# Report on Data-driven Solution for NAC

Petar Paskalev
Student number: 232725
Date: 26/01/2024

DISCOVER YOUR WORLD

Academy/Department

Saturday 01 January 2022

Breda University OF APPLIED SCIENCES

# Index

# 1 Introduction

1 Introduction
This project aims to address the challenge of acquiring new players to elevate the performance of the NAC Breda team. The primary focus lies in identifying potential anomalies in player performance and optimizing the selection process by recommending cost-effective players or those with soon-to-expire contracts.

1.1 Objective
The main objective of this analysis is to enhance the capability of the player scouting process, utilizing anomaly detection to identify standout performers, and suggest affordable player options. Using the statistics of 3 other clubs that got promoted in the 2022/2023 season and addressing what NAC Breda can do to close the gap and potentially win a promotion.

1.2 Approach
To achieve the project goals, I have adopted a multifaceted approach that incorporates anomaly detection techniques for player performance assessment. This method ensures a comprehensive evaluation of potential recruits, allowing for a nuanced understanding of their strengths and weaknesses and also by comparing with the clubs that have already won promotion.

# 2 Exploratory Data Analysis

## 2.1 Dataset Overview

The dataset consists of player statistics of potential players for the NAC Breda football team. The team can consider both established players and young talents from lower football divisions. It includes various features such as player demographics, performance metrics, market values, and contract duration in their respective clubs as of the end of 2022. The dataset comprises 45 Excel files from European leagues, compiled into one data frame using Python. Feature types include 105 numerical values related to on-field performance and 9 categorical values, providing information on players' names, birth countries, and the teams they played for.

## 2.2 Data Preparation

Handling Missing Values:

For missing values, I employed a simple function to identify columns with missing data. Subsequently, I removed rows with insufficient player statistics, considering factors such as playing time and no enough data to include them. in total deleted 237 rows of data. Notable null values were found in the "Weight," "Height," and "Market Value" categories.

Data Transformation Techniques:

Normalization and Ridge techniques were applied to certain data aspects. To get a better understanding which features are the most important for Central defenders when they have to make a defensive action. Also I used encoding to determine each players position. After this I used pandas to make separate data frames for each position. Again after all of the separation I distributed into more small data frames containing only the clubs I want to make analysis. Them being Heracles finishing second PEC Zwolle (the champions of the Eerste Divisie in season 2022/2023) and Almere city qualified during playoffs which NAC Breda played in and lost in the 2nd round to versus FC Emmen.

## 2.3 Summary Statistics

I used measure of mean tendency when had to adjust and compare with visuals grouped bar chart labels and to have a better understanding where the main areas are that team of NAC Breda lags behind the competition and dispersion (variance, standard deviation) for numerical features. This is all the data question I found about the summary statistics of the data :
- What is the range of Market Value of the players?

minimal value- 0

maximum value- 60000000

- identify the country of the youngest player in the dataset ?

there are four players that are the same age 4233 Estonia

4294 Estonia

7977 Iceland

8013 Iceland

Name: Birth country, dtype: object From Iceland and Estonia

age - 15.0

- 1 What is the average age of players in the dataset?

Average age of players in the dataset – 25

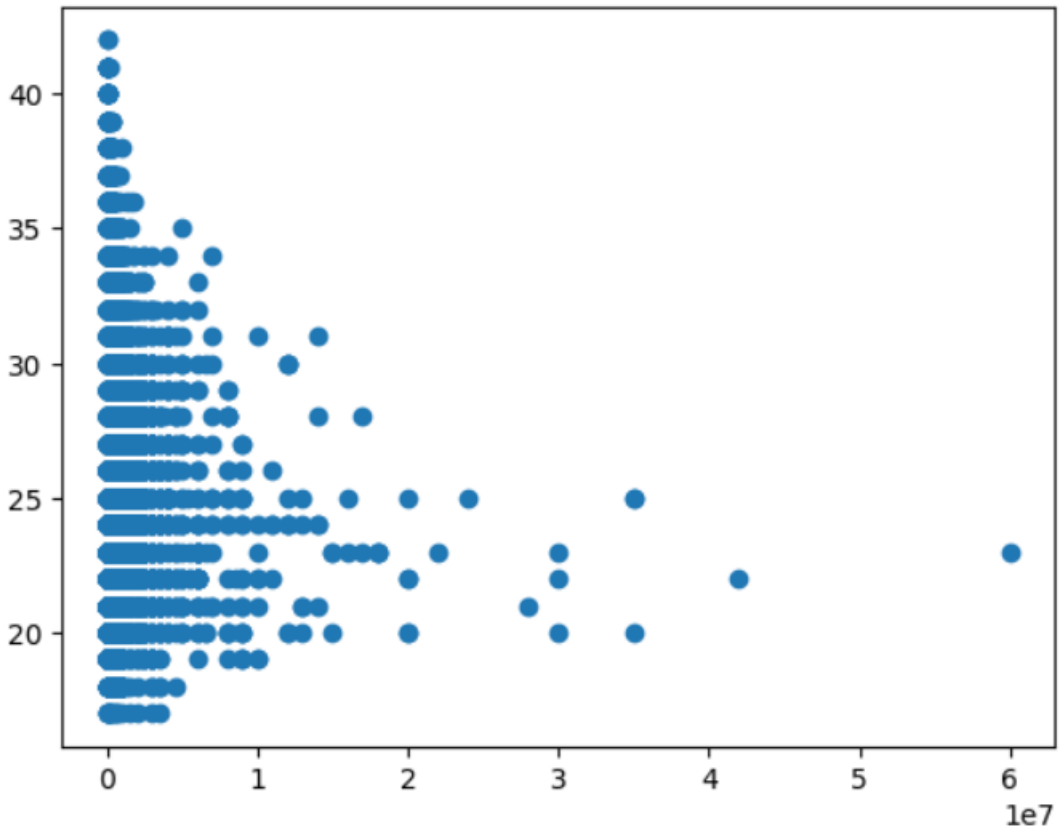
- 2 Which team has the highest market value on average?

The team with the highest average market value is: Ajax

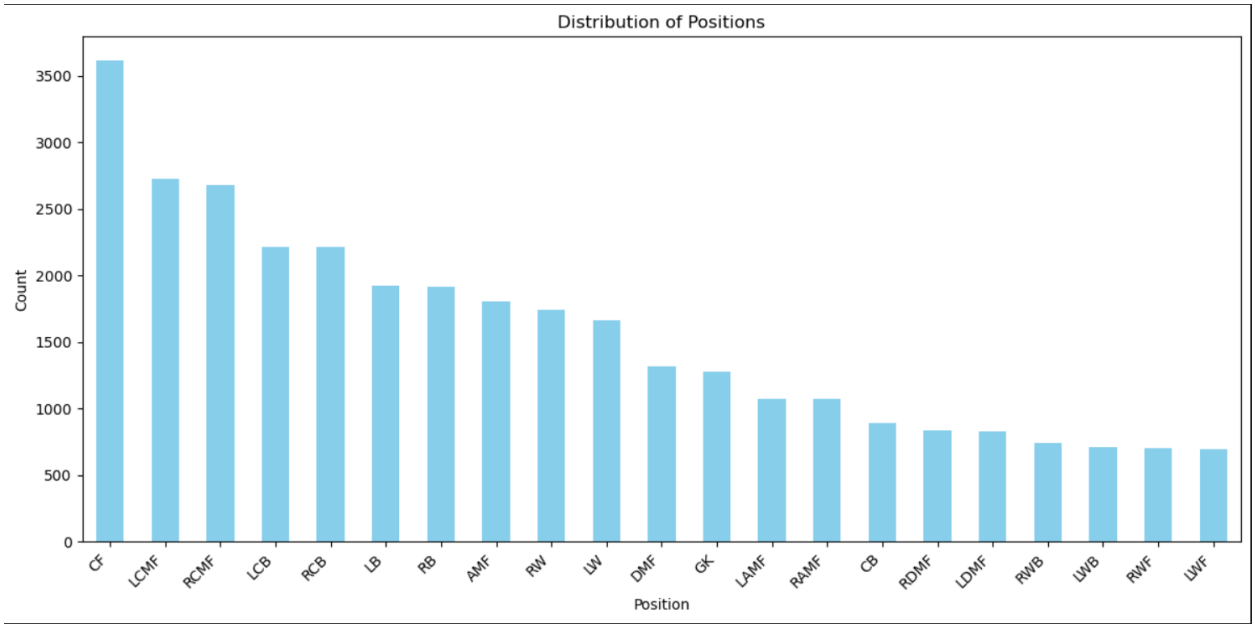
Note: Most of the other colleagues have as an answer Liverpool because they used the column „Team“, but I used 'Team within selected timeframe', because it looks for the palyers taht were gathered during the data filling.

- 3 How does the market value of players correlate with their age?

This is the correlation coefficient -0.0633773301925498 and this is the plot plot how it distributed:



- 4 What is the distribution of players' positions across different teams?




This graph shows all the player distributions in different positions.

- 5 Which country has the highest representation in the dataset in terms of player birthplace?

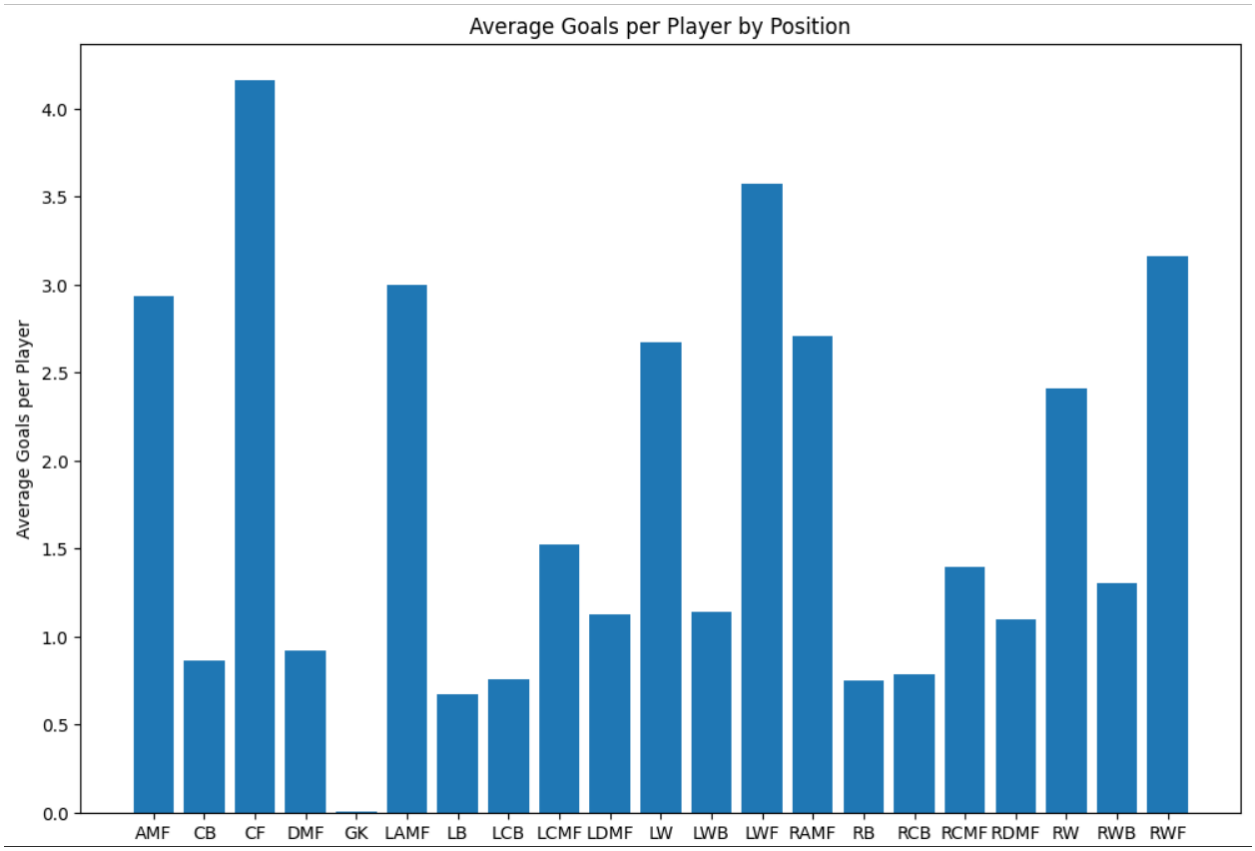
The team with the highest average market value is: Italy

- 6 Is there a correlation between a player's height and weight and the number of goals scored?

The correalation between goals scored and Height -0.06499178577864116

The correalation between goals scored and Weighty -0.06356411494504173

- 7 How does the number of goals per player vary across different positions?




This graph show all the different positions and how much they scored on average

- 8 What is the average number of matches played by players in different age groups?

This is the average number of matches per every age:

Age

15.0 1.500000

16.0 5.375000

17.0 10.174825

18.0 15.103614

19.0 17.162664

20.0 18.589626

21.0 19.861135

22.0 20.621525

23.0 21.077085

24.0 21.772763

25.0 22.290058

26.0 22.045887

27.0 21.862687

28.0 22.475057

29.0 22.172798

30.0 21.815278

31.0 22.273264

32.0 21.971781

33.0 22.230594

34.0 22.558480

35.0 21.804020

36.0 22.626761

37.0 22.790123

38.0 24.214286

39.0 19.809524

40.0 25.692308

41.0 21.600000

42.0 32.000000

- 9 Which players have the highest ‘xG (Expected Goals)' value and how does it compare with actual goals scored?

Player with the highest expected goals compared to how much he actually scored:

Player M. Dugandžić

xG 23.23

Goals 22

- 10 What is the average contract duration left for players in each team?

1860 München 277

ADO Den Haag 326

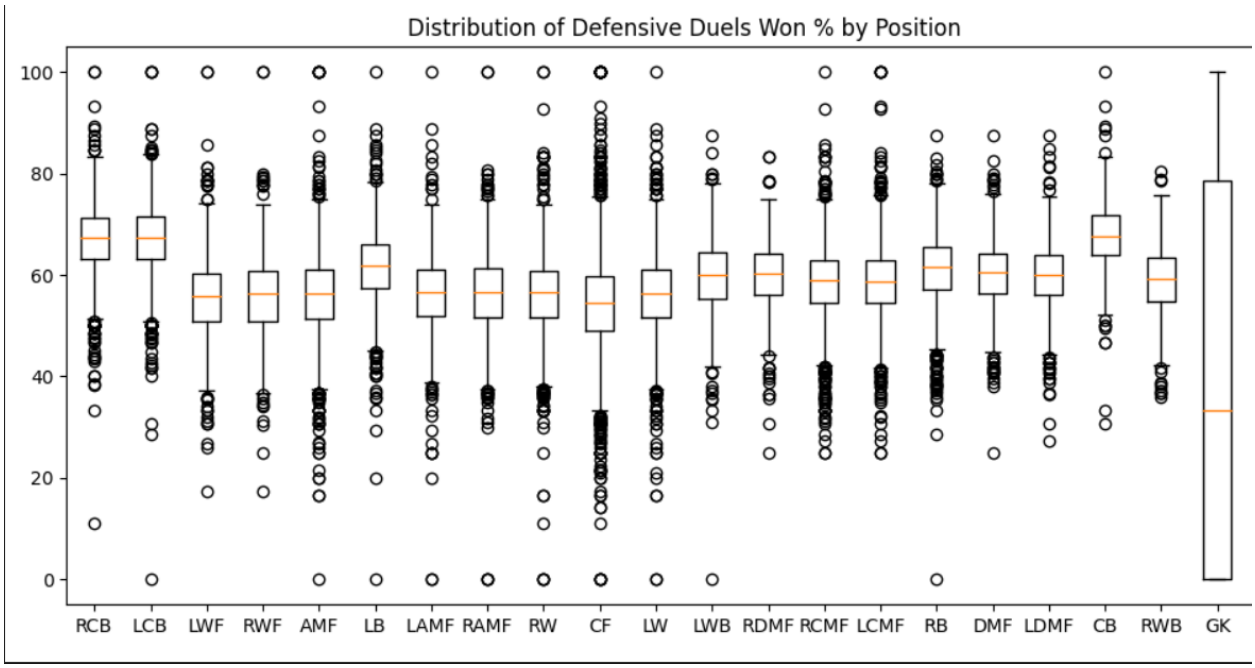
AEK Athens 660

AEK Larnaca 281

AEL 226

Note: This is just 5 of the clubs in alphabetical order.

- 11 How do ‘Duels won %' and ‘Aerial duels won %' vary by position?




I used boxplot to visualize this question.

- 12 Is there a significant difference in ‘Successful defensive actions per 90' between players on loan and permanent players?

T-statistic: -4.364167834953592

P-value: 1.283924879518197e-05

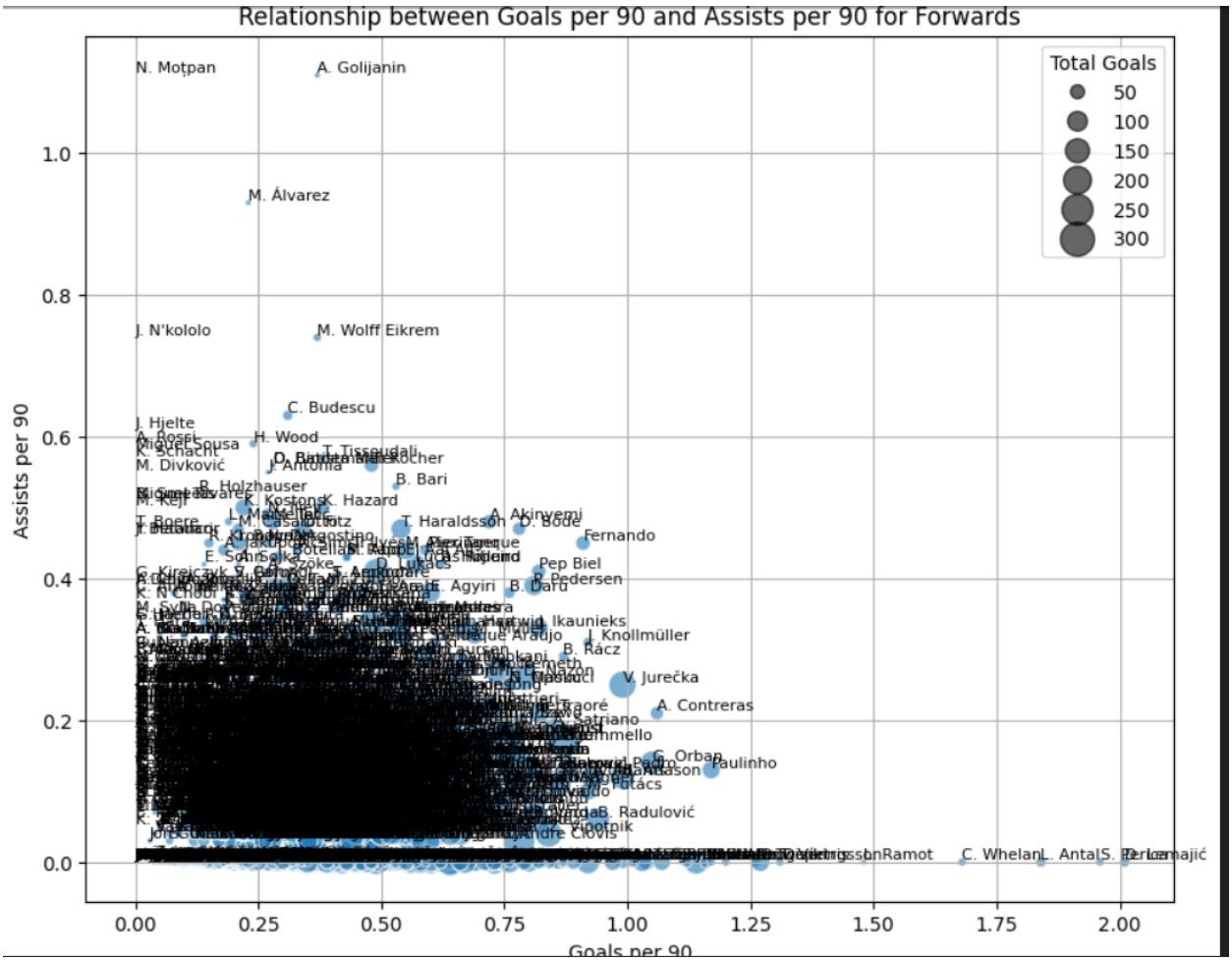
There is a significant difference.

- 13 Which players have the highest ‘Successful attacking actions per 90' and which position do they play?

Player with the highest Successful attacking actions per 90 and his position(s):

Player C. Madueke

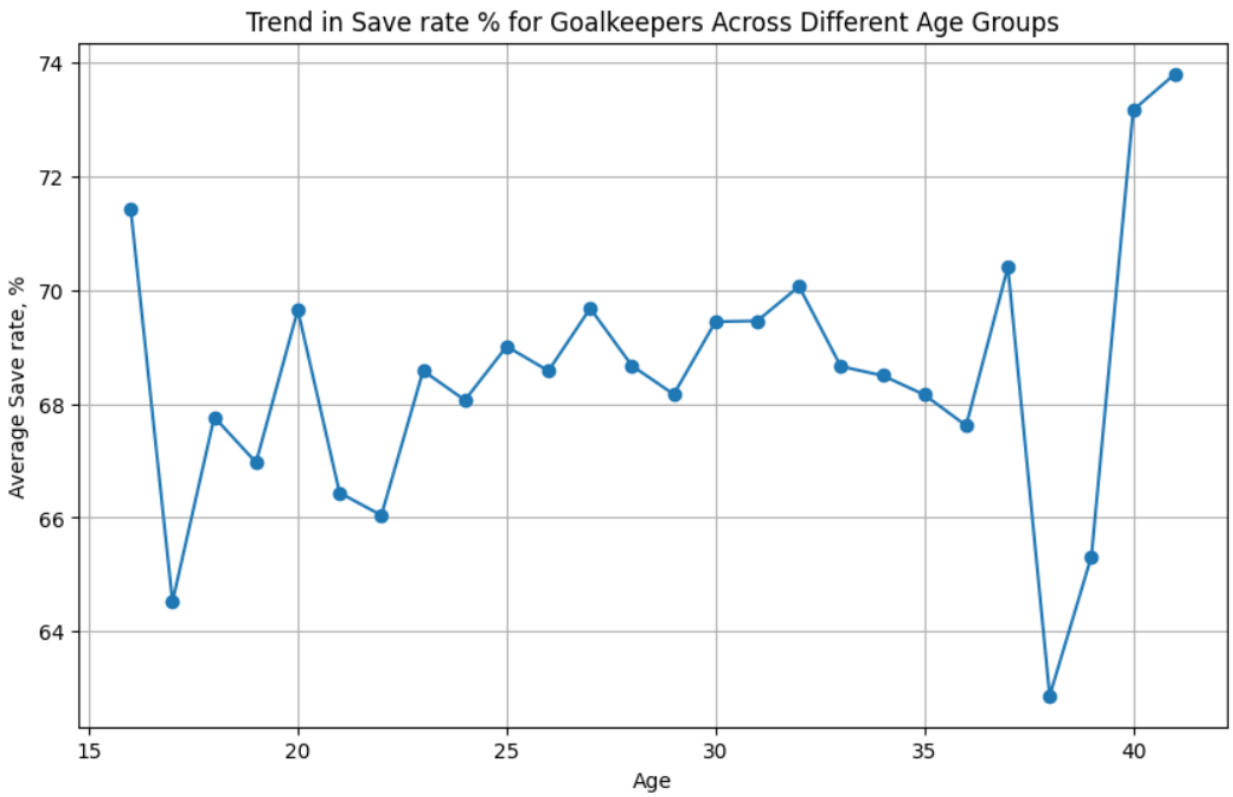
Position RAMF, RWF




- 16 How do ‘Passes per 90' and ‘Accurate passes %' correlate for midfielders?

Correlation between 'Passes per 90' and 'Accurate passes %' for midfielders : 0.641059692480672

- 17 Is there a trend in the ‘Save rate %' for goalkeepers across different age groups?




- 18 What is the distribution of ‘Yellow cards per 90' and ‘Red cards per 90' across different positions?

this is the yellow per 90Position:

AMF 1783

CB 886

CF 3572

DMF 1310

GK 1278

LAMF 1064

LB 1911

LCB 2212

LCMF 2705

LDMF 828

LW 1648

Note: This is not all of the statistic just small number of them.

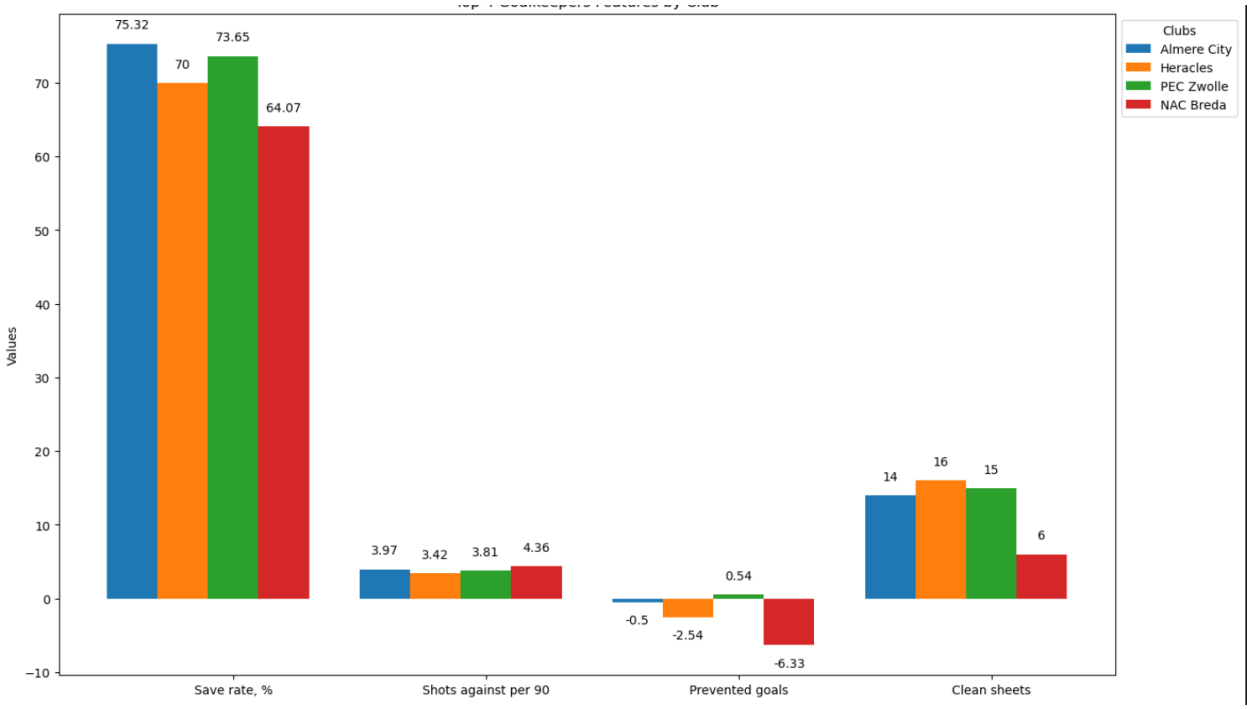
- 20 Which players have the highest ‘Penalty conversion %' and what are their overall shooting statistics?

The player with the highest Penalty conversion is 'Paulinho'.

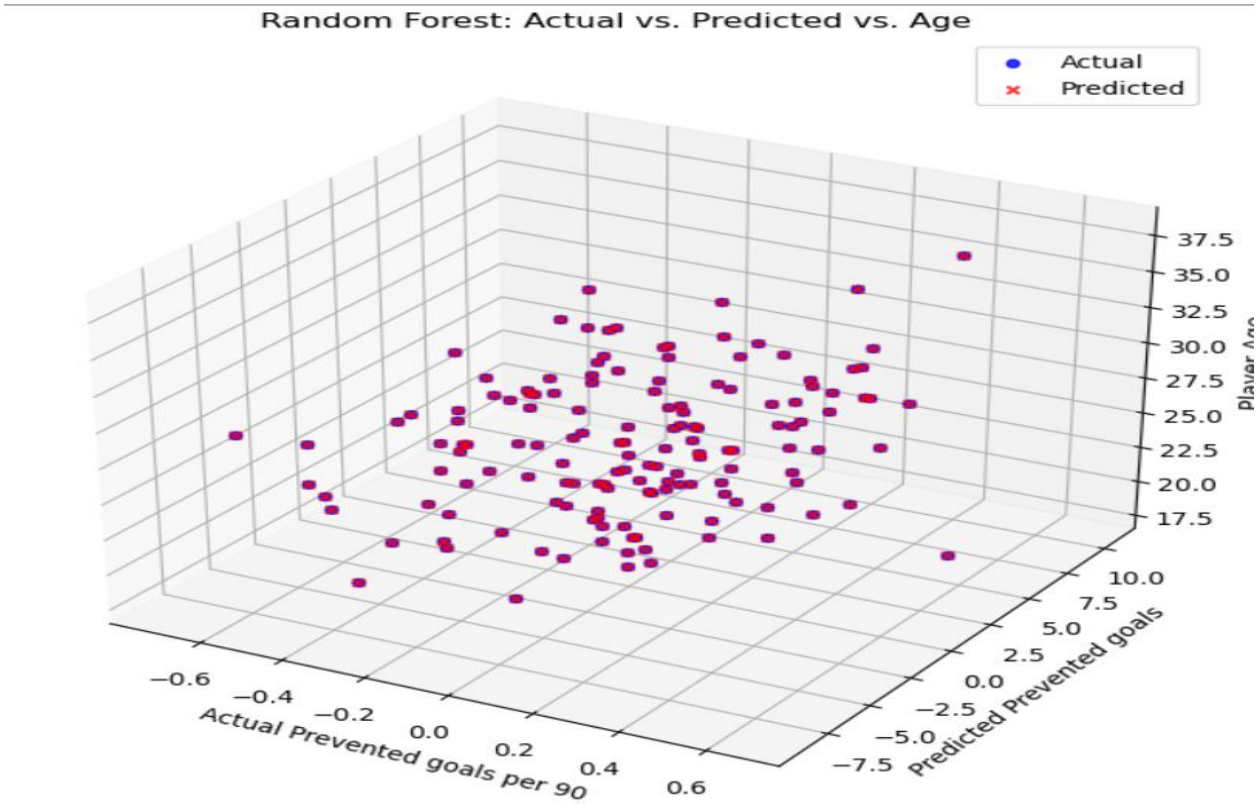
2.4 Visual Techniques

I have used different types of visuals in my models I used a heatmap to correlate for each position which feature for a given position has strong correlation with any other given feature. Like this example

The main visual I used were label grouped bar chart with labels as the example here:



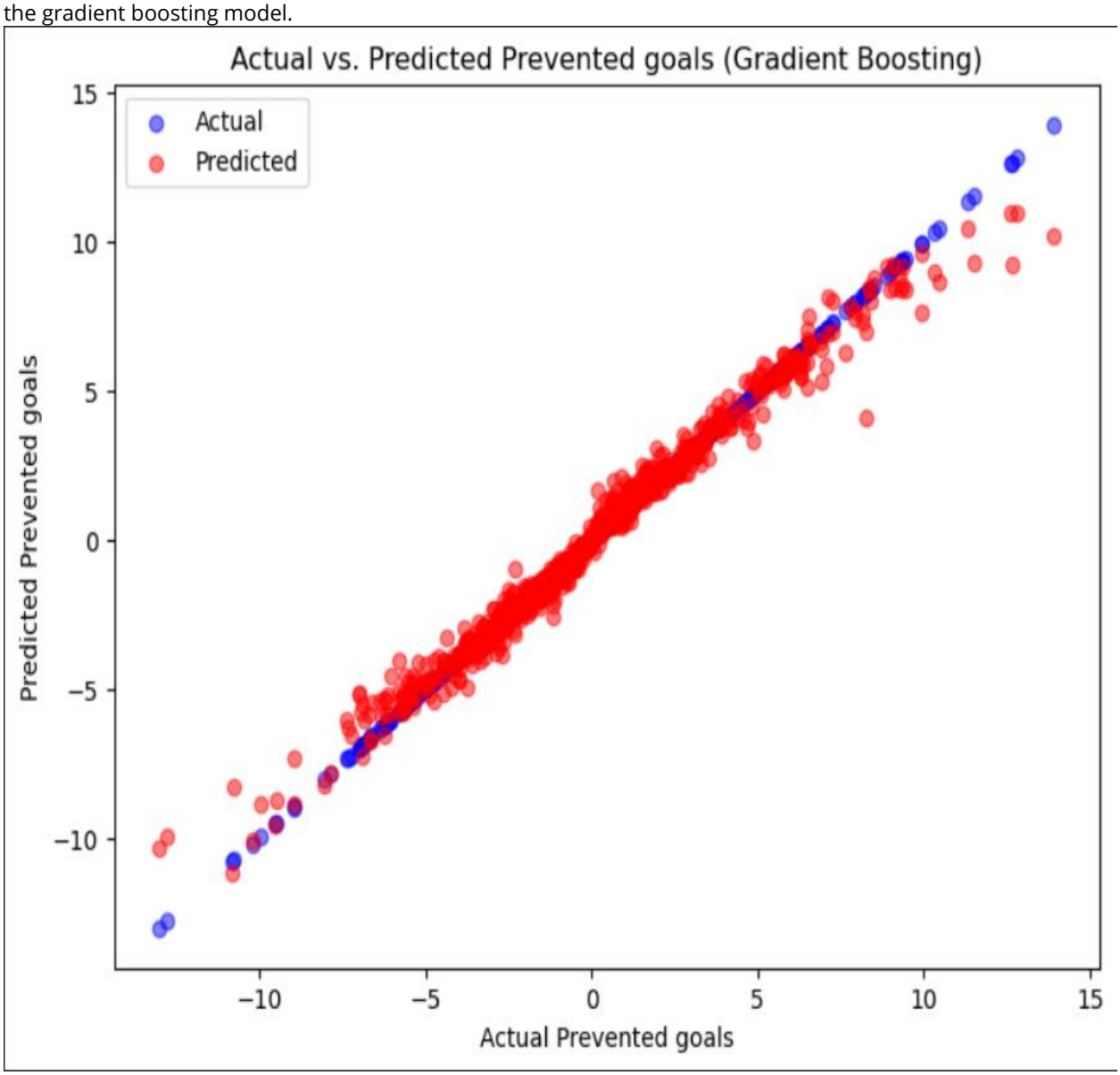

This bar chart really well showed how far behind the competition is NAC Breda in terms of goalkeepers. That is why my main shift focused on recruiting a goalkeeper for NAC. I also used 3d scatter plots to see if the data was overfitting or not for an example this is my Random Forest classification here is my scatter plot:



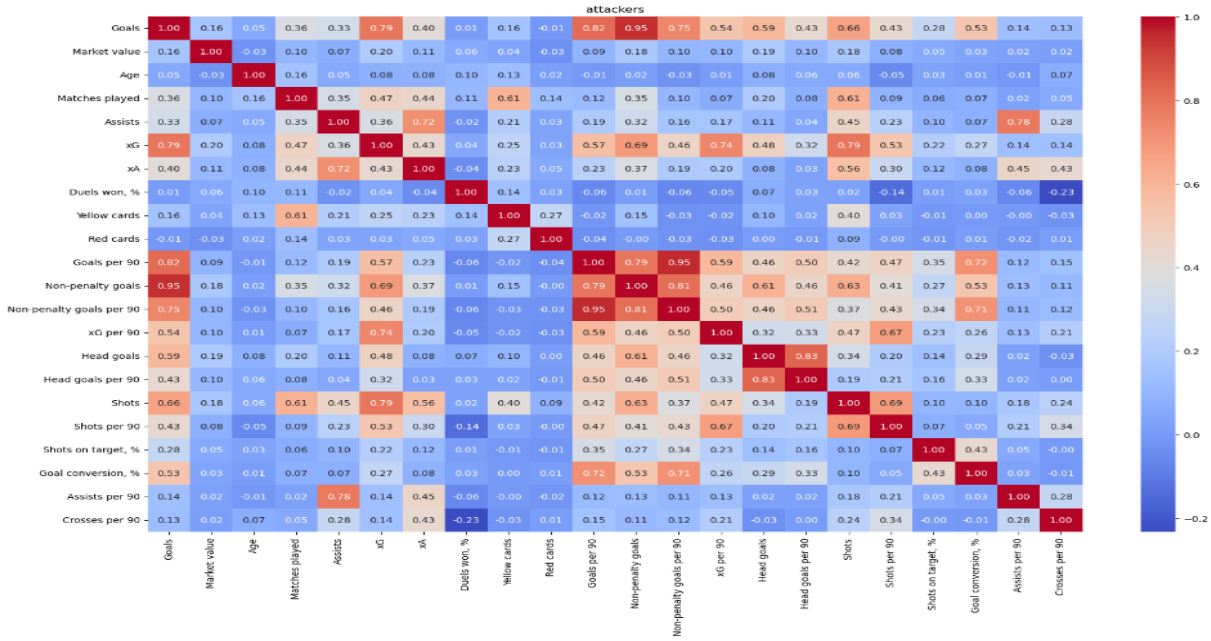

In the example I shared earlier, you can observe a clear issue with data overfitting, undermining the overall value of the analysis. The problem arises when the model excessively tailors itself to the training data, capturing noise or random fluctuations instead of genuine patterns. This overfitting, in turn, hampers the model's ability to perform well on new, unseen data, making the insights derived from the analysis less dependable and, as a result, diminishing their practical value. It emphasizes the need to incorporate techniques like regularization or thoughtful model selection to address overfitting, ensuring a more accurate representation of the true underlying patterns in the dataset and bolstering the credibility of the findings.

I employed a scatter plot specifically for the gradient boosting model, aiming to visually convey that the data exhibited characteristics indicative of a well-generalized model, thus confirming the absence of overfitting. The scatter plot served as a valuable tool to illustrate the model's predictive performance on both training and testing data points. By showcasing consistent and accurate predictions across different subsets, it provided assurance that the model's insights were not overly tailored to the training data, reinforcing the reliability of the findings. This visual verification adds an extra layer of confidence in the robustness and generalization capability of

the gradient boosting model.




2.5 Examining Relationships Between Variables




This heatmap I incorporated into the analysis provided valuable insights into the interconnections among various variables. Notably, it revealed pronounced associations, such as the correlation between Expected Assists (xA) and actual goals, shedding light on the players' playmaking capabilities and their impact on goal-scoring opportunities. Additionally, the heatmap highlighted a noteworthy relationship between Expected Goals (xG) and assists, emphasizing the interconnectedness of these performance metrics.

Another intriguing observation was the discernible correlation between assists and crosses per 90. This suggests that players who actively contribute to goal creation (assists) are also frequently engaged in delivering crosses during matches. The heatmap effectively visualized these patterns, enabling a comprehensive understanding of how different aspects of player performance correlate with one another on the pitch.

2.6 Key Findings

Amidst the comprehensive analysis, a pivotal revelation unfolded, casting a spotlight on NAC Breda's overarching business objective. A meticulous examination of the statistics brought to the forefront a crucial necessity – the immediate requirement for new goalkeepers within the team. This discovery accentuates the imperative nature of fortifying the goalkeeping department to augment the team's overall performance. Additionally, a noteworthy finding materialized concerning the forwards, specifically those occupying the Central Forward (CF) position. Although their goal-scoring figures initially appeared modest, a deeper exploration uncovered a significant overperformance relative to the anticipated data. However, this overachievement was intricately linked to challenges encountered in the team's build-up play.

# 3 Machine Learning

## 3.1 Method

Type of Algorithm: Linear Regression
Rationale: Linear Regression is a fundamental algorithm that assumes a linear relationship between the independent and dependent variables. I opted for this model initially due to its simplicity and interpretability, providing a baseline understanding of how individual features contribute to goalkeeper performance.

Type of Algorithm: Random Forest Classifier
Rationale: Random Forest is an ensemble learning technique that combines multiple decision trees to capture non-linear relationships and complex interactions in the data. I chose this model to handle potential non-linear patterns and to assess the importance of different features.
python

Type of Algorithm: Gradient Boosting Regressor
Rationale: Gradient Boosting is an ensemble technique that builds decision trees sequentially, with each tree correcting the errors of the previous one. This model provides high predictive accuracy and emphasizes feature importance. It emerged as the model of choice for its comprehensive understanding of feature relevance.

## 3.2 Model evaluation

Linear regression

The R-squared value, also known as the coefficient of determination, provides an indication of how well the model explains the variance in the dependent variable. An $R^2$ value of 0.68 indicates that the model explains about 68% of the variability in the target variable. In other words, the model performs reasonably well in capturing the variation in the data.

Mean Absolute Error: 1.5437553442716256

R-squared: 0.6768064864936028

Random Forest Classifier:
the MAE of approximately 0.29 suggests that, on average, the model's predictions deviate by around 0.29 units from the actual values.
A high $R^2$ value of 0.98 indicates that the model explains about 98.32% of the variability in the target variable. This suggests an excellent fit to the data.
With performing cross validation on the model I got a score of 1.5934383999627246. MAE measures the average absolute difference between the predicted and actual values. In your case, a Cross-Validation MAE of 1.5934 suggests that, on average, the model's predictions deviate by approximately 1.5934 units from the true values.
Gradient boosting
The R-squared value, which is close to 1 (0.996 in this case), indicates a very high goodness of fit. It signifies that the Gradient Boosting model explains a substantial portion of the variance in the Prevented goals data. Essentially, the model is highly accurate in capturing the patterns and trends within the data.
In summary, the Gradient Boosting model demonstrates outstanding performance with a very low MAE, indicating minimal prediction errors, and an exceptionally high R-squared value, showcasing the model's ability to explain and predict Prevented goals effectively.

### 3.3 Model improvement

Gradient Boosting Hyperparameters
Gradient Boosting Best Hyperparameters: {'learning_rate': 0.1, 'max_depth': 5, 'n_estimators': 150} Gradient Boosting Cross-Validation Mean Absolute Error: 0.27204819349905934

In summary, the chosen hyperparameters, including a moderate learning rate, limited tree depth, and enough estimators, contribute to the model's effective performance, as evidenced by the low cross-validation MAE. These hyperparameters strike a balance between model complexity and the ability to generalize well to new data.

# 4 Ethical Considerations

In summary, the ethical guidelines for NAC Breda's player recruitment process involve a comprehensive assessment of the organization's ethical capacity, emphasizing corporate governance, integrity, transparency, and social responsibility. The adoption of an ethical decision-making framework, compliance with GDPR regulations, and adherence to ethical statistical practices are crucial elements in ensuring ethical standards throughout the project. Identified ethical concerns, such as privacy issues in handling player statistics and potential biases in recruitment, were addressed with measures to safeguard sensitive information and mitigate biases.

Notably, the commendable intervention of fans prevented a potential deal with the City Football Group, showcasing the importance of ethical considerations in strategic partnerships. The club's past practice of facilitating ethical player-trading through loan deals contributed to talent development and mutual benefits within the football community. The recent success of Girona, affiliated with the City Football Group, underscores the positive outcomes that can emerge from ethical decision-making and collaborative expansion in football. Recommendations for NAC Breda include strengthening corporate governance, enhancing transparency, and increasing community engagement to further bolster ethical guidelines in player recruitment.

# 5 Recommendations

Leveraging the analytical insights derived from both the Linear Regression and Random Forest Classifier models, NAC Breda can strategically enhance its player recruitment process. The following goalkeepers have been identified based on their predicted prevented goals, providing valuable options for reinforcing the team's defensive capabilities.

Linear Regression Model:

![Recommended Goalkeepers - Linear Regression]

Recommended Goalkeepers:

| | Player | Team within selected timeframe | Age | Market value | Contract expires | Matches played | Minutes played | Clean sheets | Prevented goals | Save rate, % | Predicted Prevented goals |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 11943 | V. Černiauskas | Panevėžys | 34.0 | 200000 | 2023-12-31 | 28 | 2705 | 21 | 7.04 | 86.30 | 7.719118 |
| 11894 | L. Paukste | Šiauliai | 24.0 | 300000 | 2023-12-31 | 22 | 2114 | 12 | 3.79 | 75.95 | 4.661955 |
| 16378 | A. Fagerström | Västerås SK | 31.0 | 175000 | 2023-12-31 | 21 | 2082 | 12 | 3.38 | 78.67 | 4.299383 |
| 7883 | Á. Ólafsson | Stjarnan | 32.0 | 50000 | 2023-11-16 | 25 | 2418 | 9 | 5.76 | 75.00 | 3.724005 |
| 12141 | L. Wahlstedt | Odds | 24.0 | 1400000 | 2023-12-31 | 17 | 1658 | 8 | 3.06 | 75.00 | 4.451844 |
| 16327 | R. Wallinder | Gefle | 24.0 | 225000 | 2023-12-31 | 20 | 1984 | 5 | 9.92 | 79.31 | 6.373220 |
| 1068 | S. Lammens | Club Brugge II | 20.0 | 600000 | 2023-06-30 | 11 | 1050 | 3 | 4.72 | 79.10 | 5.387587 |

Selection Criteria:

Predicted Prevented Goals: Prioritize players with higher predicted prevented goals for a robust defensive strategy.
Contract Expiry: Focus on players whose contracts have already expired to explore cost-effective acquisitions.
Random Forest Classifier Model:

![Recommended Goalkeepers - Random Forest Classifier]

| | Player | Age | Market value | Contract expires | Matches played | Minutes played | Predicted Prevented goals RF |
|---|---|---|---|---|---|---|---|
| 16327 | R. Wallinder | 24.0 | 225000 | 2023-12-31 | 20 | 1984 | 8.7757 |
| 4182 | O. Forsman | 35.0 | 50000 | 2023-12-31 | 26 | 2499 | 7.6878 |
| 11943 | V. Černiauskas | 34.0 | 200000 | 2023-12-31 | 28 | 2705 | 6.3689 |
| 7883 | Á. Ólafsson | 32.0 | 50000 | 2023-11-16 | 25 | 2418 | 5.8017 |
| 1068 | S. Lammens | 20.0 | 600000 | 2023-06-30 | 11 | 1050 | 4.1232 |
| 11894 | L. Paukste | 24.0 | 300000 | 2023-12-31 | 22 | 2114 | 4.0039 |
| 16378 | A. Fagerström | 31.0 | 175000 | 2023-12-31 | 21 | 2082 | 3.5209 |
| 12141 | L. Wahlstedt | 24.0 | 1400000 | 2023-12-31 | 17 | 1658 | 3.2223 |
| 8098 | D. Lyness | 32.0 | 150000 | 2023-11-30 | 27 | 2647 | 3.0134 |

Selection Criteria:

Predicted Prevented Goals: Emphasize players with significant predicted prevented goals for reliable defensive contributions.
Contract Expiry: Prioritize players with expired contracts for strategic and cost-efficient acquisitions.
Next Steps:

In-Depth Player Evaluation: Conduct thorough assessments of each recommended goalkeeper, considering additional performance metrics, playing style, and compatibility with team tactics.

Contract Negotiation: Initiate negotiations with identified players, leveraging the advantage of their expired contracts to secure favorable terms.
Seamless Integration: Upon successful acquisitions, strategically integrate the new players into the team through well-planned training sessions and onboarding.
Future Considerations:

Extend the analysis to other playing positions, such as midfielders and forwards, to achieve a holistic improvement in team dynamics.
Regularly update predictive models to adapt to evolving player performances and market dynamics.
By adopting a data-driven approach to player recruitment, NAC Breda can make informed decisions, ensuring a resilient and competitive team in the dynamic landscape of football. Composition and elevate its competitive standing.

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