed model.
- 6. The tool helps to solve the problem about missing similarity information.

The evaluation of the tool is summarized on five main evaluation points, evaluated by the scouting team at San Lorenzo, and represented in a spider chart in Figure 5.1. The scale from 1-5 represents the added value of the tool on these main aspects and is the result of the stakeholder opinions on the main evaluation topics as described above.

For efficacy, quality and similarity, the following scale is used:

1- A big decrease in efficacy/quality/similarity
2- A small decrease
3- No change
4- A small increase
5- A big increase

For completeness and ease of use, the following scale is used:

1- Not completer/easier to use at all
2- Not really completer/easier to use
3- No change
4- A little completer/easier to use
5- A lot completer/easier to use

These five evaluation points are related with the requirements presented above. Efficacy relates to the improvement in efficacy in the scouting process as in requirement 3. Similarity relates to the similarity information of the players as in requirement 6. The completeness relates to the inclusion of the filters and the perspectives of the results as in requirements 1 and 5. The quality relates to the technically founded results from the clustering models as in requirement 4. Finally, the ease of use is related with requirement 2. Because of the reorganisation and limited people with specified knowledge and use of the scouting process, the direct evaluation of the tool in relation to the scouting process is based on only the main stakeholder (the main scout and supervisor) of San Lorenzo. Because of this, it is only supposed to provide an indication of the effect the tool can have within the scouting process of San Lorenzo.

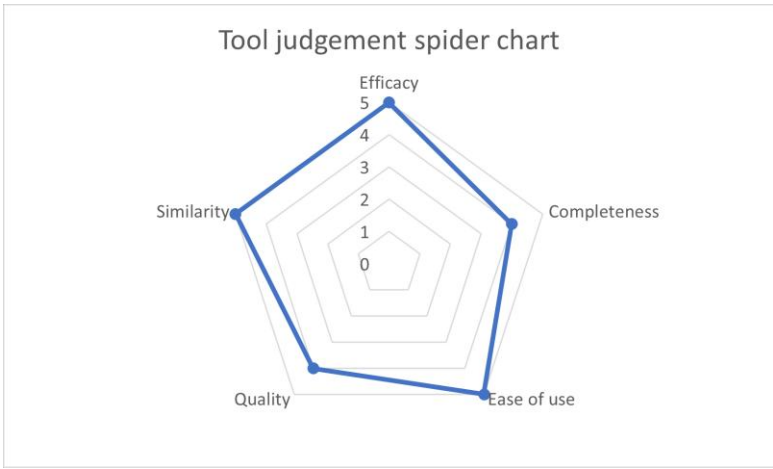

Figure 5.1: Tool evaluation spider chart on the five main evaluation points.

The improvements of the scouting process with the tool can be further evaluated. This is because these elements are evaluated based on the change after using the tool. As we can see, the efficacy is expected to increase heavily through the possible use of the tool in the initial shortlisting of players in the scouting process. The efficacy is expected to be improved as currently many football matches are watched without resulting interesting players, which can be mostly prevented using the tool. As the

tool already provides a shortlist of interesting players, the scouting team can focus watching their games on these specific players instead of on the entire player pool in the South American competitions.

The completeness of the scouting process is expected to improve a bit and is based on the completeness of information and analysis on the players that comes forward within the scouting process. With the use of the tool, the similarity information is added as well as a quick and easy overview of the players' most important statistics. Currently, the scouting process already includes a lot of information about the players, which means that even though the additional similarity information is valuable, the relative improvement on completeness of information is limited. Further possible additions in terms of tactical analysis or data about the mental side of the players would further improve the completeness of the information. This is further discussed in Chapter 6.

The ease of use in the scouting process is improved a lot compared to the old situation with the use of the tool. Because watching the games of the South American competitions is a very long and tedious process, being able to focus on an initial shortlist of players and focus on watching specific games, makes the scouting process a lot easier. The easiness of the tool provides the scouting team a quick and intuitive overview of the to be compared and analysed players including an overview of their data and similarity information.

The quality of the scouting process is expected to undergo a small increase using the tool. The addition of the similarity information provides an increase in the expected possible judgement on the fit of a player to the team, seen as the quality of the scouting process. A bigger increase in quality can be had with, similar as in the evaluation of the completeness of the scouting process, be improved through the addition of more tactical or mental analysis of the players.

The knowledge on the similarity between players is expected to gain a big increase using the tool in the scouting process. In the current situation, there is no knowledge on the similarity between players based on data. Because of this, the technically founded results that result from using the tool give a new perspective and thus a big increase in knowledge on the similarity.

5.2 Implementation and deployment tool

After evaluation of the tool, the issue arises of implementation and deployment of the tool. The tool should fit within the current scouting process as explained in Chapter 1.2.2.

The levels of automation as shown in Table 4.1 should be considered when fitting the tool within the scouting process. The tool can be imbedded within the current way of working as a supportive tool within the making of the shortlist of players that should be scouted further and more in-depth. The tool is not to be used as a decision maker, but only as a tool to give information on the similarity between players and can thus be part of the first step within the scouting process.

- Implementation results from qualitative research

The tool can be implemented in different stages within the scouting process. The tool can be used before making the shortlist of players to scout in-depth. With the tool, the scouting team will have an initial shortlist of players that could be interesting to analyse. The tool can also be used after the shortlist is already made if the scouting team cannot agree on which player to decide to approach or to analyse more in-depth. The extra information resulting from the tool can thus give a suggestion on which player is more likely to fit the team, or who is more likely to play in a specific playstyle that is desired.

In this way, the scouting team can focus their scouting on specific players, and thus they can watch games of a narrower list of players which are expected to fit within the current squad at San Lorenzo. This tool is thus to be implemented to improve the effectiveness of the making of the shortlist of scouted and potentially to be contracted players, and as a supportive tool in decision making which players to analyse more in-depth.

The tool is developed locally and can be standardized to be usable by the scouting department of San Lorenzo after deployment. This process is considered as a recommendation for the scouting department and the deployment itself is outside of the scope of this thesis. With the evaluation and consideration of the requirements of San Lorenzo, the tool will be fully functional and reusable because of possible deployment. In the recommendation section the usage of the tool will be discussed.

5.3 Conclusion

In this chapter, the designed tool from Chapter 4 is subjectively evaluated by the stakeholders of San Lorenzo. The evaluation is done to conclude the applicability and value of the tool in the scouting process. The tool is found valuable in the search for potentially to be scouted players, for example as replacements of a to be replaced player in the current squad of San Lorenzo. The tool is not only functional in finding replacements for current squad members, but also for finding players that are similar but more accessible than interesting players from other clubs based on their playstyles. The use of the tool to find the initial shortlist of players is found as an expected improvement in the efficacy of the scouting process, as the initial shortlist of players narrows down the search for players.

The next chapter will provide conclusions on the whole thesis and the main objectives. The research process is analysed, and the main suggestions and recommendations are formulated.

6 Conclusions & future research

Within this thesis, clustering and dissimilarity methods were researched and developed. To improve the scouting process at San Lorenzo, these models were implemented in a tool to build a framework for identifying possible future to be scouted players based on similarity with current players. The different chapters gave answers to the sub-questions formulated in Chapter 1.4.2. Together, this built a foundation to answer the main research question:

How can player data be used to derive similarity information for specific player positions?

The answer to this question is discussed in Chapter 6.1 structured based on the sub-questions. Possibilities for future research to build on this thesis are discussed in Chapter 6.2. Finally, recommendations for San Lorenzo are provided in Chapter 6.3.

6.1 Conclusion and discussion

To provide a structured overview of the conclusions based on the research process, the sub-questions are concluded followed by an overall conclusion based on the main objectives of the thesis. The sub-questions are denoted followed by the sub-conclusions and discussions first.

1. **What similarity measurement and clustering models come forward in literature and can be useful for deriving similarity information for the football player dataset?**

Clustering and dissimilarity algorithms that are used within the final tool are valuable in solving the main problem and improving the efficacy of the scouting process at San Lorenzo. The methods are initially found based on literature search and are evaluated based on the applicability in the thesis based on the requirements and objectives. These methods are further evaluated in the design of the tool, to be discussed below, and are used to conclude the similarities between players. Knowledge about clustering algorithms and dissimilarity metrics is found together with the applicability on different types of datasets and projects, which can be used to further analyse in the design of the method for application with the objectives of the thesis. The clustering algorithms for example include the Gaussian Mixture Model and the DBSCAN methods, where applicable dissimilarity metrics include for example the Euclidean and Manhattan distance measures. These algorithms are further analysed to conclude the most applicable within the final to be designed tool.

The current explored options in terms of clustering algorithms and dissimilarity metrics provide options which can be assessed to find an applicable method to be implemented within the designed tool. An even wider and deeper analysis of these types of algorithms, although out of the scope of this thesis, can always provide more options and potentially even better fitting models through combinations of multiple methods for example.

2. **How can the CRISP-DM methodology be further specified and designed for application on the football scouting dataset and objectives?**

To use player data in a way to derive similarity information for specific player positions and increase the expected probability of a potentially to be scouted player fitting the squad, a method was designed for San Lorenzo. The CRISP-DM methodology applicable for data mining projects is adapted, with inclusion of the clustering methods and dissimilarity metrics studied in literature, for application on the football player datasets and the requirements and objectives of the thesis. The requirements found are based on the objectives of the thesis and follow the context of the football club. This for example comes forward with applicability of the method based on the transfer philosophy, and the ease of use within the scouting process to allow for implementation and improvements in efficacy. Relevant steps are added to allow for a complete method that can be used to find the results with the

deliverable as the data scouting tool. The focus is on the implementation of dissimilarity metrics and clustering algorithms to allow for technically founded results and usable results within the designed tool. Applying this method made it possible to develop the tool conforming the requirements of San Lorenzo.

The newly designed method includes all aspects of a data-mining project focused on the football player dataset provided. To allow for even more accurate results, deeper analysis can be useful in for example the application of the weights on the attributes for the different position groups. Where currently rounding is used to distinguish the importance between the attributes, a method could be applied to make the weights more specific based on the importance in the principal component analysis as conducted. This can be done in further research on the topic to improve the academic background and accuracy of the redesigned method.

3. **What are the characteristics of the available dataset, and how can this be used and prepared for deriving similarity measures?**

An exploratory data analysis was conducted to provide an overview of the dataset and its characteristics. The player dataset is high-dimensional, requiring the application of dimensionality reduction in the preparation of the dataset through principal component analysis. Together with domain knowledge, importance in attributes could be discovered to allow for weighting of attributes in preparation of the applied clustering and dissimilarity algorithms in the designed tool. The dataset is cleaned and transformed to make the algorithms applicable and allow for presentation of the results as required. The similarity between players playing in different positions, together with domain knowledge, allowed for the formulation of different position groups with players that have specific characteristics and gave the possibility to analyse the differences between players even more in-depth.

Within this thesis, the exploratory data analysis is focused on the provided dataset with the players from the South American competitions. An even more complete analysis with inclusion of other datasets from other competitions could have allowed for interesting discoveries of for example differences in types of players between different continents. The used dataset is complete for the application of the tool for the scouting process of San Lorenzo, as the focus is on the players playing in the South American competitions. However, in case the scouting tool would be used in other continents as well, differences in types of players could be different and would for example lead to different weighting of attributes in the position groups.

4. **How can a tool be designed to derive similar players per position?**

Based on clustering methods and dissimilarity metrics, a tool could be created to derive the similarity information between players. Based on the requirements and characteristics of San Lorenzo, this tool was developed to achieve the goal of this thesis, increasing the probability of a newly scouted player to replace a to be replaced player within the squad.

The tool is developed to be used by the scouting process and provides the user information on the similarity between players and suggests a shortlist of potentially to be scouted players based on the player that is to be replaced. The transfer philosophy of San Lorenzo can be conformed by using filter options within the tool, and the results following from running the tool are statistically founded based on the clustering and dissimilarity models derived from literature.

Based on evaluation of the tool by the scouting department, the usability of the tool can be concluded. The scouting team at San Lorenzo can use this information to see which players can fit their current

team, where the players with the lowest dissimilarities suggest a most fitting and similar style of play based on attributes. A player that is deemed to have an interesting playstyle can also be compared to find players that have a similar playstyle and can thus play in a way which is deemed fitting to the playstyle of San Lorenzo by the scouting team. Furthermore, the clustering graph gives an overview of players that are close to each other, which makes it possible to find potentially valuable players (with a current low value) in case they are close in terms of statistics to players with a higher value. The most effective clustering model (GMC) is concluded through an analysis of the average and standard deviation of the silhouette scores, as shown in Table 6.1, which shows the clustering model with the highest and most consistent separability. Weighting of attributes is chosen because of the, on average, higher and thus better silhouette coefficient of -0.087 versus -0.204 in case no weights are used.

Table 6.1: Summarization of silhouette scores per clustering model.

Clustering model	**Cure**	**DBSCAN**	**GMC**
Average silhouette score	-0.17686	-0.12362	-0.01467
Average standard deviation	0.1988	0.111653	0.049318

In the design of the tool based on the algorithms found within research, the completeness of analysis could potentially be further improved with analysis on other aspects of the player performance as well. Current focus is laid on the technical abilities of the players, whereas the mental side and fit in types of playstyles of the team can be further analysed in case more data would be added in the design of the tool. As further explained in the future research, more in-depth statistics could provide information about the player's consistency for example. This data was currently unavailable for the South American competitions. Furthermore, improvements in performance of the clustering algorithms could potentially be found in case an algorithm would be designed specifically for the objectives of this thesis, with focus on the specific characteristics of the dataset. This could have potentially increased the separability of the clusters and thus the silhouette scores.

5. What is the potential added value of implementing the tool at San Lorenzo?

The tool is evaluated as an asset in deriving similarity information between a to be scouted and a to be replaced player in the squad. The results from the tool solve the core problem stated in Chapter 1.3.2, which improved the expected capability of replacing a current player in the squad, formulated as the action problem in Chapter 1.3.1. This is indicated with the possible buy of young talent Player 7 at the end of season 2018-2019 (see Figure 6.1), with his increasing value and thus good fit within the transfer philosophy of San Lorenzo, which is focused on buying young cheap players with high potential. The player would have been shortlisted as a potentially to be scouted player because he is young and because of his similarity with N. Reniero. The additional value is found in the expected improved efficacy of the scouting process, and the possibility to find similar players to a highly rated player by the scouting team.

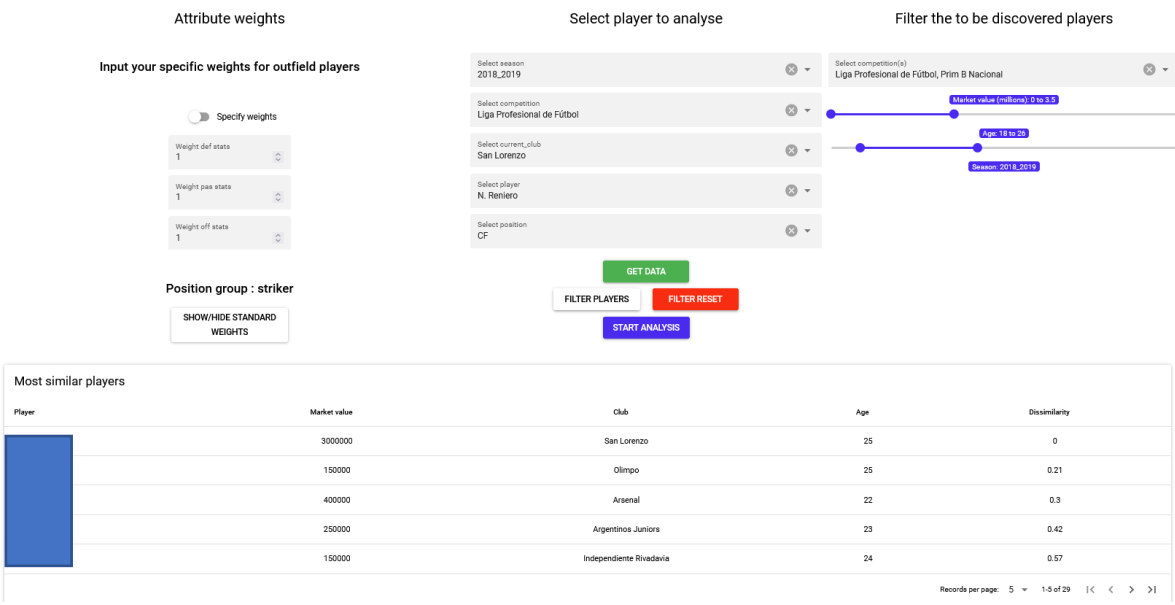

Figure 6.1: Design of the developed tool including the clustering and dissimilarity algorithms.

To further improve the evaluation of the tool in application of the scouting process, it would be useful to analyse the tool in production and see where the practical results come forward. This requires complete implementation of the tool, which is out of the scope of the thesis, but would give more validated results on the evaluation of the tool. Furthermore, the tool could have been evaluated with more football clubs and scouting teams, to look for even more suggestions and points of improvement, and have more perspectives on the quality of the tool in use in standardized scouting processes.

To summarize, player data can be used to derive similarity information between players through the application of different clustering and dissimilarity algorithms in a tool made for implementation at San Lorenzo. Through use of the tool, the scouting team can find similar players based on technically founded algorithms and improve the efficacy in the scouting process with the initial shortlist of potentially to be scouted players.

6.2 Future research

The scope maintained within this thesis, led to certain limitations and points for future research. These will be discussed in this section, together with discussion points of applied methods.

- Even though it is not entirely related to the methods applied within this thesis, the possibility of automating the quantification of the player reports from the traditional scouts is an interesting possible future project. Currently, the scouts manually annotate positive and negative points about the players while analysing the players from matches. These positive and negative points say a lot about the player quality and possible suitability to San Lorenzo. Because there are a lot of analyses on many players, it is an intensive and long process to compare all the players and review them based on the analyses made. With automatic identification of positive, neutral, or negative feedback on a certain player from a scouting report, a general overview of the players can be made more efficiently and quickly. This will make it possible to quickly identify which players have had the most positive analysis, which can be supportive in the decision-making which players are to be scouted further or to be negotiated with.
- The scope of the thesis is focused on the core problem observed within the scouting department at San Lorenzo; the lack of a data-driven analysis on the similarity between a

scouted and to be replaced player in the squad. This limits the research on the data-driven aspects of data-scouting. Even though it was not considered as the most important problem to be dealt with, it could be interesting to dive deeper into the scouting process and look at improvements in the more traditional steps of the process. This can for example be done with the first point formulated within this section on future research, focusing on the efficiency in the quantification of the player reports. Furthermore, partly in cooperation with the suggested tool, a more efficient process can be structured for the analysis of players through watching football matches. With the initial shortlisted players, a system can be made which can find recently played matches which include many players that come forward within the shortlists, and thus are interesting for the scouting department to further analyse and scout.

- As already mentioned in the conclusion and discussion, the weighting factors are currently rounded to halves. This provides a usable initial weighting list of the attributes for can be further specified upon in future research to optimize the reflected influence of attributes on the classification of the players. With this, the theoretical background can be further improved, and the results will be more specific and focused.
- The used dataset within this thesis is focused on aggregated statistics per season, per player. With the growth of the importance of data, more detailed data is expected to become available within the future. Match event datasets, which are currently mostly available for European top competitions, can be very useful in deeper analysis of the players in terms of the values of player actions or chemistry between players for example. The values of players can be extracted through analysis of the impact a certain action of a player has on the game, where for example a dribble from a low-threat position into the danger zone can be of high value compared to a simple back-pass to the goalkeeper. Similarly, the influence of combinations of players within the team in performance of the squad can give an idea on the chemistry between players and how a coach should approach the formation of the squad.
- Continuing upon the above point, the tactics of a squad can be incorporated more within the analysis, if more detailed data would be available. With the analysis on for example distance ran, pressing, and aggressiveness, the fit of a player in a high-pressing squad can be argued more in-depth. Similarly, the mindset of a player can be analysed with information per game, where a player could be more consistent if the fluctuations in his statistics are low between games or comparing won games versus lost games.
- Currently, the average gametime of a player within the games played is not considered specifically. It is probable that there will be natural differences in statistics between a player that was substituted onto the pitch ten times for 30 minutes, compared to a player that played three matches completely. Because of the extensive dataset, the influence of this phenomenon is expected to be low, as most of the statistics are based on most of the games. Because of the filters on players that played at least five games and 300 minutes, this is considered out of the scope for this thesis.
- Currently, the analysis of the players is based on the position groups where they were active in that season. This means that they will not be considered for analysis and comparison on different positions within the field. This could be possible to incorporate in future research, where a player could have statistics that would fit another type of position. A winger could be a player with high shooting accuracy and composure in front of the goal, which could be interesting attributes for a striker. In that case, they could be fit to another position, even if they have not played that position before. To incorporate this, it could be interesting to adapt the workflow to allow the option to include more initial positions within the analysis of the player dataset. Instead of thus focusing on players that have already played the position of

the to be compared player, the scout would be able to select different positions that should also be included within the analysis.

- Additions can be made to the current version of the local tool, for example a useful addition would be a football field visualisation with similar players for all positions compared to the current squad at San Lorenzo. In this way, the scouting team would be able to see quickly which players would be potential replacements in every position of the current squad.

6.3 Recommendations

Finally, recommendations will be given to San Lorenzo for using the designed tool. The practical use and added value of the tool will be described.

The goal of the thesis is the increase the probability that a newly scouted player can replace a player within the squad. With the provided tool, the scouting department can discover dissimilarity information between interesting players within the scouting process. To achieve the goal and improve the efficiency of the scouting process, the scouting department of San Lorenzo is recommended to use this tool. Through selecting a potentially to be replaced player, the scouting department can easily and quickly get an overview of players that are similar to the to be replaced player, which gives them an additional perspective on the players.

Furthermore, the list of similar players that is the result of the use of the tool, will already provide a focusing point within the scouting process. The scouting department does not have to look at all matches played within the scouted competitions, as they can directly look at the similar players and focus on the matches between teams for which these players play. It is thus suggested to use this tool as a first step of identifying possibly interesting and potentially to be scouted players.

To conform the transfer philosophy and the culture of the club, the filter options provided within the tool should be used to find players that would fit the club in terms of value and age. By filtering on the budget that is available for a set replacement, and the age of the potentially to be scouted player, the young affordable players can be discovered. Also, the clustering results shown in the tool should be looked at by the scouting team to discover lesser known and low-valued players that have similar statistics compared to higher valued players. This can be found through seeing where in the cluster higher valued players are positioned (the bigger sized data points) and identifying lower valued players (with smaller sized data points) close to these higher valued players.

Within the thesis, the tool was developed locally for evaluation and experimentation. As a recommendation for San Lorenzo, the tool can be further deployed and implemented within the scouting process, which was out of the scope of this thesis. For deployment, first, a plan can be made, including the monitoring and maintenance of the tool during the operational phase. Afterwards, the tool can be assessed for implementation. Finally, the final validated tool can be deployed for use by San Lorenzo.

In further use within or outside of the field, the scouting tool can be an interesting fit for other football clubs as well. The tool can already be easily used by other football clubs, as there is no exclusive search for players that play for San Lorenzo. With the change in used dataset, the tool can be easily used based on football players in other competitions as well. In this case, the dataset should have the same configurations (attributes). Otherwise, minor changes are required to make the tool applicable on different datasets. Interesting to think about is also the possible implementation in other sports or even other fields such as human resource. For other sports, changes are required to be made to include the probable different attributes considered for the analysis of the players. For other fields such as human resource, the principles used within the designed tool can be used to make a tool that

can find employees that are able to replace other employees that work in a particular field and have specific strong characteristics for example.

The similarity measures resulting from the application of the algorithms using the tool give a technically founded indication of the similarities between players and can be used in further scientifical research on this topic. The similarities are the result of dissimilarity metrics and are focused on the important attributes for player position groups. Within the academic field, this can be used as a baseline to define the dissimilarities between players to gain more knowledge about fitting players and playstyles of players for example. However, these can also be used outside of the football domain with some adaptation. Application of the dissimilarity measures on other fields can for example provide similarities between players in other sports, or even in general human resource as explained above.

Although the CRISP-DM method is a generalized method applicable for data-mining research, the method was adapted for application of the specific objectives of this thesis. Even though the CRISP-DM method is complete, the focus on the specific algorithms and models used is broadened with the designed method in this research. The CRISP-DM method is thus not seen as an insufficient methodology per se, but the focus on the clustering algorithms and dissimilarity metrics was desired to provide the best possible results within this thesis. The redesigned method is thus specified on data-mining projects including these types of algorithms and can be used in the future as an adaptation of the CRISP-DM method.

With the application of above formulated recommendations, the fit of potentially to be scouted players is expected to be improved. Furthermore, a lower number of players on the initial shortlist of to be scouted players will lead to an expected increase in the efficiency in traditional scouting.

References

4Q December 2021 CONMEBOL Ranking. (2021). Retrieved from https://www.kickalgor.com/2021/11/30/4q-december-2021-conmebol-ranking/

Akhanli, S. (2019). *Distance construction and clustering of football player performance data.*

Bransen, L., & Van Haaren, J. (2020). *Player Chemistry: Striving for a Perfectly Balanced Soccer Team*.

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics, 3*(1), 1-27. doi:10.1080/03610927408827101

Carpita, M., Ciavolino, E., & Pasca, P. (2021). Players' Role-Based Performance Composite Indicators of Soccer Teams: A Statistical Perspective. *Social Indicators Research, 156*(2-3), 815-830. doi:10.1007/s11205-020-02323-w

Decroos, T., Bransen, L., Van Haaren, J., & Davis, J. (2018). Actions Speak Louder Than Goals: Valuing Player Actions in Soccer.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological), 39*(1), 1-22. doi:https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

Dunn†, J. C. (1974). Well-Separated Clusters and Optimal Fuzzy Partitions. *Journal of Cybernetics, 4*(1), 95-104. doi:10.1080/01969727408546059

Ernest, M. (2018). The importance of scouting. *IFBI - International Football Business Institute*.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). *A density-based algorithm for discovering clusters in large spatial databases with noise.* Paper presented at the kdd.

Gower, J. C. (1971). A General Coefficient of Similarity and Some of Its Properties. *Biometrics, 27*(4), 857-871. doi:10.2307/2528823

Granville, V. (2019). How to Automatically Determine the Number of Clusters in your Data – and more. Retrieved from Data Science Central website:

The Growing Importance of Football Analytics. (2021). Retrieved from https://soccerment.com/the-importance-of-football-analytics/

Guha, S., Rastogi, R., & Shim, K. J. (1998). CURE: An efficient clustering algorithm for large databases. *27*(2), 73-84.

Heerkens, H., & van Winden, A. (2017). *Solving Managerial Problems Systematically*. Groningen: Noordhoff uitgevers.

Hennig, C., & Hausdorf, B. (2006). Design of Dissimilarity Measures: A New Dissimilarity Between Species Distribution Areas. In (pp. 29-37).

Hennig, C., & Liao, T. F. (2013). How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 62*(3), 309-369. doi:https://doi.org/10.1111/j.1467-9876.2012.01066.x

Hernández, C. R. (2018). This is the power that soccer has in Latin American politics. *LatinAmericanPost*. Retrieved from https://latinamericanpost.com/24303-this-is-the-power-that-soccer-has-in-latin-american-politics

Jolliffe, I. (2011). Principal Component Analysis. In M. Lovric (Ed.), *International Encyclopedia of Statistical Science* (pp. 1094-1096). Berlin, Heidelberg: Springer Berlin Heidelberg.

Kaufman, L., & Rousseeuw, P. (1990). *Finding Groups in Data: An Introduction To Cluster Analysis*.

Legány, C., Juhász, S., & Babos, A. (2006). *Cluster validity measurement techniques*.

Market values. (2022). Retrieved from https://www.transfermarkt.com/spieler-statistik/marktwertaenderungen/marktwertetop

Maymin, A., Maymin, P., & Shen, E. (2011). NBA Chemistry: Positive and Negative Synergies in Basketball. *International Journal of Computer Science in Sport, 12*. doi:10.2139/ssrn.1935972

McLachlan, G. J., & Peel, D. (2004). *Finite Mixture Models*: Wiley.

Meh, K. (2022). San Lorenzo de Almagro. Retrieved from https://en.wikipedia.org/w/index.php?title=San_Lorenzo_de_Almagro&oldid=1074221703

Milligan, G. W. (1996). CLUSTERING VALIDATION: RESULTS AND IMPLICATIONS FOR APPLIED ANALYSES. In *Clustering and Classification* (pp. 341-375).

Ng, R. T., & Han, J. J. (2002). CLARANS: A method for clustering objects for spatial data mining. *14*(5), 1003-1016.

Pedregosa, F. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*, 2825-2830.

Possession-adjusted. (2022). Retrieved from https://dataglossary.wyscout.com/p_adj/

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53-65. doi:https://doi.org/10.1016/0377-0427(87)90125-7

Saji, B. (2021). In-depth Intuition of K-Means Clustering Algorithm in Machine Learning.

Sarstedt, M. (2019). Revisiting Hair Et al.'s Multivariate Data Analysis: 40 Years Later. In (pp. 113-119).

Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O. P., Tiwari, A., . . . Lin, C.-T. (2017). A review of clustering techniques and developments. *Neurocomputing, 267*, 664-681. doi:https://doi.org/10.1016/j.neucom.2017.06.053

Sheridan, T., & Verplank, W. (1978). *Human and Computer Control of Undersea*

Teleoperators. Retrieved from Cambridge:

Smith, R. (2020, 08-10-2020). A Deep Pool of Soccer Talent Is Drying Up. Why? *The New York Times*.

Soto-Valero, C. (2017). A Gaussian mixture clustering model for characterizing football players using the EA Sports' FIFA video game system. *RICYDE: Revista Internacional de Ciencias del Deporte, 13*(49), 244-259. doi:10.5232/ricyde2017.04904

Vigneau, E., & Chen, M. (2016). Dimensionality reduction by clustering of variables while setting aside atypical variables. *Electronic Journal of Applied Statistical Analysis, 9*, 134-153. doi:10.1285/i20705948v9n1p134

Vigneau, E., & Qannari, E. M. (2003). Clustering of Variables Around Latent Components. *Communications in Statistics - Simulation and Computation, 32*(4), 1131-1150. doi:10.1081/SAC-120023882

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*.

Xu, D., & Tian, Y. (2015). A Comprehensive Survey of Clustering Algorithms. *Annals of Data Science, 2*(2), 165-193. doi:10.1007/s40745-015-0040-1

Appendix

Appendix A: Method application - figures

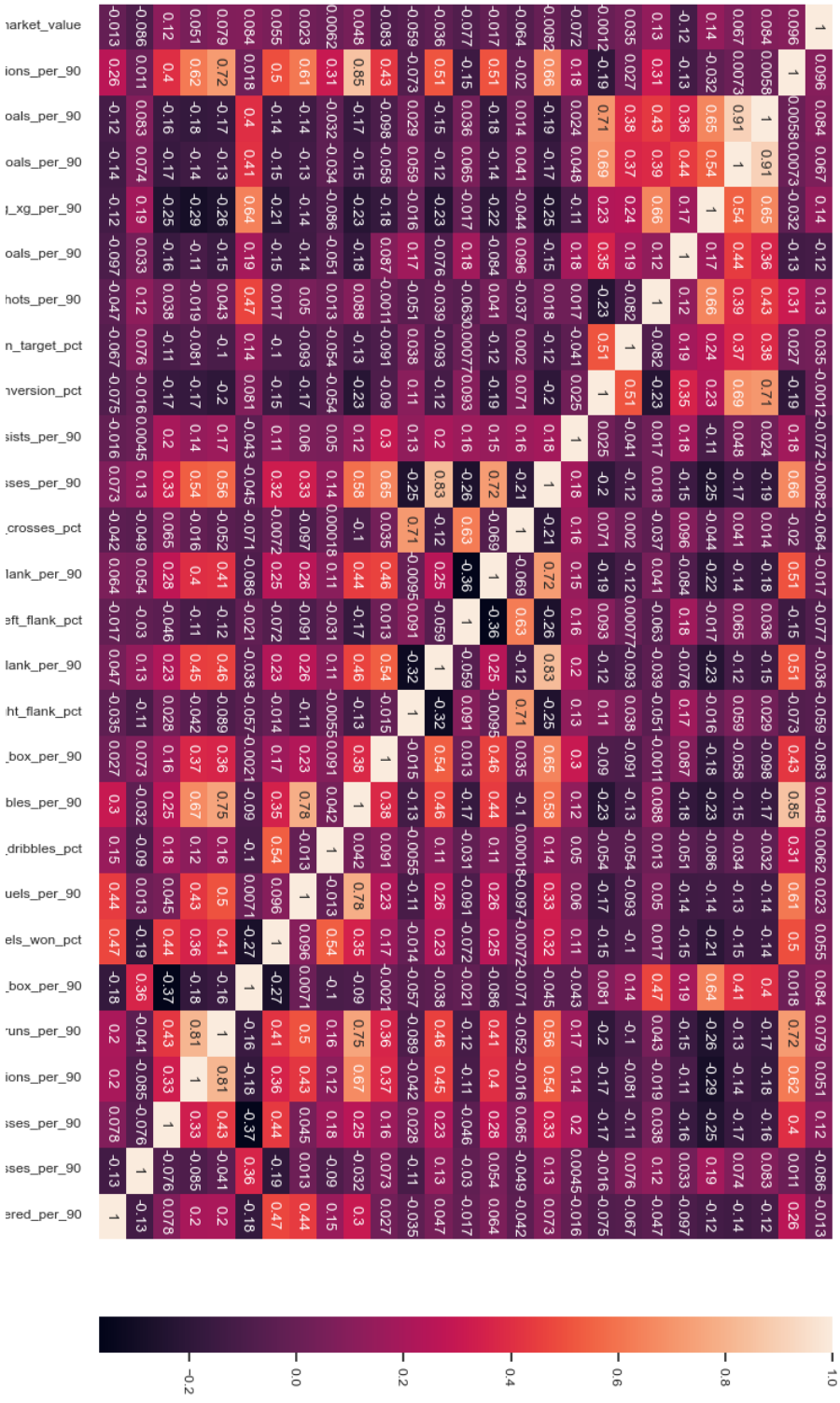

Figure A1: Correlation between attacking performance attributes of strikers.

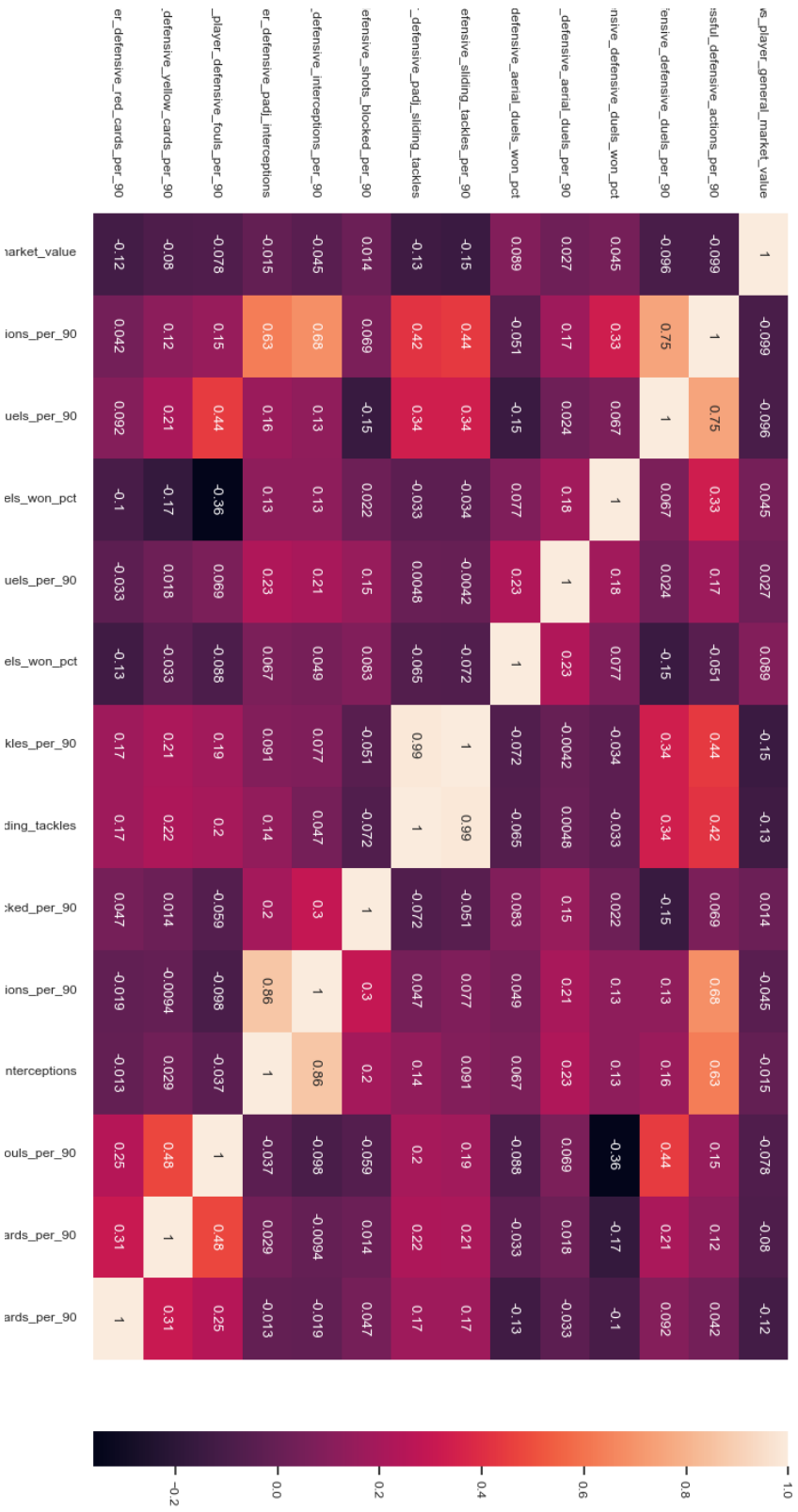

Figure A.2: Correlation between defensive performance attributes of centre-backs.

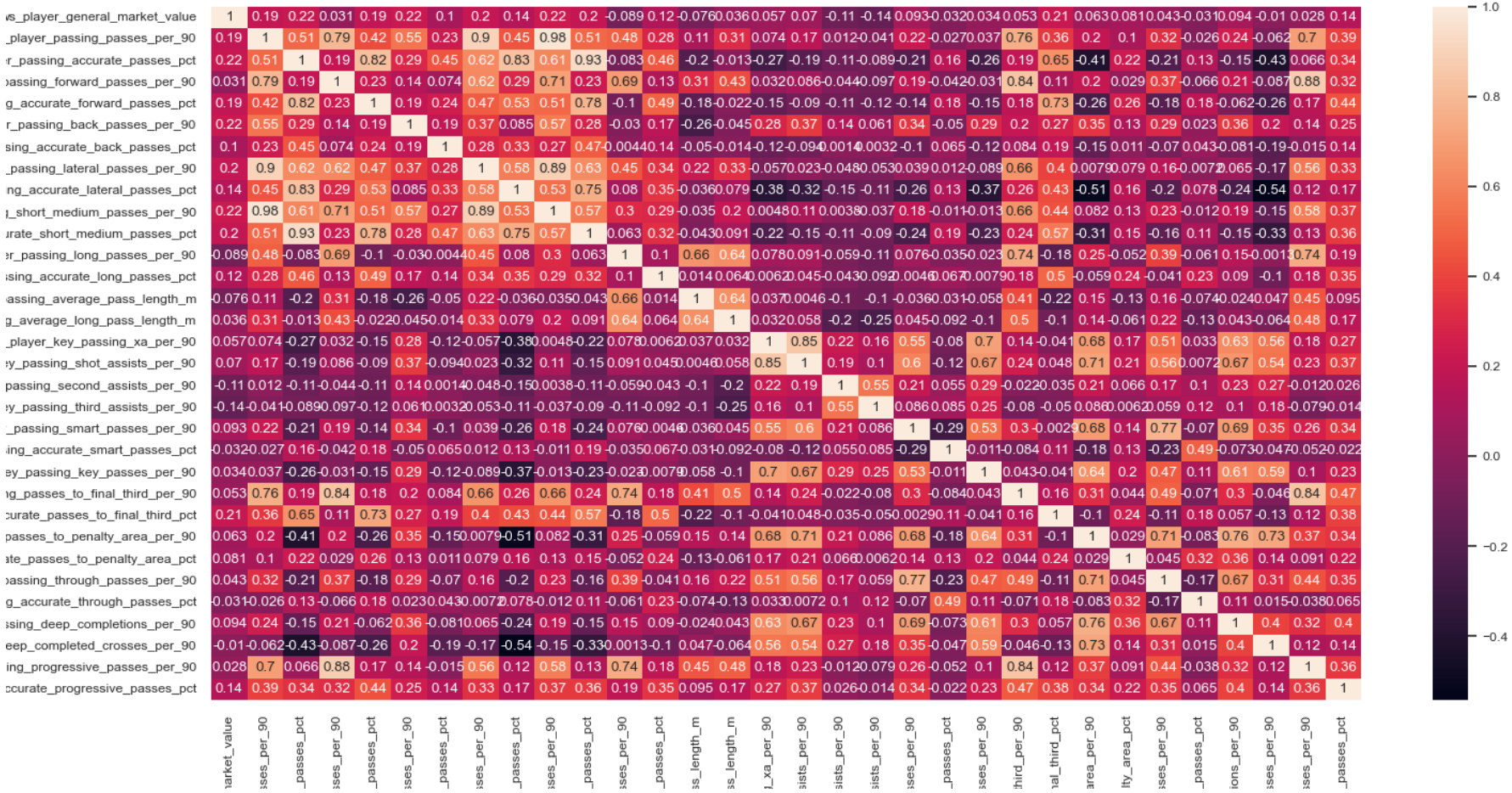

Figure A.3: Correlation between passing attributes of central midfielders.

	in...w...	ws_player_goalkeeping_conceded_goals_per_90	ws_player_goalkeeping_clean_sheets_pct	w...ti...	w...w...	w...ws	ws_pl...	w...w...	w...cl...	player	Component 1	Component 2	Component 3	dissimilarity	Segment clarans PCA
0	...	0.73000	0.61000	...	1	...	...	GK	...	J. Rojas	0.03056	-0.20988	0.11364	0.42865	3
1	...	0.89000	0.56000	...	1	...	...	GK	...	R. Quiñónez	-0.25094	0.09659	-0.06641	0.39039	5
2	...	0.75000	0.63000	...	1	...	...	GK	...	G. Viscarra	0.00838	0.12210	0.03916	0.00000	3
3	...	0.86000	0.46000	...	1	...	...	GK	...	A. Giménez	-0.26955	0.15523	0.04854	0.32044	5
4	...	0.76000	0.82000	...	1	...	...	GK	...	S. Mustafá	0.07938	-0.00133	-0.13311	0.36670	2
5	...	0.61000	0.81000	...	1	...	...	GK	...	R. Cordano	0.30216	-0.13971	0.05186	0.56829	4
6	...	0.71000	0.59000	...	1	...	...	GK	...	F. Laforia	0.03769	0.04466	0.12856	0.19616	3
7	...	0.73000	0.76000	...	1	...	...	GK	...	J. Araúz	0.10026	0.08792	-0.06719	0.23241	3
8	...	0.53000	0.90000	...	1	...	...	GK	...	E. Arauz	0.47246	-0.05235	0.05009	0.64947	1
9	...	0.80000	0.72000	...	1	...	...	GK	...	C. Franco	-0.04025	0.01353	-0.09520	0.29155	3
10	...	0.69000	0.94000	...	1	...	...	GK	...	D. Zamora	0.29284	0.19739	-0.15312	0.55203	1
11	...	0.56000	0.97000	...	1	...	...	GK	...	I. Brun	0.48580	0.19801	-0.05656	0.64904	1
12	...	0.67000	0.61000	...	1	...	...	GK	...	R. Banegas	0.12583	0.18840	0.12983	0.27443	1
13	...	0.39000	0.83000	...	1	...	...	GK	...	Jacsson	0.64325	0.14381	0.24570	0.86312	1
14	...	0.64000	0.92000	...	1	...	...	GK	...	J. Cuéllar	0.31928	-0.07724	-0.07892	0.62832	2
15	...	0.70000	0.81000	...	1	...	...	GK	...	D. Torrico	0.18006	0.07997	-0.05961	0.31258	2
16	...	0.81000	0.71000	...	1	...	...	GK	...	D. Vaca	-0.06889	-0.09678	-0.08496	0.42027	3
17	...	0.59000	0.83000	...	1	...	...	GK	...	A. Arancibia	0.33976	-0.31674	0.07637	0.80743	4
18	...	0.67000	0.81000	...	1	...	...	GK	...	L. Cárdenas	0.22897	0.10425	-0.02337	0.30098	2
19	...	0.59000	0.79000	...	1	...	...	GK	...	J. Careaga	0.33584	-0.19307	0.09379	0.69727	4
20	...	0.74000	0.61000	...	1	...	...	GK	...	A. Cousillas	0.00968	0.14851	0.06651	0.05506	3
21	...	0.26000	nan	...	1	...	...	GK	...	E. Casas	0.68726	-0.11908	0.51727	1.39817	4
22	...	0.84000	0.44000	...	1	...	...	GK	...	Santos	-0.25677	0.15616	0.08020	0.34025	5
23	...	0.84000	0.66000	...	1	...	...	GK	...	Cássio	-0.11114	0.14184	-0.07597	0.25439	5
24	...	0.89000	0.42000	...	1	...	...	GK	...	Diego Alves	-0.35472	0.04662	0.04970	0.44913	5
25	...	0.81000	0.59000	...	1	...	...	GK	...	Éverson	-0.13069	0.05016	-0.00755	0.25771	5
26	...	0.91000	0.56000	...	1	...	...	GK	...	R. Fernández	-0.28001	0.21303	-0.08433	0.50281	5
27	...	0.88000	0.50000	...	1	...	...	GK	...	Marcelo Lomba	-0.29338	-0.02628	0.01542	0.47388	5
28	...	0.93000	0.25000	...	1	...	...	GK	...	Marcelo Grohe	-0.53936	-0.13892	0.15488	0.92448	5

Figure A.4: Dissimilarity and clustering results for goalkeepers in the dataset, showing selected attributes and Principal Component values for a selected number of players in the dataset.