Universitat de Girona
Escola Politècnica Superior

Treball Final de Màster

Estudi: Màster en Ciència de Dades

Títol: Acquisition and analysis of data from futsal matches

Document: Memòria

Alumne: Llorenç Peirau Gabarrell

Tutor: Marc Comas Cufi
Departament: Informàtica, matemàtica aplicada i estadística
Àrea: Estadística i investigació operativa

Convocatòria (mes/any): Juny 2022

Universitat de Girona
Escola Politècnica Superior

TREBALL FINAL DE MÀSTER

Acquisition and analysis of data from futsal matches

Autor:
Llorenç PEIRAU GABARRELL

Juny 2022

Màster en Ciència de Dades

Tutor:
Marc COMAS CUFI

Summary

The domain of this master's thesis is futsal. More precisely, this thesis has been elaborated in collaboration wit the futsal team of Girona: *Girona Escola de Futbol Sala* (GEFS). The idea of this club is to grow as much it is possible in order to become a futsal reference. However, it must be taken into account that the growth of a sports club is closely linked to the performance of the matches. In fact, this is where the main motivation and justification for this master's thesis lies.

Each weekend, GEFS collects two sheets of paper with data about the match (papers *M* and *E*). The club collects all this data because it wants to make data-driven decisions regarding the performances of the players. So, for this thesis there have been defined these three aims:

- Digitalize existing data and propose new ways of recording match information to facilitate automatic extraction in the future.
- Structure the data for its maintenance, management, and analysis.
- Use the data to better understand the matches, investigate the performance of each player, improve collective and group performance and facilitate the preparation of the following training sessions.

To achieve them the results have been divided into two groups. On the one hand, the non-visual results are related to the first two aims. It has been proposed an automatic image process to extract the data directly from a photo. In other words, the data that GEFS collects by hand in two sheets of paper can be digitalized with only taking one photo. Also, it has been proposed a new way to collect the data in order to improve this process. Furthermore, it has been designed a relational database that stores all the collected data in a coherent way. It is divided into static and growing datasets and it is located at the Google Cloud Platform. Every time a match is played, it has been created a pipeline that digitalize the data and, afterwards, it appends it to the correspondent table from the database.

On the other hand, the visual results are linked to the third aim. There are three outputs. First, it has been produced an interactive dashboard with the requested graphs by the coach of GEFS. Secondly, it has been used the Principal Component Analysis to determine that the features *Lost*, *Tackle*, *TG_received* and *TG_scored* are the most relevant to discriminate between the performances of

the players. Lastly, to estimate the density function and to cluster the players there have been used Gaussian Mixtures. The adjusted model is an EEE (ellipsoidal, equal volume, shape, and orientation) with 7 components. Consequently, there have been created seven clusters. Each one had their own logical interpretation.

Agraïments

Per començar vull agrair molt especialment la paciència de na Carme amb la meva organització. Reconec que he de millorar la gestió del tempos. Per sort, amb aquesta entrega ja s'han acabat totes.
D'altra banda, també voldria fer especial menció al club de futbol sala que m'ha prestat les dades, el Girona Escola de Futbol Sala. Gràcies per deixar-me-les però sobretot gràcies per aquesta gran temporada. Som-hi GEFS! Som-hi Girona! Som de TERCERA NACIONAL!

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

Nowadays, the city of Girona has become a reference in terms of sports. The city has four teams in the highest national leagues of: men's soccer, men's roller hockey, and both women's and men's basketball. In addition, there are also other clubs that present very ambitious projects to reach the highest level, such as the city's futsal club: Girona Escola de Futbol Futsal (GEFS). This club works with a short-term objective of becoming the futsal reference in the province of Girona, and in the long-term to consolidate itself in the semi-professional categories of national futsal.

So, the idea of this club is to grow a lot and as fast as possible, but it must be taken into account that the growth of a sports club is closely linked to the performance of the matches. In fact, this is where the main motivation and justification for this master's thesis lies.

Each weekend, GEFS collects two sheets of paper with data about the match. On the one hand, each players minutes are collected (the minute in which a player enters or leaves each time). In futsal, changes between players are unlimited. On the other hand, the second sheet collects specific statistics about the match such as shots, goal chances, corners, etc.

The club collects all this data because it wants to perform the decision-making based on data, not on feelings or thoughts about the performances of the players. Consequently, in order to make data-driven decisions, this master's thesis aims to:

- Digitalize existing data and propose new ways of recording match information to facilitate automatic extraction in the future.
- Structure the data for its maintenance, management, and analysis.
- Use the data to better understand the matches, investigate the performance of each player, improve collective and group performance and facilitate the preparation of the following training sessions.

The first aim can be broke down in three different points.

1. **Digitalize existing data:** one of the pillars of this thesis. Being able to digitalize the data that GEFS collects in each match in an effective way is going to make the posterior analysis easier and faster. As it is going to be explained later on, this effectiveness on digitalizing the data is going to be achieved trough extracting all the information directly from only one photo per match.

2. **New ways of recording match information:** in this point it lays one of the main issues that this master's thesis has faced. As it has been said before, the data comes, chiefly, from two hand-written sheet papers that one person seating on the bench has full-filled during the match. Usually, there are a lot of typos and corrections done on it. These can be appreciated on the following Figure 1.1 composed of two images.
   On the one hand, the image on the left represents the paper sheet that collects statistics for each match. Throughout the thesis it is represented with the letter *E*. It can be divided in these four parts:

   (a) **Match report (Informe del partit):** at the top of *E* there is some basic information about the match like the league, the date, the hour the match start and GEFS's rival among others.

   (b) **Players available (Convocatòria):** just below it can be found a table of 12 rows, the maximum number of players allowed per match. For each row (i.e. each player) it is recorded the number of tackles, losses, goals, free-kicks committed, red cards and yellow cards received.

   (c) **Match situations (Situacions de partit):** going down a little bit more there are the match situations. This part is divided in two: first and second half. For each half there is a column per team. In here it is collected the number of shots independently where they go (on goal or no), chances of goal, goals, corners, throw-ins, 10m, penalties, far posts and free-kicks.

   (d) **Match Goals (Gols del partit):** finally, at the bottom of this first paper sheet, there are the goals scored. For each one it is recorded the global result, the minute when the goal was scored, the goalscorer, how the goal was scored and the five players of GEFS that were playing in that moment.

   On the other hand, the minutes that each player plays per match is recorded in a paper sheet like the one on the right of the Figure 1.1. In this thesis,

it is represented with the letter *M* and it has as many full-filled rows as players were available for that match. There are four pairs of columns for each half. The first column of the pairs represents the minute when a players gets in the match and the second column is the minute when a player is substituted.

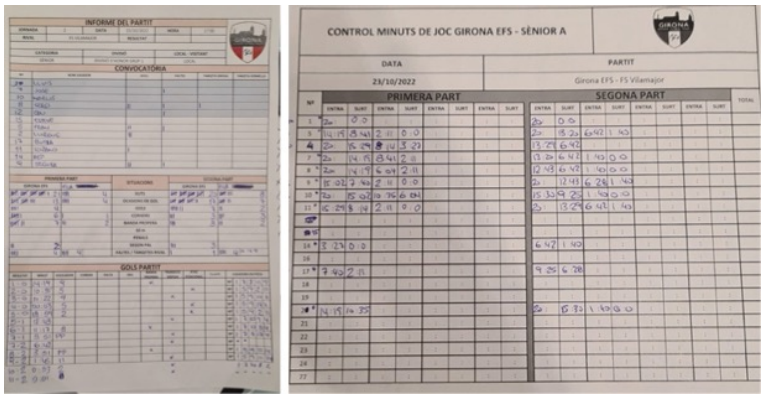

Figure 1.1: Data collected during a match

At this point, one important question is:

How with only one photo, all the information that it is in two distinct photos can be recorded?

Remember, for each match there are these two photos with a lot of typos. Consequently, if both photos are mixed and someone full-fills without the match pressure a more organised and specifically designed template to get all this information in only one photo, it solves the typos problem and generates and efficient way of digitalizing the information.
So, this is exactly what have been done. This next Figure 1.2 shows the template created in order to avoid typos and to summarise all the necessary information that can be extracted from the photos for the posterior analysis.

It must be said that this template is the fifth version and it is thought in a way that:

- Collects all the necessary information
- It is the first step of the database it is going to be created.

The firsts three tables correspond to the paper sheet *E*: players available, match situations and match goals. The next two tables are about information from the paper sheet *M*. The first one for the first half and the second one for the second half.

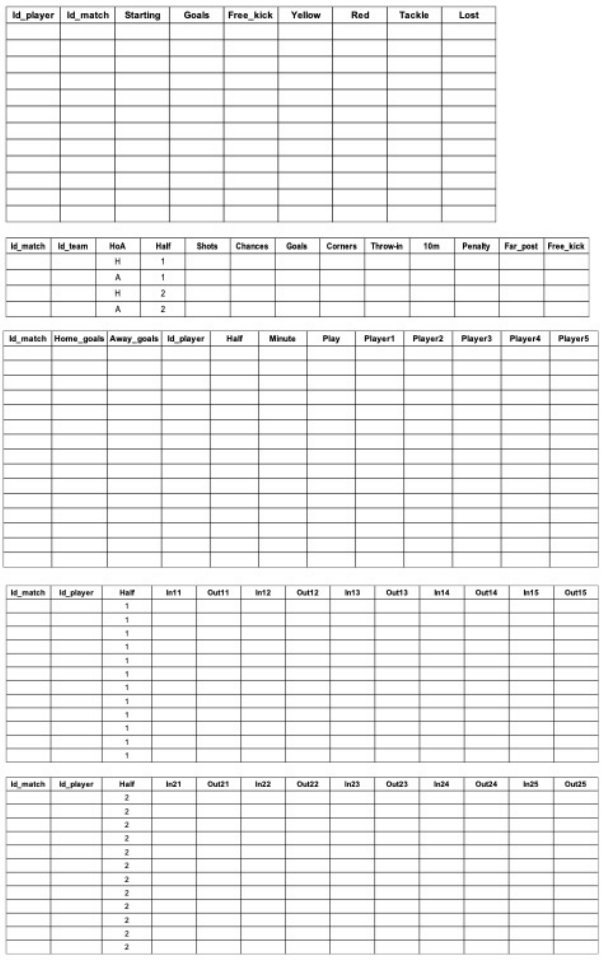

Figure 1.2: Template data collection v5

3. **Facilitate automatic extraction in the future:** finally, the last point of the first aim of this master's thesis proposes a way to automatise the extraction of information. The idea is to create a summary column next to each cell. As it can be checked on Figure 1.1, the person collecting this data annotates a stick every time the event occurs. So, next to the sticks there should be a cell with the number they are representing. Then, it will be just needed to get information from that cell.

   Another improvement that can be done to this data collection is to draw a temporal axis at the *Match situations* part of the paper sheet *E* in order to know when the events occurs. This will help to determine what is the distribution of, for example, the number of shoots in each match. Is not the same arriving at the end of the match winning than loosing. In the first scenario is not that common to shoot whereas in the second scenario you shoot, practically, every time you get close the rival's goal. The next Figure 1.3 represents the improvements mentioned.

Figure 1.3: Improvement for E

One related problem that it has been faced with this first aim, is that for a couple of matches the person in charge of collecting this data was unable to come and sit on the bench. Consequently, there were matches without any information collected by GEFS.

However, when this scenario happened due to the fact the *Federació Catalana de Futbol* publishes information about all the matches from the league each weekend and that GEFS video records all the matches, it was possible to extract information and full-fill the template shown before. It is really important to be aware that the time invested to get the information this way grows exponentially.

For the second aim of this master's thesis (**the structure of the data for its maintenance, management, and analysis**), it has been designed a relational database consisting of 8 different datasets. The content of each dataset is explained afterwards at chapter 3.

The database is conceptually divided in two groups:

- **Static information group:** it has information that does not grow as the matches are played. The information it is already known at the beginning of the season. It comes directly from the knowledge about this topic and the public information that can be found at the web page of the *Federació Catalana de Futbol*: https://www.fcf.cat.

- **Growing information group:** as the match are played more information is stored. In this second group there are the five datasets that comes directly from the full-filled templates.

The next Figure 1.4 shows the structure of the relational database created for this thesis. Each white box represents a dataset.

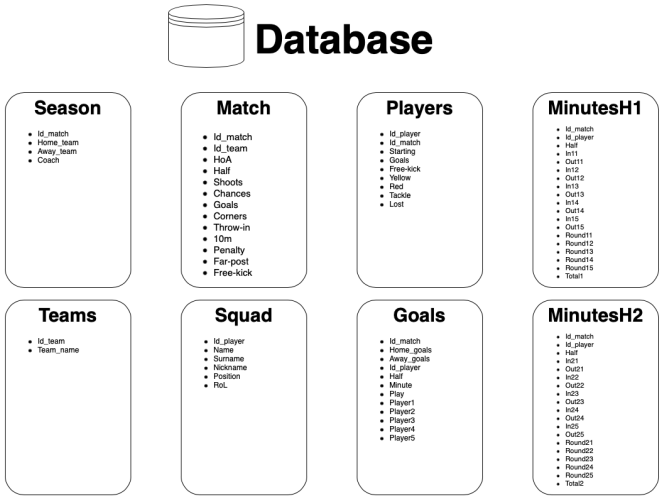

Figure 1.4: Database structure

Once this relational database was thought and created, it was saved to **Google Cloud Platform (GCP)**. More precisely, it has been used the free version of 90 days and around 350€ of credit in order to create a project. Inside this project it has been used **Cloud Storage** to create a bucket called *tfm_gefs* where the relational database created lays. Then it was used **BigQuery** in order to relate the 8 different datasets. This was a very useful tool to combine the data and to get the desired information. At chapter 4 it is explained in detail how this has been done.

Lastly, the third aim (**use of the data to better understand the matches, investigate the performance of each player, improve collective and group performance and facilitate the preparation of the following training sessions**). The order of these three aims can be understood as a waterfall. First the data is digitalized, secondly the data is structured and saved and finally the data is analysed.

So, this third aim can be broke down into two principal goals that at the same time can be interpreted as the two visual results of this master's thesis:

1. **Requested charts with Looker:** it is important not to forget that this thesis is focused in helping a futsal club to: grow, analyse the matches and improve their results. So, one output has to be something dedicated to this public. Consequently, one result of all this work of digitalizing and storing the data is showing an interactive dashboard created with Looker. It contains the three graphics that the coach of GEFS wants:

   (a) **Minutes played:** bar chart with the amount of minutes played per player.

   (b) **Goals scored:** bar chart with the total number of goals scored per player.

   (c) **Presence:** bar chart to compare the number of team goals received with the number of team goals scored per player.

   As it is going to be explained at chapter 4 it has been opted to create this dashboard with Looker because it works easily with BigQuery and the Google Cloud Platform environment generally.

2. **Performance analysis of players:** the second visual result of this thesis is the analysis of the performance of all the players that have been available at least once in this season for GEFS. The techniques chosen for this analysis are:

(a) **Principal Analysis Component**: to determine which features characterise the performance of the players.

(b) **Gaussian Mixtures**: to estimate the density function and to cluster the players by their performance.

Both techniques are detailed in chapter 5. It is important to be aware that the number of observations is small (i.e. 19 players). Consequently, the analysis proposed has to be considered as a first approach. It has to be careful with the results obtained.

About the most related literature that can be found, there is one important thing to keep in mind. This master's thesis is about an sport that, although being know mostly everywhere because somehow is football, is not that popular and there are not as many resources. This means that the amount of publications regarding this sport is much lower compared to the ones that are about ordinary football. In addition, as [Agras 2016] says, the publication in international scientific journals is scarce, in spite of the volume of high-quality studies that are published in journals of an informative nature that are closer to coaches. Also, the language most widely used is Portuguese, which hampers the international dissemination of information.

Additionally, most of the literature that is going to be referenced in the next chapter 2 is about high performance athletes and the data they collect is much more precise and detailed. Remember, that this thesis is working with data from a club that although wanting to become semi-professionals or even professional in the next years, nowadays it is not.

In fact, this is exactly what makes this thesis original and relevant. As it is said, all the literature is thought for professional clubs whereas this thesis is going to show how to enhance and take profit from the data that non-professional futsal clubs collect. It is going to be presented all the necessary steps to be taken from the very beginning such as collecting the data to how to store it, how to analyse it and how to show it.

Finally, to make this thesis reproducible all the links, code and outputs mentioned are available at this GitHub repository:

https://github.com/Lpeirau2/TFM

CHAPTER 2

State of the art

There exists two main objectives for this second chapter. Firstly, it is intended to define the framework as well as to justify why the thesis has worked in this specific direction. Secondly, to demonstrate the usefulness. While showing what is the most related literature with this thesis is going to be defended the importance of it.

First thing to be aware of and that has been mentioned previously, are the affirmations wrote by [Agras 2016] about the low amount of futsal references published in international scientific journals. This provokes a need of combining articles about futsal and related content like football.

In fact, the first important consideration to achieve the second aim proposed in this thesis (structure the data for its maintenance, management and analysis) is to remove the subjectivity on player's performances. The article published by [Kasap 2015] suggests that the database has to bring an objective point of view on the data. This implies that when storing the data no fields such as perceptions or evaluations should be considered. In addition, this points to the same direction as GEFS wants. As it has been said at the introduction, the club wants to make data-driven decisions. They do not want to perform the decision-making based on feelings or thoughts about the performances of the players. Moreover, this article [Kasap 2015] proposes to have a database containing real-time data from the matches but this is out of scope for this thesis.

Furthermore, getting to know a team's performance in an objective way it is a must to prepare future matches and to create a team identity. As [Almeida 2019] says, there are eight principles that should always be in the mentality of the coaches to prepare the matches. From these eight principles, the firsts three (i.e. the most important ones) are extremely related with the third aim of this thesis. They are:

1. Own team identity
2. Have different solutions always prepared according to the team identity and to present in different moments or as an answer to different problems
3. Always know the players and understand the income and outcome of each one

These three points can be understood as the match preparation. So, the next step should be how to go from match analysis to intervention. In other words, how the information that a team collects, in this case GEFS, should be analysed in order to improve the performance of the team and the players for the next match.

In fact, the coach of GEFS is the one who designed the paper sheets *M* and *E* where the data is collected. Accordingly to [Sarmento 2015] this should be done by all the futsal coaches, they should define a list of indicators to be observed and analysed during matches. Additionally, this article ([Sarmento 2015]) also states that the most important aspects to observe in a match are:

1. Global dynamics
2. Key moment of the match
3. Set pieces
4. Individual characteristics of players

Although GEFS and all the professional and semi-professional futsal clubs have their own indicators, the most studied one is: goals. One interesting fact that could be taken into account is the one presented by [Leite 2013]. It is affirmed that the team who scores the first goal in a futsal match is more likely to win. In addition, [Peñas 2014] presents how scoring a goal affects the development of a match. When teams are one goal up, the ball possession, the probability of reaching the final one-third of the pitch and shots on goal decrease. It can be said, that when a team is winning, players turn more defensive.

Following the same argumentative discourse, after performing a discriminant analyse of a range of matches, [Castellano 2012] concludes that the variables related to attacking are the ones that best differentiate between winning, drawing or losing. It is important to be aware that these last two articles are specifically about football.

If the focus is putted uniquely on futsal it is important to take the following points into account. These are showed by [Abdel-Hakim 2014] after analysing the Futsal World Cup played in Thailand in 2012:

- The fourth period (from minute 31:00 until minute 40:00) is when more goals are scored.
- There is no significant difference in ball possession and corners between winning teams and losing teams.

Out of curiosity, FIFA (Fédération Internationale de Football Association), the entity in charge of organising the Futsal World Cup every four years, publishes a report of it once it finishes. As an example, in this thesis the report of the Futsal World Cup celebrated in Colombia in 2016 is referenced [FIFA 2016].

Next, it is relevant to highlight that futsal is a really demanding sport and, as it is known, fatigue has a direct impact on decision-making. As [Castagna 2009] says, Futsal played at professional level is a high-intensity exercise heavily taxing the aerobic and anaerobic pathways. For this thesis, there is no data about the physical capacity of the players because this data is expensive to collect (there is need of GPS sensors) and it is not until professional levels that this is analysed.

However, although this thesis being unable to treat this class of data because it has been worked with a non-professional club, it is important to get an idea of what is the activity profile of a elite futsal player.

In futsal, [Ribeiro 2020] concludes that players increase the distance covered per minute in the second half. Also at this half stress load increases. Moreover, this article suggests that distance covered per minute (m/min), number of sprints (>18 km/h), decelerations (greater than-2 m/s), and metabolic power (W/kg) are the variables that most discriminate the load intensity of elite futsal players.

Finally, the most relevant ideas to keep in mind all along this thesis are summarised. First of all, in order to prepare a futsal match and to get a team ready to win and compete, it is a must to have a team identity ([Almeida 2019]). To create it, there has to be an iterative process of analysing and preparing the matches. Then, to analyse the matches there is need of having a list of indicators ([Sarmento 2015]) to put the focus on. This job is already done by GEFS because its coach has created the paper sheets *M* and *E*.

Moreover, it is important for the posterior analysis to be aware that in futsal matches there seem to be no significant difference in ball possession and corners between winning teams and losing teams ([Abdel-Hakim 2014]).

So, due to the fact that GEFS has already defined the list of indicators there is only need of defining the database where there must not be subjectivity ([Kasap 2015]).

CHAPTER 3

Preliminary concepts and ideas

Once it is known what are the aims and the expected results of this thesis as well as which is the most relevant literature, the next step is to go deeper one the most basic concepts and ideas.

3.1 Domain: futsal

As it has been written, this thesis is about analysing futsal matches. Consequently, the domain is: futsal. This sport although being understood as a football variant, it has much more in common with other indoor sports as basketball or handball rather than football.

For example, the futsal court has the same size as the handball court: 40m long and 20m wide. In addition, the marked lines are mostly the same except for the penalty sport and some other small differences.

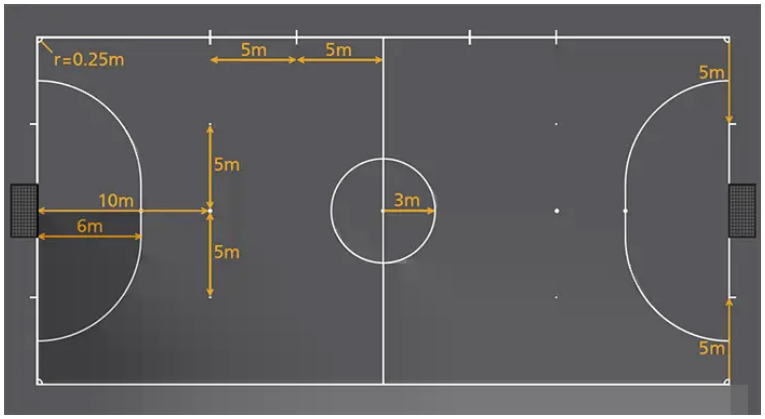

Figure 3.1: Sizes of a futsal court

As it happens in basketball, there are 5 players per team playing and there can be 7 more on the bench. This makes a total of 12 players per team per match. The substitutions are unlimited and when there is one the match keeps going unless someone gets injured.

The matches last for 40 minutes. There are two halves of twenty minutes but every time there is a free-kick, the ball gets out of bounds or it happens

something that makes the match stop, the time also stops. Moreover, each team has one time-out per half.

Another similarity with basketball is that time goes backwards as the players are competing. The time starts at 20:00 minutes and when it arrives at 0:00 minutes the half has finished. Out of curiosity, basketball players have 5 seconds to think how to serve a throw-in whereas futsal players have one second less (i.e. they have 4 seconds). The same happens with the goalkeeper, he can only have the ball for 4 seconds if he is in his half court.

Furthermore, there are 4 different positions in futsal. Usually, it is played with a goalkeeper, a defender, two wings and one pivot.

- **Goalkeeper (G):** is the player that dresses different as the others teammates and the only one who is able to touch the ball with his hands if he is inside his area. It is compulsory to have one. A team cannot play without a goalkeeper. However, sometimes the goalkeeper, in order to help his team to attack, abandons his goal. He leaves it empty to create superiority. This is a very risky play called flying goalkeeper.

- **Defender (D):** is the player that organises the defence and the one who starts the attack. He plays just in front the goalkeeper and, usually, he is in charge of the strategy as corners and free-kicks.

- **Winger (W):** frequently there is one left footed winger and one right footed winger. They play on the wing because they are fast and sacrificed players. They have the same importance while defending than when attacking.

- **Pivot (P):** is the player in charge of scoring goals and helping the team to avoid the defensive pressure of the other team. Usually, is the tall and corpulent player of the team.

As the next Figure 3.2 shows, without taking into account the goalkeeper, the other three different positions form a rhombus.

Figure 3.2: Futsal players positions

Another rule that futsal and basketball share as well as somehow roller hockey does, is the penalisation when an amount of free-kicks have been committed. In futsal, if a team has committed 5 free-kicks during one half, when they make one more, they will be penalised with a 10 meter (10m) free-kick. This is a very important rule because the 10m free-kicks can change a match direction. Most of the teams have a specialist.

Moreover, most of the goals in a futsal match come from strategy (i.e. corners or 10m) or they come from counter-attacks. Losses and tackles made during a match are really important because, most of the times, they lead to one of this two situations.

Furthermore, one common way of finishing the plays in futsal, whatever they are counter-attacks or corners, is scoring at the far-post. This means that, for example, if one play is being played on the left side of the court, it is going to be ended next to the right post. As it is going to be shown afterwards, for this thesis and according on how GEFS collects the data, to each play has been assigned a number:

1. Corner
2. Free-kick
3. 10m
4. Throw-in

5. Counter-attack
6. Positional attack
7. Flying Goalkeeper

Finally, futsal referees can show yellow or red cards to the players. With one yellow card, it is possible the keep playing but if a player receives the second yellow card or a straight red card, he has to abandon the match. The other team is going to have one more player for two minutes or less if they score a goal within these two minutes.

3.2 Setting: GEFS

In this new section it is going to be talked about what is the current situation of GEFS, the club that owns the data. Remember that GEFS stands for *Girona Escola de Futbol Sala*.

3.2.1 Matches played

Nowadays, GEFS is competing at the fifth national futsal league that is called "Divisió d'Honor Catalana" (DH). More precisely, it is at the first group out of three. It is the maximum category of the "Federació Catalana de Futbol". The next category corresponds to the "Real Federación Española de Futbol" and it is called "Tercera División". In addition, GEFS plays another competition, the cup of Girona.

The club had three main objectives for this season:

- Win the cup of Girona
- Promote to "Tercera Divisón"
- Win the league

It has to be taken into account that it is possible to promote to the next division without winning the league because the first two teams get promoted.

Moreover, at the first group of DH there are 14 teams. This means that the season has 26 matches because each team plays two times against each other (one time home and one time away). Logically, each team does not play against itself. The competition started on the 15th of October and it ended on the 28th of May.

Referring to the cup of Girona, due to the fact that GEFS is the best futsal team of the province of Girona, it accesses directly to the quarter-finals. In the best possible scenario it plays three matches: quarter-finals, semifinal and final. As the final-four was played on January, it can be celebrated that GEFS already achieved one of the main objectives: to win the cup of Girona. Winning it gives access to play the cup of Catalonia.

So, GEFS played 26 league matches and 4 cup matches because it was eliminated at the first match of the cup of Catalonia. This makes a total of 30 matches played against 15 different teams, 13 from the league and 2 from the cup.

3.2.2 Players available

As it has been said before while explaining the domain, only 12 players per team can play in a match. There are 5 players on the court and a maximum of 7 on the bench. However, a futsal team usually has around 15 players just in case there are injuries or sanctions.

In the case of GEFS, due to the fact that it is a club with formative futsal and with another senior team playing in a low category, during all the season 21 different players have been called at least once.

3.2.3 High pressure scenario

As a team that has this demanding objectives and as a club that wants to grow as fast as possible, it cannot be allowed two weeks of bad results. The exigence is so high, that the coach who started the season is no the same coach that has ended it. Just before Christmas, GEFS suffer a painful lose (7 - 2) at the court of Mataró that caused the president to opt for a change at the bench.

Moreover, this high pressure can be one of the reasons to explain the high amount of typos found in papers *M* and *E* explained at the introduction.

3.3 Background: tools used

To achieve the aims mentioned and to show the desired results there have been a combinations of tools. In this section, there is going to be a brief explanation of them. There are some that are well-known and some others that are less popular.

3.3.1 Image Processing for Data Extraction tool

To perform the image processing in order to extract the information from the full-filled template and to turn it into a CSV, it has been tried a couple of distinct open-source libraries from Python and R like:

- **Tesseract (R)**
  https://cran.r-project.org/web/packages/tesseract/vignettes/intro.html
- **Tabula (Python)**
  https://pypi.org/project/tabula-py/
- **Camelot (Python)**
  https://pypi.org/project/camelot-py/

However, it was decided to us a non open-source library from Python that is called:

ExtractTable https://pypi.org/project/ExtractTable/

Although being a non open-source library from Python it is possible to get free-credits and this is what it has been done. This library is capable of extracting tabular data either from a pdf file or an image file (png or jpg).

Every time it is used, it requires an API key and it returns the data in the desired format. It can be a DataFrame, a json, a dictionary, a csv or a xlsx file. Also, it is possible to make some corrections and to adapt the procedures for the input data.

For this thesis, to extract the information and to store it afterwards, it has been used a combination of the libraries: **ExtractTable** and **Pandas**. The first library gets the information from the image of the template (Figure 1.2) and it returns a dictionary. Then, the second library creates a DataFrame from this dictionary in order to apply some functions that are going to be explained at the next chapter 4. They are used to check that everything is correct and to calculate some extra fields such as the total minutes played per player.

3.3.2 Google Environment

The image processing for data extraction tool and the exportation of the CSVs created for each match were done inside the Google Environment. In fact, for this master's thesis there have been used the following 4 tools that Google provides.

3.3.2.1 Google Drive

A worldwide known tool that allows to store data organised in folders. It is possible to use it for free. For this thesis, a folder containing all the photos of the full-filled templates was created. In order to identify each photo the name always was "Id_match_ME.jpeg".

3.3.2.2 Google Colaboratory

It can be understood as a *Jupyter Notebook*. The library chosen before (**ExtractTable**) was executed in here. In addition, the notebook created for this thesis with this tool can be found in the GitHub repository mentioned at the end of the introduction.

Also, it was used Google Colaboratory to concatenate the 5 datasets that were generated per match. How this was done can be found in the next chapter 4.

3.3.2.3 Google Cloud Platform

Basically, there are two ways of storing the database: on the computer (local) or in the cloud. This tool, the Google Cloud Platform (GCP), provides a free version that allows to store the database in the cloud easily.

The principal reason why in this thesis is much better to store the data in GCP rather than locally is because you can easily work with **BigQuery**. Once the bucket in **Cloud Storage** is created and the different tables are linked to Drive, it is really easy to execute SQL queries inside it. In fact, due to the fact that the database created is relational and that one part of the posterior analysis is going to be focused on the players performances, there is need of executing SQL queries.

3.3.2.4 Looker

The last tool used that it is the Google Environment is called **Looker**. It allows to create and share interactive dashboards easily. Furthermore, it is possible to work with the data either stored in Drive or GCP.

3.3.3 Rstudio

Finally, for being able to perform the Principal Analysis Component and the Gaussian Mixtures on the analysis of the players performances, it has been used

Rstudio. More precisely, among others, it has been worked with the following R libraries:

- **stats**: to perform the PCA.
  https://stat.ethz.ch/R-manual/R-devel/library/stats/html/00Index.html
- **factoextra**: to show the PCA graphically.
  https://cran.r-project.org/web/packages/factoextra/index.html
- **mclust**: to deal with mixtures of Gaussian distributions.
  https://cran.r-project.org/web/packages/mclust/vignettes/mclust.html

3.4 Data: structure and explanation

Last section from this chapter of preliminary concepts and ideas is about the data. Specifically, it gets into detail about the structure and the data contained in the database.

As it has been said before, it has been stored all the information that GEFS has collected during 30 matches in a specific designed relational database structured with 8 different datasets that are divided conceptually into 2 groups.

The first group is called **static information group** because it stores data that it is known at the beginning of the season or that needs tiny modifications as matches are played. It contains these 3 datasets:

- Season
- Teams
- Squad

To explain what they content, there is a table for each one. The "Variable" column is the variable name chosen, the "Meaning" column explains what it is referring to and the "Key" column represents if it is a key column or not.

Season dataset

It contains all the 30 matches played by GEFS. It has a unique Id for each match and it gives information about which team is the home team, which team is the away team, the competition and who was the coach. Each row is a match.

Teams dataset

It has all the teams that GEFS has faced all along this season. It contains a unique Id for each team, the team name and a short version of it. As before, each row is a team.

Notation	Meaning	Key
Id_match	Unique identifier of each match	Yes
Id_home_team	Id of the home team	
Home_team_name	Name of the home team	
Id_away_team	Id of the away team	
Away_team_name	Name of the away team	
Competition	It can be "League" or "Cup"	
Coach	It can be "Castejon" or "Naranjo".	

Table 3.1: Season dataset explanation

Notation	Meaning	Key
Id_team	Unique identifier of each team	Yes
Team_name	Full name of the team	
Short_team_name	Short version of the team name	

Table 3.2: Teams dataset explanation

Squad dataset

Each row is a player. It has 22 rows because it contains the 21 different players that have been called at least once and an extra player representing own goals (i.e. those goals scored by GEFS that no one in the team did). In addition, for each player there is: the name, the surname, the nickname, his position positions and if they are right or left footed. The letters defined before (G, D, W and P) are used to determine the positions.

Notation	Meaning	Key
Id_player	Unique identifier of team member	Yes
Name	Name of the player	
Surname	Surname of the player	
Nickname	How the player is known	
Position	Futsal position	
RoL	Right (R) or left (L) footed	

Table 3.3: Squad dataset explanation

On the other hand, the second group of the database is called **growing information group** because each time a match is played more information is stored. The 5 datasets included in this second group are:

- Players
- Match

- Goals
- Minutesh1
- Minutesh2

Players dataset

It stores information about the players that the coach have called-up for each match. As it has been written there is a maximum of 12 players per match. For each player, GEFS collects data about where the player starts the match (court or bench), the goals scored, the number of free-kicks committed, the number of yellow and red cards, the number of tackles and the number of losses.

Notation	Meaning	Key
Id_player	Unique identifier of each team member	Yes
Id_match	Unique identifier of each match	Yes
Starting	Where the player started: bench (0) or court (1)	
Goals	Number of goals	
Free_kick	Number of free-kicks	
Yellow	Number of yellow cards received	
Red	Number of red cards received	
Tackle	Number of ball recuperations	
Lost	Number of ball losses	

Table 3.4: Players dataset explanation

Match dataset

This dataset contains information per match. It has four rows for each match, one row per team and per half. So, for each team and each half the information saved is the total number of: shots, chances, goals, corners, throw-ins, 10m, penalties, far posts and free-kicks.

Notation	Meaning	Key
Id_match	Unique identifier of each match	Yes
Id_team	Unique identifier of each team	Yes
HoA	Which team scored: home (H) or away (A)	
Half	Half when it happened	
Shots	The total number of shots	
Chances	The total number of chances	
Goals	The total number of goals scored	
Corners	The total number of corners	
Throw-in	The total number of throw-ins	
10m	The total number of 10m	
Penalty	The total number of penalties	
Far_post	The total number of far-posts	
Free-kick	The total number of free-kicks	

Table 3.5: Match dataset explanation

Goals dataset

It has the evolution of the score for all the different matches. For each match, it has as many rows as goals were scored (i.e. if the final score was 4 - 2, it has 6 rows for that match). In addition, it has the team who scored the goal, the player who did it, the half, the minute and the 5 players that were playing in that moment.

Notation	Meaning	Key
Id_match	Unique identifier of each match	Yes
Home_goals	Goal scored by the home team	
Away_goals	Goal scored by the away team	
Id_player	Unique identifier of each tean member	Yes
Half	Half when the goal was scored	
Play	How was to goal scored	
Player1	Player that was playing	
Player2	Player that was playing	
Player3	Player that was playing	
Player4	Player that was playing	
Player5	Player that was playing	
Minute	When the goal was scored	

Table 3.6: Goals dataset explanation

Minutesh1 dataset

One more thing that GEFS collects is the minutes that each player has played per match. This dataset only contains the minutes for the first half of each match. It has the minute a player gets in the match and the minute the players it is substituted. Moreover, there is as well the minutes of each rotation and the total minutes played.

Notation	Meaning	Key
Id_match	Unique identifier of each match	Yes
Id_player	Unique identifier of each team member	Yes
Half	When it happened	
In11	First time entering the match	
Out11	First time out of the match	
In12	Second time entering the match	
Out12	Second time out of the match	
In13	Third time entering the match	
Out13	Third time out of the match	
In14	Fourth time entering the match	
Out14	Fourth time out of the match	
In15	Fifth time entering the match	
Out15	Fifth time out of the match	
Round11	Minutes played on the first round	
Round12	Minutes played on the second round	
Round13	Minutes played on the third round	
Round14	Minutes played on the fourth round	
Round15	Minutes played on the fifth round	
Total1	Total minutes played	

Table 3.7: Minutesh1 dataset explanation

Minutesh2 dataset

It contains exactly the same information as the last dataset. Now, it corresponds to the second half of each match.

Notation	Meaning	Key
Id_match	Unique identifier of each match	Yes
Id_player	Unique identifier of each team member	Yes
Half	When it happened	
In21	First time entering the match	
Out21	First time out of the match	
In22	Second time entering the match	
Out22	Second time out of the match	
In23	Third time entering the match	
Out23	Third time out of the match	
In24	Fourth time entering the match	
Out24	Fourth time out of the match	
In25	Fifth time entering the match	
Out25	Fifth time out of the match	
Round21	Minutes played on the first round	
Round22	Minutes played on the second round	
Round23	Minutes played on the third round	
Round24	Minutes played on the fourth round	
Round25	Minutes played on the fifth round	
Total2	Total minutes played	

Table 3.8: Minutesh2 dataset explanation

CHAPTER 4

Planning and Methodology

The next step in this thesis, is to explain the work plan designed and the methodology followed in order to achieve the aims and to show the results successfully.

4.1 Planning

This thesis has to be understood as full data science project. As the aims reflects, it encompasses from the acquisition of data until the analysis. This made the elaboration encounter a wide variety of issues that has slowed down the production. However, a strong point of creating a data science project from the very beginning is that it allows to accumulate knowledge and to know all the data details.

4.1.1 Data provider

It has to be said that GEFS was not the first option. Once it was decided that this thesis had to be a data science project based on sports, it existed a contact with two professional clubs:

- Bàsquet Girona (BG)
- Futbol Club Barcelona (FCB) (futsal department)

The goal of these contacts was to be able to establish a symbiosis. They would give access to their data in exchange of analysing and getting to useful conclusions. However, it has to be taken into account that this type of data, although not being as sensible as health data, has a competitive component that makes it really difficult to share outside the club.

As an anecdote, in some scenarios not even the first coach has access to all the data about the performance of his players. When he receives the data, he gets it already filtered.

This competitive component does not allow the professional clubs to share their data because they would be sharing data about how players behave at training sessions and what is their physical condition. High professional athletes

train with a GPS sensor that measures: strength, acceleration, speed, etc. This information is inaccessible for the other clubs. In fact, the only information that they can reach, ii the one that occurs during the match.

Therefore, it is understandable that professional clubs do not want to share their data as they do not want to give any advantage to their rivals. Moreover, all this professional clubs have their own data team that is already able to achieve interesting conclusions and results.

Consequently, once professional clubs were discarded, the main focus was on getting data of a non-professional club. Luckily, Girona Escola de Futbol Sala (GEFS) gave permission on treating their data just after Christmas. It must be said that the author of this master's thesis is a player from the club and this ease the collection of data.

4.1.2 Data extraction, storage and analysis

Once it was able to collect the data from GEFS, the next step was to ask its coach which was the most interesting data. As it is written in the introduction, he wanted to visualise:

1. **Minutes played:** a bar chart with the amount of minutes played per player.
2. **Goals scored:** a bar chart with the total number of goals scored per player.
3. **Presence:** a bar chart to compare the number of team goals received with the number of team goals scored per player.

Out of curiosity, GEFS trains two times a week (Tuesdays and Thursday). Each Monday afternoon the coach makes a summary of the season so far to prepare the ahead training week. This means that each Tuesday, when the players arrive at the dressing room, they found stuck to the wall these three charts printed.

On the other side, another result for this thesis was to analyse the players performances using Principal Analysis Components and Gaussian Mixtures.

So, it started an iterative process on how to extract, storage and analyse the data. The first time GEFS provided data there were 11 league matches and 3 cup matches (i.e. 14 matches). However, due to the fact that all paper sheets either paper *M* or paper *E* have the same structure the test on how to extract the data was only performed for one match.

The match chosen was one played with the new coach because he made some little modifications on the paper **E**. Then, it was tried to extract the data directly from the photos that GEFS sent. It was impossible due to:

- The typos
- The rotation of the photo
- The low quality of the photo

Consequently, it was decided to create the template. The Figure 1.2 shown before was the fifth version of this template. It was selected this version over the others and it has not evolved for two main reasons:

1. It is able to compact all the information that GEFS collects during a match in only on sheet of paper.
2. It fits perfectly to the library chosen to extract information from the photos (**ExtractTable**).

Once it was defined the way on how to extract the data, it was time to determine how to store it. The idea of dividing conceptually the data in two groups (static and growing) was very useful. The growing group would correspond to the data obtained through the template whereas the data in the static group would be data that came directly from the author's data knowledge.

The last step of this iterative process was to check if the data stored was able to generate the demanded results. Indeed, it was very important to first determine what were the needs in order to keep them in mind and not to start the work again once it was reached this final step.

4.1.3 Cost

The cost of making this master's thesis is divided in time cost and money cost.

4.1.3.1 Time

The elaboration of this thesis can be divided in the following four epochs:

1. **First idea and availability:** this thesis started to be though before Christmas. However, it was no until after Christmas that there was some data to start thinking on how to do it. Moreover, it was the 28th of February when the proposal was accepted by the committee.
2. **Testing and state of the art**: once the data access and the acceptance were granted, it started a period of searching literature about the topic and testing what was the best way to extract, store and analyse the data. This second epoch last until after Easter. It has to keep in mind that each weekend there were two more photos to take into account.

3. **Coding, analysing and writing**: this third epoch happened between Easter and the 2nd of June. The fundamentals of the thesis were already settle and it was time to extract all the information, store and analyse it. Is the epoch where most time was invested.

4. **Final presentation**: once the all the work was done, it was only left the public defence in front of the committee the 22nd of June.

4.1.4 Money

As it has been written all along this thesis, although using an non open-source library or the GCP services, there is no money cost. It is possible to use both free versions and it works perfectly.

4.2 Methodology

In this section it is going to be explained all the followed steps from the very beginning until the end of all the way done. It is only explained the final version of this data pipeline created.

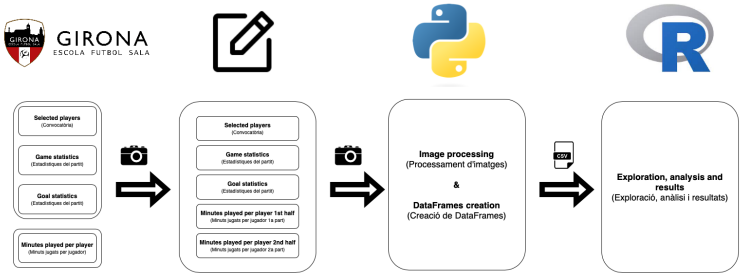

Figure 4.1: Workflow

4.2.1 From GEFS data to template data

As it is known, each weekend GEFS plays a futsal match and it collects data in two different paper sheets called *M* and *E*. Then the person of GEFS in charge of full-filling them, takes one photo per paper. These two photos are sent via Whatsapp each Monday or Tuesday and they are saved in a folder called "MEId_match", where *Id_match* is the corresponding unique identifier for each match.

Consequently, there are 30 folders, one per match, containing these two photos. The photo corresponding to paper *M* is saved as "MId_match" and the one corresponding to paper *E* is saved as "EId_match".

Afterwards, with the help of these two photos the template v5 is printed and full-filled with blue pen. As it has been said, it can be that for some matches there are not photos. However, this is not a problem because the information to full-fill the template v5 can be found combining the report that "Federació Catalan de Futbol" publishes and analysing the video that GEFS records.

When the template v5 is ready, a photo is taken and stored in a folder that contains all the others full-filled templates. The name of this folder is "Photos_ME" and it is located in Google Drive. The name for this third photo is "Id_match_ME". It is very important to keep an order and a firm structure to avoid chaos and mixing images.

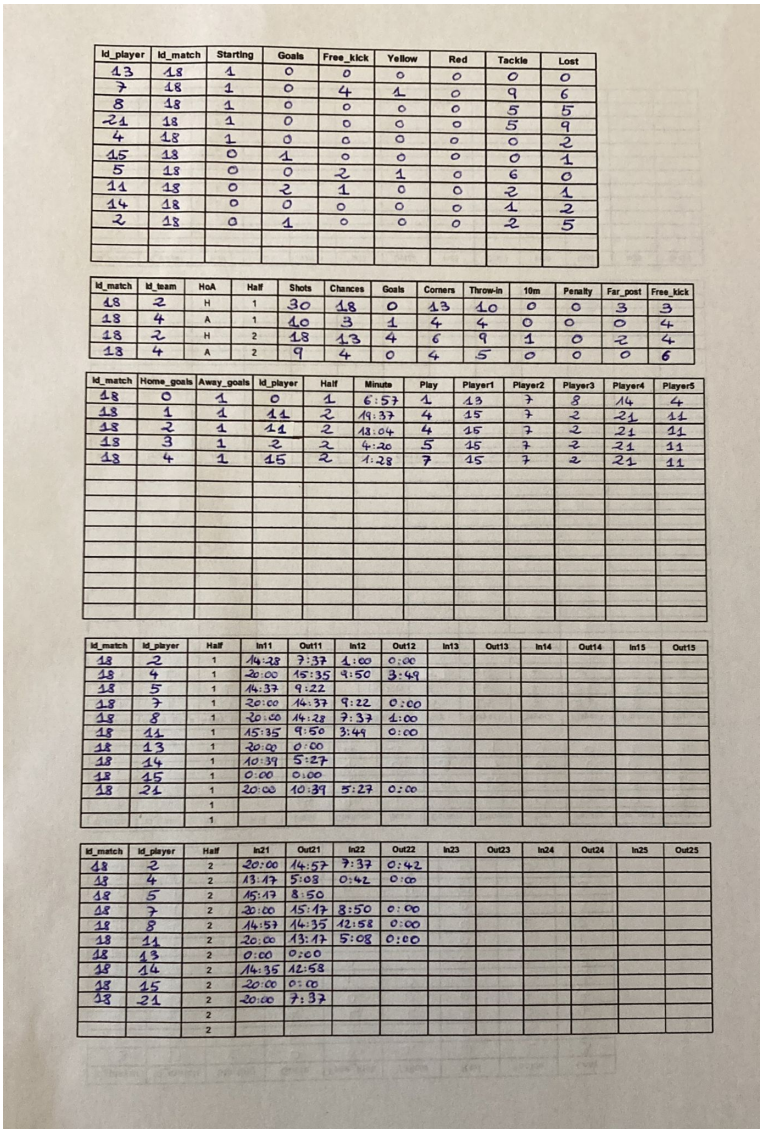

Figure 4.2: Example of a full-filled template v5

4.2.2 From template to CSV

Once the template v5 is saved in the corresponding folder at Google Drive, next step is to use Google Colaboratory. First, the two libraries needed, ExtractTable and Pandas are loaded. Then, four functions are defined. The idea is to execute them in the following order:

1. **image_show:** this first function requires an image path and it just shows with which image it is working. The image path is where from Google

Drive we can find the photo. This function is useful to visually check that it is working with the appropriate image.

2. **df_names:** once it has been checked that the image is the correct one, it is time to generate the name of the 5 DataFrames that are going to be created for each match. This function only requires the image path.
   With this information, it is able to extract the Id_match because the names of the templates v5 photos start with it. Therefore, the corresponding names are:

   (a) players_Id_match (e.g. players_5)

   (b) match_Id_match (e.g. match_5)

   (c) goals_Id_match (e.g. goals_5)

   (d) minutesh1_Id_match (e.g. minutesh1_5)

   (e) minutesh2_Id_match (e.g. minutesh2_5)

   It is already explained what each dataset contains. All of them are included in the **growing information group**.

3. **dict_creation:** it needs once again the image path and the API key in order to extract the information from the image using the library **ExtractTable** and turning it into a dictionary. This function returns a dictionary that has to be saved.

4. **df_creation**: this fourth function requires the image path and the dictionary data created on the previous function. It is going to return the 5 DataFrames created before with the corresponding function.

If all these four functions are executed in the showed order, it goes from one unique image containing all the information, to the 5 desired datasets.

4.2.3 CSVs corrections

However, sometimes the automatic image processing can be confused and make mistakes. This is why it has been created one function per dataset that corrects these possible mistakes and checks that everything makes sense. For example, for the "minutesh2.csv" that contains the minutes that each player has played in the second half, it does not allow the "In21" column to be lower than the "Out21" column.

These 5 functions can be tuned so anyone can adapt them to their hand-written. They only need one argument that it is a DataFrame and they return the corrected DataFrame. They automatically save a CSV version of it. Their names are:

1. **correction_players_dataset**
2. **correction_match_dataset**
3. **correction_goals_dataset**
4. **correction_minutesh1_dataset**
5. **correction_minutesh2_dataset**

These functions save each CSV to the corresponding folder in Drive. Remember that this is thought to work with Google Colaboratory. Now, saving each CSV to their corresponding folder in Drive means saving the goals files in a folder called goals, the match files in a carpet called match, etc. All these carpets have to be in the same path.

If after applying these functions there are some values that are still a mistake, it is recommendable to combine them with the functions *at()* and *drop()* from the library **Pandas**.

4.2.4 Concatenation of CSVs

In order to encompass all the information in the corresponding and in a unique dataset, the 5 datasets generated per match have to be concatenated.

If the steps 4.2.1 , 4.2.2 and 4.2.3 are applied to the 30 templates v5, it is generated 150 different CSVs (50 for each dataset).

Then, due to the fact that on the previous step all the corrected CSVs have been saved in one of the five corresponding folders, now it can be executed one last function called **df_concat**. This function requires a folder path and it returns the concatenation of all the CSVs files from there ignoring their index.

Afterwards, there is just need of saving this 5 CSVs into a new folder called "Global".

4.2.5 Static information group CSVs

So far, it has only been explained how to process and store the information corresponding to the **growing information group**. The other group from the

database, the **static information group**, has to be uploaded as well into the Global folder created in Drive.

The three datasets corresponding to this group, have been built by hand. They have been created with Microsoft Excel and stored as a CSV. They do not have a lot of information and thanks to the author's knowledge of the data, it did not take a lot of time.

4.2.6 Storing at GCP

Once all the 8 datasets are stored in the Drive's folder called "Global", it is time to create the bucket at the **Cloud Storage** from **GCP**. It is as easy as to create a project and inside the project to create this bucket. This bucket can be understood as the database.

Next, there is need to create the 8 different tables. Each one corresponds to a dataset. To create them, GCP allows multiple ways of importing data. The way chosen is to get the data directly from Drive because is where the 8 datasets lay.

There are two concepts really important at this point:

- When a table is created inside a bucket in GCP, the columns have to be defined. Afterwards, it is going to be impossible to modify the name or the type. If any modification has to be made on the columns, the data has to be imported again.

- The "Global" folder in Drive that contains the 8 datatsets in a CSV format, it has to contain them in a Google Spreadsheet format. This is fundamental for the next step.

4.2.7 Growing CSV

Due to the fact that each of the 8 different tables created in GCP corresponds to an specific dataset saved as a Google Spreadsheet format in the folder called "Global" from Drive, they are able to grow.

This happens because although GCP does not allow to modify the columns, it allows a file to add more rows without any problem. Because the shared file is always the same, it is possible to add more rows (i.e. more matches) to the corresponding Google Spreadsheet without issues.

When the template v5 for the first match is processed and store correctly, it is possible to proceed with the second one and to keep going until the match number 30. Basically, every time a match is added, the corresponding rows to each Google Spreadsheet should be appended and the file should not be changed. If

the new rows overwrite the existing file, then there will be need of creating a new connection between Drive and GCP.

4.2.8 Producing the results

Next step once the data have been stored correctly in GCP is to create those most useful tables to produce the results. To do it, the recommendation is to relate all the 8 different datasets making use of the column keys and the tool BigQuery from GCP.

The two expected visual results are:

- **Looker dashboard:** it has to show the three requested charts by the coach of GEFS. There is no need to wait until the end when all the data is processed and stored to create the dashboard, because Looker can take the data directly from Drive or GCP. Therefore, if the database is updated, the dashboard will be updated as well.

- **R Analysis**: in here it is included the analysis of the performances of the players using Principal Component Analysis and Gaussian Mixtures. Contrarily as before, to do this analysis there is need to work with a static file. The way this has been done, is to produce the desired table with BigQuery to afterwards export it locally as a CSV file. Then, read it inside Rstudio and work from there.
  One tip while working in this analysis is to save the query so next time there is only going to be need of executing the query and downloading the file.

4.3 Backup

Although that so far it has been said to save the data only in GCP and Drive, it is very recommendable to save a backup locally.

There is not that much information, so it does not need a lot of room. Having the data in both ensures that it will always be available.

CHAPTER 5

Methodological Contribution

Once it is known how the data has been collected, transformed and stored, it is time to explain which methods have been used in order to perform the correspondent analysis. In other words, in this chapter is going to be shown how the results have been obtained and which issues have been faced.

Furthermore, this chapter and has to be understood as a first approach on how this data can be analysed. Remember there are only 21 players and this is a really important limitation.

This chapter is divided in two sections. Firstly, it is described how the dashboard has been created. Secondly, it is going to be explicated the techniques used to analyse the players performances.

5.1 Requested charts with Looker

To illustrate the requested charts, as it has been said before, it was used Looker. This tool allows the creation of an interactive dashboard that can be shared with the squad of GEFS.

The aim of this dashboard is to show the three graphics that each week the coach of GEFS presents to their players:

1. **Minutes played**
2. **Goals scored**
3. **Presence**

5.1.1 Data selection

At the previous chapter 4, it has been explained that the data showed in Looker is directly connected with GCP. Consequently, it is used a SQL query to get the desired data.

In this case, there is need of grouping the data by player to calculate the values requested:

- The total minutes played per player can be obtained by adding the features *Total1* and *Total2* from the datasets *Minutesh1* and *Minutesh2*, respectively.
- To get the goals scored per player it has to be added the feature *goals* of the dataset *Players*.
- To know the presence of the players it has to be checked which players were playing when a goal was scored and which ones when a goal was received. This information can be found in the dataset *Goals*. It is going to be denoted the goals received by *TG_recevied* and the goals scored by *TG_scored*.

5.1.2 How to show the data

The way the data is shown in this dashboard is the same as the coach does it. It has been opt to do it like he does it because the squad is already familiarised with it:

1. **Minutes played:** a stacked bar chart with two colours. One colour represents the minutes played on the first half and the other colour the minutes played on the second half. Each column is a player and it is ordered by total minutes played.

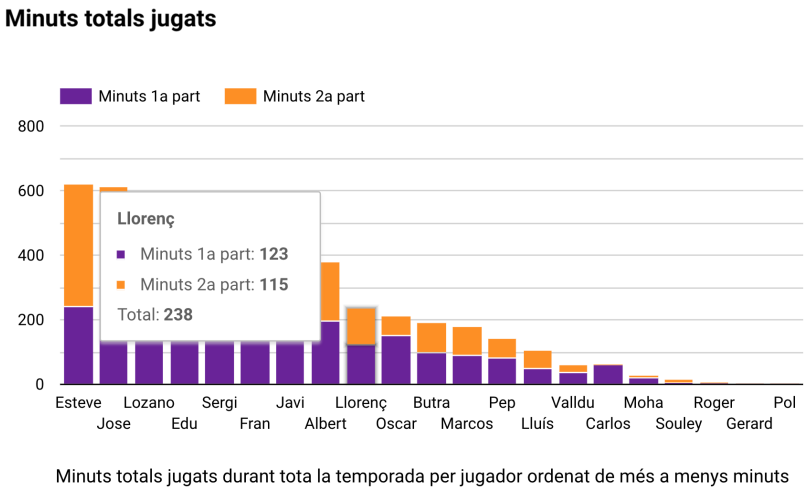

Figure 5.1: Minutes played per player at Looker

2. **Goals scored:** a bar chart where each column represents a player and the total number of goals scored.

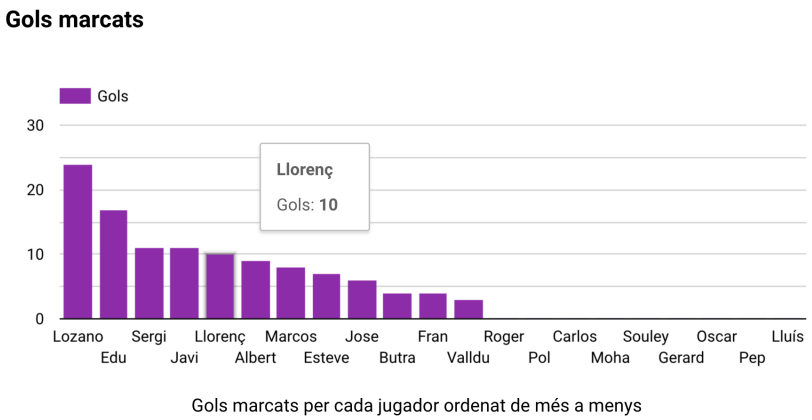

Figure 5.2: Goals scored per player at Looker

3. **Presence:** a bar chart where each player has two columns. One column are the team goals received and the other column are the team goals scored. As before, each column has a different colour and it is ordered from the player who was the highest difference to the lowest. In this case, this difference is calculated by subtracting the team goals received to the team goals scored.

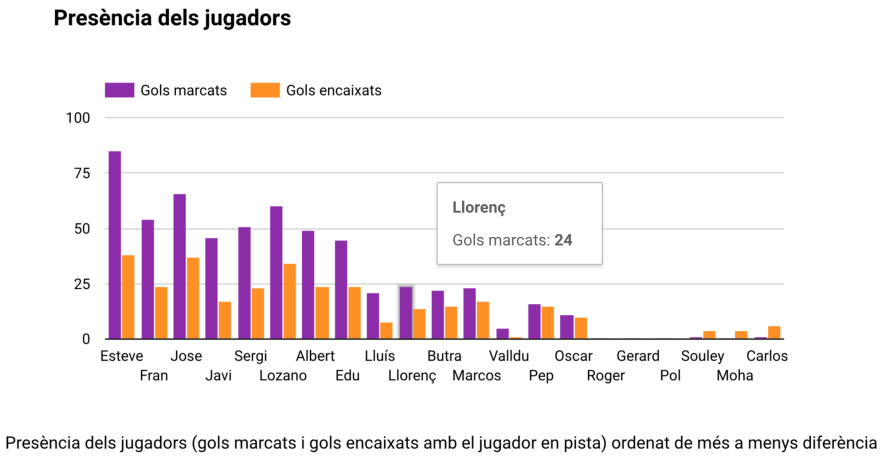

Figure 5.3: Presence per player at Looker

5.2 Analysis of players performances

As it has been written at the introduction, the second visual result of this thesis is the analysis of the performance of all the players that have been available at

least once in this season for GEFS. To do so, the tool used was R working in the RStudio environment and the techniques chosen were:

1. **Principal Analysis Component:** to determine which features characterise the performance of the players.

2. **Gaussian Mixtures**: to estimate the density function and to cluster the players by their performance.

This second section is divided in three parts. Firstly, it is explained how the data is manipulated in order to obtain coherent results. Then, Principal Analysis Component (PCA) is shown and, lastly it is explicated the Gaussian Mixtures point.

5.2.1 Data obtain and manipulation

Due to the fact that the aim is to analyse the performance of each player, the first step is to collect the most relevant data available. From the database created for this thesis, the most interesting fields are the followings.

The column *Feature* is the relevant feature selected and the column *Dataset* represents where this feature can be found.

Feature	Dataset
Nickname	Squad
RoL	Squad
Position	Squad
Goals	Players
Starting	Players
Available	Players
Free_kick	Players
Yellow	Players
Red	Players
Tackle	Players
Lost	Players
TG_received	Goals
TG_scored	Goals
Total1	Minutesh1
Total2	Minutesh2

Table 5.1: Data Manipulation to analyse players performances

To obtain them, as it has been done before while preparing the data for the dashboard, it has been used BigQuery to perform joins on the different datasets. Then, it has been downloaded a CSV file containing the data.

While analysing the performance of the players, it was detected an intuitive and logical high correlation between the minutes a player plays and its performance. The more minutes a player is on the court, the more he is likely to score goals, get a yellow card, etc. Consequently, after detecting this correlation, it was decided to work only with those variables that are numeric and related with the total minutes played.

So, the CSV downloaded was transformed with R and the final variables taken into account were:

- Goals
- Free_kick
- Yellow
- Red
- Tackle
- Lost
- TG_received
- TG_scored

All these variables were divided by the total minutes played per player. Consequently, they were transformed into values from 0 to 1. The sum of total minutes was obtained by adding *Total1* and *Total2*.

Furthermore, there was the case of a couple of players that were available for at least one match but they did not play any minute. It was decided to not take them into account because they distorted the results. They could be considered as outliers. This decision implied to reduce even more the number of observations.

Finally, the variable "Nickname" was defined as the names of the rows of the DataFrame containing the data in R.

5.2.2 Principal Analysis Components

The objective of doing this Principal Analysis Components (PCA) on the data was to determine which were the variables that most discriminate the performances of the players.

To calculate the PCA with R it was used the *prcomp* function. Moreover, it was decided not to scale the data, it was already between 0 and 1, but to centre

it. With the first two dimensions it can be explained approximately the 83.60% of the total variance. There is not a huge loss of information.

The next Figure 5.4 represents the different features taken into account plotted in the first two dimensions. The X axis is the first dimension and the Y axis is the second dimension. The percentage of the total variance explained by each dimension is 61% and 22.6%, respectively.

Additionally, the Figure 5.4 shows that the features *Lost* and *Tackle* (i.e. number of losses and number of recuperations) are the most important ones on the first dimension. On the other hand, for the second dimension the features *TG_received* and *TG_scored* are the most relevant. The other 4 features do not play an important role to characterise the performances of the players.

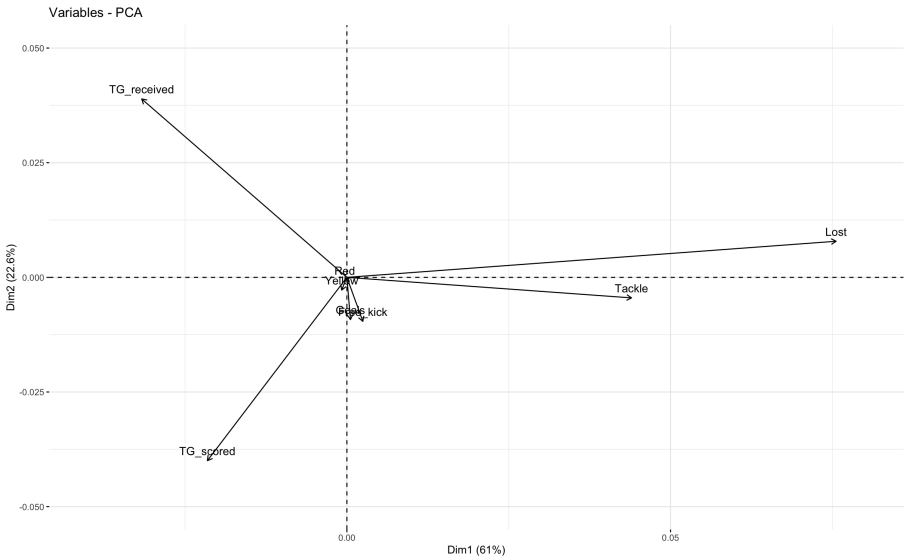

Figure 5.4: Features represented in PCA

Furthermore, if instead of plotting the features, the players are plotted, it is obtained the following Figure 5.5. It has been opted to paint the players by their positions. The goalkeepers (G) are painted with green, the defenders (D) with red, the wingers (W) with purple and, finally, the pivot (P) with blue.

It cannot be determined any relationship between the futsal position and where they are located in the plot. However, it must be said that the players that are far away from the centre like Souley, Moha, Carlos or Roger are players that did not play a lot of minutes wit the team.

For example, Roger has only played 6 minutes during all the season and he lost the ball 2 times. Consequently, the number of losses per minute is really high compared to all the other team members. The same argument could be applied

with the player Souley but this time regarding the total of goals received when he was playing.

If the focus is putted on players that had an important role during the season it can be determined that, for example, Javi's performance can be characterised by the feature *Tackle*, the number of recuperations. For all the other players it is hard to detect what is the most relevant feature about their performance.

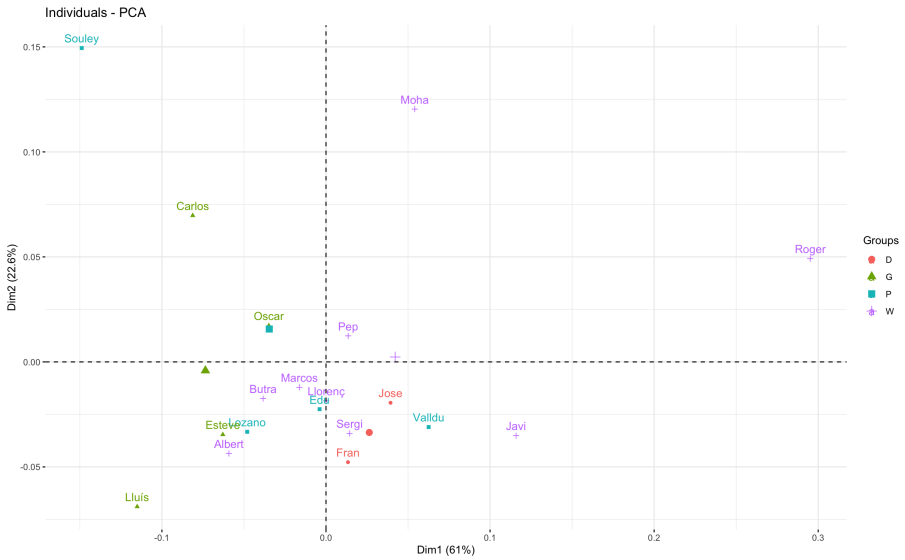

Figure 5.5: Individuals represented in PCA

The representation of the *biplot* was considered but decided not to show because it was confusing and it did not give any new point of view on how to interpret the data.

5.2.3 Gaussian Mixtures

After doing the dashboard and determining what are the features that most characterise the performance of the players, it is turn for the Gaussian Mixtures.

For this point of the thesis, the data used is the same as before while doing the PCA. From the 21 players that have been available for GEFS at least once this season, it has only been considered those who have played minimum 1 minute. Consequently, two players were discarded and the total of players was 19.

Furthermore, the aim of this final step of the analysis is to estimate the density function and, afterwards, to cluster the players according to their performances.

Before applying the mixture modelling, it was calculated the Maximum Likelihood Estimation (MLE) for the parameters of Gaussian distributions. It was

used the sample mean and covariance.

As it could be expected the Gaussian distribution could not capture any of the features. The Figure 5.6 shows how any of them can be adjusted wit the distribution that the blue line is representing. It must be said that some features might admit some discussion about if the Gaussian distribution is able to capture it or not.

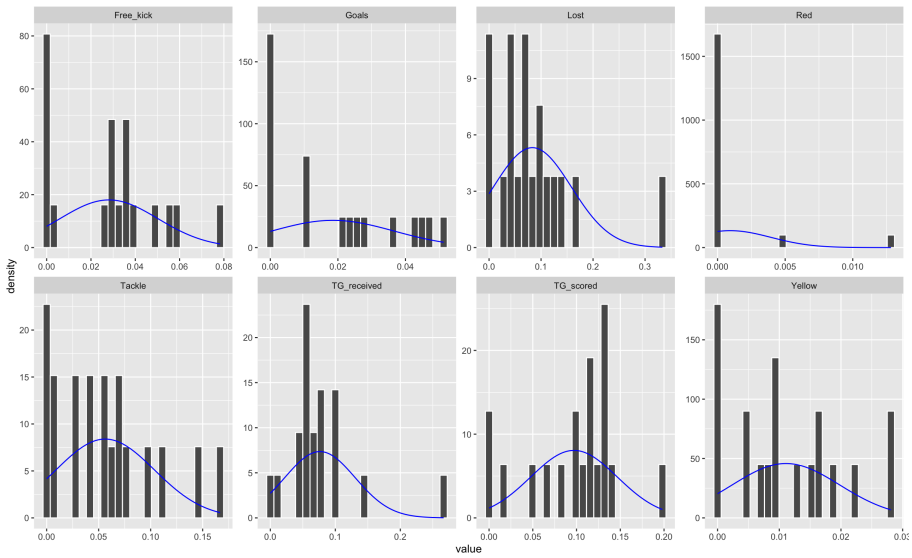

Figure 5.6: Multivariate Gaussian distribution

Then, in order to find the MLE of a finite mixture parameters it was applied the *Mclust* function. The best model chosen by this function using the BIC criterion was an **EEE** (ellipsoidal, equal volume, shape, and orientation) with 7 components. This can be checked at Figure 5.7.

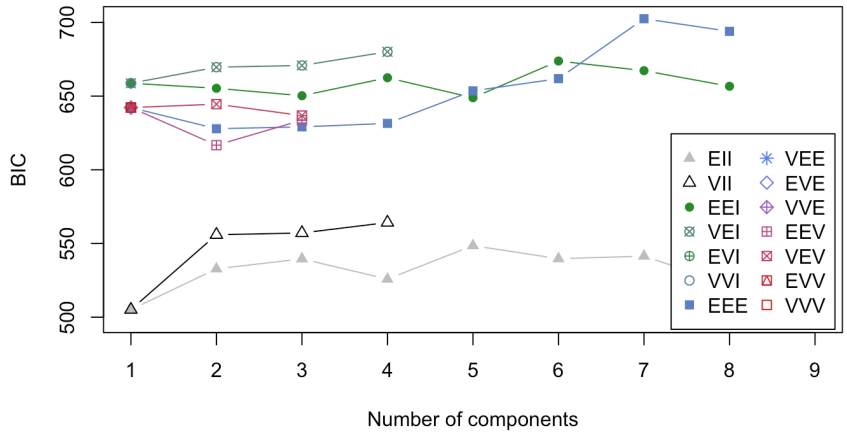

Figure 5.7: BIC obtained with different models

Finally, it was performed the clustering using the Gaussian Mixture Model adjusted. The next Figure 5.8 reduces the dimensionality while trying to preserve the Gaussian Mixture structure.

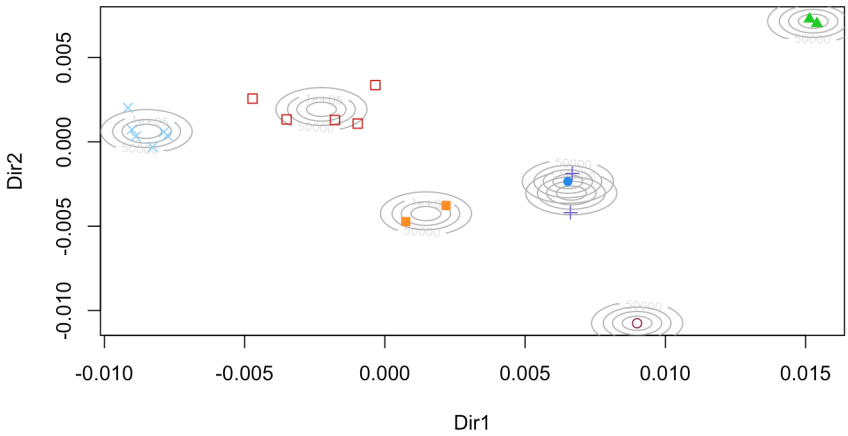

Figure 5.8: Clustering using Gaussian Mixture Model

It has been suggested to cluster the players into 7 different groups and this could not be the best idea. There exists other tools such as *combiTree* that would provide a hierarchical structure and this would allow to choose the number of clusters. However, due to the fact that this it is just a first approach on how the data could be analysed, this is left for a future work.

The following table shows to which cluster corresponds each player:

Player	Cluster
Roger	1
Valldu	2
Llorenç	2
Butra	2
Edu	2
Albert	2
Moha	3
Souley	3
Carlos	4
Oscar	4
Lluís	5
Esteve	5
Marcos	6
Javi	6
Fran	6
Jose	6
Sergi	6
Pep	6
Lozano	7

Table 5.2: Cluster assigned per player

As before, the player called Roger is a rare case. He has his own cluster for the reasons explained before at the PCA analysis. The low amount of minutes he has played and the number of losses he did provokes this player to be far away from the others.

In the second cluster there are 5 players. In fact, these are the most defensive players of the team . It makes sense to cluster them together.

Next, in the clusters 3 and 4 there are two players and two goalkeepers, respectively. These players have not played all the season with GEFS. They could be unified.

At cluster number 5 there are the two goalkeepers that have been at the squad during all the season. Although that they did not have the same role in the team, it is logic to put them together as they are goalkeepers.

Then, it comes the most populated cluster. In cluster number 6 there are 6 players, the most offensive of the team. In fact, cluster number 7 only contains one player. He was the top scorer of the team.

CHAPTER 6

Results

In the penultimate chapter of this master's thesis there are the results explained before summarised and referenced. There are two groups of results: non-visual and visual results. At least, there is one result per aim.

6.1 Non-visual results

In this first group, it can be found those results related with the collection of the data and with the structure of the database:

- **Data collection improvement:** the first result of this thesis are the improvements suggested while collecting the data. This result is related to the first aim and it is exemplified with Figure 1.2 and Figure 1.3.
  On the one hand, Figure 1.2 shows how to combine the data that comes from the papers *M* and *E* in a logical and coherent way.
  On the other hand, Figure 1.3 gives an idea on how to improve the data collection and to facilitate the future extraction. Adding a column containing the value that the sum of the sticks represents is going to ease this process. Additionally, it could be very interesting for the posterior analysis to know when the event occur. This information could be capture by using a temporal axis as it is shown.

- **Database:** the second non-visual result is related with the second aim and it is about the database and its structure. In fact, it can be considered as the pillar for this thesis.
  A lot of time was invested to think about how to create a database that could store all the information collected by GEFS. The Figure 1.4 represents the final version of it. Also, in this result it is included the automatic image process for data extraction defined. It allows to extract the data from a photo and to store it in the correspondent table of the database.

6.2 Visual

In this second group, there are the results corresponding to the third aim: dashboard with Looker, Principal Analysis Component and Gaussian Mixtures. As it has been explained all along this thesis it has to be aware that the amount of observations (i.e. players) is really low. It has been worked with only 19 players.

- **Dashboard with Looker:** this result is the one showed with Figure 5.1, Figure 5.2 and Figure 5.3. There are represented the three charts the coach of GEFS requested. It can be found in the following link:

  https://lookerstudio.google.com/reporting/c70e0f3f-6235-4e15-8114-57491b9030e9

  Moreover, it allows interaction and to filter the data by the match.

- **Principal Analysis Components:** after performing PCA on the selected data to analyse the performances of the players, it is concluded that the features that characterise the performances the most are:

  - *Lost*
  - *Tackle*
  - *TG_received*
  - *TG_scored*

  This is represented in Figure 5.4. Furthermore, the position of the players does not help to know where they are located in Figure 5.5.

- **Gaussian Mixtures:** first it was checked that the Gaussian distribution is not able to capture the features (Figure 5.6). Then, the best model selected using the BIC criterion was an EEE with seven components (Figure 5.7). Afterwards, the players where clustered (Figure 5.7) according to it and it could be concluded that it makes sense. Although the suggestion being a high number of clusters, it could be found a logical explanation for each group. Consequently, it seems that in order to find similarities between the performances of the players this technique works well.

CHAPTER 7

Conclusions and future work

This master's thesis has to be understood as a full data science project. It has gone from the very beginning (i.e. obtaining the data) to the last step (i.e. applying Unsupervised Machine Learning). In fact, this is the main value of the thesis. Being able to go through all the process allows to fully understand the project developed.

Additionally, two more components that give extra value to it are the source of data and that it has been designed a relational database from scratch. Usually, the data used to work came directly from an online repository or it was already processed. However, in this thesis the data was obtained by extracting information from a photo. In addition, as the data was being obtained it was uploaded to GCP. From there it was really easy to relate all the datasets.

Furthermore, this thesis has been able to achieve all the three aims proposed at the introduction. To do it, the results have been separated into non-visual and visual in order to match them with an aim. In fact, having this different types of results allows this thesis to be understood by different publics. It goes from basic analysis like the dashboard, that can be shared with the squad of GEFS, to the Unsupervised Machine Learning techniques that allows the scientific futsal community to go further with the analysis.

About this analysis, it has to be taken into account what it has been written all along the thesis. The conclusions reached are based on a very few amount of observations. It has only been able to consider 19 players. However, this does not have to be considered as a weakness, it has to be considered as a first approach on how to treat this type of data. The fundamentals of this thesis were to collect and to store the data.

Moreover, this type of data does not present any obvious ethical problem but it does have a competitive component which makes it difficult to share. No teams want to give advantage to his rivals.

To conclude, this master's thesis has worked with concepts that were not taught at master's lessons. In addition, it has fully accomplish his aims and it has proposed a first approach on how to analyse this data.

If someone wants to carry on with this thesis it is suggested to investigate in two directions. Firstly, to cluster the players with a hierarchical component. This

will allow to determine the number of clusters and to be more precise on their definitions. Secondly, to treat the data as compositional data. Due to the fact that for the analysis of the performances of the players, it has been decided to divide the different features by the total minutes played per player, it has been working with data carrying relative, rather than absolute, information.

Bibliography

[Abdel-Hakim 2014] Hosam Hussein Abdel-Hakim. *QUANTITATIVE ANALYSIS OF PERFORMANCE INDICATORS OF GOALS SCORED IN THE FUTSAL WORLD CUP THAILAND 2012*. Pamukkale Journal of Sport Sciences, vol. 5, pages 113–127, 2014. (Cited on pages 10 and 11.)

[Agras 2016] Haydée Agras, Carmen Ferragut and Arturo Abraldes. *Match analysis in futsal: a systematic review*. International Journal of Performance Analysis in Sport, vol. 16, 2016. (Cited on pages 8 and 9.)

[Almeida 2019] João Almeida, Hugo Sarmento, Seamus Kelly and Bruno Travassos. *Coach decision-making in Futsal: from preparation to competition*. International Journal of Performance Analysis in Sport, 2019. (Cited on pages 9 and 11.)

[Castagna 2009] Carlo Castagna, Stefano D'Ottavio, Juan Granda Vera and José Carlos Barbero Álvarez. *Match demands of professional Futsal: A case study*. Journal of Science and Medicine in Sport, vol. 12, pages 490–494, 2009. (Cited on page 11.)

[Castellano 2012] Julen Castellano, David Casamichana and Carlos Lago. *The Use of Match Statistics that Discriminate Between Successful and Unsuccessful Soccer Teams*. Journal of Human Kinetics, vol. 31, pages 139–147, 2012. (Cited on page 10.)

[FIFA 2016] FIFA. *FIFA Futsal World Cup Colombia 2016: TECHNICAL REPORT AND STATISTICS*, 2016. (Cited on page 11.)

[Kasap 2015] Suat Kasap and Nihat Kasap. *DEVELOPMENT OF A DATABASE AND DECISION SUPPORT SYSTEM FOR PERFORMANCE EVALUATION OF SOCCER PLAYERS*. International Conference on Computers and Industrial Engineering, no. 35, 2015. (Cited on pages 9 and 11.)

[Leite 2013] Werlayne Stuart Soares Leite. *The Impact of the First Goal in the Final Result of the Futsal Match*. Annals of Applied Sport Science, vol. 1, 2013. (Cited on page 10.)

[Peñas 2014] Carlos Lago Peñas and Maite Gómez López. *HOW IMPORTANT IS IT TO SCORE A GOAL? THE INFLUENCE OF THE SCORELINE ON MATCH*

PERFORMANCE IN ELITE SOCCER. Perceptual and Motor Skills: Learning and Memory, vol. 3, pages 774–784, 2014. (Cited on page 10.)

[Ribeiro 2020] João Nuno Ribeiro, Bruno Gonçalves, Diogo Coutinho, João Brito, Jaime Sampaio and Bruno Travassos. *Activity Profile and Physical Performance of Match Play in Elite Futsal Players*. Frontiers in Psychology, vol. 1, no. 1709, 2020. (Cited on page 11.)

[Sarmento 2015] Hugo Sarmento, Paul Bradley and Bruno Travassos. *The Transition from Match Analysis to Intervention: Optimising the Coaching Process in Elite Futsal*. International Journal of Performance Analysis in Sport, vol. 15, pages 471–488, 2015. (Cited on pages 10 and 11.)