Daviano Almeyda Kustaryanto - 3212610

A sign lit up at night

Description automatically generated

Prescriptive Analytics Challenge Report

Business Case : FIFA

Contents

[Introduction 2](#_Toc24842359)

[Assumption 2](#_Toc24842360)

[Data 3](#_Toc24842361)

[3](#_Toc24842362)

[Cleaning Data 4](#_Toc24842363)

[Removing and Changing Columns 4](#_Toc24842364)

[Analysis Report 6](#_Toc24842365)

[Result 8](#_Toc24842366)

[K Nearest Algorithm 8](#_Toc24842367)

[Accuracy Score and F1 Score of Potential Player and Attacking Rates 9](#_Toc24842368)

[Decision Tree Classifier Accuracy For Potential Player Accuracy 9](#_Toc24842369)

[Decision Tree Regressor For Predicting Potential Player 10](#_Toc24842370)

[Conclusion 10](#_Toc24842371)

[Appendix 11](#_Toc24842372)

# Introduction

Nowadays, Data has become the crucial part in professional world. Many company used data in order to make decision in the future therefore they won’t be making any wrong decision. Data also can be used to predict something is likely to happen in the future. Based on this statement, we believe that data can also predict which player who have high potential to be a star in the future since many club is willing to invest millions of euro in order to buy a talented player to increase their team value or to be sell into bigger club in the future. Therefore we believe that predicting player based on data is more effective rather than using traditional way.

In Business Proposal Document, we already talked about the general information of the project that we are going to do that cover purpose, benefit and goals. Meanwhile, this document is going will be explained how we do our analysis, what method that we are going to used and the result of the analysis.

# Assumption

First, we are going to assume that the data that we get from Kaggle is from an official data and we get the data from FIFA itself. Otherwise, this entire project will be illegal since we’re using data from a third-party that the source of the data is not very credible or possible a stolen data, hacked data and etc. We make an assumption that data that we get from Kaggle is legal and free to use since for us Data Ethics and Laws are very important.

Second, we are going to assume that the data that we get from Kaggle is based on real life observation from FIFA. Therefore, if there any mis value from the data that we get from Kaggle and the official data from FIFA, it’s not our fault.

Third, We are assuming that all professional football club is already a data-driven company, therefore predicting player based on data is already pretty common. Because, up until now many professional club is still using traditional way to recruit or scout a player by looking their performance directly.

Fourth, We are assuming that any result that we get from machine learning that we made, our clients will consider to recruit it and try to make their player fulfil their potential.

# Data

We assumed that the dataset is originally from FIFA that we retrieved from Kaggle. The dataset contains the attributes and profile of a player such as Name, Date of Birth, Speed and etc. The data set is contains 89 Columns and 18.159 rows however we’re not going to used all of them. The Rough data will look like the picture Below :

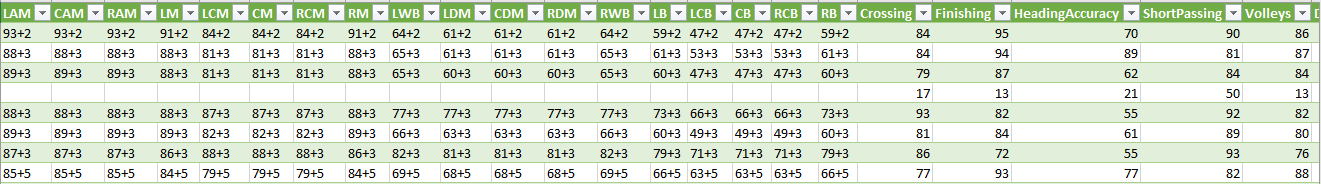
A screenshot of a cell phone

Description automatically generatedA screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generatedA screenshot of a cell phone

Description automatically generated

A screenshot of a computer

Description automatically generated

# Cleaning Data

### Removing and Changing Columns

A close up of a logo

Description automatically generatedIn this section we’re going to explain how we cleaned the dataset. Since the datasets contains many columns and thousands rows of data that we don’t need to do this project.

After we remove some unnecessary column, we are going to change the value and wage into numeric since the value from the dataset is messy.

A screenshot of a cell phone

Description automatically generated

On the dataset, we also found a anomaly column that called work rates, in order to make it more suitable we separated into two different column called Attacking Rates and defensive rates.

A screenshot of a cell phone

Description automatically generated

The last thing in this section, we are going to fix the body type in this datasets because we found some miss input in the datasets.

A screenshot of a social media post

Description automatically generated

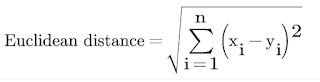
# Analysis Report

In order to complete this project, we are going to used some machine learning modules. Below is the type of machine learning that we used :

1. KNN Algorithm( K Nearest Neighbours)

The K-Nearest Algorithm is a supervised classification algorithm, It takes a lot marked points and use them to learn how to label another point. The reason we used this technique is this technique is quite common among data scientist, It is very simple to implemented and many major company used this technique to predict Something.

1. Euclidian Model



Euclidian is a method for machine learning where they spotted a common divisor of a number, where the largest number can divided without leaving a remainder. Euclidian method is pretty common when we used KNN as a method.

1. Decision Tree Classifier Accuracy

A screenshot of a cell phone

Description automatically generatedDecision Tree is a machine learning method where they break down a data set into smaller subset. We also use Decision Tree Classifier to test our accuracy score therefore we can make a comparison between accuracy score with decision tree and with KNN.

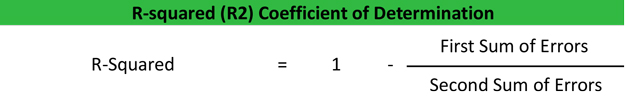
1. F1 Score

A screenshot of a cell phone

Description automatically generatedF1 Score is a tool to measure the accuracy of a test. F1 Score is often used in Information Retrieval department for measuring search, document classification, and query classification.

1. Decision Tree Regressor

R2 is a statistical Method who have a purposes to either predict a future of the outcomes or testing the hypothesis based on the machine learning. We try this method to test our hypothesis about the data



# Result

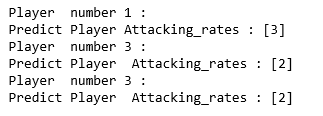
### K Nearest Algorithm

A screenshot of a cell phone

Description automatically generatedThe KNN algorithm was successfully applied in our machine learning. The algorithm show the potential overall of a player. Picture of proof are shown below :

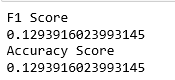
To get this result, we predict the potential Based on the age and current overall of the player and test it with dummy data. For Example : Player number 1 have overall 75 and the age is 17 so the potential he will received in the future is 88.

As a side goals of this project is to predict attacking rates of a player, we also going to shown the prediction as proof below :

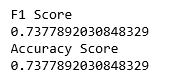


This algorithm is going to predict the attacking rates of a player, because attacking rates determine the energy of player to do an attacking move towards the enemies. The number on the picture above stands for High, Medium and Low where 3 equals to High, 2 equals to Medium and 1 equals to Low. This calculation is based on Finishing, Shot Power and Heading Accuracy. For instance : Player 1 have 85 Finishing, 76 Shot Power and 83 on Heading accuracy so he will get the High attacking rates.

### Accuracy Score and F1 Score of Potential Player and Attacking Rates



The F1 and Accuracy score that we applied for predicting potential player is both have 12,9 % accuracy score, which is quite low for an accuracy test. However predicting a player based on data is quite a challenge since there is also external factor that determined the potential of a player.



The F1 and Accuracy score that we used on Predicting attacking rates is quite goof where both of them have 73,7 % accuracy.

As you can see above the, F1 score determined that the prediction near perfection so the higher the score the higher chance of the prediction is coming true, meanwhile Accuracy score determined that that the prediction is nearly accurate therefore the higher the score the higher also it will happening in the future.

### Decision Tree Classifier Accuracy For Potential Player Accuracy

A close up of a logo

Description automatically generated

To make sure that we had the same result we going to used Decision tree accuracy test, which our prediction using this machine learning resulted 13% which is quite similar with The K Nearest Algorithm.

### Decision Tree Regressor For Predicting Potential Player

A screenshot of a cell phone

Description automatically generatedFor this section we are going to measure if there is correlation between the variable that we choose to predict the Player potential using R2.

The R2 calculation in this machine learning is 81,7% which is quite high where in the KNN and Decision Tree classifier the accuracy is rather low. We also calculated the MAE (Mean Absolute Error) where we get the calculation 2.07 which is quite good where the mean of the potential is 73 so MEA of the data is less than 10% of the actual mean.

# Conclusion

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

Based on these two diagram, I can make a conclusion that our machine learning is working perfectly since the result between KNN and Decision Tree have similar accuracy score. However the result that we get for Potential accuracy is quite disappointing since the accuracy is quite low. On the other hand our predictions of Attacking Rates show a good result since the accuracy looks as we expected.

Based on this conclusion, we want give a recommendation for FIFA to add few column as a factor to predict potential such as sport intelligence and Awareness therefore, it will help us to give them more accurate score.

# Appendix

Our complete data set will look like this :

|  |  |
| --- | --- |
| Number 17790 non | null int64 |
| ID 17790 non | null int64 |
| Name 17790 non | null object |
| Age 17790 non | null int64 |
| Photo 17790 non | null object |
| Nationality 17790 non | null object |
| Flag 17790 non | null object |
| Overall 17790 non | null int64 |
| Potential 17790 non | null int64 |
| Club 17557 non | null object |
| Club Logo 17790 non | null object |
| Value 17790 non | null object |
| Wage 17790 non | null object |
| Special 17790 non | null int64 |
| Acceleration 17790 non | null int64 |
| Aggression 17790 non | null int64 |
| Agility 17790 non | null int64 |
| Balance 17790 non | null int64 |
| BallControl 17790 non | null int64 |
| Body Type 17790 non | null object |
| CAM 15806 non | null object |
| CB 15806 non | null object |
| CDM 15806 non | null object |
| CF 15806 non | null object |
| CM 15806 non | null object |
| Composure 17790 non | null int64 |
| Contract Valid Until 15800 non | null object |
| Crossing 17790 non | null int64 |
| Curve 17790 non | null int64 |
| Dribbling 17790 non | null int64 |
| FKAccuracy 17790 non | null int64 |
| Finishing 17790 non | null int64 |
| GKDiving 17790 non | null int64 |
| GKHandling 17790 non | null int64 |
| GKKicking 17790 non | null int64 |
| GKPositioning 17790 non | null int64 |
| GKReflexes 17790 non | null int64 |
| HeadingAccuracy 17790 non | null int64 |
| Interceptions 17790 non | null int64 |
| International Reputation 15811 non | null float64 |
| Jersey Number 15800 non | null float64 |
| Joined 14629 non | null object |
| Jumping 17790 non | null int64 |
| LAM 15806 non | null object |
| LB 15806 non | null object |
| LCB 15806 non | null object |
| LCM 15806 non | null object |
| LDM 15806 non | null object |
| LF 15806 non | null object |
| LM 15806 non | null object |
| LS 15806 non | null object |
| LW 15806 non | null object |
| LWB 15806 non | null object |
| Loaned From 1258 non | null object |
| LongPassing 17790 non | null int64 |
| LongShots 17790 non | null int64 |
| Marking 17790 non | null int64 |
| Penalties 17790 non | null int64 |
| Position 17778 non | null object |
| Positioning 17790 non | null int64 |
| Preferred Foot 17790 non | null object |
| RAM 15806 non | null object |
| RB 15806 non | null object |
| RCB 15806 non | null object |
| RCM 15806 non | null object |
| RDM 15806 non | null object |
| RF 15806 non | null object |
| RM 15806 non | null object |
| RS 15806 non | null object |
| RW 15806 non | null object |
| RWB 15806 non | null object |
| Reactions 17790 non | null int64 |
| Real Face 17790 non | null object |
| Release Clause 16292 non | null object |
| ST 15806 non | null object |
| ShortPassing 17790 non | null int64 |
| ShotPower 17790 non | null int64 |
| Skill Moves 17790 non | null int64 |
| SlidingTackle 17790 non | null int64 |
| SprintSpeed 17790 non | null int64 |
| Stamina 17790 non | null int64 |
| StandingTackle 17790 non | null int64 |
| Strength 17790 non | null int64 |
| Vision 17790 non | null int64 |
| Volleys 17790 non | null int64 |
| Weak Foot 17790 non | null int64 |
| Work Rate 17790 non | null object |