Soccer Analytics Student Challenge

Pablo Soler Garcia, Paul Schlossmacher

November 2023

1 Introduction

In this project we use the data sets about the Women's World Cup 2023 and the Men's World Cup 2022 provided by StatsBomb [StatsBomb(2023)].

The goal of the project is to find an algorithm that based on data sets in a format such as above distinguishes between women's and men's games. The challenge is to do so without taking advantage of differences in physiology and training facilities. While it is always possible to argue about whether the classifiers that are being used are (in-)directly being impacted by such factors, we make a conscious effort to explain our choices on the following pages.

While working on this project we could also draw on the results of [Pappalardo L(2021)]. While their work is based on a different data set, it still at times gave us clues as to which direction to go in as far as our classifiers that are directly derived from event data are concerned.

To work on the data we used a combination of the software *Knime* and *Python*. As can be seen in our Knime workflow, we processed the data via Knime nodes and then went on to do the analysis in a *Python script* node. This node finally outputs a graphic that shows the accuracy of our model when splitting the data into a training- and a test-set with a 30-70 split. As can be seen later, we observe a 100% accuracy already with very few features.
In the spirit of transparency we upload both the Jupyter Notebook file on which we worked and the Knime workflow. However both work independently and should yield the same results.

2 Feature Extraction

2.1 Event Data

We start off by quite simply calculating the time passed in each period in seconds to use it instead of the normal timestamp. This allows us to calculate the exact frequency of features taking into account added time in each period.

Based on the event data, we then extract a set of features. The approach is to get as many features as possible and then analyse the correlation and classification capability of those features. With the knowledge of which features are important, we apply feature selection taking into consideration that no classifier related to differences in physiology and training facilities

should be used.

The first type of feature is the frequency of an event happening. This is calculated using the time difference between events. We then not only calculate the average time between events happening but also the standard deviation to that average time, as this might also contain important information to the nature of the event. A total of 55 events, such as shots, possession duration and free kicks, are taken into account, which yields a total of 110 features for the mean and standard deviation. The features analysed on their appearance frequency are:

Tabelle 1: Features based on appearance frequency

Pass	Ball Receipt	Pressure
Low Pass	Ground Pass	High Pass
Carry	Clearance	Ball Recovery
Shot	Block	Goal Keeper
Miscontrol	Duel	Interception
Foul Committed	Dispossessed	Dribble
Injury Stoppage	Substitution	Bad Behaviour
Player Off	Own Goal	50/50
Offside	Possession duration	Time waste
Throw-in	Corner	Goal Kick
Recovery		
Right Foot Pass	Left Foot Pass	Keeper Arm Pass
Header Pass	Drop Kick Pass	No Touch Pass
Other Passes		
Free Kick	Half Volley Shot	Normal Shot
Volley Shot	Backheel Shot	Diving Header
Lob	Overhead Kick	
Regular Play	From Kick Off	From Throw In
From Counter	From Corner	From Keeper
From Goal Kick	From Free Kick	

Of course some of the variables in table 1 - such as the variable *Overhead Kick* - will definitely not go into our final model, because of the low frequency at which they were observed.

In addition to this simple yet important feature extraction technique, we also investigate the position where events occur. For every event type listed in table 2 and for every match, we calculate the mean and standard deviation of where these events took place.
We can see an example of this in graphic 1, where for every match the average location of dribbles in this match is shown.

Tabelle 2: Features based on location of appearance

Pass	Shot	Offside
Dribble	Clearance	Pressure
Interception		

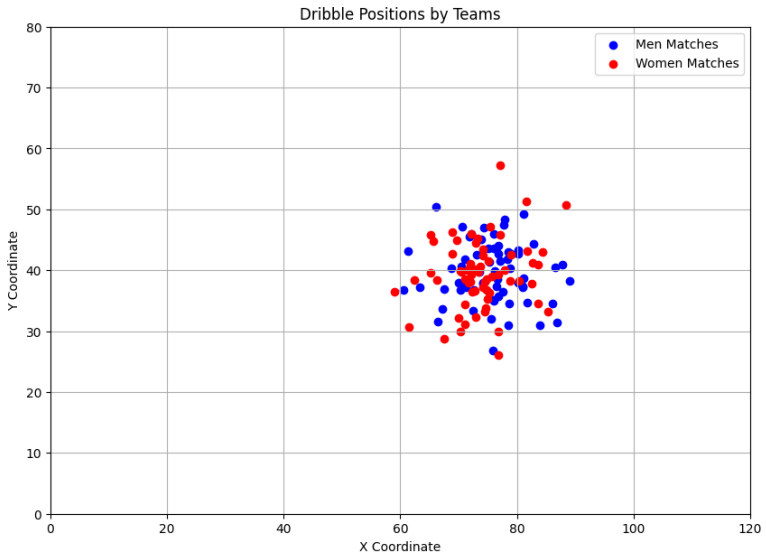


Abbildung 1: Average Dribble Position per Match

2.2 Event360 Data

Event360 is also investigated as a source of features that yield good classification results. The data consists of positional information of all visible players at multiple frames during the match. Therefore more complex feature extraction methods need to be used. To visually explain the concepts behind the features, we created plots based on the game between Argentina and France in the final of the Men's World Cup 2022. The scripts can be found in the accompanying Jupyter notebook - which we also uploaded - and are based on the same function that extract the features.

2.2.1 Pressure on the attacker

In the first of these methods we calculate the number of defending players around the attacker whenever the ball is in the attacking third of the field. We do this by creating circles with different radii around the player holding the ball and then counting the number of defending players inside this circle.
We test several different radii separately to determine the optimal value for this feature and again calculate the mean and standard deviation for each match. The intent of this method is to construct a better measure of pressure on the *actor* than the one provided through the `pressure` variable in the data. We can see an example of this in figure 2.

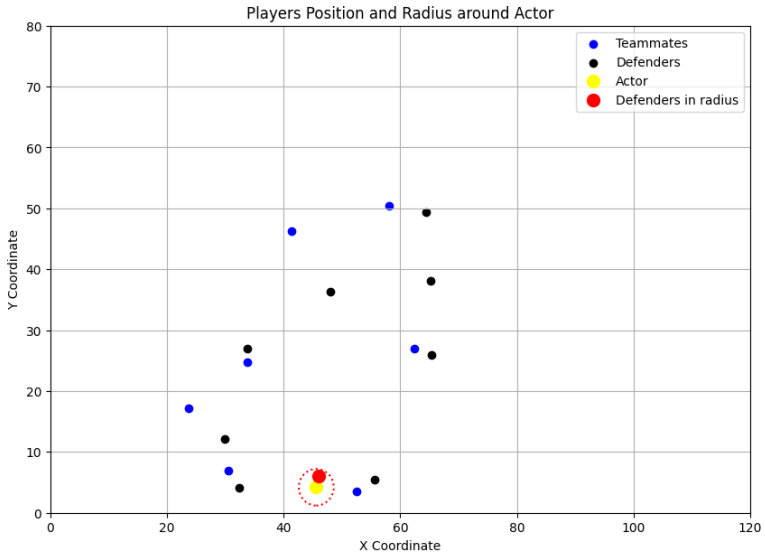


Abbildung 2: Defending Players in a 3 Meter Radius around Ball

2.2.2 Area covered

The second data extraction technique intends to extract structural information from the teams at each frame. The area covered by each team is calculated as a measure of how much of the field the players cover. This is achieved by creating a convex hull around the players and calculating its area. In addition to the area, the width and height of the teams is calculated at each frame by looking at the players on the edges of the field. While calculating these features for each match, we also differentiate between the defending and the attacking team, since we expect the structures to be vastly different in attack and defence. On figure 3 one can observe how exactly the are is calculated.

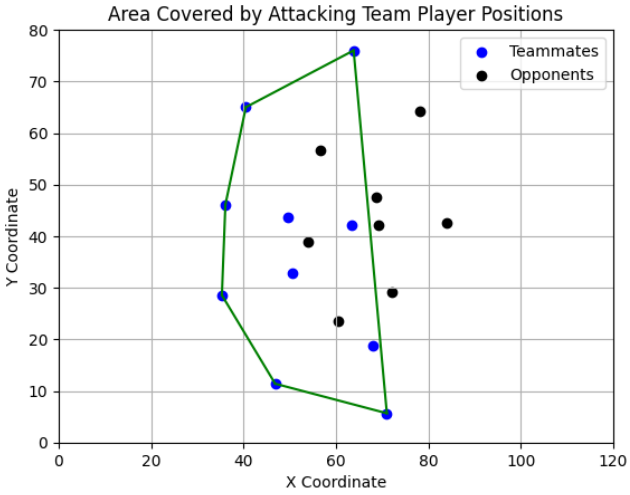


Abbildung 3: Team Area Coverage

2.2.3 Heterogeneity and compactness

In a further attempt to investigate the organisation of the teams, we calculate heterogeneity and compactness of both attacking and defending teams.

We define the **compactness** of a given player as the minimal distance between the player and another teammate. We then aggregate over every player in the team and all available time frames by taking the mean, to calculate the overall compactness of a team in a given match. Note that at all times the goalkeeper is left out of such calculations and that we perform these calculations seperately for both the attacking and the defensive phase.

While we are aware of idiosyncratic cases, where teams would not look compact to the eye, but would yield high marks in this metric (e.g. if 2 groups of 5 players stood far apart), we think that this metric gives us a fine account of a teams's compactness in practice.

In order to calculate the **heterogeneity** of a team, we use a similar set-up as for calculating compactness. However instead of calculating the mean of players' individual compactness in every moment, we calculate the standard deviation between them. We then go on to calculate the mean of these standard deviations for each time frame.
Both these metrics can be seen in figure 4.

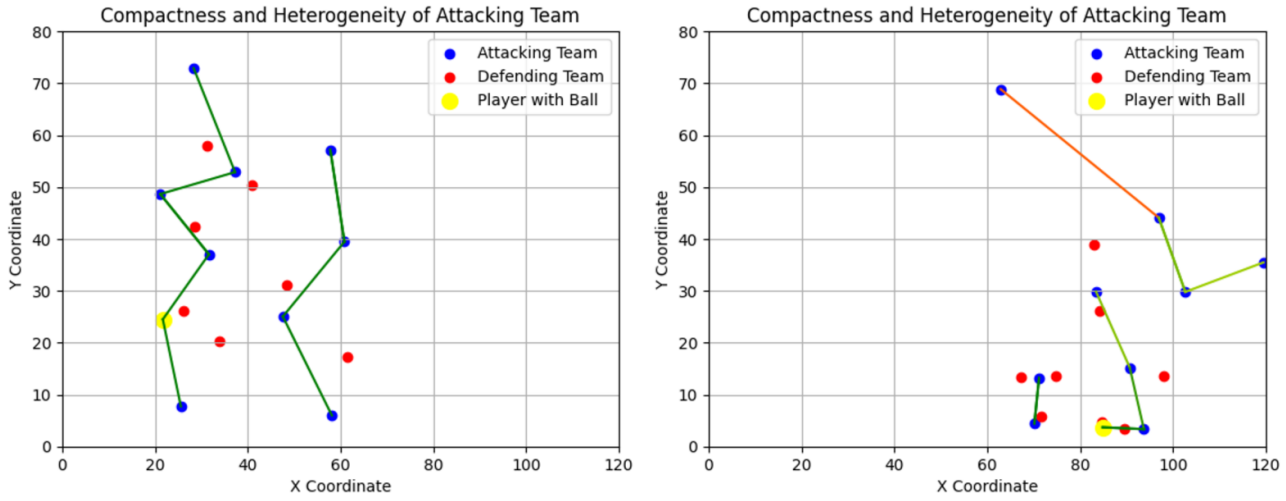


Abbildung 4: Closest Player Distance Variance: Homogeneous vs. Heterogeneous

2.2.4 Passing lanes and defensive proximity

Finally, the last features related to the 360 events is the number of direct open passing lanes and the defensive proximity.

To calculate the metric for passing lanes, we draw lines from the player with the ball to all his teammates that he could pass to and the ones where the defender blocks the pass are subtracted. This can nicely be observed in figure 5

To calculate defensive proximity a circle of reach is created around defender's positions that describes the area of their reach, from where they could still intercept the pass.
Together, these two measures intend to capture how the attacker organises to create more direct passing options and how well the defending players block those lanes. Instead of blocking passing lanes, a team can defend by marking each attacker individually. By calculating the average distance of the closest defender for every attacker past the mid-field, one establishes a

measure of defensive proximity, which intends to recognise if a team is marking each attacker individually as can be seen in figure 6.

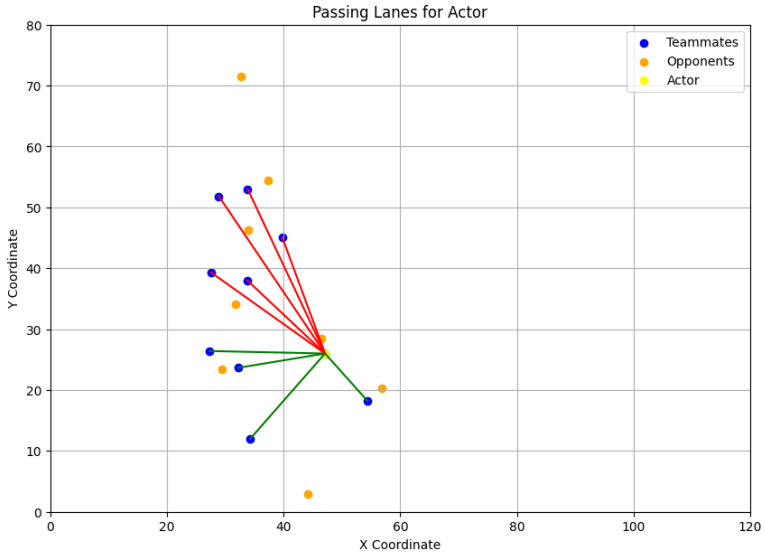


Abbildung 5: Open and Blocked Passing Lanes

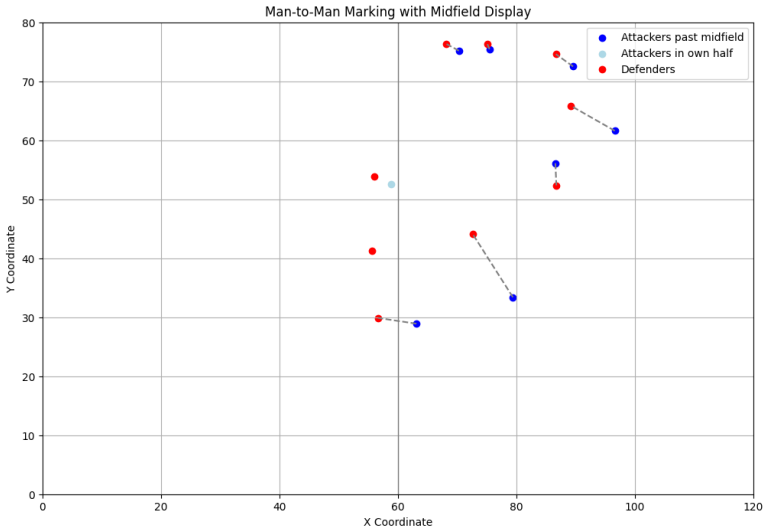
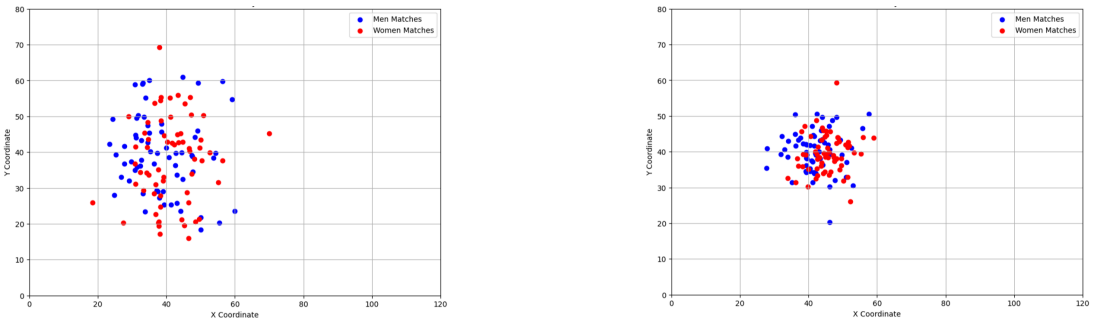


Abbildung 6: Man to Man marking for players past Midfield

2.3 Deciding between Mean and Median

While reading this article, you may have noticed that we use both the mean and the standard deviation very frequently. However we did not take the decision whether to use the mean and the standard deviation as opposed to the median and the median absolute deviatoin lightly. Since the median is less sensitive to outliers, it appears reasonable to think that given the small amount of samples, 128 matches, the median would be a more robust approach. All features were tested on both measures and the values between the two vary significantly, which is most noticeable in the Interception locations that are depicted in figure 7. In the end using the mean



(a) Median Interception Positions

(b) Mean Interception Positions

Abbildung 7: Comparison of Median and Mean

and standard deviation of the different variables resulted in better accuracy for our models when it came to prediction.
Therefore we decided to go for the mean and standard deviation even though it is the less robust approach.

3 Model used

3.1 Feature Selection

For our first screening of possible explanatory variables, the correlation matrix in figure 8 was crucial.

Here, positive values are features that are positively correlated to a women's match and negatives values are values that are usually smaller than in men's football matches. For example the average time between ground passes is highly correlated to women's football. The highest correlation to women matches is the number of defending players in a 3 meter radius of the player with the ball when the attacker is in the final third. This implies that the defensive strategy in women's teams focuses more on pressuring the player with the ball. An example of a negative correlation to women's teams is pass accuracy.

The correlation matrix in figure 8 was crucial for our first screening of possible explanatory variables. This information, combined with the information gain of each feature and the importance of each feature when training a random forest algorithm, ultimately lead us to a set of 10 variables.
In the end we decided to choose the following features for our final model. In the following we list them and justify why we don't think they are (in-)directly caused by physical differences or training facilities:

- Ball_recovery_mean: Our model utilises that ball recoveries on average occur with a different frequency in men's than women's games. In principle one may very well say this is due to physical differences by arguing that men can run more and sprint more and therefore recover more balls. We see however in our data that in fact per minute there are more recoveries in women's games than in men's games. Therefore we argue that this difference is in fact due to either tactical or technical differences between women and men.

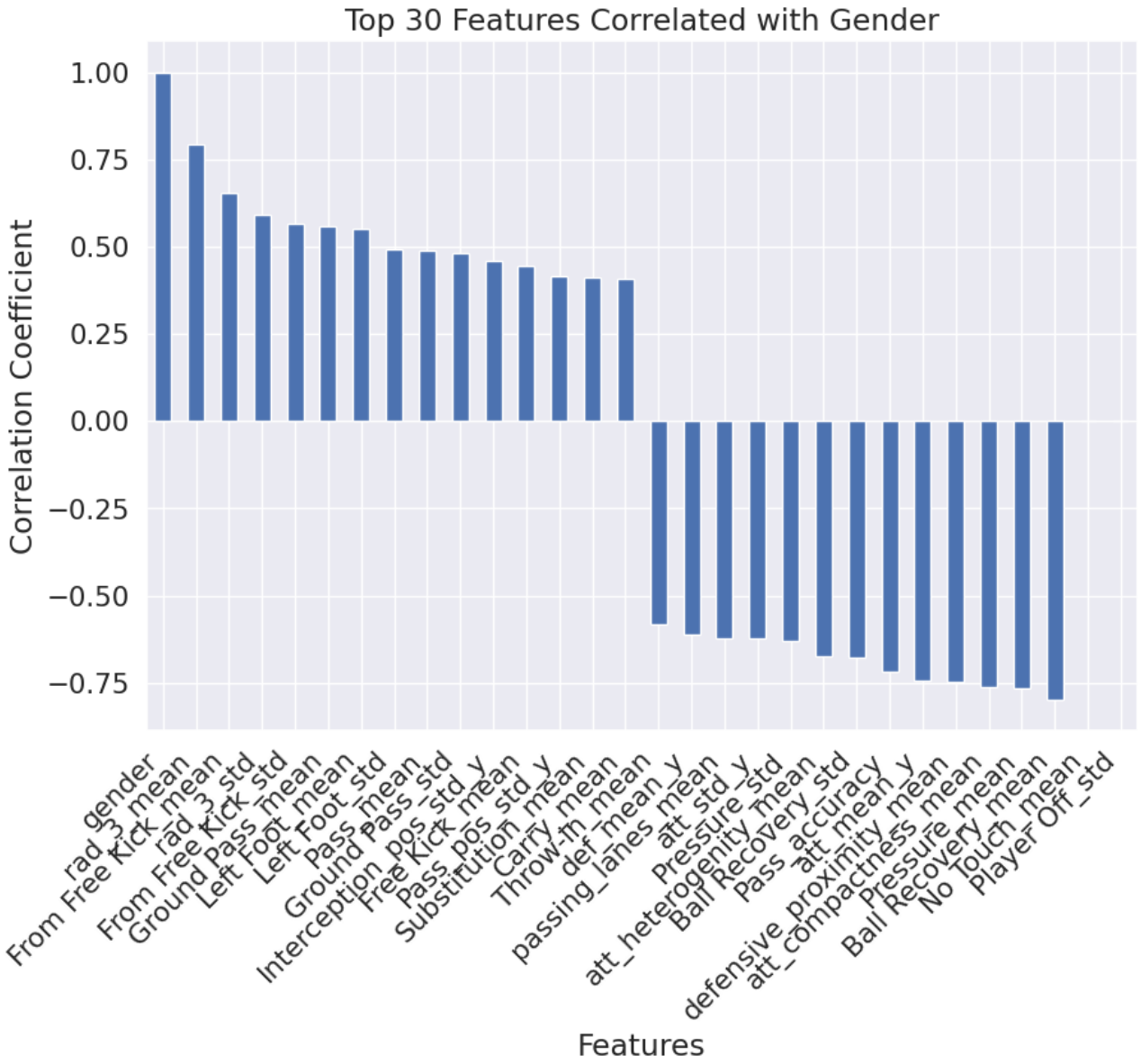


Abbildung 8: Correlation Matrix to Women Matches

- Ball_recovery_std: A similar argument holds, when it comes to this feature. We observe that the variation in time between ball recoveries is higher for men than for women. This might be due to women being quicker at recovering the ball in general, which we think - as outlined above - is due to either tactical or technical differences.
- Pressure_mean: Obviously ball recoveries and pressures being applied are closely related. So it is no surprise that we also see more pressures per minute for women than for men. We argue analogously to above that this can not be due to physical differences.
- defensive_proximity_mean: We defined defensive proximity as the average distance between every attacker and their closest defender, whenever they're in the attacking half. Therefore one might interpret higher values here as defending teams applying zonal marking compared to man-marking for lower values. Hence this in our opinion definitely is a tactical phenomenon.
- From_Free_Kick_mean: This means that there is a difference in the frequency of events taking places following a free kick. We don't see a reason why this could be down to physical differences.
- From_Free_Kick_std: Here the same applies as for the mean.
- Throw-in_mean: Again, a higher frequency of throw-ins seems to us to be completely unrelated to physical differences.
- rad_3_mean: This value describes the number of defenders that are on average in a 3 meters radius of the attacker holding the ball. We observe that on average there are many more players in this radius for women than for men. Therefore the similar argument to the bullet point *Ball_recovery_mean* holds.
- rad_3_std: We see that the standard deviation is slightly higher for women than for men (0.69 vs 0.64). This is not a big difference but makes sense in light of women having a higher mean as well (as outlined just above)
- Pass_mean: The time between passes being played is lower for men than for women. This makes sense in light of the fact that there are more pressures being applied in women's games. Either way we definitely see this as a tactical choice.

3.2 Sequential Feature Selector

We input our variables into the `Python` function `SFS` from the `mlxtend` package. When splitting the data into a 30% training set and a 70% test set, we already achieve 100% accuracy from using 4 features onwards as can be seen inf Fig. 9. This dataset split ensures there is no overfitting of the data and proofs the robustness of the classification. Other algorithms such as softmax and random forest were also applied and achieved similar results, therefore proving the classification capabilities of the selected features.

Literatur

[Pappalardo L(2021)] Natilli M Cintia P Pappalardo L, Rossi A. Explaining the difference between men's and women's football. *PLoS ONE*, 2021. doi: 10.1371/journal.pone.0255407.

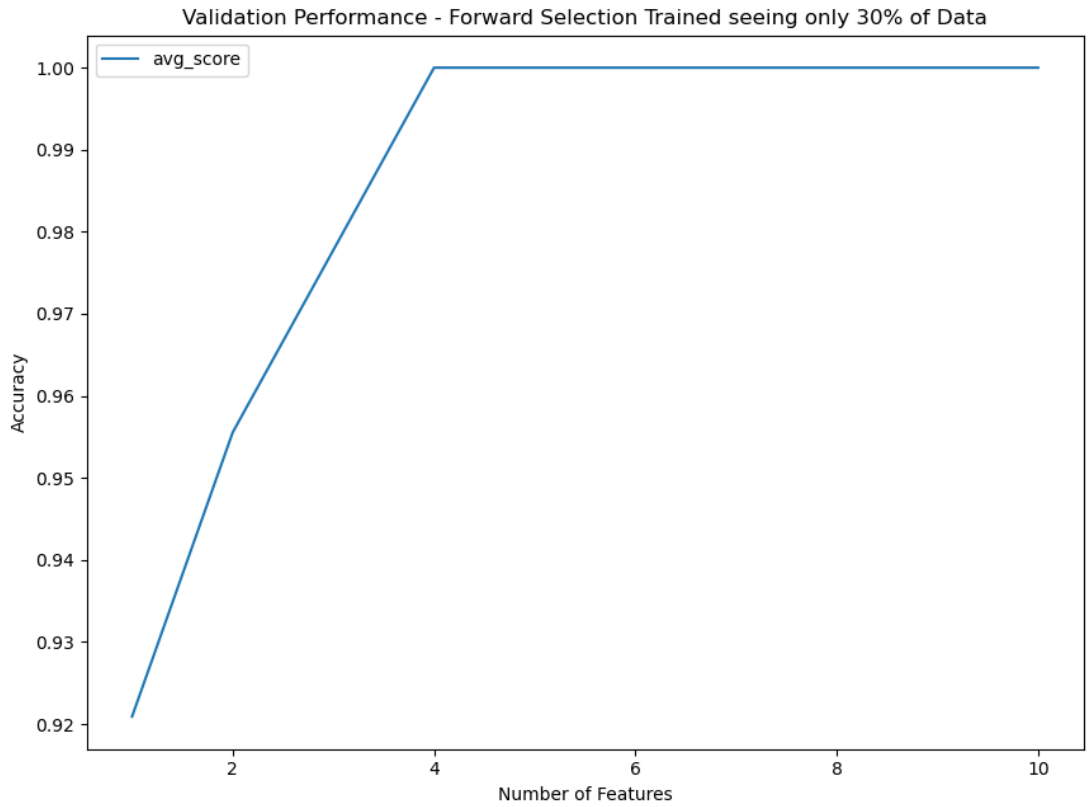


Abbildung 9: Forward Selection Regression Algorithm

[StatsBomb(2023)] StatsBomb. Open data, 2023. URL https://github.com/statsbomb/open-data.