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: Resum

Alumne: Llorenç Peirau Gabarrell

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

Convocatòria (mes/any): Juny 2022

# Summary

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.

In order to achieve them, the results have been divided into two groups: non-visual and visual. On the one hand, the non-visual results are related to the first two aims. On the other hand, the visual results correspond to the third aim.

Moreover, to elaborate this results, it has been used a wide range of tools. Firstly, to extract the data from the photo it has been worked in Google Colaboratory. More precisely, it has been used a combination of two libraries: Extract-

Table and Pandas. In fact, most of the tools used are from the Google environment. For example, it has been used Google Drive to save all the images and the CSV files created and the Google Cloud Services (GCP) to store the database on the cloud. Inside GCP, it has been used BigQuery and Cloud Storage as well. One more tool used from the Google environment has been Looker to share an interactive dashboard.

Outside the Google environment, it has been worked with R in Rstudio to perform Principal Analysis Component and Unsupervised Machine Learning (Gaussian Mixtures) on the data.

Inside the first group of results, the non-visual results, there are three main outputs, two corresponding to the first aim and one corresponding to the second aim.

The first output it is just an improvement on how to collect the data. The person in charge of full-filling the two paper sheets that GEFS collects, writes a stick every time an event occurs. So, to facilitate the data extraction it was suggested to add a cell next to each event that would contain the number that the sticks represent.

The second output is the pipeline created in order to be able to digitalize the data with only taking one photo. At this point it was checked that due to how the photos sent by GEFS were taken and due to the amount of typos in them, there was need of creating a template that could summarise all the information. The last version of this template was the fifth and it worked perfectly with the ExtractTable library. Once all the data was summarised in template v5, a photo was taken and uploaded to Drive. From there with Google Colaboratory and specific functions created for this thesis, the data was extracted and saved in the correspondent dataset. There were as well functions to calculate some extra features, to correct some possible mistakes and to check that the content made sense.

Once the data was obtained, it was turn to create the different datasets that form the relational database created for this thesis from scratch. This database is formed by two groups of datasets. The first group is called static information group because the information does not grow as the matches are played. In this group there are three datasets that contains:

1. **Squad:** the team members of GEFS.

2. **Teams:** the teams that GEFS have played against.

3. **Season:** the matches that GEFS have played

On the other side, the second group is called growing information group and it is formed by 5 datasets. Contrarily as before, they save more information as matches are played. They contain:

1. **Players:** the players from GEFS available in each match.
2. **Match:** the statistics from the each match such as the total number of corners or free-kicks.
3. **Goals:** the goals scored in each match and the GEFS players that were playing when a goal was scored among others.
4. **Minutesh1:** the minutes played per match and player at the first half.
5. **Minutesh2:** the minutes played per match and player at the second half.

It has to be taken into account that at the beginning there was an iterative process between these second and third outputs to design the best template and the best structure for the database. In fact these outputs could be considered as the most important ones from the thesis.

In addition, the pipeline created allowed the data to be appended to the correspondent dataset every time a photo was processed. So, the workflow was:

- During the weekend GEFS plays a match and the data is collected in two different paper sheets.
- At the begning of the week, GEFS sends two photos, one per paper.
- The template v5 is full-filled.
- The template is processed. Consequently, the data is obtained.
- The data is automatically stored in the correspondent dataset from the database.
- Another weekend has arrived and the proces goes back to the first point.

For the second group of results, the visual group, there were also three outputs. The first output was a dahboard created with Looker that had the information that the coach of GEFS requested: the total minutes played per player, the number of goals scored per player and the presence of each player.

The second and third output of this group had to be considered as a first approach because the amount of observations was very low. There were only

19 players available to analyse their performance. Consequently, the conclusions reached had to be interpreted cautiously.

This second output consisted on performing Principal Component Analysis on the performances of the players. To get the desired data to be able to evaluate it, there were used SQL queries in BigQuery. The aim was to determine which features characterise the most among the performances of the players. The results showed that the most important features are:

- **Lost:** the number of losses per match.
- **Tackle:** the number of recuperations per match.
- **TG_received:** the number of team goals received when the player was on the court.
- **TG_scored:** the number of team goals scored when the player was on the court.

For the third output there were two aims: to estimate the density function and to cluster the players according to their performances. The data used was the same as before (i.e. only taking into account 19 players). After checking that the Gaussian distribution was unable to capture all the features, the best model according to the BIC criterion was an EEE (ellipsoidal, equal volume, shape, and orientation) with 7 components. Then the cluster was made based on it. There were 7 groups suggested and it could be found a logical explication for each. Consequently, it seemed that in order to find similarities between the performances of the players this technique works well.

Additionally, it is important to comment that there is no money cost for this thesis. All the tools mentioned are free or they have a free version that suits perfectly for this scenario.

To conclude, 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.

In addition, 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. Separating the results into two groups has eased this achievement. 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.

Additionally, 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.

Finally, this master's thesis has worked with concepts that were not taught at master's lessons. Moreover, 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.

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

https://github.com/Lpeirau2/TFM