DEPARTMENT OF INFORMATICS, SYSTEMS AND COMMUNICATIONS

MASTER DEGREE IN DATA SCIENCE

The Italian women's national volleyball team: building the dream for the 2025 world championship

DATA MANAGMENT REPORT

Aurora Castelnovo

Nicole Gemelli

Daria Marinucci

Academic Year 2024/2025

Contents

1. Introduction

The Italian Women's National Volleyball Team stands among the most dominant and respected teams in international volleyball. Over the years, they have built a reputation for excellence, consistently ranking among the world's top teams. As **r**eigning Olympic champions from the Paris Games, they have showcased their superior skill, exceptional teamwork, and unwavering determination in numerous prestigious competitions, including the European Championships, World Cup, and World Championships.

What sets the Italian women's team apart is not just their impressive array of titles but their style of play. Renowned for their dynamic, fast-paced game, they combine technical precision with powerful attacks, creating a mesmerizing spectacle for volleyball fans. This unique playing style, marked by fluidity and intensity, has earned them respect and admiration across the volleyball world. Their performances are a testament to their relentless drive for excellence, and they continue to inspire countless fans and future players.

The Italian volleyball league, or Serie A1, is also among the most competitive and widely followed leagues in the sport, attracting top talent from all over the world. Playing in this league is considered a pinnacle for many players, as the level of competition is high, and the passion for the sport runs deep. Almost every player dream of being selected to play for one of the league's prestigious clubs, knowing that it provides the opportunity to showcase their skills on the world stage and further develop their careers. The excitement and fervor surrounding the league contribute to Italy's long-standing volleyball legacy, making it a major force in both the European and global volleyball scenes.

As part of our ongoing efforts to improve team selection for major tournaments, we aim to create and integrate a database of Italian players and their performance statistics. The objective is to analyze these data points and establish clear criteria for player selection. This will allow for a more informed decision-making process, ensuring that the final team chosen for the tournament is not only well-rounded but also composed of the players who offer the best fit based on performance metrics. In order to do this, we will follow the data science pipeline going through these steps:

- *data acquisition*: collection of data using web scraping
- *data storage*: storage of data on a persistent database
- *data integration and enrichment*: to get a better understanding and a more valuable dataset
- *data quality*: evaluate and improve the different quality dimension of data
- *exploratory data analysis*: query data in order to discover insights
- *predictive analytics*: to propose a team for a future tournament

2. Goal: A Data-Driven approach to team selection

As volleyball continues to evolve, so must the methodologies used for player evaluation and selection. Traditional scouting methods, while valuable, often rely on subjective judgments and qualitative assessments. This project aims to introduce a data-driven approach that ensures fairness, accuracy, and efficiency in identifying the most suitable players for major international tournaments.
By gathering data from various sources, primarily web pages and official league statistics, we aim to construct a detailed and well-rounded player profile that considers both individual and team contributions.
This approach aims to achieve the following key objectives:

- **Improve accuracy** in player evaluations by utilizing performance data from actual matches.
- **Reduce selection bias**, ensuring players are chosen based on merit and data-driven insights.
- **Identify rising talents** who may not yet have widespread recognition but demonstrate strong potential.
- **Optimize team composition**, selecting players who complement each other's strengths.
- **Enhance Italy's chances of continued success** in international competitions.

To achieve this, we will collect, process, and analyze performance data from multiple sources, including match statistics, league records, and historical trends. The insights gained will be a powerful tool for coaches, analysts, and selectors, helping them make well-informed decisions when assembling the national team.
A crucial aspect of this project is establishing a reasonable threshold for selection. Since players may have varying numbers of games played, we must determine a minimum number of matches or sets required to ensure a valid evaluation. Without a sufficient sample size, performance metrics could be misleading. Thus, we will implement data filters to include only players who have participated in a reasonable number of matches to be considered for selection.
Another key aspect of this approach is role-specific evaluation, recognizing that different positions require distinct skills:

- **Middle blockers (centrali)**: primarily evaluated on blocking efficiency, attack effectiveness, and serving impact, while reception is less relevant.
- **Outside hitters (schiacciatori/opposti)**: their evaluation considers both offensive (attack success rate, serving) and defensive (receiving, digging) capabilities.
- **Setters (palleggiatori)**: assessed based on assist accuracy, strategic decision-making, and distribution effectiveness.
- **Liberos**: since they do not attack or serve, their evaluation focuses entirely on defensive capabilities, reception quality, and dig efficiency.

3. Tools

The development of the project has involved different tools through the stages. Each tool was chosen based on its strengths in handling specific tasks, ensuring efficiency, accuracy, and scalability. In this paragraph we will just provide a brief introduction of the instruments that we have chosen, further explanation will be given later.

Below is a brief introduction to the technologies we utilized, with further details provided in later sections.

- Data Acquisition: web scraping was performed using Python with the *BeautifulSoup* library, allowing us to extract structured data from reliable online sources.
- Data Storage: given the need for a flexible and schema-less model, we opted for MongoDB, a NoSQL document database that efficiently stores structured player statistics.
- Data Processing and Enrichment: the data processing and feature engineering tasks were carried out using Python (pandas, NumPy) and MongoDB. KNIME, a powerful data analytics platform, was used for data integration and preprocessing.
- Data Exploration and Visualization: to enhance understanding and generate meaningful insights, we used Tableau, a data visualization tool that helped us create clear and interactive visual reports.
- Predictive Analytics and Machine Learning: the final stage of analysis was conducted using KNIME, a leading machine learning platform. This allowed us to apply predictive models to evaluate player potential and optimize team selection.

Data acquisition

Data integration and enrichment

Data quality

Data storage

Data integration and enrichment

Data quality

Prediction

4. Data acquisition

Data acquisition is one of the most important steps of the data science pipeline because it represents the phase in which we search for data and get them. Unfortunately, we couldn't find a single dataset that met all of our requirements, so we had to consult multiple sources. We looked for the most reliable sources, starting with the official website of Lega Volley Femminile. Additionally, we needed to find a website for the player data of each tournament; this source was challenging because, initially, we only found online articles. Ultimately, we decided to use Wikipedia.

We explored various acquisition techniques, but we had to exclude API because none of the websites we consulted allowed us to connect. However, all of them support web scraping quite easily. Web scraping is an acquisition technique that allows access to websites and extract data from them. We have considered different ways to perform it but at the end we chose Python because it offers several libraries that are well-suited for handling web scraping. Among these, we specifically considered Selenium first, and, then, *Beauifulsoap*. *Selenium* is often praised for its ability to handle dynamic content rendered by JavaScript. However, we encountered a key limitation: Selenium's performance can be slow because it controls an actual browser to retrieve and parse the content, which introduces a layer of overhead. For a large-scale scraping project, this can significantly reduce the efficiency of the task, especially when dealing with numerous pages or extracting large volumes of data.

On the other hand, *BeautifulSoup* is a library specifically designed to parse HTML and XML documents. It is lightweight, fast, and does not require the overhead of simulating browser actions. This made it the more suitable option for our project, where speed and efficiency were paramount. *BeautifulSoup* excels in scenarios where the data we are scraping is already present in the static HTML source code of the webpage. Since we were primarily gathering statistics and player data from structured pages, *BeautifulSoup*'s ability to parse and navigate HTML documents quickly allowed us to automate the extraction process without the additional complexity that comes with using *Selenium*. Another significant factor in our decision was the nature of the data we needed to extract. In this project, we were dealing with relatively straightforward, tabular information — player statistics, personal information, and tournament rosters — that could be easily accessed from the raw HTML. *BeautifulSoup* is particularly effective at working with this kind of data, as it enables easy navigation through tags, attributes, and nested elements to pull out the necessary information. Furthermore, the library integrates seamlessly with other Python tools like requests and pandas, which made the data processing pipeline smoother and more efficient.

Lastly, the need for scalability played a role in our decision. *BeautifulSoup* is more lightweight and less resource-intensive than *Selenium*, which made it better suited for handling large amounts of data over extended periods without consuming excessive system resources or causing delays. Given that we needed to scrape data from multiple seasons and player datasets, the efficiency and speed of *BeautifulSoup* aligned better with the project's requirements.

In summary, we opted for *BeautifulSoup* because it offered faster performance, lower resource consumption, and simplicity in handling the static data we needed. While *Selenium* is a powerful tool for more dynamic, interactive web scraping, the straightforward nature of the datasets we were targeting made *BeautifulSoup* the more efficient and effective choice for our goals. By choosing it, we ensured that the scraping process would be automated, efficient, and scalable without unnecessary complexity.

To summarize, we used the following sources:

- *statistics athlete* section of Lega Pallavolo italiana for statistics broken down by season, athletes, etc.
- *athlete* section of Lega pallavolo italiana for players profiles.
- wikipedia for team turnaments

4.1 Statistics

Lega Pallavolo Italiana (Italian Volleyball League) plays a central role in providing the key statistics that help us understand player performance in Italy's top volleyball league. These statistics are not only crucial for assessing individual players but also for evaluating team dynamics and overall league performance. The statistics collected from this source serve as the foundation of our dataset, providing quantitative insights that are essential for making data-driven decisions when selecting players for the national team.
In the world of sports, statistics are vital for assessing player performance. They allow coaches, analysts, and selectors to make objective decisions rather than relying solely on subjective observations. In volleyball, specific statistics such as attack success rate, serving effectiveness, reception accuracy, blocking efficiency, and digs are key to evaluating the impact of each player on the court. This is particularly important in a team sport like volleyball, where different roles and positions require distinct skill sets.

The dataset we collected from *Lega Pallavolo Italiana* included various statistical categories that were essential for a comprehensive analysis.

A brief explanation of the columns included in the statistics dataset:

- **Athlete:** name and surname of the player.
- **Total number of matches played:** how many matches an athlete has played.
- **Total number of sets played:** these are the individual segments of a match, and they can vary between 3 and 5 for each match.
- **Total number of points:** points scored by the players.
- **Winning points:** points scored from actions starting with the opponent's serve.
- **Break point (bp):** points scored during the break phase, which is when the play starts with the team's own serve.

Serve: the fundamental action to start the game.

- **Total:** the total number of serves made.
- **Ace:** a serve that the opponent cannot return or touch, resulting in an immediate point.
- **Errors:** mistakes made during serving, such as incorrect serves or faults.
- **Aces per set:** calculated as the ratio of the total number of aces to the total number of sets played.
- **Efficiency:** calculated as the ratio of the difference between perfect serves and errors to the total number of serves made.

Reception: this is the action of intercepting the opponent's serve and trying to send the ball to the target area, which is the area of action for the setter.

- **Total:** the total number of receptions made.
- **Errors:** mistakes made during the reception.
- **Negative:** the number of receptions that were either erroneous or poor.
- **Perfect:** the number of perfect receptions.
- **Perfect%:** the percentage of perfect receptions relative to the total number of receptions made. These are receptions directed to the target area that allow the setter to play any offensive solution.
- **Efficiency:** calculated as the ratio of the difference between perfect and negative receptions to the total number of receptions made.

Attack: any technical movement with which the player sends the ball into the opponent's court; the primary attack is the spike, but there are others, such as the dink or the push shot.

- **Total:** the total number of attacks made.
- **Errors:** mistakes made during the attack phase.
- **Blocked:** the number of attacks that were blocked by the opponent at the net.
- **Perfect:** the number of winning attacks.
- **Perfect%:** the percentage of winning attacks relative to the total number of attempts.
- **Efficiency:** the ratio between the difference of winning attacks and errors to the total number of attacks made.

Block: The action by net players to intercept the opponent's attack.

- **Invasion:** the number of faults committed during the blocking phase, such as crossing the center line or touching the net.
- **Perfect blocks:** the number of successful blocks, i.e., blocks that land in the opponent's court.
- **Points per set:** the number of points scored per set.

To ensure accuracy and consistency in our results, we divided the dataset by season. This segmentation allows us to focus on the performance trends over time and ensure that the most recent data is available for evaluation. Additionally, we added a *Role* column to each dataset, linking each player to their respective position on the court. By organizing the data season by season, we could analyze the performance trajectory of players and better understand how their contributions vary in different seasons.

4.2 Personal information

The *Personal Information* section is crucial for understanding who the players are outside of their performance metrics on the court. While the statistics tell us about their skills and contributions in matches, personal information gives us a deeper context about their careers, nationality, and eligibility for the national team.

One challenge we faced was distinguishing between players who are eligible for the Italian national team and those who are not. Since the statistics dataset included players from all nationalities participating in the Italian league, we needed to filter out only the Italian players, as they are the ones eligible to represent Italy in international competitions.

To extract this information, we turned once again to web scraping. This allowed us to gather the necessary details about the players' backgrounds, including:

- **Name:** the first name of the athlete
- **Surname:** the surname of the athlete
- **Role:** the role in which they play, it can be: Setter, Middle Blocker, Libero, Opposite Hitter or Spiker
- **Nationality:** the athlete's nationality; this was essential for determining whether the player could potentially be called up to the national team.
- **Height:** ow tall they are
- **Birth Year:** birth year of the athlete; provides insight into the player's age and potential for future involvement with the team.

We immediately notice that there are some discrepancies with the first dataset, such as the format of the name. Before the integration we need to perform the needed changes.

4.3 Volleyball team

The *Volleyball Team* dataset presented a unique challenge because finding reliable and comprehensive information about the national team players, especially regarding their participation in different tournaments, was not as straightforward. Most of the data available were scattered across various sources, including news articles, online lists, and official team rosters. However, these sources often lacked sufficient details, such as the player's role, and didn't provide the full context we needed for the project.

The difficulty stemmed from the fact that the team rosters for different tournaments were often not available in structured formats. We found lists of names, but they typically lacked the roles (positions) of the players, and there was often little to no information about their performance or specific contributions to the team. Additionally, some tournament rosters were only available in the form of online news articles, which did not have the standardization we needed for our analysis.

To overcome this, we used **Wikipedia** as the primary source to extract the data. While Wikipedia is not always 100% reliable, it often provides a decent starting point for gathering data, especially for well-known entities such as national teams. The official team rosters for major tournaments like the VNL,World Cup, European Championship, and the Olympics are typically updated on Wikipedia pages dedicated to those events, and they usually include player names, positions, and relevant details.

We have removed the *trainer* because it was not necessary.

5. Data storage

After the data acquisition phase, the next crucial step is data storage. The choice of a storage solution is pivotal as it directly impacts the efficiency, scalability, and flexibility of the data processing pipeline. In our case, we were dealing with several different datasets, each with varying characteristics. The statistical dataset, for example, contained a large number of columns with complex attributes, whereas personal data or tournament-related datasets were simpler and smaller in size. These differences in the types of data required us to select a storage model capable of handling diverse structures and large volumes of data.
Initially, we considered using a **relational database model**, which is the most commonly used solution for structured data. Relational databases use tables with fixed columns and rows, enforcing relationships between different data entities through foreign keys and indexes. While SQL-based systems are excellent for handling structured, normalized data, they became unsuitable for our project for several reasons.

5.1 Why not SQL?

One of the primary limitations of relational databases is their rigid structure. They require a predefined schema that does not easily accommodate variations in the data. In our case, some datasets, such as the statistics, contained a large number of columns with varied data types (e.g., numeric, text, percentages), making them ill-suited for relational tables. Additionally, our data was fragmented across multiple sources, requiring a more flexible storage model that could handle these differences without the need for complex schema modifications.
Relational databases also impose strict relationships between tables, which can create challenges when integrating data from disparate sources with different structures or relationships. Our project required the ability to store semi-structured and unstructured data, such as player profiles, match statistics, and tournament details, in a scalable manner that could evolve as new data was added. Thus, a more adaptable storage model was essential.

5.2 Why MongoDB

To address these challenges, we opted for a NoSQL model, which offers greater flexibility and scalability than relational databases. NoSQL databases are designed to accommodate a wide range of data models, including document, key-value, column-family, and graph-based structures. These databases are optimized for large-scale, distributed environments, making them ideal for managing vast datasets spread across multiple servers.
Among the various NoSQL options, we chose a document-based database, specifically **MongoDB**. MongoDB stores data in JSON-like documents, allowing us to incorporate nested structures, arrays, and other complex data types. This flexibility was crucial for our project as it allowed us to store athletes' personal data, match statistics, and tournament details in a format that was easy to manage and manipulate. We thought about a NoSQL models, that were first proposed by Carlo Strozzi in 1998, this kind of models are designed to handle a wide variety of data and scales horizontally across multiple servers. NoSQL models follow the CAP theorem or the BASE principle. The first says that it’s impossible for computer systems to provide at the same time all three guarantees of consistency, availability and partition tolerance. The second one provides guidelines that describe the behaviour of these models: Basic Availability, Soft State and Eventual Consistency.

MongoDB's schema-less structure made it an ideal choice for this project. In a document-based database, each document can store data in a semi-structured format, enabling us to handle a variety of data types and structures without the need to adhere to a fixed schema. This was particularly important for managing athletes' names, as some players had single names, while others had multiple names or compound surnames (e.g. Monica De Gennaro). MongoDB allowed us to store these variations without imposing rigid naming conventions or formats.
Additionally, MongoDB supports horizontal scaling, meaning that as our dataset grows over time, we can distribute the data across multiple servers. This ensures that the system remains fast and responsive, even as the volume of data increases. Horizontal scaling was especially important as we anticipated the continual expansion of the dataset with new seasons, players, and tournaments.
To interface with MongoDB, we utilized the *pymongo* library in Python, which provided an efficient and user-friendly way to interact with the database. This library allowed us to store, retrieve, and manipulate data directly from Python scripts, ensuring seamless integration between the database and the data analysis pipeline.

In summary, adopting MongoDB enabled us to build a flexible, scalable, and efficient data storage system capable of managing diverse datasets with varying formats and sizes. This decision laid a solid foundation for the data analysis pipeline, allowing us to store large amounts of data in an accessible and organized manner while supporting future growth and complexity as the project continues to evolve.

6. Data profiling

After data acquisition, it is pretty important to conduct a preliminary examination of the data to better understand its properties and characteristics. Here comes data profiling, this activity aims to provide a quick overview of the content, allowing us to find the issues or inconsistencies that could create problems in future steps as integration.

6.1 Statistics dataset

Starting from *statistics* dataset, we wanted to check the percentage of missing values and the type of attribute contained. All the statistics data frames have the same structure, for this reason we have decided to take the last one as example to perform a preliminary exploration of the data contained in it. Firstly, we wanted to know which types contained the columns and the total number which is equal to 25.

[Athlete, Role] = String

[Match_played]= int64

[Set_played, Pt_Tot, Winning_Pt, BreakPt, Serve_Tot, Ace, Serve_Err, Ace_per_set, Serve_eff, Rice_Tot, Rice_Err, Rice_Neg, Rice_Prf, %Rice_Prf, Rice_Eff, Att_Tot, Att_Err, Att_Block, Att_Prf, %Att_prf, Att_Eff, Block_Inv, Block_Prf, Block_Pt_Set] = float64

We wanted also to know if there were missing values, most of attributes had them. The columns with the hights percentage of missing are:

- Block_Inv 89.316239
- Rice_Prf 50.427350
- %Rice_Prf 50.427350
- Rice_Err 42.735043
- Rice_Eff 42.307692

For a better understanding of the data, we use the function *describe* to do it.
Here is a sample:

	Match_played	Set_played	Pt_Tot	BreakPt	Serve_Tot	Att_Tot	Ace
count	234.000000	199.000000	166.000000	164.000000	174.000000	163.000000	153.000000
mean	18.636752	52.572864	121.313253	48.140244	153.775862	243.171779	9.398693
std	7.602835	28.412074	113.824419	41.740197	112.927115	252.873794	8.210503
min	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
25%	18.000000	30.000000	20.500000	10.750000	42.000000	42.500000	3.000000
50%	23.000000	55.000000	92.000000	39.500000	155.500000	164.000000	7.000000
75%	23.000000	79.000000	178.500000	73.000000	264.750000	359.500000	13.000000
max	24.000000	97.000000	506.000000	200.000000	378.000000	1082.000000	52.000000

Figure 1. Summary of the main statistics.

6.2 Personal information dataset

Moving to *personal information* dataset, we found out immediately that it was smaller than the previous one. It has only 5 columns, with these types:

[Athlete, Role, Nationality, Height] = string

[Birth Year] = int64

There aren't missing values. The oldest player seems to be the spiker Aguero Taismary, from 1977. The youngest players are from 2008 and they are: Arcangeli Chiara [Middle Blocker], Monti Martina [Spiker], Monti Noemi [Spiker], Zeni Beatrice [Libero].

Here is a visual representation of the distribution of the athletes per role, relative to the 2024/2025 season. As shown in the graph, the most represented role is spiker, followed by middle blocker. The number of liberos and setters is slightly lower, while the opposite hitters have the least representation.

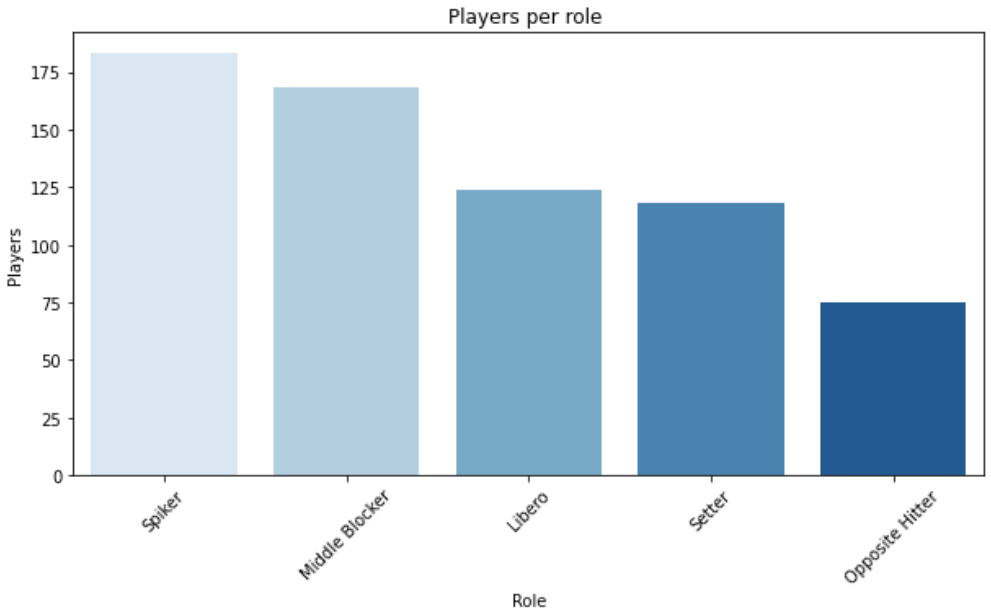

Figure 2. Role distribution, season 2024/2025.

6.3 Team dataset

About *team* dataset, we have first checked the attribute types.

> *[Player, Role] = string*
>
> *[Year, EU23, WO22, EU21, VNL24, VNL23, VNL22, VNL21] = int64*

We noticed that there aren't missing values.

Based on this graph, we can make the following observations:

- The Spiker role is clearly the most common in all leagues, with a particularly high peak in the Wo22 league.
- The Libero role is the least represented in all leagues: this is probably caused by the fact that this role is very specialized, so it's less common.
- Each league has a slightly different distribution of roles, suggesting possible differences in team strategies or regional preferences.

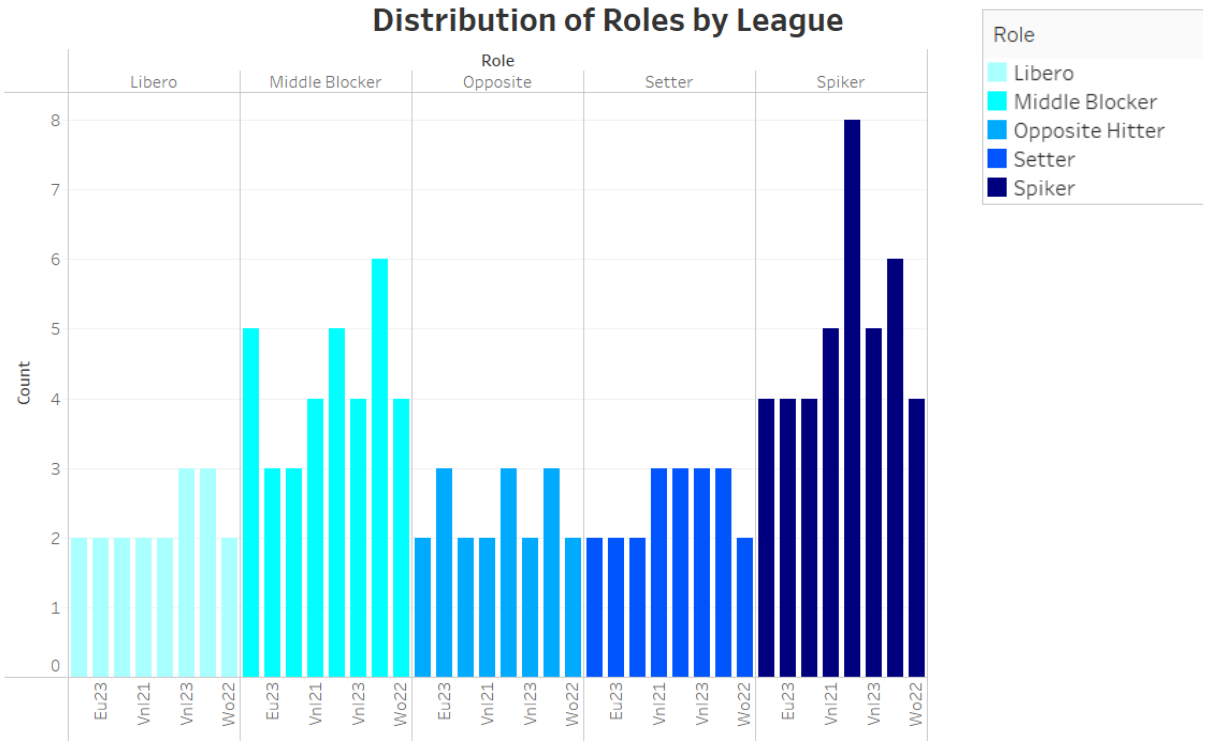

Figure 4. Role distribution by tournament.

7. Data preparation

This phase is divided in two parts: data cleaning, data integration and enrichment.
Data cleaning is the process of identifying and correcting errors, inconsistencies, and missing values in a dataset to improve its quality and reliability. It involves removing duplicates, handling missing data, correcting inaccuracies, and standardizing formats to ensure accurate analysis and decision-making.
Data integration consists in the process of combining data from multiple sources into a single unified view, this makes the data more understandable and easier to analyze. Data enrichment is the process of adding additional context or information to the data. This operation makes them more useful and valuable.

7.1 Data cleaning

The data cleaning process was carried out using several tools: Python, Knime, and MongoDB. Below is a summary of the steps taken, particularly focusing on Python.
In Python, one issue we encountered was that some tools had trouble recognizing the decimal part when numbers were written with commas, while others specifically required dots. To address this, we created a function that automatically detects and replaces commas with dots.
Additionally, we needed to reformat the *Height* column in the Italian player dataset. The values were originally in the format "180 cm" as a string. We wrote a function to extract the numeric value and remove the text, resulting in a clean, numeric attribute.
Furthermore, we had to reformat the *birthdate* field in the tournament dataset. The birthdate was initially in a parsed format, but we only needed the year. To achieve this, we wrote a function that extracted and stored just the year.
When working with MongoDB, we faced the challenge of handling missing values. Removing rows with missing values would have been a significant mistake, as the percentage of missing data was over 5%, which would compromise the dataset's representativeness. Although we considered replacing missing values with a fixed quantity (such as the mean), we determined this could distort predictions. For instance, *Monica De Gennari*, a libero, had null values in the *Ace_per_set* column because it's not her role to serve. In other cases, missing values were due to players who hadn't played enough to accumulate meaningful statistics. We wrote the following query to address these missing values appropriately.

```
use Volleyball; #selects the database
const collections = db.getCollectionNames(); #gets all the collections
#for each collection we ask to iterate each document, checks if any value is equal to "NaN" and
eventually transform it into " "
collections.forEach((collectionName) => {
  print(`working with collection n. : ${collectionName}`);
    db[collectionName].find().forEach((doc) => {
    let updates = {};
    Object.keys(doc).forEach((key) => {
      if (typeof doc[key] === "number" && doc[key] !== doc[key]) {
        updates[key] = "";
      }
    });
```

```
#if there are updates execute them
    if (Object.keys(updates).length > 0) {
      db[collectionName].updateOne(
        { _id: doc._id },
        { $set: updates }
      );
      print(`Updated  id: ${doc._id}`);
    }
  });
});
```

As we had the column *Height* we have decided to change the name into *Height_in_cm* in order to be more understandable.

```
db.athletes.updateMany(
  {},
  { $rename: { "Height": "Height_in_cm" } }
)
```

7.2 Data integration and enrichment

The data integration process was carried out following the steps outlined below:
Schema Transformation: this step was partially explained in the previous section. Additionally, we created a new *Athlete* column in every dataset to enable proper matching.
Schema Matching: we identified all possible matches between datasets.
Schema Integration: this involved combining schemas and resolving conflicts through transformations.
One of the first challenges we encountered was the different formats used to identify athletes across datasets. These changes were made in Python. As previously mentioned, each dataset had a different format for identifying players, and we decided to standardize it according to the format used in the statistics dataset. We stored the athlete's full name under the *Player* column (last name, first name). In the *Italian Players* dataset, the first and last names were stored in separate columns, so we wrote a function to concatenate these two strings. Additionally, the *Tournaments* team dataset used a different format (first name, last name), which we converted to match our chosen format.
Another issue was related to the *Role* column. In the statistics dataframe, we manually added a *Role* column with values like Middle Blocker, Libero, Opposite Hitter, Setter, and Spiker. In the *Italian Players* list, we needed to translate role names from Italian to English, for which we implemented a translation function. The *Team* dataset had a similar issue, using only the first letter of the Italian role name (e.g., 'C' for *Centrale*, meaning Middle Blocker), which we resolved by expanding it into the full role name.
We carefully examined all the datasets to identify similarities and differences. We found that the *Athlete*, *Role*, and *Birth Year* columns were the only common attributes across the datasets, so we decided to use these as the key for matching the datasets.

During the integration process, we encountered various types of conflicts that required adjustments. The first issue was a structural conflict due to the different formats used for the athlete identifiers. We also faced descriptive conflicts, as the *Role* column was written in different formats.

Deduplication was performed on the *Italian Players* dataset. During the scraping phase, we implemented a function that downloaded the data for each season, inserted it into the dataset, and removed duplicates. For example, if Paola Egonu appeared in multiple seasons, we ensured she was listed only once to avoid data distortion.

Our integration primarily focused on combining multiple data sources. Starting with the *total_ds* dataset, we combined the statistics dataset with the *Italian Players* dataset. We first concatenated all the role-related datasets, adding a *Year* column to indicate the season, and then merged them together. Since we were only interested in Italian players, we performed an inner join to enrich the data with personal details such as *Height* and *Birth Year*.

Further integration was done with the *Team Tournament* datasets. Each tournament had its own dataset, but to streamline the data, we decided to merge them into a single dataset. We concatenated all the tournament data and added a binary column for each tournament, where 1 indicated that an athlete was part of that team, and 0 indicated they were not. The resulting dataset included *Name*, *Role*, *Birth Year*, and the teams for each player. We were aware of some changes that needed to be applied here as well.

Lastly, to obtain a comprehensive view of the players and their 2024-2025 season statistics, we decided to merge all the role-related datasets and join them with the *Italian Players* dataset. This decision was made to determine which players could potentially be selected for international tournaments. This final integration step was performed using MongoDB.

```
db.stat_set_2425.aggregate([ #starting from the setter dataset
  {
    $unionWith: {
      coll: "stat_opp_2425" #combine with the opposite hitter statistics
    }
  },
  {
    $unionWith: {
      coll: "stat_mb_2425" #union with the middle blocker
    }
  },
  {
    $unionWith: {
      coll: "stat_lib_2425" #union with the libero
    }
  },
  {
    $unionWith: {
      coll: "stat_spi_2425" #union with the spiker
    }
  },
#consider the new collection
  {
    $lookup: {
      from: "athletes", #collection to consider for join
      localField: "Athlete ", #attribute from the first collection
```

```
foreignField: "Athlete", #attribute of the athlete collection
            as: "Athlete2425" #final document
        }
    },
    {
        $match: {
            "Athlete2425": { $ne: [] } #consider only the matching documents
        }
    },
#unwind in order to transform the array in a single document
    {
        $unwind: "$Athlete2425"
    },
    {
        $out: "stats2425" #output file
    }
]);
```

8. Data quality

Data quality is important because it directly impacts the decisions made based on data. It affects both rows and columns within a dataset, and poor data quality can lead to incorrect analyses and flawed conclusions. Conversely, high-quality data provides a solid foundation for decision-making. Unfortunately, data integration and enrichment can affect the level of quality, but we have worked to manage and resolve these issues.
The three most important and relevant dimensions for your project are **accuracy, completeness, and timeliness**.
Here's why:

1. Accuracy

Accuracy is crucial because your goal is to select the best players based on performance data. If the data does not accurately reflect reality, you may make incorrect decisions, such as selecting players who appear strong in the data but are not in reality, or excluding talented players due to inaccurate statistics. Since most of the data comes from the Lega Pallavolo, ensuring its correctness and integrity is essential for reliable player evaluations.

2. Completeness

Completeness is particularly relevant because missing data can distort analyses. In volleyball, some statistics may be absent due to the player's role (e.g., liberos do not serve) or because they have played very few matches. Ensuring that the dataset is as complete as possible allows for fair comparisons between players and prevents potential biases in the selection process. Implementing clear rules on handling missing data (e.g., setting thresholds for the minimum number of matches played) is crucial for meaningful evaluations.

In evaluating the completeness of our project's three primary datasets, we observed a notable disparity in missing values. Specifically, our analysis revealed the following:

- **Completeness (total_ds):** 75.21%
- **Completeness (athletes):** 100%
- **Completeness (team_turn):** 100%

These findings indicate that while the athletes and *team_turn* datasets are entirely complete, the *total_ds* dataset contains a significant proportion of missing values.

3. Timeliness (Currency)

Timeliness ensures that the data used is up to date, which is critical for evaluating current player performance. Since form and fitness levels fluctuate over time, outdated data may lead to misleading conclusions. By updating the dataset regularly (e.g., weekly), the model remains relevant and accurately reflects each player's recent performance, increasing the reliability of the team selection process.

By prioritizing these three dimensions, your project will ensure that the data used for selecting the national volleyball team is reliable, fair, and relevant.

9. Exploratory data analysis

During the EDA, we analyzed the general statistics of the *total_ds* dataset for each variable using the *Statistics* node in KNIME.
To gain an overall understanding of the characteristics of Italian athletes, we examined the distribution of age and height for each role. The following boxplots illustrate this distribution.

The chart below shows the distribution of players' ages by role. There don't appear to be significant differences between the roles, as the age range is similar across most of them, except for the *Spiker* role, which seems to have a wider range.

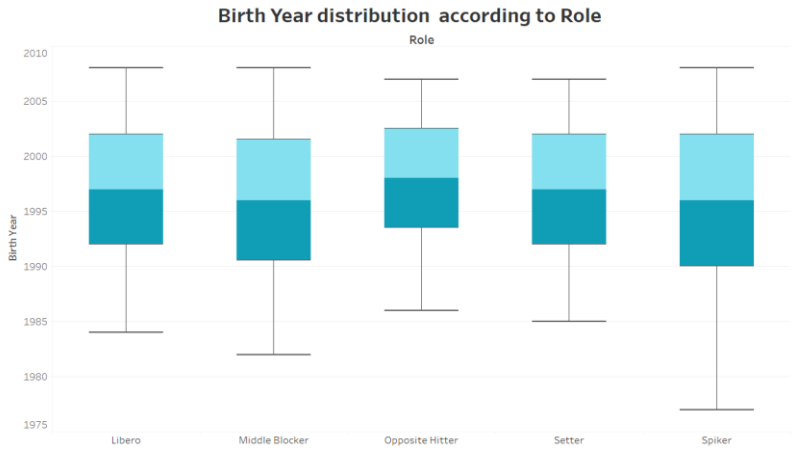

Figure 5. Birth year distribution according to role.

The following chart displays the distribution of players' heights by role. It is evident that *Middle Blockers* and *Opposite Hitters* have a wider height range compared to the other roles, suggesting that height can be more crucial and relevant for these positions. On the other hand, the *Libero* role shows a lower distribution, highlighting that height is less important for this position.

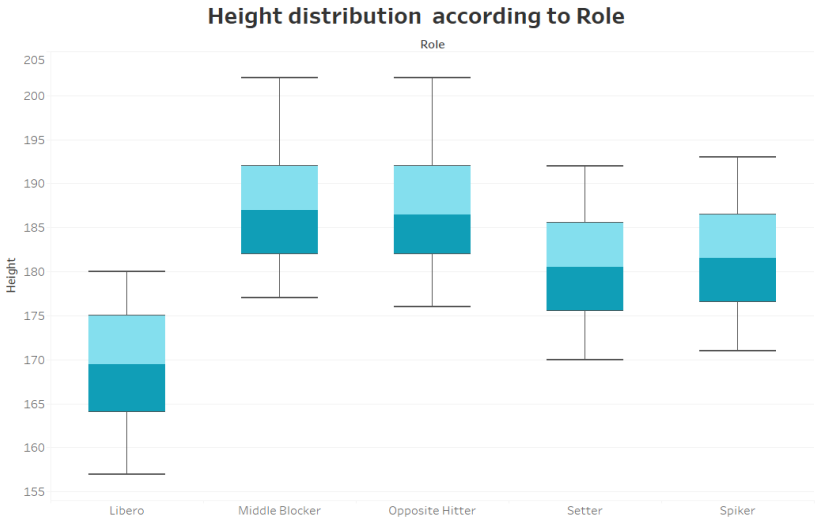

Figure 6. Height distribution according to role.

The chart below shows the distribution of roles by season in the *total_ds* dataset. It is evident that the number of players has decreased from one season to the next, with particular attention to the *Setter* role, which has seen a more significant decline compared to the others.

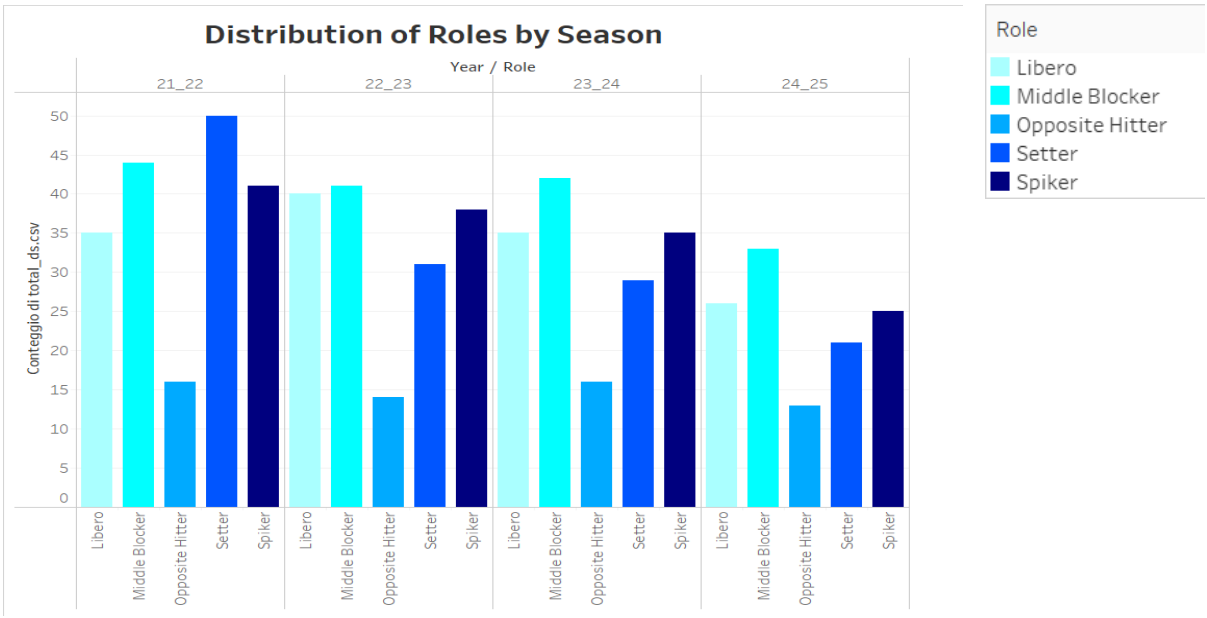

Figure 7. Role distribution by season.

10. Prediction

Using KNIME, we decided to cluster the athletes' performance data across various seasons to identify the players with the best performances and evaluate their potential selection for the 2025 World Championships.

10. 1 Data collection and preparation

- Data Connection: We integrated KNIME with MongoDB using the MongoDB Connector node, enabling us to retrieve and process the necessary performance data.
- Data Transformation: Initially in JSON format, the data was converted into a tabular structure. Non-numerical values were then transformed using the "String to Number" function to facilitate analysis.
- Data Normalization: To ensure fair comparisons across different athletes and metrics, we normalized the dataset, scaling values to a standard range.
- Data Segmentation: The dataset was divided into five subsets, each corresponding to a specific player role (e.g., setter, spiker, etc.), allowing for role-specific performance analysis.

10.2 Performance analysis and athlete selection

Initially, the 'total_ds' dataset was partitioned into five subsets, each containing observations pertaining to a unique role. The "SelfOrganizingMap" node, implementing Kohonen's Self Organizing Map algorithm for unsupervised clustering, was applied to each subset. Subsequently, the “Weka Cluster Assigner” node was utilized to assign the imported data to the corresponding clusters based on the cluster model generated previously.

For each subset, only the most relevant variables for each role were employed in the clustering model.

Upon identifying the clusters exhibiting the best performance for each role, the focus was narrowed to the 2022-23 and 2023-24 seasons to pinpoint the players associated with these top performances. Subsequently, the required number of players per role was selected (e.g., two players were selected for the libero role). In instances of performance ties, the player exhibiting the strongest performance in the current 2024-25 season was chosen.

To validate the selected players, a check was conducted to ensure their participation in the current season, thereby excluding players who are no longer active.

Finally, a comparison was made between the selected players and those who participated in recent European Championships, World Championships, Olympics, and the last two VNLs. This comparison provides insights into the efficacy of our selection process and identifies which and how many players could be classified as "emerging talents."

By following this rigorous process, we aimed to identify the best-performing athletes for the 2025 World Championships, ensuring a fair and data-driven selection process.

To form the team, we selected:

- 2 liberos
- 2 setters
- 2 opposites
- 3 middle blockers
- 4 outside hitters

The players called up for the upcoming World Cup are the following:

Row ID	S Athlete	S Role
Row0	De Gennaro Monica	Libero
Row1	Panetoni Sara	Libero
Row2	Cambi Carlotta	Setter
Row3	Orro Alessia	Setter
Row4	Mingardi Camilla	Opposite Hitter
Row5	Egonu Paola	Opposite Hitter
Row6	D'Odorico Sofia	Spiker
Row7	Perinelli Elena	Spiker
Row8	Sorokaite Indre	Spiker
Row9	Sylla Myriam	Spiker
Row10	Lualdi Giuditta	Middle Blocker
Row11	Gray Anna	Middle Blocker
Row12	Sartori Benedetta Maria	Middle Blocker

10.3 Players selected for the upcoming World Cup

Finally, we compared our selected players with their participation in major recent tournaments, specifically the latest Olympics, World Cup, European Championships, and the last two editions of the VNL. The following observations were made:

- Monica De Gennaro, Alessia Orro, Myriam Sylla, Paola Egonu, and Carlotta Cambi participated in the most recent World Cup.
- Monica De Gennaro, Alessia Orro, Paola Egonu, Carlotta Cambi, and Myriam Sylla participated in the latest Olympics.
- Alessia Orro, Paola Egonu, and Myriam Sylla participated in the most recent European Championships.
- Sara Panetoni, Sofia D'Odorico, and Myriam Sylla took part in the penultimate VNL.
- Monica De Gennaro, Alessia Orro, Myriam Sylla, Paola Egonu, Camilla Mingardi, and Carlotta Cambi participated in the most recent VNL.

In conclusion, among the 13 players selected, 8 have demonstrated their capabilities in previous high-level competitions, while the remaining 5 can be classified as promising emerging talents.

10.4 Why clustering and not classification?

We opted for clustering instead of classification due to a significant amount of missing data in the primary performance dataset. Approximately 25% of the data was missing, leading to several challenges that would have negatively impacted the accuracy and reliability of a classification model:

- Reduction in model accuracy: Classification models assume complete and accurate data. Missing values, particularly when they are widespread or systematically distributed, can cause models to overlook key patterns, resulting in less precise predictions. In cases where missing values are not randomly distributed, they may bias the model, causing it to learn incorrect relationships.
- Potential model distortion: When a large portion of the dataset is missing, especially if it affects key features, the resulting model may not fully reflect the underlying data distribution. This can lead to a distortion of the model's behavior, causing it to make incorrect generalizations and unreliable predictions. The issue becomes more pronounced in supervised learning tasks like classification, where the model learns from labeled examples. Missing data can interfere with the training process, reducing the model's ability to generalize effectively to unseen data.
- Interpretation challenges: A dataset with a high rate of missing values complicates the interpretation of results. Missing data can create ambiguity in the analysis, making it difficult to draw meaningful conclusions or trust the outputs. For example, in classification, the presence of missing values may obscure patterns that would otherwise be clear, leading to a lack of confidence in the predictions. This can pose challenges for decision-making or further analysis, especially in fields where the accuracy and reliability of results are critical.

Given these challenges, clustering was chosen as a more suitable approach. Unlike classification, which requires labeled data and can be heavily impacted by missing values, clustering algorithms are typically more robust to incomplete datasets. They can identify inherent groupings within the data without relying on complete data points, making them better equipped to handle missing values. Therefore, clustering offered a more practical solution in this scenario, enabling meaningful insights to be drawn from the available data while minimizing the impact of missing information.

10.5 Why not a Graph-Based model?

Instead of using a graph-based model, we opted for clustering techniques for several reasons. Firstly, clustering is simpler to implement and more adaptable to our dataset, which contains diverse and sometimes incomplete data. Clustering allowed us to automatically identify meaningful groups without requiring pre-defined labels, making it a more flexible approach given our constraints.
Graph-based models, on the other hand, tend to be computationally intensive and require complex algorithms to analyze relationships between data points. Given our dataset and project scope, this level of complexity was unnecessary. Additionally, clustering models are easier to interpret, particularly when evaluating performance metrics and making data-driven decisions.
By leveraging clustering techniques, we achieved a more efficient and effective analysis of athlete performance, ensuring a fair and transparent selection process for the 2025 World Championships. This approach not only enhances our ability to identify top-performing athletes but also streamlines decision-making, ultimately contributing to Italy’s competitive strength on the global stage.

11. Conclusion and future development

The rapid evolution of technology has paved the way for a more data-driven approach to evaluating and selecting athletes. While traditional scouting methods remain important, they often rely heavily on a coach's subjective judgment. A more effective solution lies in blending both approaches: combining the objectivity of data analysis with the insight of human expertise. By starting with a selection process rooted in statistical analysis, we gain objective insights into player efficiency and performance. Once a pool of candidates is identified, coaches and evaluation teams can apply human-driven criteria—such as experience, effort, and consistency—to finalize the selection.

Leveraging performance data from official league records and match statistics, this project provides a comprehensive and unbiased assessment of players, ensuring a more accurate and equitable selection process for major international tournaments. This approach enhances transparency, helps uncover emerging talent, and optimizes team composition—ultimately strengthening Italy's standing in global competitions.

The final team we have selected consists of: De Gennaro Monica (libero), Panetoni Sara (libero), Cambi Carlotta (setter), Orro Alessia (setter), Mingardi Camilla (opposite hitter), Egonu Paola (opposite hitter), D'Odorico Sofia (spiker), Perinelli Elena (spiker), Sorokaite Indre (spiker), Sylla Myriam (spiker), Lualdi Giuditta (middle blocker), Gray Anna (middle blocker), and Sartori Benedetta Maria (middle blocker), and we are confident it closely aligns with the actual selection.

Looking ahead, there are several avenues for further refining and enhancing our approach. One key improvement would be integrating an updated dataset that includes information on injuries and recovering players. This would allow us to account for temporary performance fluctuations and assess long-term availability. Additionally, incorporating video analysis could offer deeper insights into player performance, capturing nuances beyond traditional statistics.

Another valuable enhancement would be developing an intuitive user interface for coaches, enabling them to easily track player statistics and trends. If paired with a real-time system, this could enable immediate performance assessments, empowering coaches to make data-driven decisions on the spot.

By continuously refining our methodology and embracing new technologies, this project has the potential to transform how volleyball talent is evaluated, ensuring Italy remains at the forefront of global competition.

In conclusion, by combining the precision of data analysis with the expertise of coaches, we create a more reliable, fair, and transparent player selection process. This not only enhances team performance but also lays the foundation for sustained success on the international stage, keeping Italy competitive at the highest levels.