



# Data Mining Project

A.A. 2021 / 2022

*Mariagiovanna Rotundo*

*Nunzio Lopardo*

*Renato Eschini*



# Introduction

We work on all the assigned four tasks included the **optional** ones.

From task 1 to 3 the goal was to work on **tennis matches** and **players dataset**.

We tested and analyzed clustering and classification algorithms on players profiles.

For the task 4 we choose the **analysis** of the **time series** about the temperatures of the cities.



# Notebooks

The created notebooks are 6.

- Task 1
  - Data\_understanding
  - Data\_preparation
  - Functions\_understanding
- Task 2
  - Clustering\_analysis
- Task 3
  - Predictive\_analysis
- Task 4
  - Time\_series\_analysis

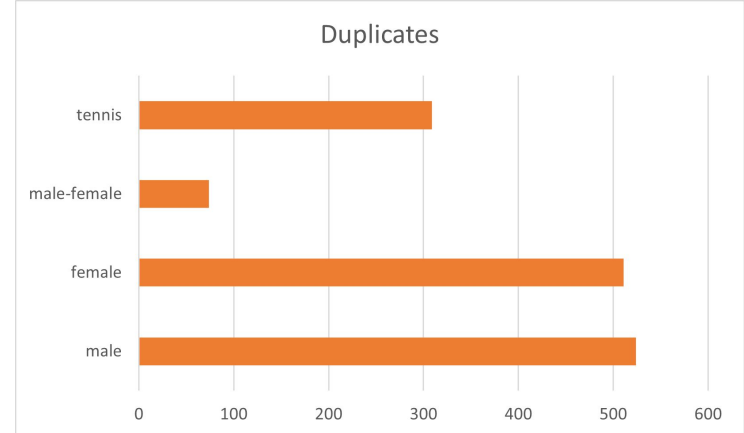
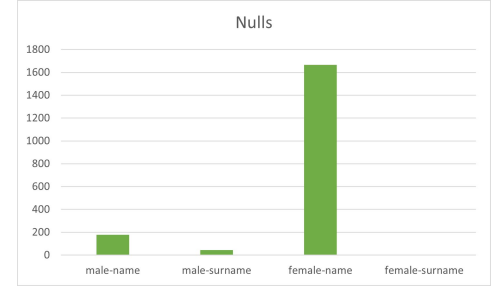
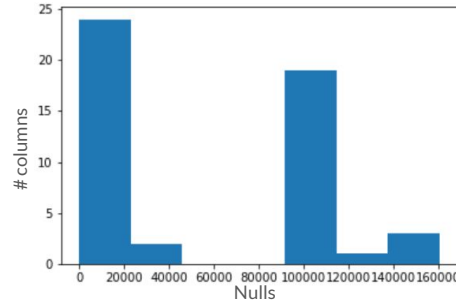
# Task 1

## Understanding

Analysis of the provided dataset to understand the context and to find errors and missing values in the data. Problems founded:

- many null values on the match statistics;
- duplicates in all three datasets;
- inconsistent data on match information

During the work, we use **2 external CSV** files: one with the country IOC codes and one with some people's names



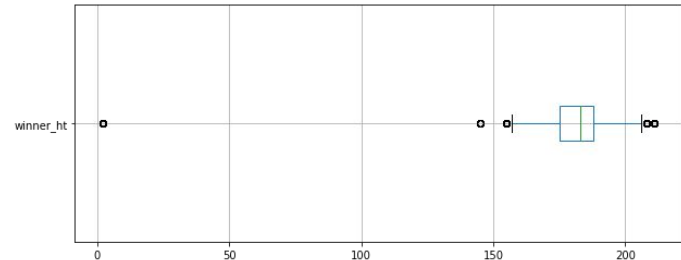
# Main dataset errors and outliers

## Errors:

- None unique ID for players;
- Best of values and scores;
- Best hand not always the same.

## Outliers:

- Player tall 2 cm;
- Number of served points over 1500;
- Too much older tennis player.





## Preparation

Dataset correction and feature extraction for tennis players. We have created a new dataset containing all the **players' information** that has been used in the next phases.

## Correlation

We have analyzed correlation both **before and after the changes** to the tennis dataset to see if changes introduce correlation between some features. We have noticed that some features are correlated both before and after the changes, but others only in one of this moments.



## Task 2: Clustering

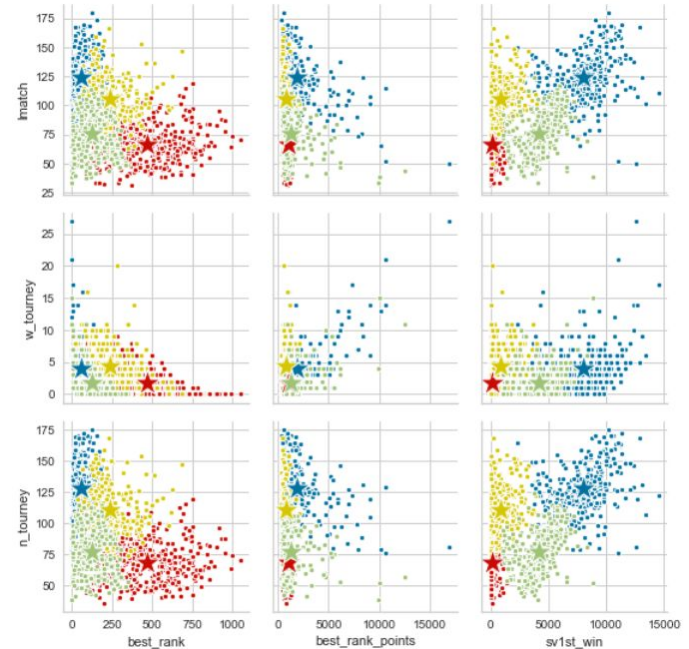
Used algorithms on the players dataset.

- K-Means
- DB - Scan
- Hierarchical - Agglomerative Clustering
- X-Means
- G-Means

# Results

Analyzing the clustering results we noticed that the best results, evaluated with elbow method, came from **K-Means** with four clusters. Indeed using 4 centroids we get an optimal balancing between players separation, SSE and Silhouette Score.

## K-means

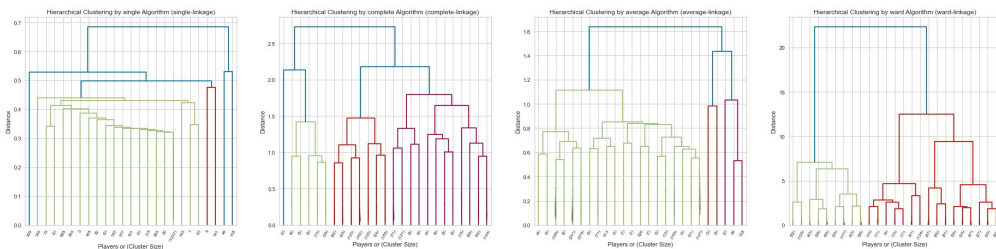




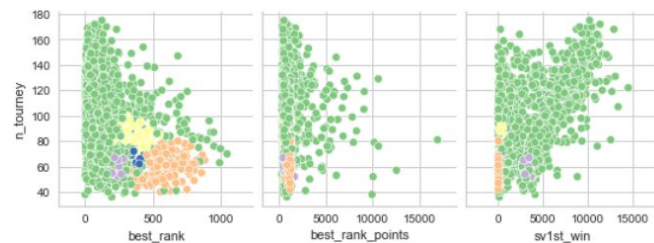
For **DBSCAN** (with *StandardScalar* and *MinMaxScalar*) and **Hierarchical clustering** (*single*, *complete*, *average*, *ward*) in most cases we get poor results due to the very small distances between the players.


Evaluated by Silhouette, Davies–Bouldin and Calinski-Harabasz scores for Hierarchical and with grid search for *eps* and *min\_samples* in DBscan.

## Hierarchical



## DBscan

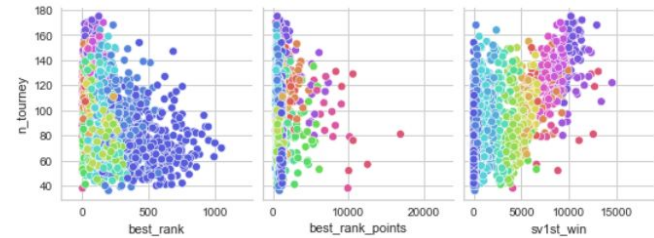




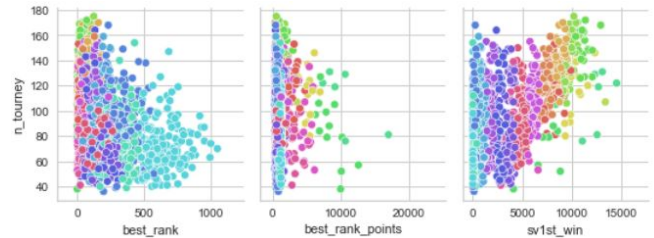
The behavior of optional clustering algorithms (**X-means** and **G-means**) is very similar. For both the algorithms, the computation to find the optimal number of clusters stops to 20 clusters, the number indicated as maximum.

We evaluated the two algorithms by comparing SSE, which was very similar.

### X-means



### G-means





## Task 3: Classification

### Classifiers

- Decision Tree (DT)
- SVM
- Rule Based (RB)
- AdaBoost (AB)
- Random Forest (RF)
- KNN
- Neural Networks (NN)
- Gaussian Naive Bayesian (GNB)

### Approaches

1. ignore the imbalance in data
2. assign **weights** to the low and high classes
3. use of oversampled training set using **SMOTE**



## Classification analysis

We have trained classifiers to predict if a player has a high or low rank:

- $best\_rank \leq 50$ , high rank has label 1;
- $best\_rank > 50$ , low rank has label 0.

For the NN and KNN classifiers we have used normalized datasets.



<b>F1-score - High Rank class</b>	DT	SVM	RB	NN	KNN	GNB	AB	RF
Train - Original	0.93	0	0.69	0.80	1	0.59	0.99	1
Test - Original	0.84	0	0.69	0.81	0.71	0.60	0.85	0.91
Train - Weights	0.85	0.61	0.72	0.81	-	-	-	1*
Test - Weights	0.82	0.63	0.70	0.85	-	-	-	0.85*
Train - SMOTE	0.92	0.59	0.87	0.90	0.99	0.58	0.99*	1*
Test - SMOTE	0.87	0.61	0.81	0.84	0.75	0.60	0.84*	0.87*

\* Data not present in the notebook



## Task 4: Time series analysis

We have chosen this task because, for us, it's the most interesting and it was an opportunity to go into detail of the topic (looking at it with a practical point of view).

We consider as **similar the trends** of 2 cities if the shape of the time series is similar. We consider them similar also in the case they are translated in time and they shows the same behavior but with different amplitudes of the temperature's oscillations.

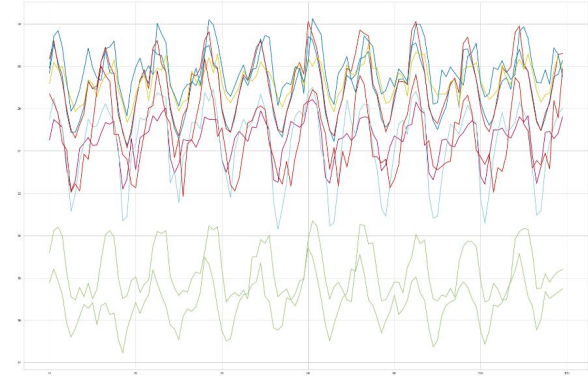
Instead, for example, 2 trends are considered **different** if one always oscillates around a certain value and the other shows a growth of temperatures in time.

## K-Means

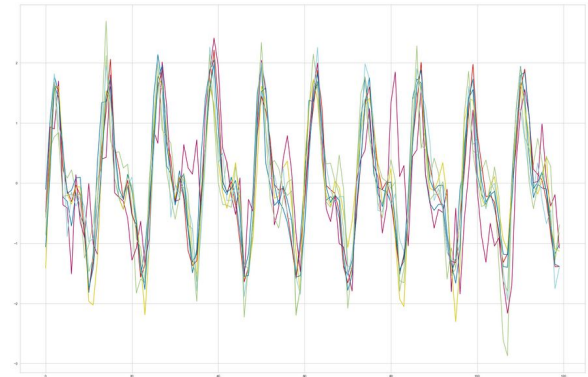
For find the best the number of clusters for the K-means, we used the elbow method. We used 2 kind of distances:

- **Euclidean distance.** For use this distance we have corrected the distortions of offset translation, amplitude scaling.
- **Dynamic Time Warping.**

Euclidean distance



DTW



## Hierarchical clustering

For the hierarchical clustering we use the **Euclidean distance** and, for the distances between "partial" clusters we use an **average** method. For the number of cluster we choose the value that gives us the best clusters.







**Thanks for your attention**

**Questions?**