PREDICTING THE POTENTIAL ABILITY OF FOOTBALL PLAYERS IN THE FOOTBALL MANAGER GAME

MASTER'S THESIS

WILLIAM VAN WIJK

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

STUDENT NUMBER

260640

COMMITTEE

Paris Mavromoustakos Blom
dr. Michal Klincewicz

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

June 14, 2022

WORD COUNT

8072 words

PREDICTING THE POTENTIAL ABILITY OF FOOTBALL PLAYERS IN THE FOOTBALL MANAGER GAME

MASTER'S THESIS

WILLIAM VAN WIJK

Abstract

This work aims to accurately predict the potential ability of football players, making use of the Football Manager 2020 database. Therefore, we want to answer the following main research question: To what extent is it possible to accurately predict the potential ability of football players in the Football Manager game? There is no previous work on predicting players' potential, making use of this database. The prediction task is done using Decision Trees, Logistic Regression, and Support Vector Machine models. For this analysis, 56,562 football players aged 21 and under were considered, of which 6% were considered high-potential players. The results show that the SVM model has the highest accuracy of finding talents and the LR model gives the highest recall. After balancing classes, the DT model performed the best with an accuracy of 0.91. When focussing on player attributes, balance turned out to be the most influential attribute for high-potential players. Furthermore, in a comparison between the top five and bottom five leagues, it turned out that talented players from the top five leagues could be predicted more accurately. The models proved to have a value in real life, as the predictions consisted of a few football players for which their value grew substantially in the past two years. Football clubs can make use of these results to improve their scouting, and for example focus on the important attributes, like the balance of football players.

1 INTRODUCTION

The Football Manager (FM) game tries to imitate the professional career of a manager in football. This is done in tactical aspects of managing a team and its economic and human aspects like building a relationship to manage with the press, players, agents, or club boards (Hocquet, 2016). It

contains a global database of football players from more than 50 countries and incorporates a set of 58 attributes that describe a player's in-game ability. All of these attributes are determined based on scouting reports from professional scouts employed by FM. They are divided into three different categories: technical, mental, and physical, as can be seen in Figure 1.

Figure 1: Football Manager attributes (Sports Interactive, 2020).

The FM database consists of a few hidden attributes, including current ability (CA) and potential ability (PA). Hidden attributes are not available to the FM game player but are affecting the ability of the in-game football players. CA describes what level the football player is currently at, and PA is the maximal potential level the player could grow to. The PA is a fixed value and the only value in the FM database which cannot develop over time. The focus of this study is on predicting the potential ability of football players in the FM database. We want to predict if a player falls into the category of high potential or not. The potential ability will be used as ground truth in this prediction task.

When looking at the interpretation of the FM attributes in real life, mental attributes could be very interesting. According to Mills, Butt, Maynard, and Harwood (2012), the awareness of football talents emerged as a fundament of change that drives the development of young football players. Awareness appeared to be a stimulant for developing behavior like coping with setbacks, professional attitude and emotional intelligence.

1.1 *Motivation*

The motivation for this research topic is firstly the curiosity to spot upcoming stars in the database of Football Manager. Next to this, predicting the potential ability is interesting for football clubs because players with a high potential could be a good investment. A real-life example is Italian football club Udinese, who went from the Italian second tier to the top of Italian football after selling Alexis Sanchez with a profit of 30 million euros. (Bleaney, 2017)

1.2 *Relevance*

FM is known for its scouts around the world, which gather useful information about the players in their database. This information is trusted to the extent that soccer clubs sometimes use this database to scout potential players (Walker-Emig, 2017). Sports Interactive, the company behind FM, have over 1300 scouts around the world to gather information for their database (Bleaney, 2017).

A benefit of FM database is that it provides easy access to lower league football in harder-to-access places of the football world – places where not every club can afford to send their own scouting staff (Stuart, 2020). One might ask how accurate the data from these countries are, which makes it is interesting to compare the accuracy of predicting high potential ability between the most popular leagues and the lesser ones. Lower leagues may have worse infrastructure and therefore, players with high potential might not get the attention they deserve. A study like this could help to detect these players and help players in lesser known countries. This is due to the game database possibly containing insightful information about upcoming football players that are not getting enough exposure because of the league they participate in. Lower leagues do not participate in top-tier competitions, like the Champions League, and therefore the players may not get the recognition they deserve.

1.3 *Problem statement*

In this thesis, the aim is to accurately predict the potential ability of football players, making use of the Football Manager 2020 database. Therefore, we want to answer the following main research question:

To what extent is it possible to accurately predict the potential ability of football players in the Football Manager game?

To answer the main research question, we want to answer the following research questions:

RQ1 *Which model can most accurately predict the potential of football players in Football Manager?*

RQ2 *Which (publicly available) attributes are the most important indicators of a high potential player?*

RQ3 *How does the prediction accuracy in the top five leagues compare to the lesser known leagues?*

1.4 *Data/code statement*

The dataset is obtained from Kaggle, including hidden attributes. The dataset is available under a CC0 license (Oktanto, 2021). The dataset is retrieved from the FM 2020 game in september 2019.

The algorithms for this thesis are used in Python. In Python, the library Pandas is used for data handling and transformation (development team, 2020). For the decision tree, logistic regression and support vector machine models, the library scikit-learn is used (Pedregosa et al., 2011). For oversampling, SMOTE is used from the Imbalanced Learn Python package (Lemaître, Nogueira, & Aridas, 2017).

2 RELATED WORK

In this section, we discuss two separate aspects. First, we will be looking at the research that has been done around the game Football Manager. Next to this, sources from talent prediction in the world of football will be analyzed.

2.1 *Football Manager*

First, it is important to mention that there is a limited offer of research papers on the game Football Manager. This is one of the reasons why the work can be considered novel. On the other hand, this means there is no benchmark to compare the results to and there are no methods that have been proven to work for this task.

The first work by Lima, Tertuliano, Aroni, Machado, and Fischer (2018) focuses on the development of young players in relation to their current and potential abilities using the simulator. The study simulated four seasons in the game and compared the ability of two groups of young players (16-year-olds and 20-year-olds). The results showed a significant difference

between the two groups in favor of the 20-year-olds because these players were closer to their peak at 27 years old. The conclusion of this work is that digital games like FM can be an important tool for information about athletes, which could support professionals working in the field. The described work is similar to our work, as they simulated the game to predict the future ability of these players. Our work, however, will make use of Machine Learning models to do this.

Next to this, the paper by Yaqin, Ramadhan, Jauhari, and Humami (2019) focuses on the prediction of players' positions in Football Manager 2018, based on attributes in the database. The study predicts the position of four players using a Naive Bayes model. After that, it compares their performance in the simulation on these positions to the performance on their default positions. The results show an 8.7% improvement in performance in terms of player grade in the simulated matches on the predicted positions. Although the sample size in this study was only four players, it is an interesting finding that a model can recommend the player position to the extent that the performance improves notably. This could be an indication of the public player attributes being capable of recommending player positions, which could mean that their system can play the game better than default. This is an interesting finding, and it inspires to see if the same can be done for predicting the potential ability of players.

The final research paper on the topic of Football Manager, by Yiğit, Samak, and Kaya (2019), is about assessing players' value using machine learning techniques. The goal is to support transfer decisions of football clubs. Value assessment models like lasso regressions, random forests and gradient boosting are conducted and compared to the widely trusted database of Transfermarkt.com. The attributes of 5316 players from 11 different major leagues are used as the base value. The gradient boosting model proved to be the most accurate model for predicting players' values, with a Mean Squared Error of 0.170. This is a different approach than the other two papers, as Transfermarkt.com was used as the ground truth value. It is interesting to see the connection to real-life data and the accuracy with which the FM database can predict the market value of football players.

2.2 *Talent prediction*

The second part of this section consists of research papers focusing on the influential factors for talent development. This is a very interesting area because young talents have big potential values. Therefore, success can be achieved by scouting the right talents. This also means a lot of research is

done by private companies, not for academic purposes. However, several resources can help with this research; these are described in this section.

The paper by Sarmento, Anguera, Pereira, and Araújo (2018) is a systematic review of talent identification and development in the world of football. 2944 Papers were reviewed around this topic, and 70 of these are analyzed in detail. This paper emphasizes the need for coaches and scouts to consider tactical and technical skills in combination with their body size and measures. But there is a complex relationship between these factors. The conclusion is that talent identification and development programs should be dynamic, providing the possibility of having changing evaluation parameters in the long term. However, this paper does not detail outlying factors, but the following work focuses more on these details. The main conclusion taken away from this research paper is that the combination of many factors is vital for talent development.

The work by Mills et al. (2012) examines the factors influencing the development of youth football players based on the development theory by Gagné (2004). In this work, ten experts were interviewed. These interviews revealed six primary factors that influence player development, which are: (1) awareness, (2) flexibility: handling setbacks and having an optimistic attitude (3) goal-directed attributes: determination and professional mindset, (4) intelligence: sporting intelligence and emotional resilience, (5) sport-specific attributes: coachability and competitiveness and finally: (6) external factors: family and home environment. Out of these six factors, awareness proved to be the most important factor. It would be interesting to check if attributes similar to awareness are important in the development of talents in FM. Therefore, we will be especially focusing on mental attributes, like determination and decision-making.

Finally, Beckmann and Beckmann-Waldenmayer (2019) suggest that the current capacity of young football players is not a prerequisite for a high future potential. According to their work, physiological factors like motivation, determination, and coachability are very important. To grow these attributes, it is important that not only talent development is undertaken, but expanded to a level of 'life coaching'. This paper motivates us even more to look into the details of the impact of mental attributes on player potential in the FM database.

This section has found that mental factors are the most important when developing talent. Therefore, the assumption can be made that the mental attributes are influential in predicting the potential ability in the FM database. Attributes like determination and decision-making will be assumed to be important, and this assumption will be tested in research question 1. However, it should not be forgotten that the total package remains the most important. A combination of innumerable factors must

be fulfilled for potential players to grow into excellent football players (Mills et al., 2012). So, in addition to the mental attributes, the technical and physical attributes are also important.

3 METHOD

In this section, we describe the methods for predicting the potential ability of football players in the Football Manager game. An illustration of the methods is provided in Figure 2. This section will start with a brief explanation of this flowchart.

The process in the flowchart starts with retrieving the dataset. After retrieving the dataset, the exploratory data analysis will begin. Afterwards, the preprocessing of the data takes place so that it can be used as input for the models. During the preprocessing phase, features and players are selected from the dataset to remove irrelevant data points. To train the model, the data is first split into train and test data.

The classification task will be done using three different models: decision tree, logistic regression and support vector machine. After the models are trained, they are evaluated on the test data. Based on the evaluation results, hyperparameters can be tuned to improve performance. To ensure class balance, oversampling will take place. After this, the models will be trained again. The important features can be extracted when the oversampling is done. The rest of this section explains each part of the flowchart more extensively.

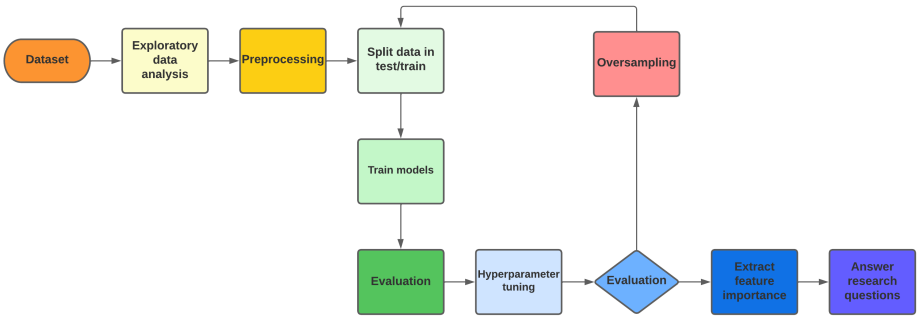

Figure 2: Flowchart of the research methodology

3.1 *Exploratory Data Analysis*

The Football Manager 2020 dataset contains more than 130,000 football players. In this section, we explore the dataset.

First, we will be looking at what league most footballers play in. In Figure 3, the ten most frequent competitions are plotted. As can be seen in this plot, there are no leagues that stand out in terms of the number of players. The Argentine first league and Italian first league contain the most players. This is followed by the English second league and the fourth Brazilian league. Of the top 20 leagues with the most players, there are 8 leagues that are not the highest league in the respective countries. This could mean that there are many players from these leagues with low potential, as these leagues tend to consist of lower quality players.

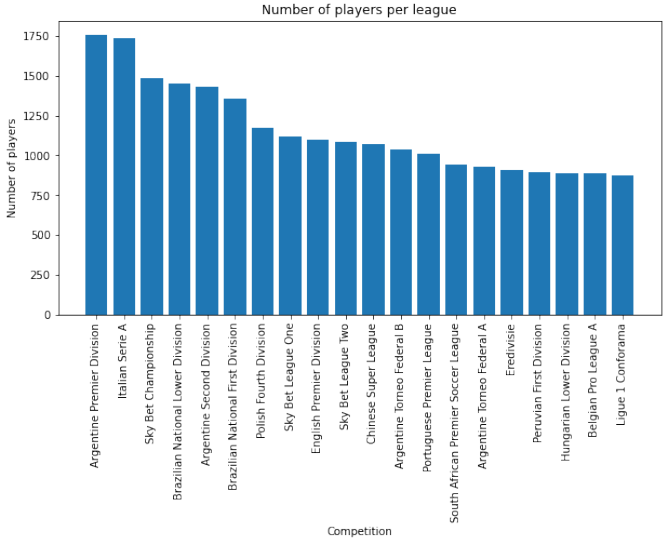

Figure 3: Number of players plotted per competition

Next to the competitions, we look into the most frequent nations. In Figure 4, the ten most frequent nationalities of players are plotted. As shown in Figure 4, Argentina, and Brazil are the most common nationalities in the dataset. Together, they account for more than 10 percent of the entire dataset. Following this are only European countries, with China filling out the top 10.

Finally, the current and potential ability of all players is explored. As can be seen in Figure 5, the peak of current ability is just under 100, and the peak of potential ability is just above 100. It is remarkable to see that there are very few players with a current and potential ability of above 150.

3.2 *Preprocessing*

The dataset consists of 64 columns of information about the players. The first six columns consist of practical information, like the players' names,

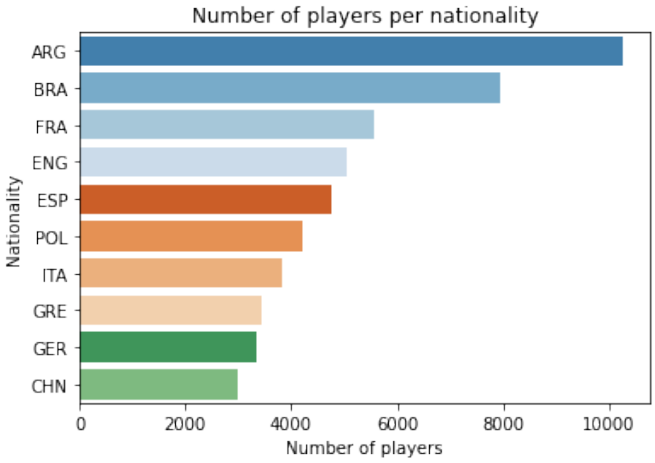

Figure 4: Number of players plotted per nationality

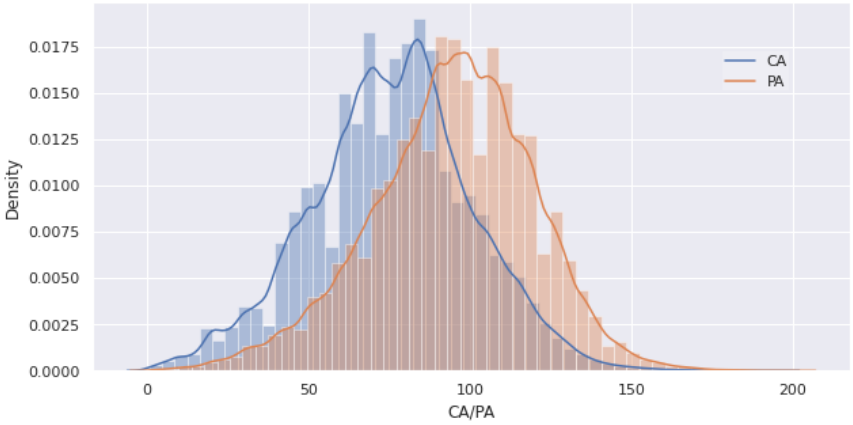

Figure 5: Density plot for current and potential ability

position, and division they play in. All the remaining columns contain numeric attributes containing information about the players' skill. The attributes can be directly used as features in the models, so there is no data transformation needed.

During the preprocessing phase, a selection is made of which features are considered suitable for the analysis. First, we decide to only use the numeric values, as the non-numeric values will not be used in the models. Next to this, the player value is not available as a fixed value in the game and depends on the club's managed stature. The wage has nothing to do with the talent and pure skill of the players and is therefore considered irrelevant in this analysis. In addition, we have the CA, which is present

in the dataset as a hidden attribute and is therefore not used. Finally, we have the goalkeeper attributes, these are stand-alone attributes and have no overlap with the attributes of other players. Goalkeepers are not included in this analysis because only 10% of the dataset consists of goalkeepers, and they have different attributes than the rest of the dataset.

Attribute	**Description**	**Reason for omitting**
Name	Name of player	Irrelevant for players' skill
Position	All possible positions of a player	Irrelevant for players' skill
Club	Club of player	Irrelevant for players' skill
Division	Division of player's club	Irrelevant for players' skill
Based	Nation of player's club	Irrelevant for players' skill
Nationality	Nationality of player	Irrelevant for players' skill
Age	Age of player	Irrelevant for players' skill
Preferred foot	Preferred foot of player	Irrelevant for players' skill
Best position	Best position of player	Irrelevant for players' skill
Best role	Best role of player in position	Irrelevant for players' skill
Value	Current value of player (in euro)	Not available as fixed value in the game
Wage	Wage of player in club (in euro, per week)	Irrelevant for players' skill
CA	Current Ability of player (hidden attribute)	Hidden attribute in the database
GK attributes	Goalkeeper attributes (10 in total)	Goalkeepers are filtered out

Table 1: Removed attributes from the dataset, with the reason for omitting

The above methods result in a total of 40 features. All these features are listed in the appendices, including explanations, as appendix A. Apart from the height and weight, all attributes are on a scale from 1 (lowest) to 20 (highest).

3.2.1 *Players' selection*

The task is to predict high potential players from the dataset. Therefore, age is the first factor to consider when selecting players from the dataset. All players aged 21 or younger are included in this prediction task. As can be seen in Figure 6, the peak of high-potential players (with a PA of 130+) are at age 21, making it a suitable cutoff. The age of 21 is also in most countries the maximum age where football players are allowed to make appearances in non-first teams of professional football clubs. From the 144,750 players in total, 56,562 players remain after the 21 years old and younger are selected. This equals to about 40% of the whole dataset.

The next step is to select the threshold for Potential Ability. We are dealing with a binary classification task, so a boundary is needed where players are considered high potential and low potential. The PA attribute is numerical, on a scale from 0 to 200. We consider the top 5% as high-potential players. 5% Is a measurement of 2 standard deviations from the mean in a normal distribution, and therefore chosen to be an appropriate cut-off for talented players. As can be seen in the plot in Figure 7, the

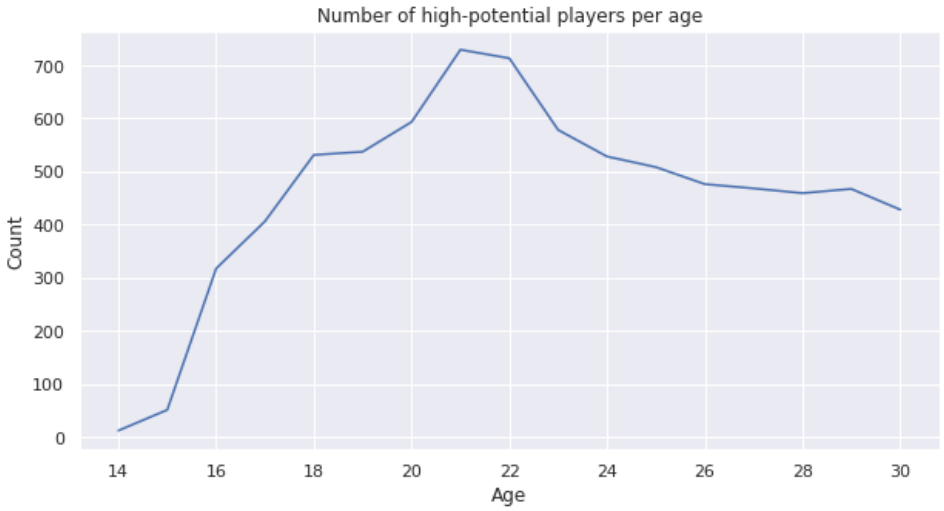

Figure 6: Number of players per age with high potential

PA attribute is normally distributed over the players because the sample quantiles show an almost straight line.

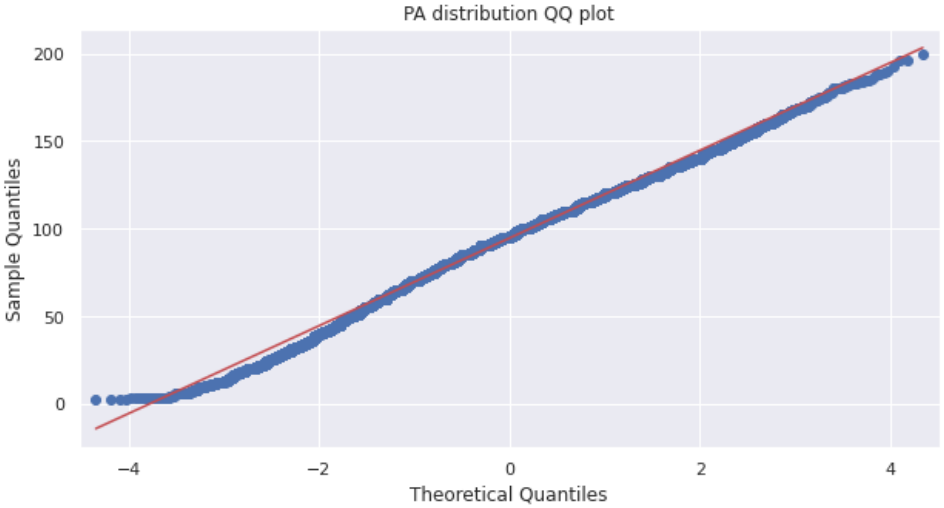

Figure 7: QQ plot PA distribution over all players

The top 5% of this dataset has a PA of 130. Therefore we settle with this as the threshold value. This means that all players with a PA lower than 130 are considered as low-potential players, all players with a PA higher than 130 are considered high-potential players. When selecting all players with 130+ PA, in total, 6% of the dataset are identified as high-potential and the other 94% are considered as low-potential.

For the third research question, a comparison will be made between the top five leagues and lower leagues regarding prediction accuracy. According to the Union of European Football Associations, the highest rated leagues in the world are England, Spain, Italy, Germany, and France (UEFA.com, 2022). According to Ackerson (2022), the five lowest leagues (excluding lower divisions from big competitions) are from South Africa, Australia, China, Sweden, and Norway. The top five league consist 5.7% of the total dataset, the bottom five leagues make up 2.6% of the total.

3.3 *Model preparation*

In this section, we discuss the train/test split of the dataset, hyperparameter tuning of the models, oversampling of the dataset and feature importance.

3.3.1 *Train/test split*

After the data is preprocessed, the model can be prepared by splitting the data into training and test parts. To do this, 'the train_test_split' function from scikit-learn Pedregosa et al. (2011) is used. The test size is set up to 20%, which means the training data is 80% of the total dataset. In Table 2, the proportion of training and test data is displayed.

high-potential?	**train**	**test**	**total**
0	37902	9459	47361
1	2347	604	2951
	40249	10063	50312

Table 2: Train/test split proportion over both classes

3.3.2 *Hyperparameter tuning*

In this phase, the optimal parameters are chosen to let the model perform the best way possible. For the decision tree, the parameter 'max_depth' will be used. The goal is to maximize the information gain and make the best split at every decision tree node (Singh, 2020). Gini index and entropy are used to calculate the information gain. A low max_depth setting prevents the model from potentially overfitting.

For the Logistic Regression and Support Vector Machine models, grid search will be used to optimize the parameters. For LR, different solvers, penalties, and penalty strengths will be used as parameters to optimize the model. Different kernels and penalty strengths will be used for model optimization for the SVM model.

3.3.3 *Oversampling*

Initially, the three models are implemented with only 6% of high-potential players. Afterwards, oversampling will be applied to possibly improve the model performance. With oversampling, new samples are created in the minority (high potential) class. Oversampling will be done in addition to the original models, as it is tough to duplicate footballers, as each footballer has unique attributes. Therefore, we will look at the results look of the models with and without oversampled data. In Table 3, the proportion of training and test data after oversampling is displayed.

high-potential?	**train**	**test**	**total**
0	37839	9522	47361
1	37938	9423	47361
	75777	18945	94722

Table 3: Train/test split proportion after oversampling

As the minority class of data consists of only 6% of the dataset, the high-potential data points are multiplied about 16 times to get the same amount of data. This means a lot of synthetic data is created, and non-existing football players are analyzed, which could make the results of this analysis less valuable. In other work, Mohammed, Rawashdeh, and Abdullah (2020) show accurate oversampling results while multiplying the minority class by a factor of 8. Next to this, Kovács (2019) uses a lot of high-performing sampling methods with oversampling methods with the multiplier of 16 and higher. Therefore, we trust the method of oversampling the minority class by a factor of 16.

To perform oversampling, SMOTE is used Chawla, Bowyer, Hall, and Kegelmeyer (2002) from the Imbalanced Learn Python package (Lemaître et al., 2017). SMOTE works by performing a k-nearest neighbor algorithm to create synthetic data, and is less prone to overfitting than random oversampling.

3.3.4 *Feature importance*

After the models have been implemented, the feature importance will be determined to highlight the most important attributes during classification. For the DT and LR algorithm, the 'StandardScaler' function from scikit-learn Pedregosa et al. (2011) is used. In this way the coefficient values per feature are extracted, and it becomes clear which features are the most important for the prediction. For the SVM, permutation importance is calculated. The permutation importance is the difference between the baseline

metric and metric from permutation per feature column (Pedregosa et al., 2011).

3.4 *Models*

In this study, a binary classification task will be performed that predicts from soccer players whether they are high-potential players or not. Three different models will be used to do this, and their results will be compared.

The first model that will be used is the Decision Tree (DT). A DT is a graphical representation of all possible solutions to a decision based on certain conditions. Every step of the DT forms a condition on the features to separate all classes in the dataset (Roy, 2021), an example can be seen in Figure 8. In this DT, there are three different nodes with features. For example, the first node separates the data by the threshold of 'Bal' being less than or equal to 10.5. The Gini value varies between values 0 and 1, where 0 expresses the purity of classification and 1 indicates the random distribution of elements across the classes.

The goal of a DT is creating a model that predicts the class of the target variable (in this case: high potential) by learning simple rules derived from the training data. These rules are based on the features in the training data like in Figure 8, the model will determine which value per attribute is the most useful classification to differentiate high potential and low potential players. Decision trees are a relatively simple and transparent way of classification, although the calculation can be sometimes overly complex and time-consuming (Ke et al., 2017).

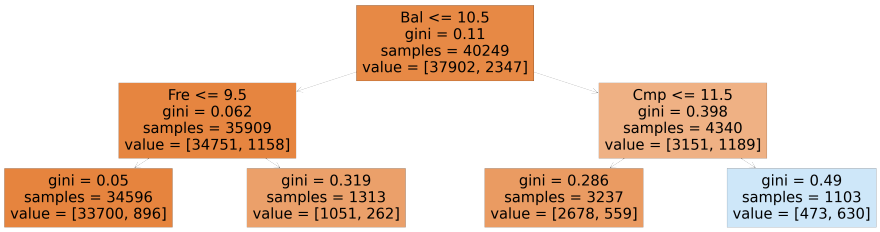

Figure 8: An example of a DT for classification.

The next model that will be used is Logistic Regression (LR). To look into the functionality of LR, we first take a look at Linear Regression. Linear Regression is a regression model that estimates the relation between two variables, and aims to do this using a straight line (Cook, 1977). LR is different to Linear Regression in the sense that the predicted value must be categorical. A plot of the LR classification with labeled sample data can be seen in Figure 9. The LR algorithm provides probability scores for

different features, but has a poor performance with a lot of irrelevant and correlated features (Bonney, 1987).

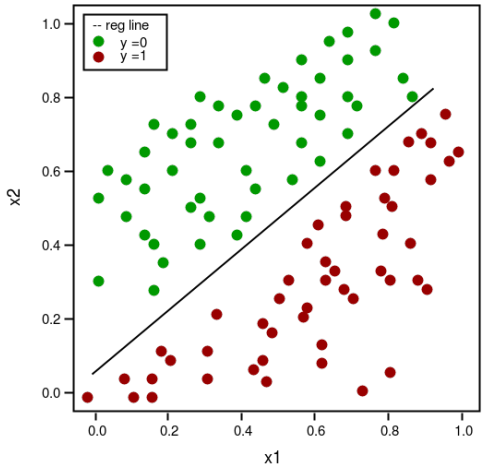

Figure 9: A plot of the LR classification with labeled sample data (Kumar, 2022).

Finally, the Support Vector Machine (SVM) is a linear model that creates a decision boundary to separate data into classes, as shown in Figure 10. It is similar to LR, but SVM tries to find the best margin that separates the classes and this reduces the risk of error on the data, while LR does not. An additional benefit is the risk of overfitting being less likely with SVM, compared to LR (Bassey, 2019). According to Bassey (2019), it is recommended to use and evaluate logistic regression. If it does not perform well enough, the next step is to try using SVM, since it is expected to improve performance. Logistic regression and SVM have similar performance, but depending on the features, one may be more efficient than the other.

3.4.1 *Alternative models*

In this section, alternative models will be discussed for this classification task. First, neural networks are powerful models for a lot of Machine Learning tasks, including binary classification. Neural networks are estimated to be unsuitable for this task because they tend to lack transparency, and therefore it is difficult to explain which features of the model are important for prediction. When transparency can be achieved for neural networks, they could be used to improve the results of this work in the future.

Next to neural networks, XGboost was considered for this task. XGboost uses the gradient boosting decision tree algorithm and generally achieves good results with it. Thus, it is a more elaborate and less transparent version of the decision tree. As a result, it can achieve better results, but it

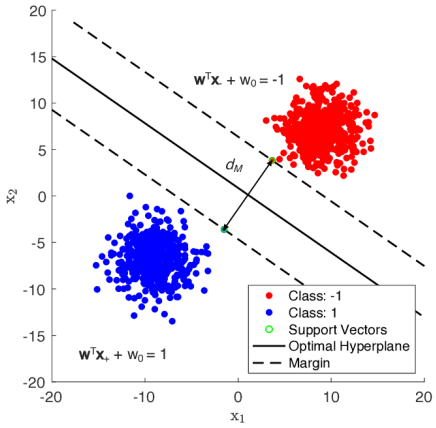

Figure 10: SVM example for binary classification (Bianco et al., 2019).

is less easy to interpret. In future work, XGboost could be used to improve the results of this work.

3.5 *Evaluation*

The model will deal with an imbalanced dataset, as we have considered only 6% of the dataset as high potential players, so the most common prediction will be 'low potential'. The baseline of the model results will be the prediction of the majority class. This will mean that the model outperforms the baseline when it achieves an accuracy of above 94%. When oversampling is used, the baseline is 50% as the classes are balanced with approximately the same amount of data.

To fully evaluate the effectiveness of a model with imbalanced classes, it is important to use precision and recall instead of just accuracy (Tharwat, 2020). Precision, also called Positive Predictive Value (PPV), is an evaluation method which is calculated based on the number of true positive instances out of the samples predicted as positives (true positives and false positives). Precision is not affected by many negative class instances. Therefore, it is an appropriate method to use when evaluating intent predictions. Recall, also known as True Positive Rate (TPR), is an evaluation method which is calculated based on the number of true positive instances out of the actual positives (true positives and false negatives). The f1-score is a metric which takes precision and recall into account, and calculates a weighted ratio between these two metrics. Therefore, the f1-score will also be used as evaluation method.

4 RESULTS

This section presents the analysis results, aiming to address the research questions. The first part focuses on model comparison with the original data and afterwards with oversampled data to balance classes. Next, the most important features of the best-performing model are extracted. Finally, a comparison between the top five and bottom five leagues is made.

4.1 *Model comparison*

In this section, the first research question will be answered, namely:

RQ1 *Which model can most accurately predict the potential of football players in Football Manager?*

To do this, we compare the classification accuracy of three different models, namely DT, LR and SVM. All the models aim at predicting high-potential football players in the FM 2020 database. The models are set up with a train/test split of 80%, all 39 features (see Appendix A) and will be evaluated using accuracy and f1-score.

In Table 4, the DT model results are presented. The majority class (0: low potential) constitutes 94% of the dataset, which results in a baseline (that always picks the majority class) of 0.94. Based on the accuracy score, the model marginally exceeds the baseline by 1%. The model is set up with the maximum depth of 5, as this gave the best results during the parameter tuning (see Figure 11). The run-time of the experiment is 0.45 seconds.

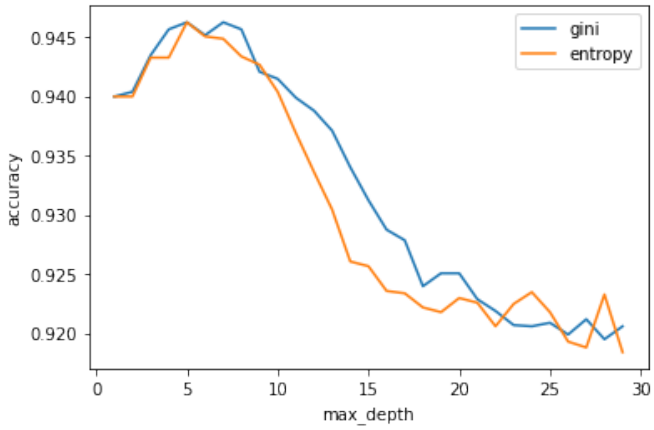

Figure 11: Parameter tuning DT model

The model performs well on the majority class, reaching a 0.97 f1-score. The performance is clearly worse in the minority class, reaching a 0.32 f1-score when predicting the 604 players in the test set.

	precision	**recall**	**f1-score**	**support**
0	0.95	0.99	**0.97**	9459
1	0.66	0.22	**0.32**	604
accuracy			**0.95**	10063

Table 4: DT model results

Furthermore, the LR model results are presented in Table 5. The model marginally exceeds the baseline, by 1%, based on the accuracy score. The model is set up with a C value of 0.1. This C value has a default of 1. a lower value tells the model to give less weight to the training data. The penalty method of 'l2' is set up for the model, which is the default option. Finally, the solver method 'newton-cg' is used, which is compatible with the penalty method. The grid-search gave the best results with these parameters, although the difference in accuracy was only 0.002 between the best and worst parameters. The run-time of the experiment is 10.7 seconds, which is considerably more than the DT model but still acceptable.

The model performs well on the majority class, reaching a 0.98 f1-score. The performance is clearly worse in the minority class, having a 0.49 f1-score when predicting the 604 players in the test set. The f1-score of 0.49 is noticeably higher compared to the 0.32 performed by the DT model.

	precision	**recall**	**f1-score**	**support**
0	0.96	0.99	**0.98**	9459
1	0.76	0.36	**0.49**	604
accuracy			**0.95**	10063

Table 5: LR model results

Finally, in Table 6, the SVM model results are presented. Based on the accuracy score, the model also marginally exceeds the baseline by 1%. The model is, just like the LR model, set up with a C value of 0.1. The linear kernel is used; this is an effective kernel when the data is linearly separable. The grid-search gave the best results with these parameters, although the difference in accuracy was less than 0.01 between the best and worst parameters. The run-time of the experiment is 388 seconds, which is a lot more than both the DT and LR model and reduces the usability of the SVM model.

The model performs well on the majority class, reaching a 0.97 f1-score. The performance is clearly worse in the minority class, having a 0.35 f1-score when predicting the 604 players. The f1-score of 0.35 is similar to the 0.32 performed by the DT model in the test set. Although, the precision in the minority class increased from 0.66 when using DT to 0.90 when using

SVM. This means there is a higher percentage of actual high-potential players in all predicted high-potential players.

	precision	**recall**	**f1-score**	**support**
0	0.95	1.00	**0.97**	9459
1	0.90	0.22	**0.35**	604
accuracy			**0.95**	10063

Table 6: SVM model results

To finalize this section, we observe that the LR model has performed best when predicting high-potential football players, with an f1-score of 0.49 over the predicted high-potential players. However, SVM shows the highest precision and lowest recall. It means that SVM model does not predict many high potential players, but it is accurate when it does.

4.1.1 *Oversampling*

This section presents the model results with oversampling of the data. The number of data points in each class is equalized by oversampling, resulting in a baseline accuracy of approximately 50%. To do this, SMOTE is used. It performs a k-nearest neighbor algorithm to create synthetic data.

Table 7 shows the DT model results with oversampled data. The classes are balanced, and therefore we have a baseline of about 0.5 to compare the results to. As you can see, the classes are not exactly balanced on 50% (9522/9423), this is a limitation of the oversampling method.

The model is set up with a maximum depth of 18, as this gave the best results during the parameter tuning (see Figure 12). The information gain (based on Gini and entropy) was almost on the highest level, and picking in a lower maximum depth will avoid overfitting.

The results of the high-potential class (1) improved a lot compared to the original model (0.91 vs. 0.32) and the low-potential class had a slight decrease in f1-score, when comparing the non-balanced to the balanced dataset experiment using DT (0.91 vs. 0.97). The accuracy score is reduced by 4% compared to the non-oversampled data, but is a lot higher than the baseline of 0.5. The run-time of the experiment is 4.29 seconds.

	precision	**recall**	**f1-score**	**support**
0	0.94	0.87	**0.90**	9522
1	0.88	0.95	**0.91**	9423
accuracy			**0.91**	18945

Table 7: DT model results with oversampled data

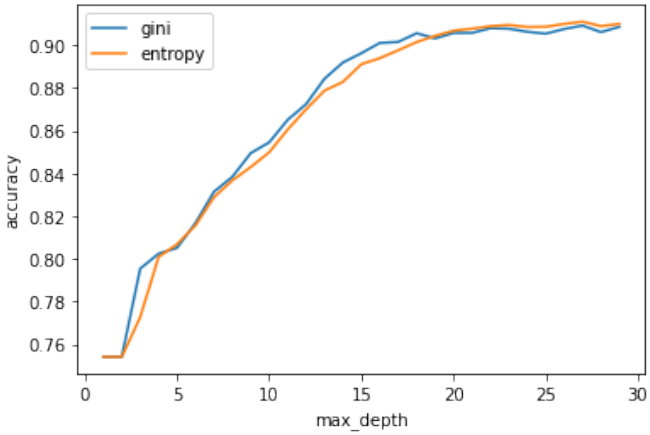

Figure 12: Parameter tuning DT model with oversampled data

Table 8 shows the LR model results with oversampled data. The model is set up with the same parameters as for the non-balanced dataset experiment because this gave the best results during grid-search. The f1-scores of both 0.82 is an improvement on the minority class but a decrease in the majority class. The overall accuracy also decreased from 0.95 to 0.82. Oversampling was not as effective for this model, compared to the performance of the DT. The run-time of the experiment is 4.29 seconds and the same as the DT model.

	precision	**recall**	**f1-score**	**support**
0	0.83	0.81	**0.82**	9522
1	0.82	0.83	**0.82**	9423
accuracy			**0.82**	18945

Table 8: LR model results with oversampled data

Table 9 shows the SVM model results with oversampled data. The model is set up with the same parameters, as this gave the best results during grid-search. The f1-score of 0.78 and 0.77 is an improvement on the minority class, but a decrease in the majority class. The overall accuracy also decreased from 0.95 to 0.78. Oversampling was just like LR not as effective for this model, compared to the performance of the DT. The run-time of the experiment is more than 3600 seconds (one hour), which is proportionally more than the other models. This could suggest that there is a lot of noise in the data, not allowing the model to distinguish between the two classes.

In the first part of this section, three models outperformed the baseline, but scored low on the f1-score. After oversampling, all the models

	precision	**recall**	**f1-score**	**support**
0	0.77	0.78	**0.78**	9522
1	0.78	0.77	**0.77**	9423
accuracy			**0.78**	18945

Table 9: SVM model results with oversampled data

improved on the baseline. Especially, the DT model performed well after oversampling, and has proven to be the best model to predict high-potential players in the FM database with oversampled data. However, we acknowledge that the balanced dataset contains 44,410 synthetic data points of the total number of 94,722, which may be biasing the classification result. This limitation should not be overlooked and will be discussed in more detail further on.

4.2 *FM attributes*

In this section, the second research question will be answered, namely:

RQ2 *Which (publicly available) attributes are the most important indicators of a high potential player?*

To do this, the important features (in-game attributes) are analyzed. We chose to explore the feature importance of the balanced DT model, as it was the best performing model in the previous section. The attributes are visualized in Figure 13.

As can be seen, there is one attribute that dominates the others in terms of importance. This attribute ('Bal') represents balance: how well a player can stay on his feet, both on and off the ball. The second most important attribute is technique, followed by bravery (how committed a player is, putting themselves into risky situations), long shots and composure (players' steadiness of mind and ability to make intelligent decisions with and without the ball).

It is interesting to see that balance stands out as the most important attribute, and that the coefficient value is more than 4x higher than for the second most important attribute. Balance is the only attribute in the top 5 features in the category of physical attributes, technique and long shots are part of the technical attributes. Bravery and composure are both part of the mental attributes category.

4.3 *Division comparison*

In this section, the third research question will be answered, namely:

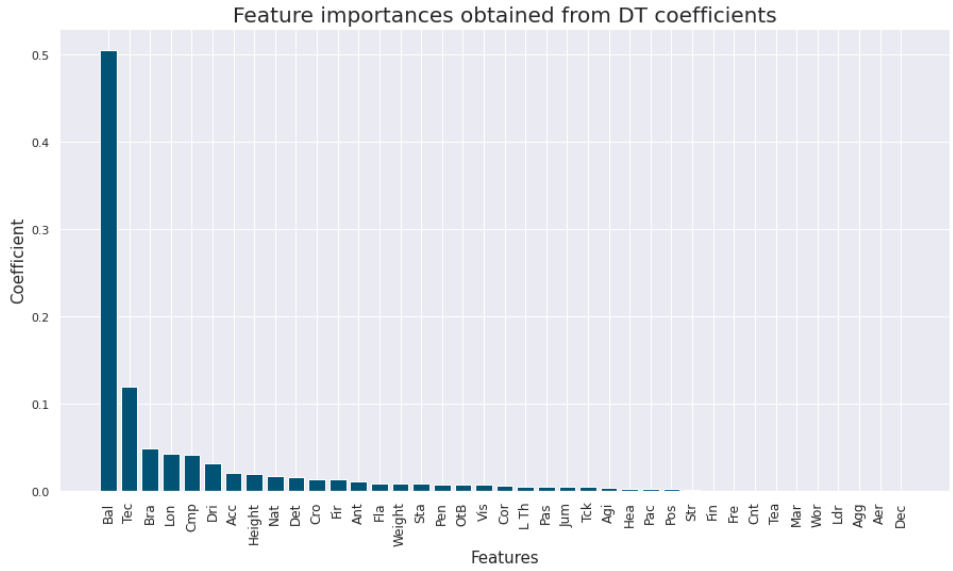

Figure 13: Feature importance coefficients from the DT model with oversampled data

RQ3 *How does the prediction accuracy in the top five leagues compare to the lesser known leagues?*

To do this, a selection of the top five and bottom five leagues will first be made. The bottom five leagues are from the following countries: South Africa, China, Australia, Sweden, and Norway. The top five leagues for this comparison are selected from the following countries: England, Spain, France, Germany, and Italy. Only the first league from these countries are included in this analysis.

4.3.1 *Bottom five leagues*

The prediction accuracy of the bottom five leagues subset will be assessed using SVM, because this model performed the best for this data subsection, as can be seen in Table 10. The overall accuracy is 0.96, and the f1-score is 0.25 for the minority class.

		DT	**LR**	**SVM**	**support**
f1-score	0	0.97	0.98	**0.98**	255
	1	0.12	0.13	**0.25**	14
accuracy		0.95	0.95	**0.96**	269

Table 10: Model comparison bottom five leagues

In Table 11, the bottom five leagues results are presented. The majority class (0: low potential) consists of 94% of the whole database, this gives us a baseline of 0.94 to compare the results to. The model outperforms the baseline by 0.02 based on the accuracy score of 0.96.

The model performs well on the majority class, reaching a 0.98 f1-score. The performance is worse in the minority class, and the 0.25 f1-score is not accurate when predicting the 14 players in the test set.

	precision	**recall**	**f1-score**	**support**
0	0.96	1.00	**0.98**	255
1	1.00	0.14	**0.25**	14
accuracy			**0.96**	269

Table 11: Bottom five leagues' prediction results using SVM model

In Figure 14, we look at the permutation importance per feature of the SVM model. High permutation values mean a positive contribution to the prediction, and negative values negatively correlate with the attribute. As can be seen, only one attribute (heading) has negative importance in the graph. The most important attributes of this model are dribbling, tackling and vision. Dribbling and tackling are technical attributes, and vision (the ability to see a potential opening and spot an opportunity another player may not have seen) falls into the category of mental attributes.

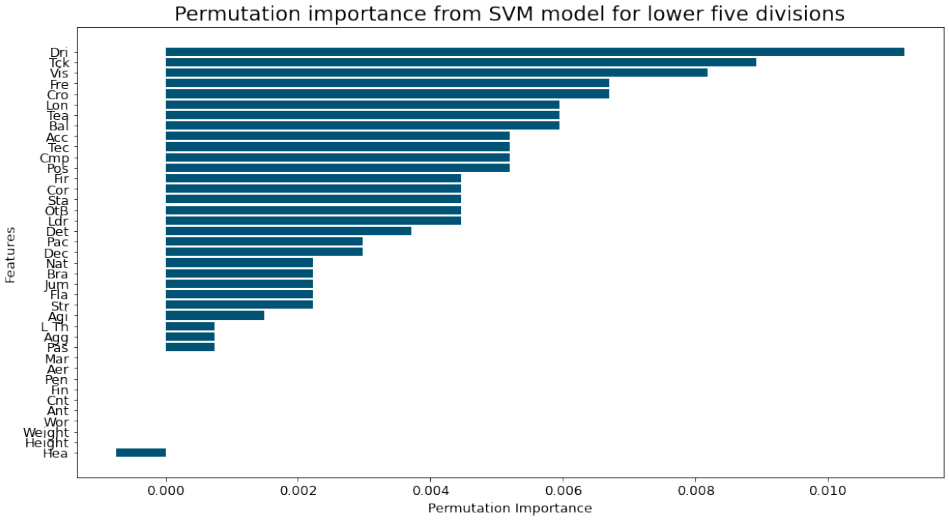

Figure 14: Permutation importance coefficients from the SVM model for bottom five divisions

4.3.2 *Top five leagues*

The prediction accuracy of the top five league subset will be assessed using SVM, because this model performed the best for this data subsection, as can be seen in Table 12. The overall accuracy is 0.89, and the f1-score is 0.79 for the minority class.

		DT	LR	SVM	support
f1-score	0	0.92	0.93	**0.93**	417
	1	0.77	0.80	**0.79**	153
accuracy		0.87	0.90	**0.89**	570

Table 12: Model comparison top five leagues

In Table 13, the top five leagues results are presented. The majority class (0: low potential) consists of 64% of the whole database, and this gives us a baseline of 0.64 to compare the results to. The model outperforms the baseline by 0.25 based on the accuracy score of 0.89.

The model performs well on the majority class, reaching a 0.93 f1-score. The performance is worse in the minority class, but the 0.79 is still good when predicting the 153 players in the test set.

	precision	recall	f1-score	support
0	0.91	0.95	**0.93**	417
1	0.85	0.73	**0.79**	153
accuracy			**0.89**	570

Table 13: Top five leagues' prediction results using the SVM model

In Figure 15, we look at the permutation importance per feature of the SVM model. In contrast to the values at the low divisions, there are many attributes here with negative importance values. For example, the attribute Lon (long shots: prowess when shooting from outside the penalty area) has the most negative importance in the top five divisions and the sixth-highest importance in the lower divisions.

In the comparison between the top five and bottom five leagues, a big difference in f1-score can be observed. Whereas the top five leagues have an f1-score of 0.79 on the minority class, the bottom five leagues have an f1-score of only 0.25. Although it is difficult to make a comparison here, because the classes are more balanced in the top five divisions.

Overall, it is interesting to see that the feature importance graph (Figure 15) of the top five leagues looks completely different from the bottom five divisions (Figure 14). In the bottom five divisions, almost all attributes have positive importance values, whereas the top five divisions only has positive importance values for 7 attributes.

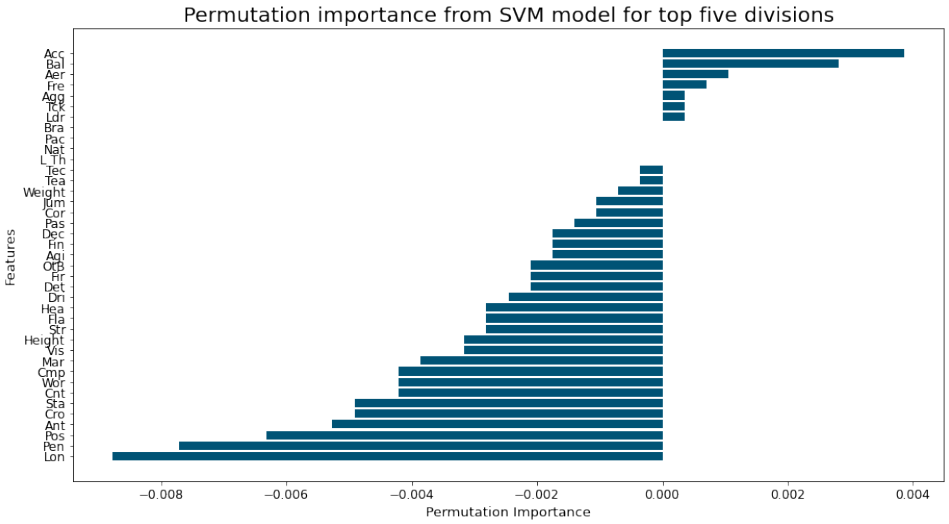

Figure 15: Permutation importance coefficients from the SVM model for the top five divisions

Although, there are also a few similarities, in both graphs, acceleration (Acc), balance (Bal), free kick tacking (Fre) and tackling (Tck) have high importance values. The other three attributes with high importance value in the top five divisions are: aerial reach (Aer), aggression (Agg) and leadership (Ldr). Aerial reach and aggression even appear at the bottom half of the graph of importance of the bottom five divisions, leadership is in the middle of this list. Vision (Vis) and crossing (Cro) are important attributes for the lower five divisions, but appear as attributes with negative importance in the top five divisions.

5 DISCUSSION

The goal of this work is to predict high-potential players in the Football Manager game. We did predict these players using three different models, including oversampling of the data. Afterwards, the most important attributes were extracted from the best performing model, and the top five and bottom five leagues' accuracy was compared. In this section, the results will be discussed.

5.1 *Model comparison*

In the model comparison, we observed that the LR model had the best performance when predicting high-potential football players, with an f1-score of 0.49 over the predicted high-potential players.

However, SVM shows the highest precision and lowest recall. Precision is the number of correct predicted high-potential players from the total predicted high-potential players. This means that if the model predicts that a player has high potential, the probability is 90% that the player has high potential. This also means that the model will miss players who have high potential but who are not marked as such by the model.

This is an interesting trade-off. When scouts are searching for a big pool of possibly high-potential players, LR would be the right model to use. In the case of looking for more certainty of finding actual talents, the SVM model would be preferred.

5.2 *Predicted high-potential football players*

We inspected the predicted football players of the SVM model and selected the players with low potential according to the dataset, and high potential according to our model prediction. Table 14 shows three interesting examples of football players from our model prediction who proved to have high ability in real life. These three players have a PA lower than 130 in fm20, and their PA score improved by respectively 32, 37 and 18. Pedro Gonçalves' earned a transfer to Sporting CP, one of the top clubs in Portugal, and his value rose to 35 million euros, according to transfermarkt.com (Transfermarkt, 2022).

Name	**PA fm20**	**PA fm22**	**value (€) sept 2019**	**value (€) may 2022**
Samuele Ricci	128	160	150k	18m
Pedro Gonçalves	125	162	800k	35m
Daryl Dike	128	146	50k	10m

Table 14: Inspection of interesting high-potential predicted (and low actual) players, according to transfermarkt.com (Transfermarkt, 2022)

These examples show that the model is able to predict football players to have high potential in real life, and can be valuable to clubs and scouts when investing in talent. It is impressive how a player can increase in value in two and a half years, which makes the analysis relevant, and the FM database also proves to be a valuable source for talent scouting like Yiğit et al. (2019) already suggested earlier. Investing in football players before they prove to have high ability is a challenge, and the FM database and our models can help with this challenge.

5.3 *Balanced classes*

After oversampling, and hence balancing the classes, the DT, LR and SVM models outperformed the baseline considerably. With a baseline of approximately 0.5, the LR and SVM model reached an accuracy of around 0.8. The DT model was the best scoring model after oversampling, with an accuracy of 0.91.

The DT model performed this accuracy after setting up the max_depth of the tree to 18. This is quite a high number and might cause overfitting. Next to this, it is difficult to validate the reliability of this prediction. 94% of the data is synthetic and thus does not contain attributes collected by scouts of real soccer players. Each soccer player is unique, and therefore it is difficult to create synthetic data, even though it is only about attributes on a scale of 1 to 20.

5.4 *Most important attributes*

The balanced DT model was used to extract the most important attributes. The results showed a domination in importance of the 'balance' attribute. According to the model, the most important attribute for high-potential players is to have a high balance. The balance attribute represents the ability of how well a player can stay on his feet, both on and off the ball. The second most important attribute is technique. This is an interesting observation, because the model suggests that as long as a player is able to stay on his feet and has good aesthetic technical qualities, then they have a high potential ability. It would be interesting to test this hypothesis in the future, and run models with only the most important features.

The balance and technique attributes fall in the category of technical attributes, this is in contrast to the suggestion of Beckmann and Beckmann-Waldenmayer (2019) and Mills et al. (2012) that mental attributes are the most important for high-potential players. The fact that balance is an important attribute for a player's potential ability does not mean that this is the only important factor in the total skill set of a football player. The multidimensional nature of talent development is the determining factor, and a combination of innumerable factors must be fulfilled for potential players to grow into excellent football players, as already suggested by Mills et al. (2012).

5.5 *Division comparison*

The top and bottom five leagues show big differences in prediction accuracy. The bottom five leagues' prediction outperforms the baseline marginally,

but reaches a low f1-score in the majority class. However, it reached a precision score of 1.00, which means that if the model predicts that a player has high potential, the probability is 100% that the player actually has high potential. This is a good result, but the low recall suggests that the model also misses a lot of high potential players in the prediction, and the low number of players in this class makes it difficult to generalize.

One of the players with high predicted potential in the bottom divisions is Jens Petter Hauge, playing for FK Bodø/Glimt in 2019, being worth €400k around that time. This player earned a transfer to the top five leagues in 2020 and is currently playing for AC Milan, being worth €8 million.

The top five leagues' prediction outperforms the baseline by a greater margin than the bottom five. The f1-score is high over both classes and suggests that the model is well capable of predicting high-potential players in the top five divisions, a lot better than the bottom five divisions.

The most important attributes differed vastly for the top and bottom five divisions. This could suggest that the scouting information of the bottom five divisions is not as accurate as in the top five divisions, but it is difficult to conclude this based on this study. It could also mean that the players in the bottom five divisions have different qualities, and that is why the difference arises. The sample size is also quite small, so it is difficult to generalize the difference in attributes.

5.6 *Limitations*

The biggest limitations of this work have to do with the class imbalance. The selection of high-potential players consisted of only 6% of the total players in the dataset, which made it difficult for the models to outperform the baseline beyond a marginal difference.

After oversampling, the high-potential class of football players consisted of approximately 94% synthetic data. Every football player has its own ability, and corresponding attributes in the FM database, which makes it difficult to work with synthetic data.

It was possible to have chosen a different threshold for PA, but this could obscure the meaning of high potential and therefore did not seem like a good choice. However, in practice, the threshold of PA value could be handled more flexibly. The skill level of players needed for a club depends on the stature and quality of the club.

5.7 *Suggestions for future research*

In future research, the results of this work could be improved in various ways. In this work, the choice was made to perform a binary classification task, and we proved to be able to accurately predict the potential ability of players. The results could be made more valuable in future work by using the PA value as the prediction target, without using a threshold. The goal will then become to predict the exact PA value, where the deviation from the target value will determine the model performance. The suggested future research could be done using regression models.

There could also be more focus on feature selection in future work. In this analysis, almost all FM attributes were used to see which ones were the most important for model prediction. These attributes could be used to see if an accurate prediction could be made with minimal information. The assumption can be made that with only balance and technique as attributes, the models can also predict the potential ability accurately.

6 CONCLUSION

The DT, LR and SVM classification models are capable of predicting players having high potential in the Football Manager 2020 database. The SVM model proved to be the most accurate on the original dataset, and the Decision Tree was the most accurate after balancing classes. After the extraction of feature importance, the balance attribute turned out to be the most important attribute for the prediction. The prediction accuracy in the top five leagues was higher than in the bottom five leagues, and different attributes were important for these predictions. The class imbalance was the biggest challenge during these predictions, and in future research, this problem could be avoided by performing a regression task instead of classification.

REFERENCES

Ackerson, K. (2022). *Football League Rankings.* Retrieved from https://www.globalfootballrankings.com/

Bassey, P. (2019, 09). *Logistic Regression Vs Support Vector Machines (SVM).* Retrieved from https://medium.com/axum-labs/logistic-regression-vs-support-vector-machines-svm-c335610a3d16

Beckmann, J., & Beckmann-Waldenmayer, D. (2019). Talent development in youth football. In *Football psychology* (pp. 283–296). Routledge.

Bianco, M. J., Gerstoft, P., Traer, J., Ozanich, E., Roch, M. A., Gannot, S., ... Li, W. (2019). Machine learning in acoustics: a review. *arXiv preprint*

arXiv:1905.04418.

Bleaney, R. (2017, 03). *Football Manager computer game to help Premier League clubs buy players.* Retrieved from https://www.theguardian.com/football/2014/aug/11/football-manager-computer-game-premier-league-clubs-buy-players

Bonney, G. E. (1987). Logistic regression for dependent binary observations. *Biometrics*, 951–973.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321–357.

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, *19*(1), 15–18.

development team, T. P. (2020, February). *pandas-dev/pandas: Pandas.* Zenodo. Retrieved from https://doi.org/10.5281/zenodo.3509134 doi: 10.5281/zenodo.3509134

Gagné, F. (2004). Transforming gifts into talents: The dmgt as a developmental theory. *High ability studies*, *15*(2), 119–147.

Hocquet, A. (2016). Football manager: Mutual shaping between game, sport, and community. *Journal of media studies and popular culture= Revue d'études des médias et de culture populaire*, *6*(Special Issue).

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, *30*.

Kovács, G. (2019). An empirical comparison and evaluation of minority oversampling techniques on a large number of imbalanced datasets. *Applied Soft Computing*, *83*, 105662.

Kumar, N. (2022, 04). *Understanding Logistic Regression.* Retrieved from https://www.geeksforgeeks.org/understanding-logistic-regression/

Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, *18*(17), 1-5. Retrieved from http://jmlr.org/papers/v18/16-365

Lima, E. M. R., Tertuliano, I. W., Aroni, A. L., Machado, A. A., & Fischer, C. N. (2018). Saga football manager na gestão esportiva: uma ferramenta tecnológica para monitorar jogadores promissores. *Lecturas: Educación Física y Deportes*, *23*(239), 2–12.

Mills, A., Butt, J., Maynard, I., & Harwood, C. (2012). Identifying factors perceived to influence the development of elite youth football academy players. *Journal of sports sciences*, *30*(15), 1593–1604.

Mohammed, R., Rawashdeh, J., & Abdullah, M. (2020). Machine learning with oversampling and undersampling techniques: overview study

and experimental results. In *2020 11th international conference on information and communication systems (icics)* (pp. 243–248).

Oktanto, C. (2021, Jul). *Football manager 2020 dataset.* Retrieved from https://www.kaggle.com/ktyptorio/football-manager-2020

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... others (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research*, *12*(Oct), 2825–2830.

Roy, A. (2021, 12). *A Dive Into Decision Trees - Towards Data Science.* Retrieved from https://towardsdatascience.com/a-dive-into-decision-trees-a128923c9298

Sarmento, H., Anguera, M. T., Pereira, A., & Araújo, D. (2018). Talent identification and development in male football: a systematic review. *Sports medicine*, *48*(4), 907–931.

Singh, B. D. (2020, 06). *Decision Trees: Parametric Optimization - Towards Data Science.* Retrieved from https://towardsdatascience.com/decision-trees-parametric-optimization-344354bd3db5

Sports Interactive. (2020, 07). *Developing Mental Attributes in Young Players | Wednesday Wisdom.* Retrieved from https://www.footballmanager.com/the-byline/developing-mental-attributes-young-players-wednesday-wisdom

Stuart, K. (2020, 04). *Why clubs are using Football Manager as a real-life scouting tool.* Retrieved from https://www.theguardian.com/technology/2014/aug/12/why-clubs-football-manager-scouting-tool

Tharwat, A. (2020). Classification assessment methods. *Applied Computing and Informatics*.

Transfermarkt. (2022, 05). *Pedro Gonçalves - Player profile 21/22.* Retrieved from https://www.transfermarkt.com/pedro-goncalves/profil/spieler/426620

UEFA.com. (2022, 02). *Country coefficients | UEFA Coefficients.* Retrieved from https://www.uefa.com/nationalassociations/uefarankings/country/#/yr/2022

Walker-Emig, P. (2017, 05). How real life teams are using Football Manager to target their next star player. *pcgamer*.

Yaqin, M. A., Ramadhan, M. Z., Jauhari, A. F., & Humami, A. G. (2019). Optimasi pemilihan posisi terbaik pemain muda pada game football manager 2018 dengan metode naïve bayes. *Prosiding SENIATI*, 59–65.

Yiğit, A. T., Samak, B., & Kaya, T. (2019). Football player value assessment using machine learning techniques. In *International conference on intelligent and fuzzy systems* (pp. 289–297).

Appendices

A FM ATTRIBUTES

Attribute	Explanation
Height	Height of player (in cm)
Weight	Weight of player (in kg)
Work Rate	Player's willingness to work to his full capacity
Vision	Player's ability to see a potential opening and spot an opportunity
Technique	The aesthetic quality of a player's technical game
Teamwork	Player can follow tactical instructions whilst working for and alongside his team-mates
Tackling	Player's ability to win the ball cleanly without conceding foul in such situations
Strength	Player's ability to exert his physical force on an opponent to his benefit
Stamina	Player's ability to endure high-level physical activity for extended periods of time
Positioning	Player's ability to read a defensive situation and position themselves accordingly
Penalty taking	Player's ability from the penalty spot
Passing	Player's ability to successfully find a team-mate with the ball
Pace	Player's top speed both on and off the ball
Off the Ball	Player's ability to move when not in possession of the ball
Natural fitness	How well a player stays fit when injured or not in training
Marking	Player's ability to stick close to his direct opposition in defensive situations
Long throws	Player's ability to throw the ball long, often in attacking situations
Long shots	Player's prowess when shooting from outside penalty area
Leadership	Player's ability to influence player around them on the pitch
Jumping reach	The highest point that a player can reach with his head
Heading	Player's ability to head the ball
Free kick taking	Player's ability to strike a dead ball
Flair	Player's natural talent for the creative and unpredictable
First touch	Player's ability to control the ball immediately as it is passed into feet
Finishing	Player's ability to putt the ball in the back of the net when presented with a chance
Dribbling	Player's ability to run with the ball and manipulate it under close control
Determination	Player's commitment to succeed and do his very best on and off the pitch
Decision-making	Player's ability to make the correct choice both with and without the ball
Crossing	Player's ability to cross the ball accurately from wide areas
Corner taking	Player's ability to accurately take a corner
Concentration	Player's mental focus and attention to detail on an event basis
Composure	Player's steadiness of mind and ability to make intelligent decisions
Bravery	How committed player is, often putting themselves into risky situations
Balance	How well a player can stay on his feet, both on and off the ball
Anticipation	Player's ability to predict and react to events going on around them
Agility	How well player can start, stop and move in different directions at varying speed levels
Aggression	Player willingness to get stuck in, perhaps at expense of giving away more fouls
Aerial reach	Player's physical ability to challenge in aerial situations
Acceleration	How quickly a player can get to top of speed from standing start

Table 15: All attributes, used as features, from the dataset