# Navigating Ethical Waters: Establishing a Framework for Ethical Data Usage by NAC

## And the road to a perfect model

DISCOVER YOUR WORLD

1st Year Block B/ADSAI

Friday 26 January 2024

Breda University OF APPLIED SCIENCES

# Index

## Table of Contents


**Index 1**

| | | |
|---|---|---|
| **1** | **Introduction** | **3** |
| 1.1 | Discussion | 3 |
| 1.2 | The ethical standards | 3 |
| | **1.2.1. Decision making** | **3** |
| | **1.2.2. Ethical approaches** | **3** |
| 1.3 | The model approach | 3 |
| | **1.3.1. Correlation** | **3** |
| | **1.3.2. Underlying potential** | **3** |
| | **1.3.3. Better decisions** | **3** |
| **2** | **Exploratory Data Analysis** | **4** |
| 2.1 | Dataset overview | 4 |
| | **2.1.1. Size and scope** | **4** |
| | **2.1.2. Source and collection** | **4** |
| | **2.1.3. Feature types** | **4** |
| 2.2 | Cleaning before analysis | 4 |
| | **2.2.1. Handling missing values** | **4** |
| | **2.2.2. Working with outliers** | **4** |
| | **2.2.3. Data transformation** | **4** |
| 2.3 | Statistics of the data | 5 |
| 2.4 | Visuals | 5 |
| 2.5 | Relationships | 6 |
| 2.6 | Key findings | 6 |
| **3** | **Machine Learning** | **7** |
| 3.1 | Method | 7 |
| | **3.1.1. The ML algorithm used** | **7** |
| | **3.1.2. The rationale behind it** | **7** |
| | **3.1.3. The first accuracy** | **7** |
| 3.2 | Model Evaluation | 7 |
| | **3.3.1. Metrics used** | **7** |
| | **3.3.2. Handling** | **7** |
| 3.3 | Model Improvement | 8 |
| | **3.3.1. Key hyperparameters** | **8** |
| | **3.3.2. Hyperparameter optimization** | **8** |
| | **3.3.3. Challenges encountered** | **8** |
| | **3.3.4. Performance after Tuning** | **8** |
| **4** | **Ethical Considerations** | **9** |
| 4.1 | The three elements | 9 |
| | **4.1.1 Ethical company** | **9** |
| | **4.1.2 Ethical process & tools** | **9** |
| | **4.1.3 Ethical people (employees and clients)** | **9** |
| 4.2 | The responsible parties | 9 |

| | | |
|---|---|---|
| 4.3 | The findings | 10 |
| 4.4 | Proof in terms of GDPR | 10 |
| 4.5 | Identified ethical problems. | 10 |
| 4.6 | Possible improvement | 10 |
| **5** | **Recommendations** | **11** |
| 5.1 | Problem solving | 11 |
| 5.2 | The next adaptations | 11 |
| 5.3 | Final word | 11 |
| **6** | **References** | **12** |
| 6.1 | Programs used. | 12 |
| 6.2 | Scholarly source | 12 |
| 6.3 | Websites | 12 |

# 1 Introduction

## 1.1 Discussion

What ethical standards should NAC adhere to in order to be considered as an ethically responsible user of the dashboard?

## 1.2 The ethical standards

### 1.2.1. Decision making

In the data-driven landscape of professional sports, ethical considerations are paramount for organizations like NAC. This analysis explores and dissects the ethical standards guiding NAC in deploying its dashboard. Employing the five phases of critical thinking, it aims to develop a nuanced understanding of the implications tied to data usage decisions.

### 1.2.2. Ethical approaches

Beginning with identifying and scrutinizing dashboard information, the analysis progresses to evaluating data relevance and reliability. Ethical dimensions, including the impact on stakeholders and considerations of data privacy, security, and transparency, are crucially examined. The final phase emphasizes continuous reflection to refine ethical practices in the evolving data utilization landscape. This concise exploration equips NAC with a robust framework for navigating ethical complexities in data-driven decision-making.

## 1.3 The model approach

### 1.3.1. Correlation

In the pursuit of enhancing player selection strategies, my model aims to delve into the intricate dynamics of predicting optimal acquisitions based on a player's goal-scoring performance and preferred foot. By analyzing the correlation between a player's goal-scoring prowess and the dominant foot they employ, I aim to paint a comprehensive picture of which players excel in terms of their primary foot usage.

### 1.3.2. Underlying potential

The significance lies in not only identifying top performers but also uncovering potential opportunities for improvement. The model will explore the possibility that some players might unlock their full potential by honing their skills with their non-dominant foot. This raises intriguing questions about the malleability of player performance—could there be hidden gems among those who, with training, could significantly enhance their goal-scoring capabilities using their less utilized foot, or perhaps even both feet?

### 1.3.3. Better decisions

The exploration of such nuances not only contributes to a more nuanced understanding of player dynamics but also holds the potential to revolutionize player development strategies. The insights gained from this analysis could inform decisions on player transfers, training regimens, and ultimately lead to a more refined approach to team composition. Through this model, the objective is to go beyond merely identifying top performers and delve into the realm of unlocking untapped potential, shaping a new paradigm in the realm of football analytics and player development.

# 2 Exploratory Data Analysis

## 2.1 Dataset overview

### 2.1.1. Size and scope

The dataset comprises 16,535 rows and 144 columns, encompassing a total of 1,624,263 numerical variables and 139,989 categorical variables. There are no duplicate rows present in the dataset. And prior to cleaning, the dataset contains a total of 120,738 missing values.

### 2.1.2. Source and collection

The dataset is a compilation of 35 different CSV files, merged using the Pandas functions 'pd.append' and 'pd.concat'.

### 2.1.3. Feature types

The dataset encompasses a total of 3 columns with incorrect data types. Among the columns, 9 are classified as "objects" (categorical), 16 as "ints" (integer), and the remaining are labelled as "floats" (numerical).

## 2.2 Cleaning before analysis

### 2.2.1. Handling missing values

The way I handled missing values was like this:

- The column team had missing values, these were replaced by the team they had played in. And the other way around.
- The column position / foot / passport country / Birth country got its missing values turned into "unknown".
- The columns that end with "per 90" were turned into 0.0 because these are float columns.
- The columns weight / height / age got their missing values, turned into the mean of all the values with the NaN values excluded.
- The missing values from the column contract expires got turned into "1900-01-01".
- There were also rows with almost every value missing, these were a total of 232. I removed these rows.

### 2.2.2. Working with outliers

There were almost no outliers, only the 0 values that we replaced instead of all the missing values. So, I kept it this way, and in the end, it worked out fine and I had no problems with it.

### 2.2.3. Data transformation

There were 3 columns that were not the correct data type, I changed them with the following methods:

- I turned the column age from "float" to "int" with the ".astype" method.
- The column on loan was automatically turned into "bool" (Boolean) after changing the words in the column from "yes/no" to "True/False" with the ".replace" method.
- The column contract expires was a bit harder, but in the end, we found a method. We turned it into a "datetime" type which was also useful when using it later. We used the method "pd.to_datetime".

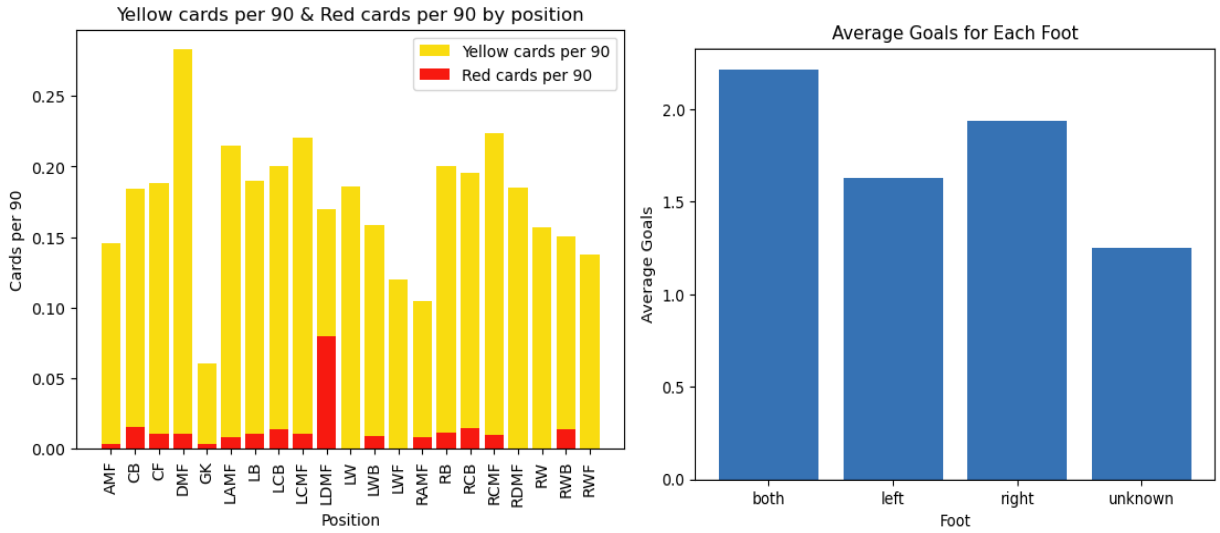
## 2.3 Statistics of the data

When searching for any notable patterns or frequency counts, I applied the method 'value_counts()' per column to determine if there were any popular values. This was not the case. Only the height and weight exhibited a high spike in the middle. This is, of course, because of the cleaning process where I replaced NaN values with the mean.

One thing to note is that there were 0 values in the market value, which I did not expect. This must mean that there are young players who have no value yet. The market value was an interesting column because the values were all over the place. There was only 1 player with a market value of 60,000,000, and the one below that was 42,000,000. Quite a difference.

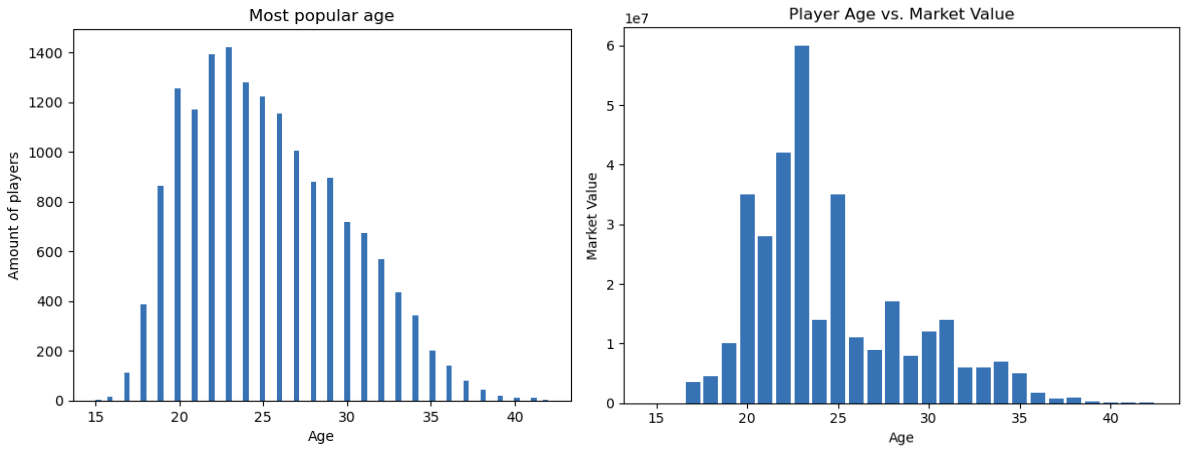
## 2.4 Visuals

In my days of exploration, I had to create a lot of different visuals. I made use of a bar chart, histogram, 3D scatter plot, stacked bar chart, and scatter plot (with a trend line). These were all very useful to gain knowledge about the connection between certain columns in the data, mostly regarding which foot is better to score with and which positions get the most red or yellow cards per 90 minutes.



## 2.5 Relationships

As you can see there are in this dataset indeed correlations:




These graphs are “left-skewed” and have almost the same heights. This tells us that market value does correlate with age. But there are also a lot more players between the ages of 18 and 30.

## 2.6 Key findings

- The market values exhibited considerable variability, with instances like a player valued at 60,000,000 and the next at 42,000,000, suggesting diverse valuations within the dataset.
- The presence of players with a market value of 0 suggests that young players, possibly newcomers or those with limited playing time, may not yet have an assigned market value. This hypothesis could impact how market values are determined, especially for emerging talents.
- The observed variability in market values hints at potential non-linear relationships. Considering this, more advanced machine learning models, beyond traditional linear regression, may be explored for predicting market values.

# 3 Machine Learning

## 3.1 Method

### 3.1.1. The ML algorithm used

I employed a random forest classifier model, a type of ensemble learning algorithm based on decision trees. Random forests are renowned for their efficacy in classification tasks, leveraging the aggregation of predictions from multiple decision trees. In this approach, each decision tree is constructed using a subset of the training data and a random subset of features, contributing to the model's robustness and ability to generalize well to unseen data.

### 3.1.2. The rationale behind it

The rationale behind choosing the random forest model was rooted in its outstanding performance during testing, exhibited the highest accuracy among the various models I evaluated. This ensemble learning method not only demonstrated superior predictive power but also showcased resilience to noise and outliers in the dataset. Moreover, the model strikes an optimal balance between simplicity and complexity, making it suitable for the characteristics of the given task.

### 3.1.3. The first accuracy

The accuracy score of 76.85% on the test set underscores the effectiveness of the random forest classifier in providing reliable predictions for the classification task at hand.

## 3.2 Model Evaluation

### 3.3.1. Metrics used

When assessing the performance of the model, I employed a comprehensive set of evaluation metrics, with a focus on RMSE (Root Mean Squared Error) and accuracy score. RMSE served as a valuable metric to quantify the magnitude of prediction errors, particularly beneficial in classification tasks, providing a clear measure of the model's predictive accuracy. Simultaneously, accuracy score was leveraged to gauge the model's overall correctness in classifying instances, offering insights into its classification performance.

### 3.3.2. Handling

To enhance the robustness of the evaluation process, I explored the impact of handling unknown values. By removing these unknown values from the dataset, I observed an improvement in the model's performance. This step was particularly valuable in enhancing the model's ability to make accurate predictions, underscoring the importance of data preprocessing in refining the model's effectiveness.

## 3.3 Model Improvement

### 3.3.1. Key hyperparameters

The key hyperparameters focused on during tuning were ‘max_features’ and ‘n_estimators’. ‘max_features’ represents the maximum number of features considered for splitting a node, and’ n_estimators’ is the number of decision trees in the random forest.

### 3.3.2. Hyperparameter optimization

The optimization process involved using Grid Search, a systematic approach that explores a predefined grid of hyperparameter values. In this case, ‘max_features’ was varied between 1 and 5, and’ n_estimators’ between 8 and 20 with a step size of 8. Grid Search exhaustively evaluates all possible combinations within this parameter grid.

### 3.3.3. Challenges encountered

One challenge encountered during tuning was a data type issue. The model didn't run until y_test was converted to float. To address this, a data type conversion was performed (y_test = y_test.astype(float)), ensuring compatibility with the model's requirements. This step resolved the issue and allowed the tuning process to proceed smoothly.

### 3.3.4. Performance after Tuning

The hyperparameter adjustments, as identified by the Grid Search, resulted in the selection of the best parameters as {'max_features': 4, 'n_estimators': 16}. This combination demonstrated a significant impact on model performance, with the model achieving a score of 83.13%. This represents a notable improvement from the initial accuracy score of 76.45%, indicating that the tuned hyperparameters enhanced the model's ability to generalize and make accurate predictions.

# 4 Ethical Considerations

## 4.1 The three elements

### 4.1.1 Ethical company

In the context of the project, NAC Breda can be considered an ethical organization if it incorporates policies such as GDPR compliance and demonstrates ethical behaviour towards its employees and external stakeholders.
NAC Breda should ensure that its data management practices align with GDPR regulations to safeguard individual privacy and data rights. By adhering to GDPR standards, NAC Breda demonstrates a commitment to ethical data handling.
Additionally, fostering a diverse team, promoting transparency, and actively soliciting and incorporating feedback from employees and external stakeholders contribute to the ethical culture of the organization. A diverse team ensures different perspectives are considered, transparency builds trust, and feedback mechanisms allow for continuous improvement in ethical practices.

### 4.1.2 Ethical process & tools

The project's ethical process can be evaluated based on how ethics are incorporated into the development processes of data analysis and machine learning models.
Ensuring that ethical considerations, such as fairness, transparency, and respect for laws, are integrated into the entire process of data analysis and model development is crucial.
For example, during data cleaning and preprocessing, ensuring that handling of missing values, outliers, and data transformations align with ethical standards and legal requirements demonstrates ethical process implementation.
Similarly, using machine learning algorithms and tools that prioritize fairness, transparency, and accountability contributes to ethical process and tool adoption in the project.

### 4.1.3 Ethical people (employees and clients)

Ethical behaviour of professionals involved in the project, including employees and clients, is essential for maintaining ethical organizational capacity.
Employees should demonstrate ethical behaviour towards customers, stakeholders, society, and the environment throughout the project lifecycle.
This includes respecting individual privacy rights, ensuring data security, and using data responsibly and transparently.
Clients, such as stakeholders within NAC Breda or external parties, should also act ethically by adhering to ethical guidelines, respecting data privacy, and promoting ethical data usage within their organizations.
Additionally, raising awareness about ethics, morals, and responsible professional behaviour among employees and clients is crucial for fostering a culture of ethical conduct within the organization and its ecosystem.

## 4.2 The responsible parties

- Fairness: The responsibility for ensuring fairness in data handling and decision-making lies with the data management and analytics teams, along with input from coaching and management staff.
- Transparency: Stakeholders such as executives, management, and communication teams play a pivotal role in ensuring transparency in decision-making processes, financial decisions, and overall management strategies.
- Respect for Laws: Legal and compliance teams within NAC bear the responsibility for ensuring adherence to laws and regulations governing data usage, contracts, and overall operations.

## 4.3 The findings

NAC is diligent in considering ethical elements in its operations. Fairness is evident in unbiased data handling, transparency is maintained in decision-making processes, and respect for laws is a commitment embedded in the organization's ethos. The findings reveal a comprehensive approach, integrating ethical considerations into the core of NAC's operations.

## 4.4 Proof in terms of GDPR

The critical thinking process, involving the five phases, was applied to evaluate ethical standards for NAC's use of the dashboard. This framework includes identifying information, analyzing data relevance and reliability, assessing the impact of decisions, exploring ethical considerations, and continuous reflection. Additionally, considerations of GDPR and Ethical Guidelines for Statistical Practice are woven into the analysis, ensuring a comprehensive and ethically sound decision-making process.

## 4.5 Identified ethical problems.

While NAC exhibits a commitment to ethical standards, potential challenges and dilemmas in balancing these standards are recognized. The dynamic and competitive environment of professional sports poses ethical considerations that need careful navigation. Ensuring fairness and transparency without compromising the competitive edge presents an ongoing challenge.

## 4.6 Possible improvement

Recommendations for NAC involve the establishment of clear policies and procedures aligned with ethical standards. Active promotion of a culture prioritizing fairness, transparency, legal compliance, honesty, data privacy, and ownership is essential. Regular training programs and effective communication strategies should be deployed to fortify these ethical standards throughout the organization, fostering a culture that values integrity, respects individual rights, and embraces ethical considerations in the handling of player data.

# 5 Recommendations

## 5.1 Problem solving

As questioned in the discussion: What ethical standards should NAC adhere to in order to be considered as an ethically responsible user of the dashboard?

After discussing with my fellow peers, we ended up in a great answer. Which is:

“Data should be gathered by mutual agreement within boundaries Supporters of NAC Breda should aim for the same ethical standards. NAC Breda should adhere to GDPR and terms and conditions. Implement an ethical, diversity, transparency and accountability.”

## 5.2 The next adaptations

- NAC Breda should ensure that data gathering is conducted with the mutual agreement of all relevant stakeholders, including players, staff, and supporters. Establishing clear boundaries and obtaining consent for data collection will uphold ethical standards and respect individual rights.

- Encourage a collective commitment to ethical standards among the supporters of NAC Breda. Promote awareness and understanding of the ethical considerations related to data usage, fostering a community that values and upholds ethical principles.

- Strict adherence to the General Data Protection Regulation (GDPR) and other relevant terms and conditions is imperative. NAC Breda should ensure that all data collection, storage, and usage align with legal frameworks to safeguard individual privacy and comply with international data protection standards.

- Implement a robust framework encompassing ethical practices, diversity, transparency, and accountability in all aspects of data management. This includes transparent decision-making processes, diversity considerations, and mechanisms for accountability to ensure responsible data use.

- Establish ongoing training programs for players, staff, and relevant stakeholders to enhance awareness of ethical considerations in data usage. Regular communication on the importance of ethical standards and updates on compliance measures will reinforce a culture of responsibility within the organization.

- Integrate diversity and inclusion considerations into the ethical framework. Ensure that data collection and decision-making processes are sensitive to diverse perspectives, avoiding biases and promoting inclusivity within the organization.

- Foster a culture of continuous evaluation and improvement in ethical practices. Regularly review policies, procedures, and the impact of data usage on stakeholders. Seek feedback from supporters and stakeholders to identify areas for enhancement and ensure that ethical standards evolve with changing dynamics.

## 5.3 Final word

By adopting these recommendations, NAC Breda can not only meet but surpass ethical standards, establishing itself as a paragon of responsible and conscientious data usage within the realm of professional sports.

# 6 References

## 6.1 Programs used.

OpenAI. ChatGPT. Source text was written by me and then summarized. Prompt: ‘Here I made a report can you do a spelling check and make it a small bit more formal!', (25-01-2024)

## 6.2 Scholarly source

*Browse by Document Type - D-Scholarship@Pitt*. (n.d.). https://d-scholarship.pitt.edu/view/type/

## 6.3 Websites

Twin, A. (2023, 17 March). *Business Ethics: definition, principles, why they're important*. Investopedia. https://www.investopedia.com/terms/b/business-ethics.asp

Games

Leisure & Events

Tourism

Media

Data Science & AI

Hotel

Logistics

Built Environment

Facility


Mgr. Hopmansstraat 2
4817 JS Breda

P.O. Box 3917
4800 DX Breda
The Netherlands

**PHONE**
+31 76 533 22 03
**E-MAIL**
communications@buas.nl
**WEBSITE**
www.BUas.nl


DISCOVER YOUR WORLD

Breda University
OF APPLIED SCIENCES