YILDIZ TECHNICAL UNIVERSITY
FACULTY OF CHEMICAL AND METALLURGICAL ENGINEERING
DEPARTMENT OF MATHEMATICAL ENGINEERING

BACHELOR THESIS

EXAMINATION OF FIFA 18 DATA WITH DATA MINING TECHNIQUES

Project Supervisior: Associate Prof. Dr. Nilgün GÜLER BAYAZIT

13053001 Selchuk HADZHAAHMED
(Turkish Name: Selçuk HACIAHMET)

İstanbul, 2018

CONTENTS

Page

LIST OF FIGURES

PREFACE

In this study, the process of extracting meaningful information from data sets, which has become an indispensable part of today's world, has been processed with data mining techniques. By using many techniques in data mining, meaningful information was tried to be revealed by using the data of the football player and football player characteristics in the video game called FIFA 2018.

I would like to thank my project and mental advisor, Assist. Assoc. Dr. Nilgün GÜLER BAYAZIT.

I would like to thank my dear family and Öge Tuana KALAYCIOĞLU for their financial and moral support during my studies.

İstanbul, 2018 **Selchuk HADZHAAHMED**

ABSTRACT

In this thesis, data mining techniques, which are widely used in data mining, have been implemented using the Python 3.6 programming language. The concept of "Video Games Analysis", which analyzes the video games in thesis, which techniques produce more efficient and meaningful information and which gradually enters our life in today's world, has been examined.

Different data mining techniques and the differences between these techniques were examined in relation to the average score of the FIFA 2018 video game data set previously obtained with the help of the libraries that the Python 3.6 programming language included data mining techniques and the game scores in the game.

1. INTRODUCTION

Data mining is the search for correlations that can provide meaningful information from large-scale data or to make predictions about the future from large data piles. Computer programs or programming languages are used to process this big data. In order to use raw information in data mining, it needs to go through some steps. These steps can be data cleaning, data integration, data transformation, data selection, and pattern evaluation.[1] In this study, some of these data cleaning steps and the Python 3.6 programming language were used in the data processing part. The data mining techniques used are linear regression, polynomial regression, LASSO, Support Vector Regression (SVR).

In this study, the techniques performed on a dataset and the differences of these techniques, the inference of meaningful information on the average scores and transfer values of football players on the dataset of the FIFA 2018 game, and the estimation of the transfer value of the player with any average score based on these inferences with different data mining techniques. is being examined. In addition, information about the world applications of data mining on video games, Python 3.6 programming language and data mining techniques used are given.

2. VIDEO GAMES ANALYSIS AND APPLICATIONS

Video game analysis is the extraction of meaningful information or information from the dataset of the game to be analyzed. This analysis can be done by anyone, from the video game player to the employees of the company that created the video game. Video game analysis is usually done using data mining techniques. It is generally referred to as "Game Data Mining" or "Data Mining for Games" in the literature.

When we look at the past in video game analysis, there are not as many examples as in other fields. Leonardo Cotta, Pedro O.S. Vaz de Melo, Fabrício Benevenuto and Antonio A.F. In the study titled "Football Analysis using the Data of the FIFA Football Video Game" prepared by Loureiro, the differences between the Brazilian and German National Teams in 2014 and the different football style of the FC Barcelona team from the other teams were examined.[2]

In 2015, Filip Johansson and Jesper Wikström's thesis named "Result Predictions of Replay Mining in DOTA 2", which emerged in the joint work, was studied on predicting which team will win in the next matches by using data mining techniques over the replays of the matches played in the DOTA 2 game. [3]

In 2017, Peter Braun, Alfredo Cuzzocrea, Timothy D. Keding, Carson K. Leung, Adam G.M. In the study titled "Game Data Mining: Clustering and Visualizing Online Game Data in Cyber-Physical Worlds" jointly conducted by Padzor and Dell Sayson, game data was visualized and some inferences were made with the help of clustering, which is one of the data mining techniques of the game called Overwatch.[4]

3. DATA MINING TECHNIQUES USED ON FIFA 2018 DATA

3.1. Linear Regression

Regression analysis is the process of explaining the relationships between a dependent variable and an independent variable (simple regression) or more than one independent variable (multiple regression) with mathematical equations. In regression analysis, if the relationship between the variables is linear, it is called linear regression, if not, it is called nonlinear regression. [5]

Linear regression analyzes the relationship between the observed responses in the dataset and the predicted responses with the linear approximation. [6]

An example of a linear regression model is shown below.

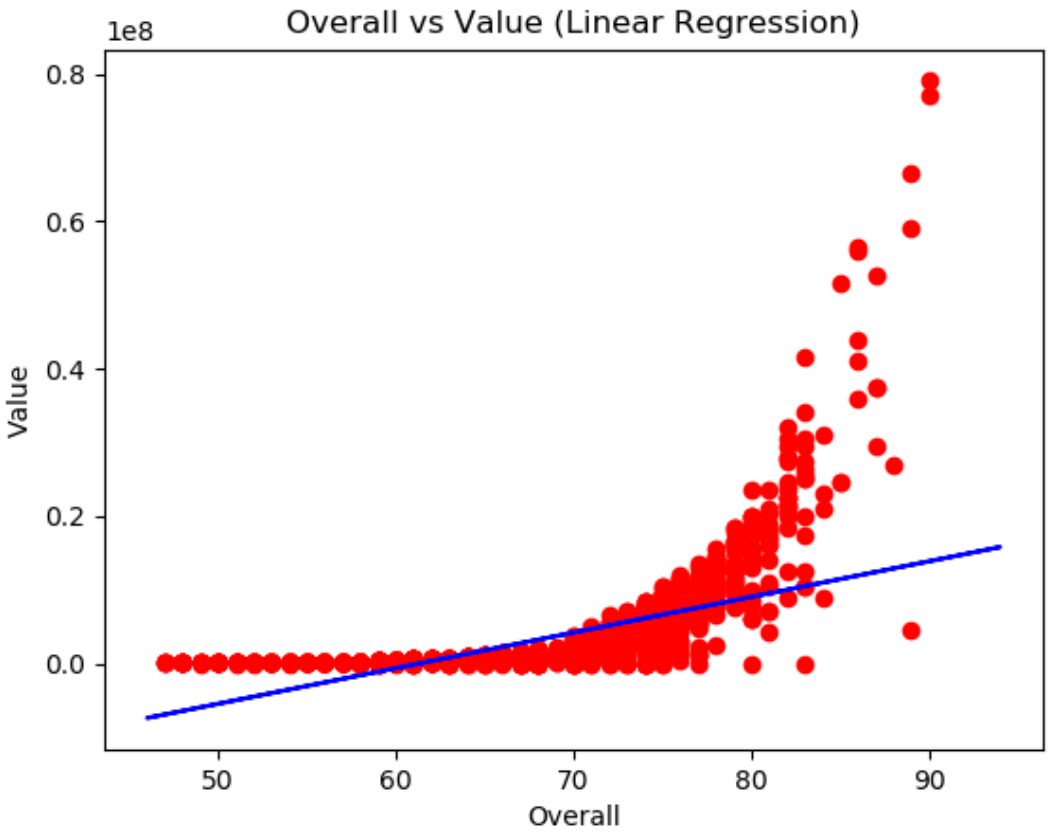


Figure 1: Linear Regression Model Comparing Player Average Score to Player Market Value on FIFA 2018 Dataset

3.2. Polynomial Regression

This regression is a regression often used in machine learning. The nonlinear function is trained with the data it has and creates a polynomial curve. This regression works just as fast as linear regression, allowing it to be trained and studied over a much wider range of data.[7]

An example of a polynomial regression model is shown below:

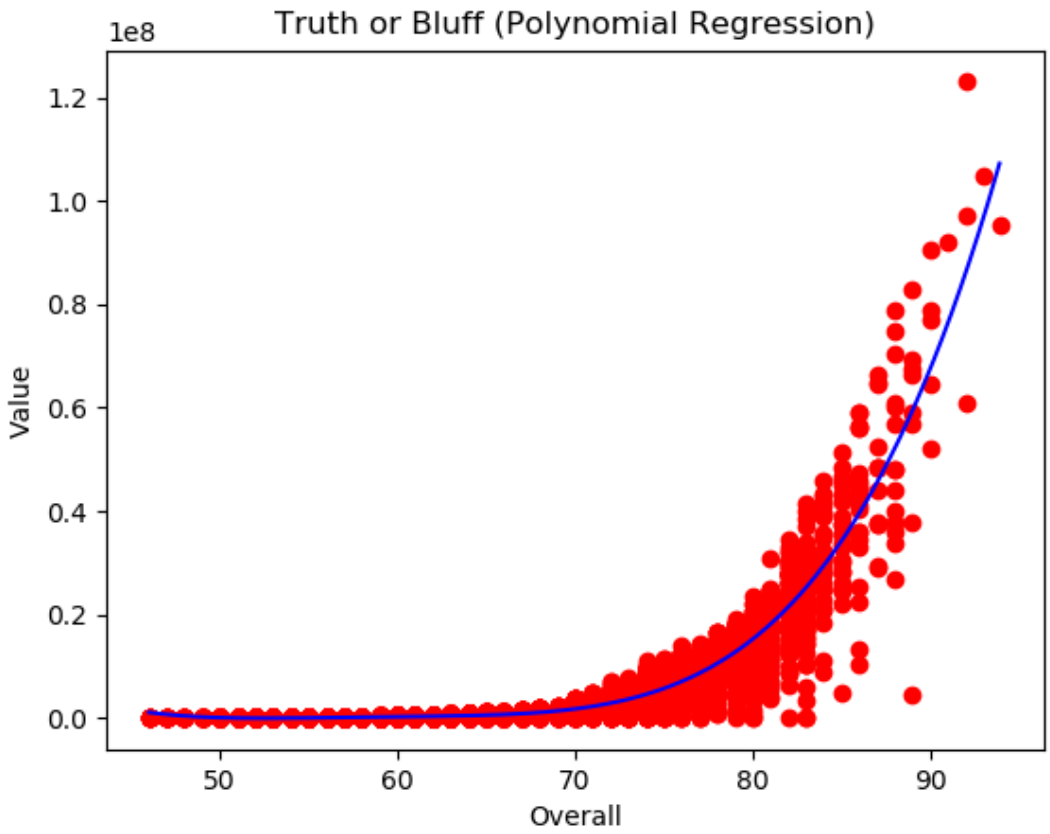


Figure 2: Polynomial Regression Model Comparing Player Average Score to Player Market Value on FIFA 2018 Dataset

3.3. LASSO

Lasso is a linear model that estimates sparse coefficients. In some contexts, it can be made more useful by effectively reducing the number of variables on which the given solution depends, due to the tendency to prefer solutions with fewer parameter values. Therefore, Lasso and its derivatives have a very important place in the compressed sensing field.[8]

The visual model of Lasso is shown below.

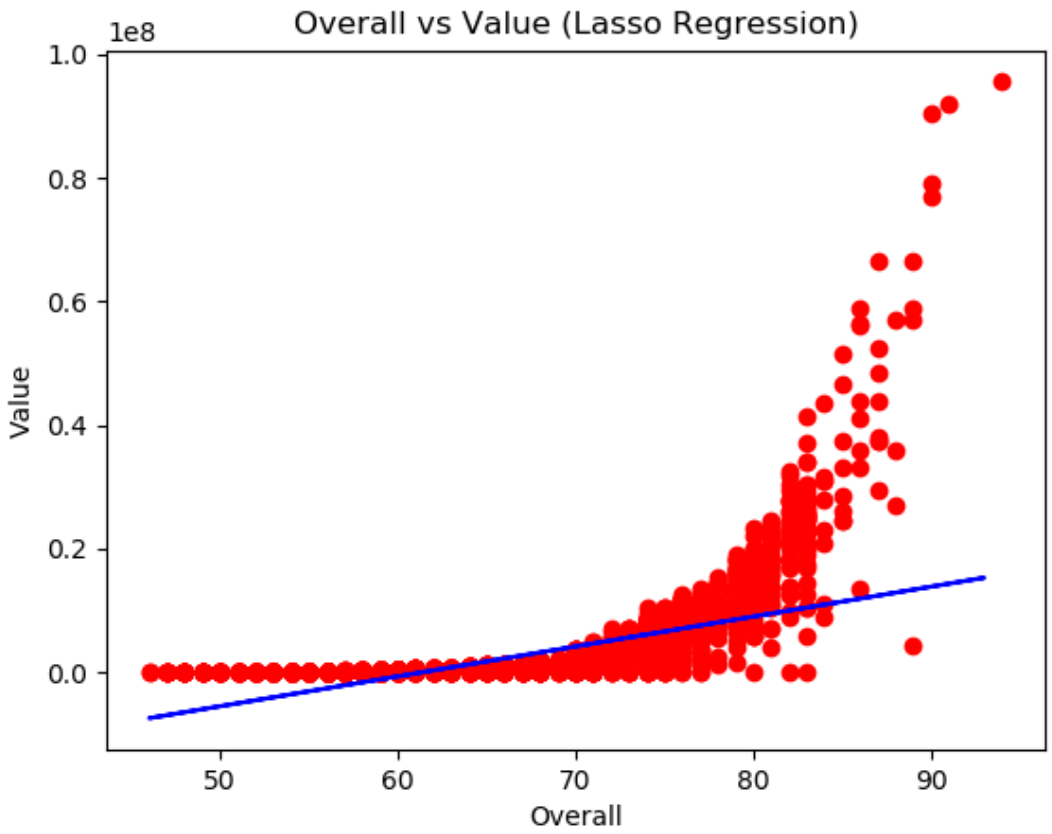


Figure 3: Lasso Regression Model Comparing Player Average Score to Player Market Value on FIFA 2018 Dataset

3.4. Support Vektor Regression (SVR)

The Support Vector Classification method has been extended to solve regression problems. This expansion method is called Support Vector Regression.

The model produced by support vector classification depends only on a subset of the training data. Because the cost function, which plays a role in the creation of the model, is not concerned with the training points below the linear line. Similarly, the model produced by Support Vector Regression depends only on the data of the training set. Because the cost function used in the creation of the model ignores any training data close to the model estimation.[9]

An example of a Support Vector Regression model is shown below.

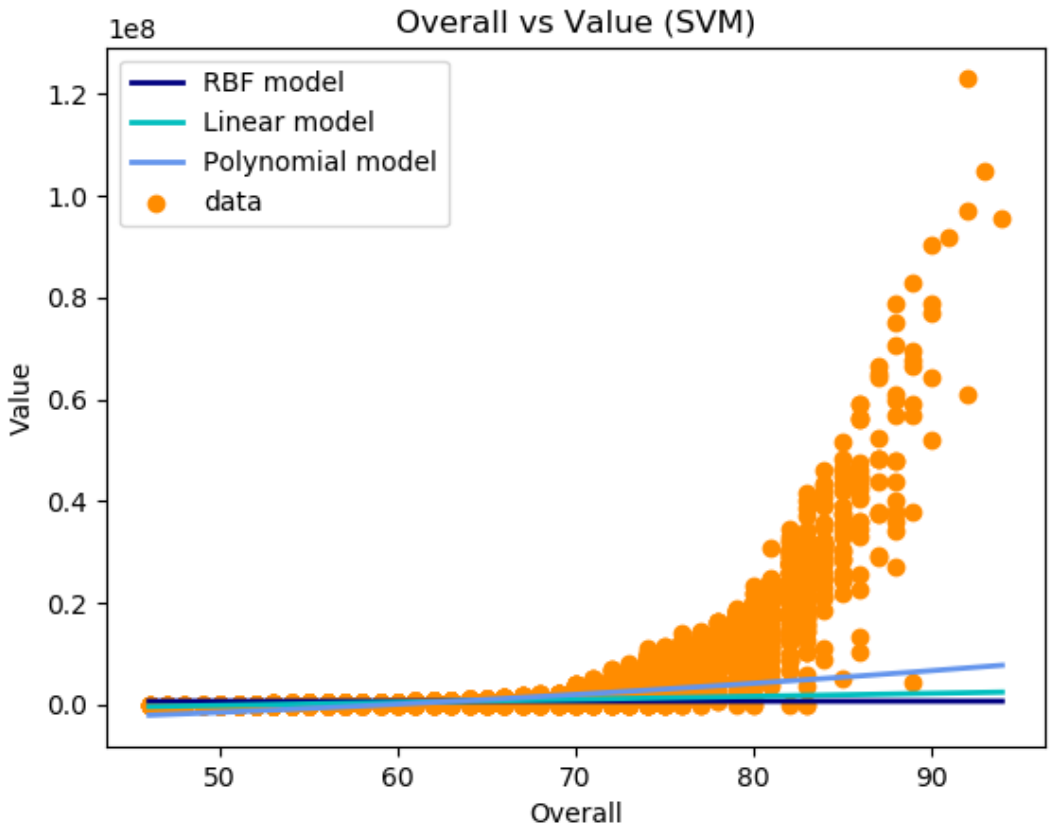


Figure 4: Support Vector Regression Model Comparing Player Average Score to Player Market Value on FIFA 2018 Dataset

4. APPLICATION OF ANALYSIS OF FIFA 2018 DATA WITH DATA MINING TECHNIQUES

4.1. What is Python?

Python is a programming language developed by Dutch programmer Guido Van Rossum. Python, the development of which started in 1990; Compared to programming languages such as C and C++, results such as:

- Python is easier to learn.
- It shortens the program development process. In other words, it is written faster than other programming languages.

- Unlike other programming languages, it does not need any compiler programs.
- Compared to other programming languages, it has a more readable and cleaner coding syntax.

Python programming language is a programming language that has increased its users in recent years, due to many other items such as the above items. In addition, Python is integrated into Java and .NET virtual machines as well as running on Windows, Linux/Unix and Mac-OS. Another important feature of Python is that it is free and free software with open source code license. It is a programming language that supports multiple programming paradigms, such as object-oriented programming with Python, functional or structural programming. However, it is also used for different purposes, from programs that can be packaged as executable files to writing scripts.[10]

Figure 5: Python Logo

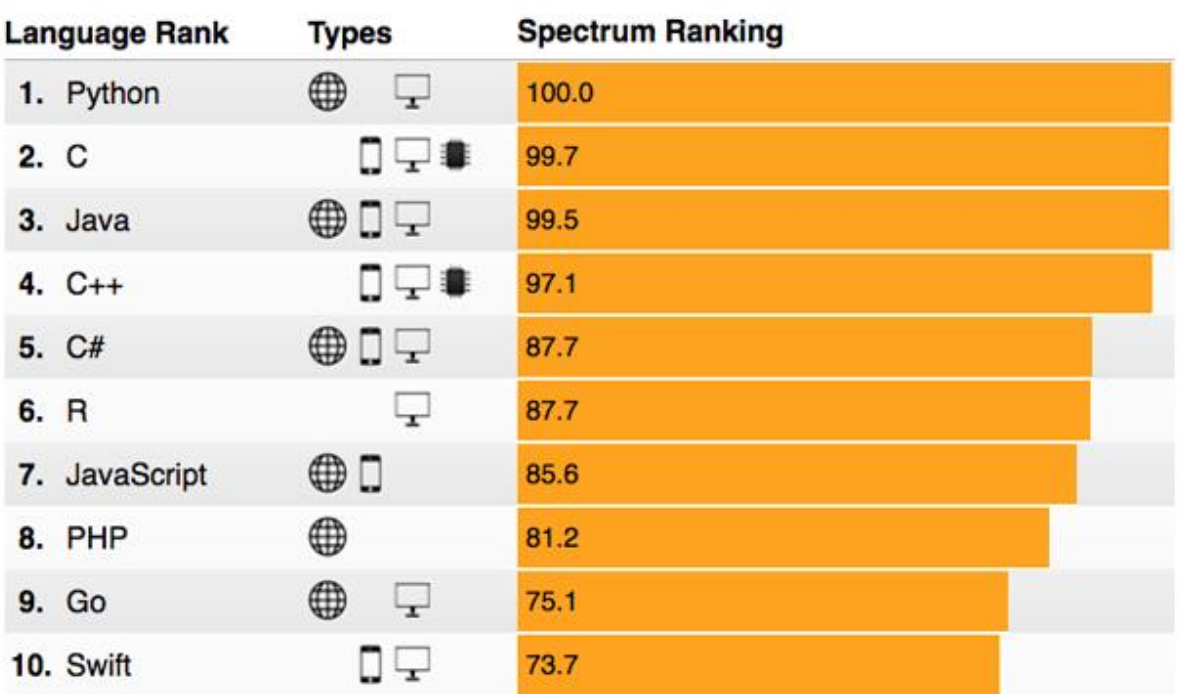


Figure 6 : The Most Popular Programming Languages of 2017[11]

4.2. THE IMPORTANCE OF PYTHON PROGRAMMING LANGUAGE IN DATA MINING

Python programming language has found its place among analysis languages and tools such as R, MATLAB, SAS, Stata in data analysis, interactive information processing and data visualization. The biggest factor in this is the support provided to libraries (such as pandas and scikit-learn) that have facilitated data analysis in recent years. The combination of these libraries and the ease of Python's software syntax make it an excellent choice as a primary language for building data applications.[12]

4.2.1. BASIC PYTHON LIBRARIES USED IN DATA MINING APPLICATIONS

4.2.1.1. NumPy

NumPy, short for Numerical Python, has long been the cornerstone of numerical computing in Python. It contains the data structures, algorithms, and libraries required for most scientific applications involving numerical data in Python. In addition to the above, NumPy includes the following items;

- ndarray as a multidimensional array object is fast and very efficient.
- Calculations are made with the elements in the indexes. In addition, mathematical operations can be performed between sequences.
- It has the means to read array-based datasets and write these data to disk.
- Linear algebra operations, Fouier transformations and random number generation can be done.[12]

4.2.1.2. Pandas

The Pandas library provides high-level data structures and functions designed to enable working with structured or tabular data quickly, easily and effectively. Since its emergence in 2010, it has made Python a powerful and productive data analysis environment.

The Pandas library blends NumPy's high-performance array-computation with the flexible data manipulation capabilities of spreadsheets and relational databases (such as SQL). It provides an advanced indexing functionality that makes it easy to reshape, slice, dice, perform sets, and select subsets of data. The Pandas library is one of the main focuses of data analysts, as data manipulation, preparation and cleaning is so important in data analysis.[12]

4.2.1.3. Matplotlib

Matplotlib is the most popular Python library for generating plots and other two-dimensional data visualizations. This library was originally created by John D. Hunter and is currently under development by a large team of developers. It is designed to create suitable space for visualization. While there are other visualization libraries in Python, Matplotlib is the most common. The reason for this is that it can integrate well with other libraries in general. [12]

4.2.1.4. SciPy

SciPy is a collection of packages that cater to scientific computing. With this library, the NumPy library provides a reasonable and efficient environment for many traditional scientific computing applications.[12]

4.2.1.5. Scikit-Learn

Since the beginning of this library project in 2010, Scikit-Learn has become the leading general purpose machine learning tool for Python programmers. Sub-modules and algorithms prepared by 1500 developers in 7 years are listed below;

- *Classification: SVM (Support Vector Machine), Nearest Neighbors, Random Forest, Logistic Regression, etc.*
- *Regression:* Lasso, Rigde Regression, etc.
- *Clustering:* K-Means, Spectral Clustering, etc.

- *Size Reduction:* PCA, Feature Selection, Matrix Factorization etc.
- *Model Selection:* Grid Search, Cross-validation, Metrics
- *Preprocessing:* Feature Extraction, Normalization

Along with the Pandas, Statsmodels, and IPython libraries, the Scikit-Learn library is critical for Python to be an efficient data science programming language.[12]

4.3. CLEARING FIFA 2018 DATA

The player data of the FIFA 2018 game is taken from the "FIFA 18 Complete Player Dataset" of the profile named AkshatUppal from kaggle.com. There are 17 000 pieces of data in the approximate dataset. In this dataset includes;

• Features of all players in the FIFA 2018 game,
• This is more than 70 attributes of each player,
• Flag pictures of players and countries,
• Position data played by the player,
• All game style attributes of players (such as Acceleration Goalkeeper Skill)
• Players' personal data (such as Nationality, Photo, Club Played for, Age, Transfer Value)
Features.

In this study, inferences were made with the average scores determined by the game of the player called the overall and the transfer values of the player called value.

While the player's overall score has a numeric value, since the value attribute contains both numeric and special characters in the dataset, it must be converted to a numeric value in order to be able to operate with this data.

In the image below, there is a limited section of the data showing the attributes of the players:

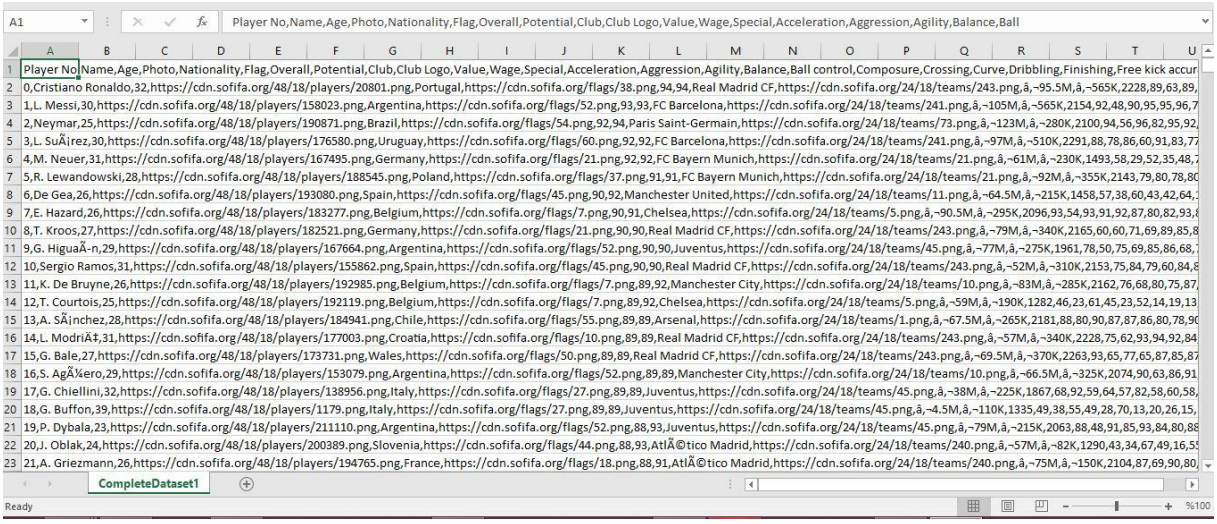


Figure 7: Specific Part of FIFA 18 Dataset ('Complete Dataset')

Since the data of the "Value" and "Wage" attributes in the 'Complete Dataset' file is not a numeric result in some players, these data have been replaced with an empty value with the following lines of code.

```
dataset["Value"] = dataset.Value.str.replace('[^\x00-\x7F]','')
dataset["Wage"] =  dataset.Wage.str.replace('[^\x00-\x7F]','')
```

Figure 8 : Replacing non-numeric data with a null value

```
dataset.loc[dataset.Value.str.contains('M') == True, "Value"]
dataset.loc[dataset.Value.str.contains('K') == True, "Value"]
dataset.loc[dataset.Wage.str.contains('K') == True, "Wage"] =
```

Figure 9 : Equations used when clearing 'M' and 'K' letters in Value and Wage values from data

```
= pd.to_numeric(dataset.loc[dataset.Value.str.contains('M') == True, "Value"].str.replace('M',''), errors='ignore')*1000000
= pd.to_numeric(dataset.loc[dataset.Value.str.contains('K') == True, "Value"].str.replace('K',''), errors='ignore')*1000
pd.to_numeric(dataset.loc[dataset.Wage.str.contains('K') == True, "Wage"].str.replace('K',''), errors='ignore')*1000
```

Figure 10 : The process of replacing the letters 'M' and 'K' in Value and Wage values with spaces and multiplying the numbers in front of them with the specified numbers

The rows shown in Figure 9 and Figure 10 are respectively synchronized with each other. The values in the 'Value' and 'Wage' columns in the dataset are not numeric.

In certain data mining techniques, data must be numerical in order to be applied. At the end of the data in the 'Value' and 'Wage' columns there are values according to the player's value. The letter 'M' with the number in front of it represents 'Million' and the letter 'K' with the number in front of it represents 'Thousand'. The above method was used to transform these data into numerical data.

After the data is converted into numerical form, a new file is created and saved. This registration method is as follows.

```
dataset.to_csv('CompleteDataset_updated.csv')
```

Figure 11: Saving the cleaned data to the new file

4.4. PROCESS OF SEPARATION OF DATA INTO TRAINED AND TEST SET

In this part of the study, all the data in the data set are divided into trained and test sets. This separation is done by machine learning. It is used to estimate how much a player's transfer value is based on any of their overall points, as it will also be kicked in the following divisions. In this study, the data is trained and divided into test sets in Python programming language as follows.

```
X = dataset.iloc[:, [6]].values
y = pd.to_numeric(dataset.loc[:, 'Value'].values)

# Splitting the dataset into the Training set and Test set
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.4, random_state = 0)
```

Figure 12: Separation of Data into Trained and Test Sets

In the above lines of code, it has been decided first of all what the x and y parameters will be. The 'Overall' value of all players in the 7th column in the x parameter, and the data in the 'Value' column whose data is cleared in the y parameter are integrated.

All the data in these x and y parameters are divided into clusters as X_train, X_test, y_train, y_test with the train_test_split method in the Scikit-Learn library. The data were clustered so that 6 out of 10 of all data were in the trained and 4 out of 10 in the test set.

4.5. APPLYING LINEAR REGRESSION TO THE DATA

After the data were divided into both clusters, the 'Linear Regression' method was applied using the trained x and y parameters as follows.

```
from sklearn.linear_model import LinearRegression
regressor = LinearRegression()
regressor.fit(X_train, y_train)
```

Figure 13: Application Linear Regression

After applying linear regression, the estimation mechanism of this end of regression is as follows.

```
y_pred = regressor.predict(X_test)
```

Figure 14: Prediction Process

In this estimation process, the y parameter, that is, the 'Overall' score, estimates how much the transfer value of a football player should be. The test set data of the parameter x was used in the estimation.

```
print(regressor.predict(94))
```

Figure 15: Line of code that outputs an estimated transfer value of a player with 94 'Overall' points

The code lines compiled after these estimation code lines make an estimation as follows.

```
[15802619.50682703]
```

Figure 16: Prediction Result

According to the linear regression, a football player with an 'Overall' score of 94 has a transfer value of approximately 16 million Euros.

The representation of the separated data sets and linear regression on the coordinate axis is as follows.

```
plt.scatter(X_test, y_test, color = 'red')
plt.plot(X_train, regressor.predict(X_train), color = 'blue')
plt.title('Overall vs Value (Linear Regression)')
plt.xlabel('Overall')
plt.ylabel('Value')
plt.show()
```

Figure 17: Representation of Linear Regression as Figure

The output of the above lines of code is as follows.

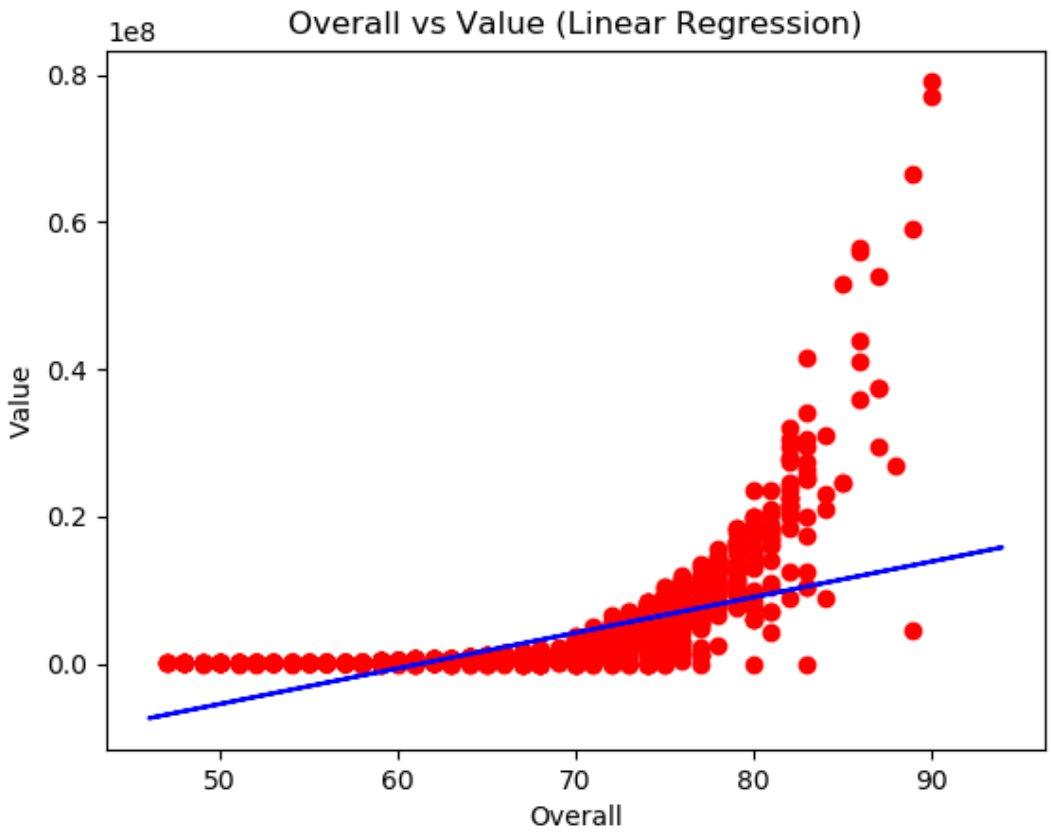


Figure 18: Linear Regression Figure

4.6. APPLYING POLINOM REGRESSION TO THE DATA

It is a regression performed using data that is trained and divided into test sets. Again, this regression is performed with the PolynomialFeatures method from the Scikit-Learn library.

```
from sklearn.preprocessing import PolynomialFeatures
poly_reg = PolynomialFeatures(degree = 4)
X_poly = poly_reg.fit_transform(X)
poly_reg.fit(X_poly, y)
lin_reg_2 = LinearRegression()
lin_reg_2.fit(X_poly, y)
```

Figure 19: Application of Polynomial Regression

The figure of the polynomial regression with a method similar to the visualization method used in linear regression is as follows.

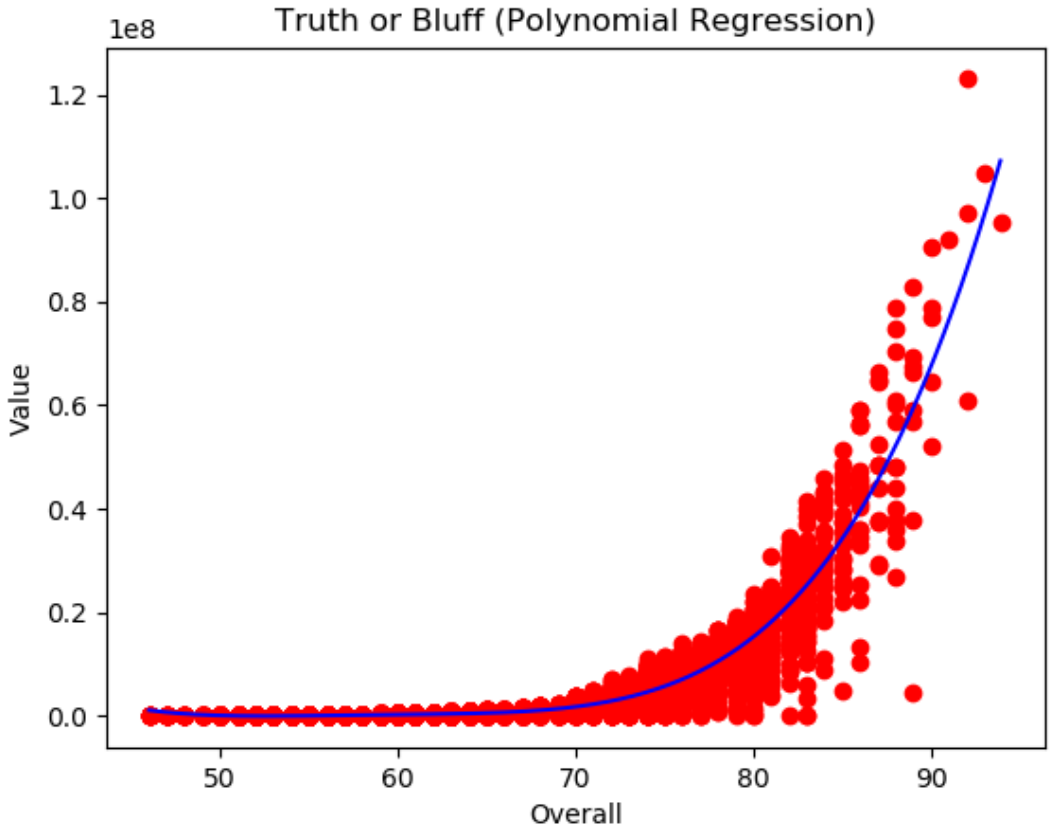


Figure 20: Polynomial Regression Figure

As will be noticed in this figure, the regression line is in the form of a curve formed by the data. For this reason, the estimation performance in polynomial regression is quite good, but it also gives results close to reality. The player with the highest 'Overall' score in the data in the FIFA 2018 data set is 'Cristiano Ronaldo' and has 94 points. The transfer value is 95.5 million Euros. In the estimation made in polynomial regression, this result is as follows. #

```
[1.08421296e+08]
```

Figure 21: Predicted transfer value of the player with 94 'Overall' points (108,42 Million Euro)

4.7. APPLICATION OF LASSO METHOD TO THE DATA

In the application, there is no obvious difference in visualization and coding as in the two regressions above. In this method, trained and test data of x and y parameters are used.

The difference between them is the values of the trained data and the result of the estimation. The figure of the LASSO method is as follows.

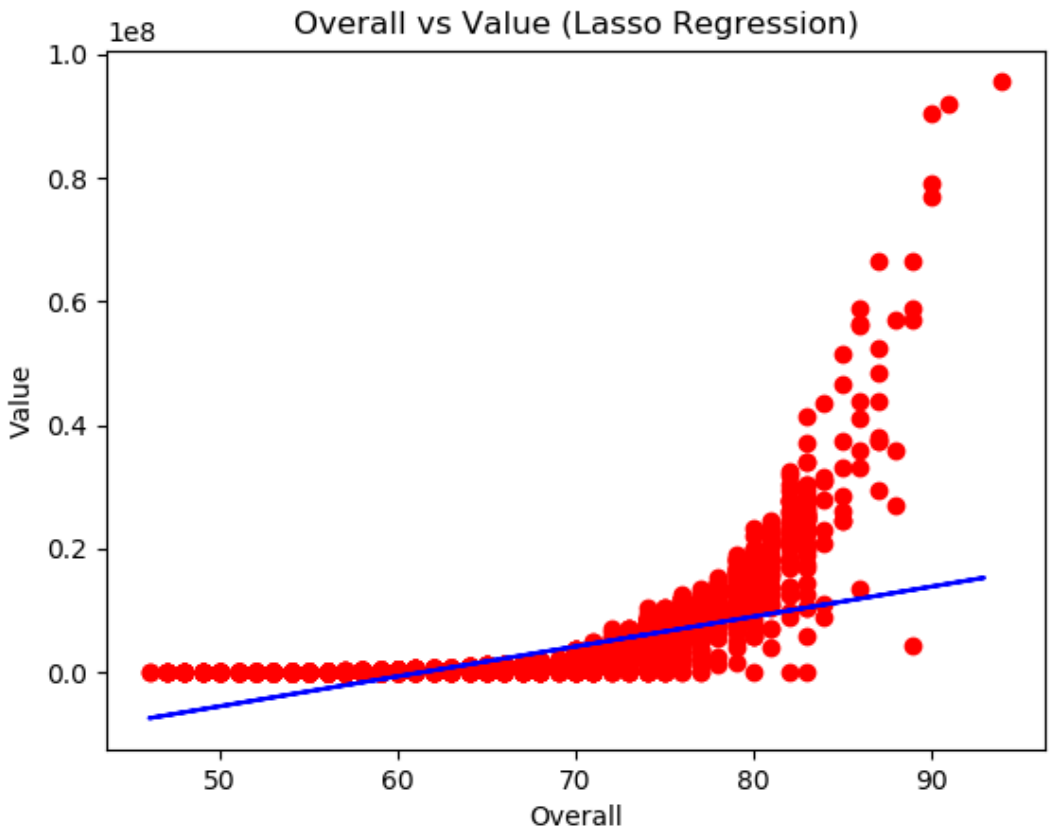


Figure 22: LASSO Method Figure

```
[15802619.45017958]
```

Figure 23: Prediction of the transfer value of the player with 94 points of the LASSO Method

4.8. APPLYING SVR TO THE DATA

It works on the data and subsets of these data, which are mostly trained in the SVR method. For this reason, it makes an estimation different from the estimations of the above regressions and its visualized version is different from the others. In the figure below, there is the 'Ridge Regression' line, linear line and polynomial curve used in data mining on SVM.

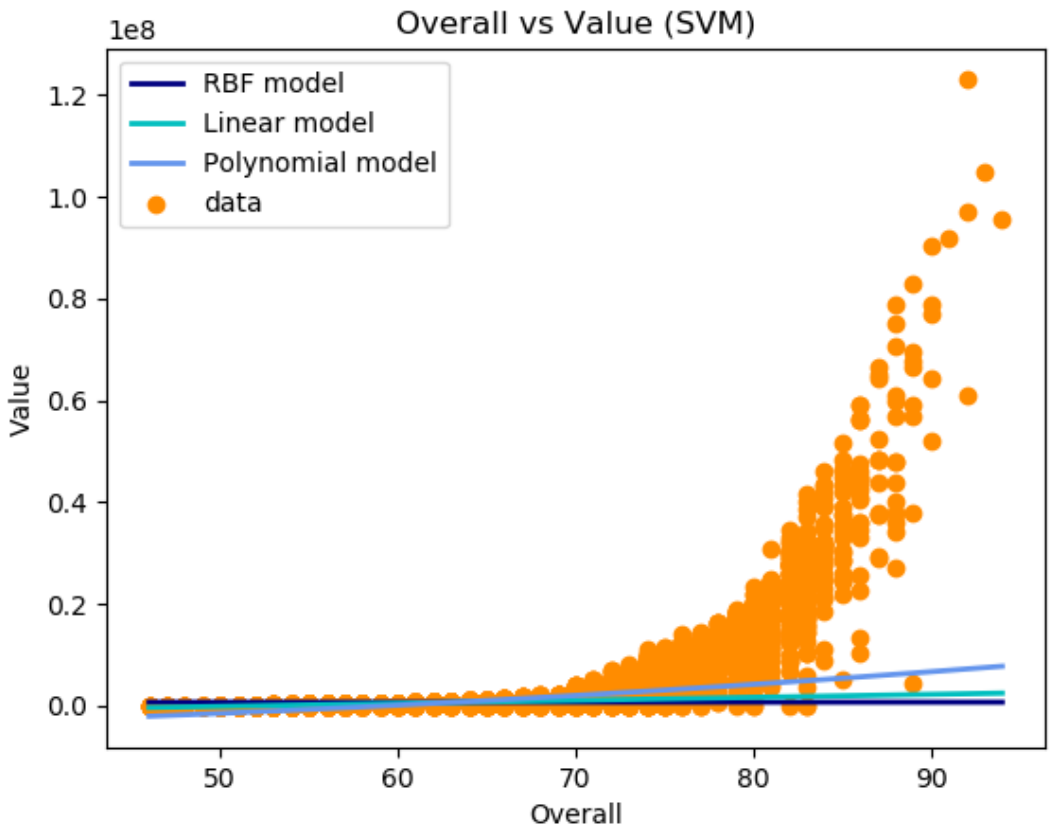


Figure 24: SVM Figure

The transfer value estimated by SVM for the 94-point player is as follows.

```
SVR Tahmin(94)= 699518.57680388
```

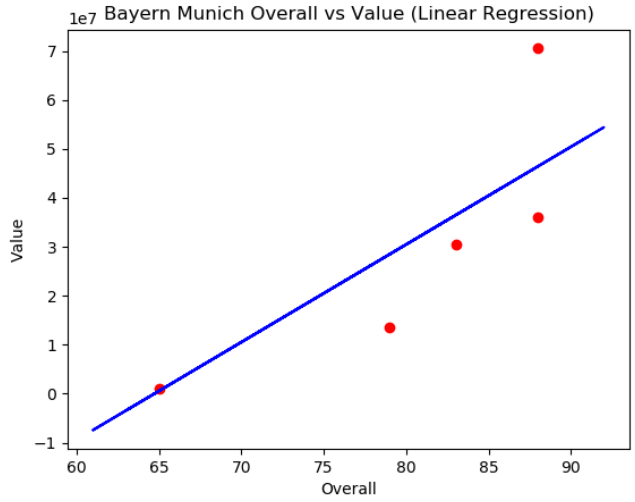
Figure 25: SVR Prediction

The data of the six big teams of Europe were analyzed and visualized studies were carried out. Linner and polynomial regressions were used in these six sets of data and estimations were made on these regressions. Images of these works can be found in the 'Appendix' section. It is possible to make some inferences from these images and predictions. As can be seen from the images, it is easy to find answers to questions such as which team has more transfer budget and which teams have players with higher scores.

5. CONCLUSION

Within the scope of the study, information about video game analysis was given, case studies in the world were mentioned, important points about data mining and Python programming language were mentioned, how Python libraries are used in data mining was understood and appropriate, as a result of these applications, "Data Mining Application on FIFA 2018 Data". In this project, some of the data mining techniques are mentioned, the definitions of the techniques used on this project and the differences between them are explained and the results obtained by using these techniques in the project are compared. In addition, this project will be continued in the future and different inferences will be obtained by using different data mining techniques on the same dataset.

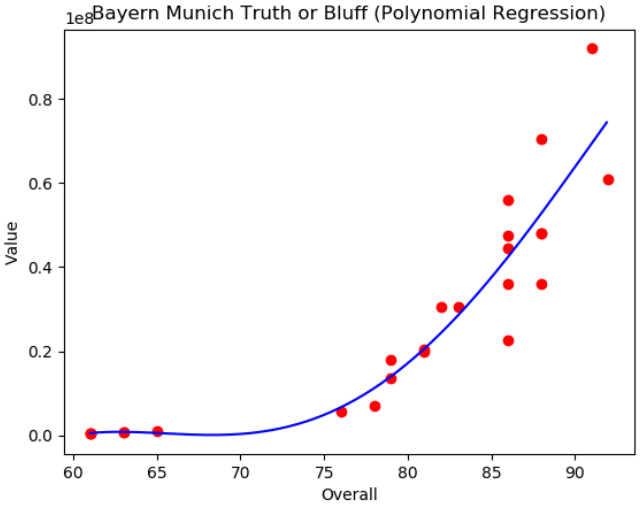
ATTACHMENTS



Attachment 1: Linear Regression Figure of Bayern Munich

```
[58407093.84460142]
```

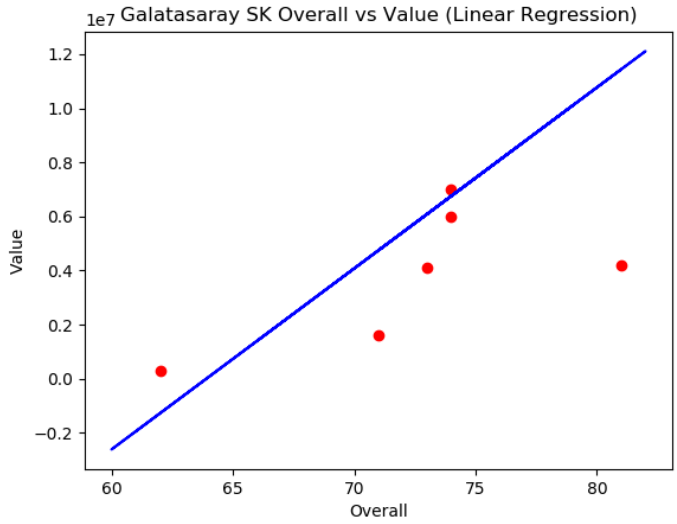
Attachment 2: Linear Regression Estimation of Bayern Munich Team Data (58.4 million Euro)



Attachment 3: Polynomial Regression Figure of Bayern Munich

```
[85919188.25260162]
```

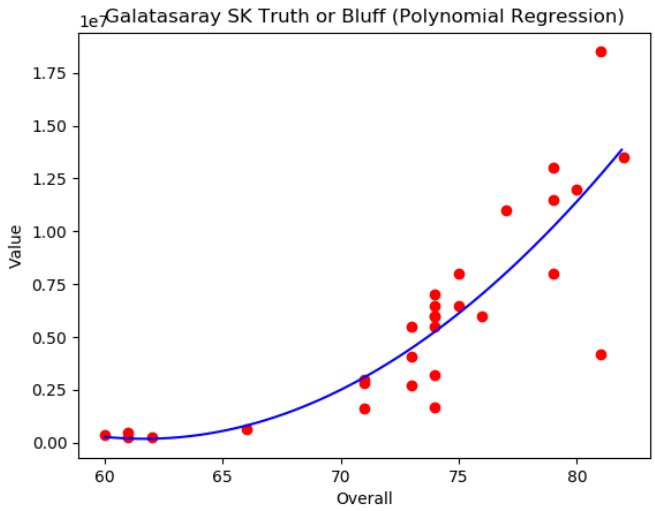
Attachment 4: Polynomial Regression Estimation of Bayern Munich Team Data (85.9 million Euro)



Attachment 5: Linear Regression Figure of Galatasaray

```
[16121500.8699212]
```

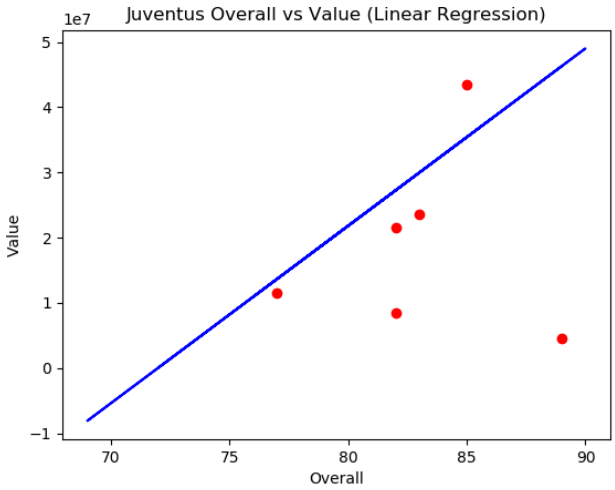
Attachment 6: Linear Regression Estimation of Galatasaray Team Data (16.1 million Euro)



Attachment 7: Polynomial Regression Figure of Galatasaray

```
[35129466.23903748]
```

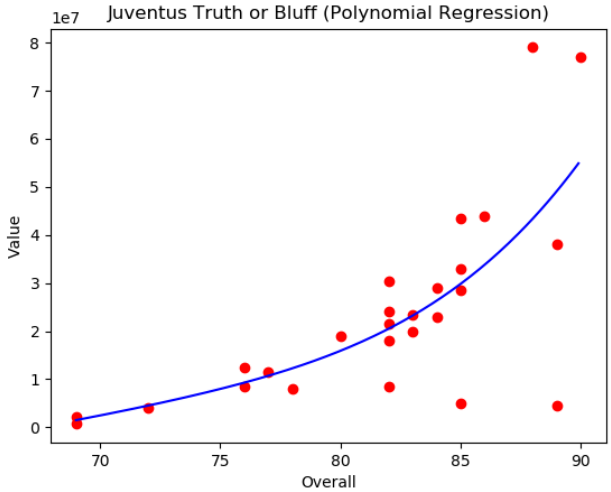
Attachment 8: Polynomial Regression Estimation of Galatasaray Team Data (35.1 million Euro)



Attachment 9: Linear Regression Figure of Juventus

```
[59882431.39979157]
```

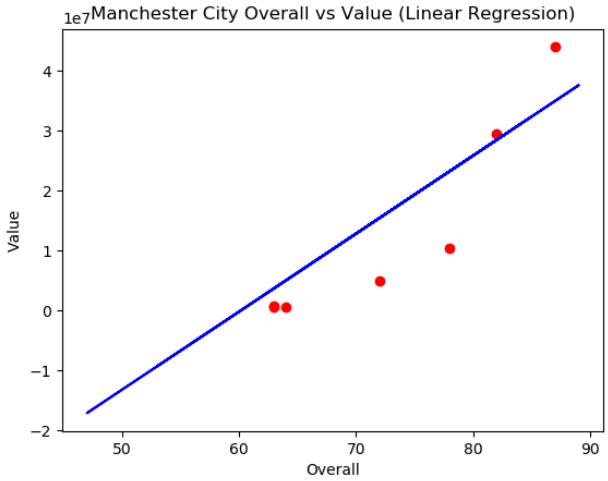
Attachment 10: Linear Regression Estimation of Juventus Team Data (59.8 million Euro)



Attachment 11: Polynomial Regression Figure of Juventus

```
[89962317.36191034]
```

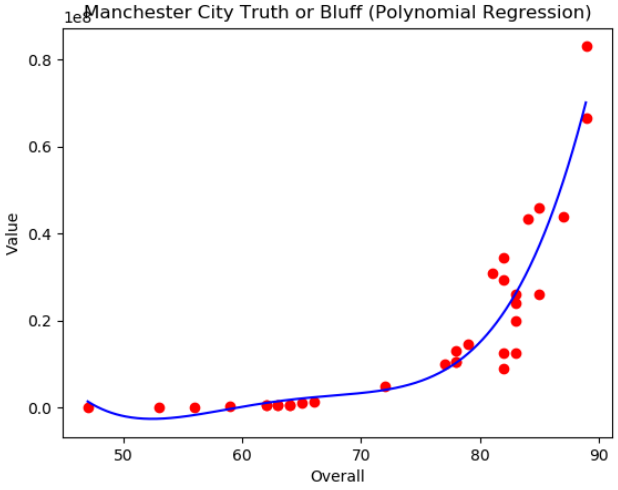
Attachment 12: Polynomial Regression Estimation of Juventus Team Data (89.9 million Euro)



Attachment 13: Linear Regression Figure of Manchester City

```
[44140073.18019313]
```

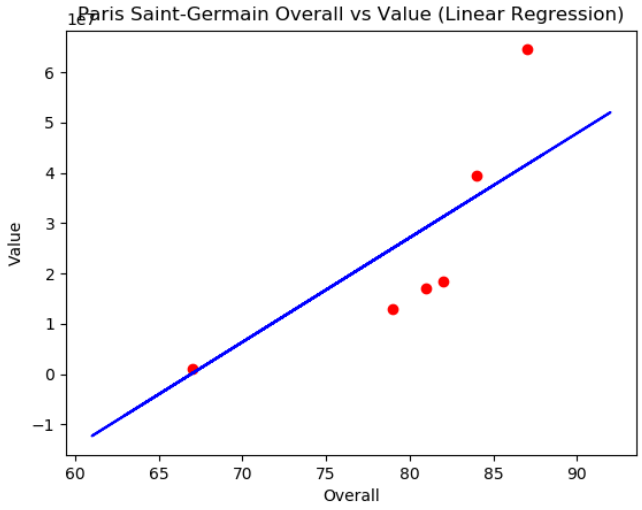
Attachment 14: Linear Regression Estimation of Manchester City Team Data (44.1 million Euro)



Attachment 15: Polynomial Regression Figure of Manchester City

```
[1.43511926e+08]
```

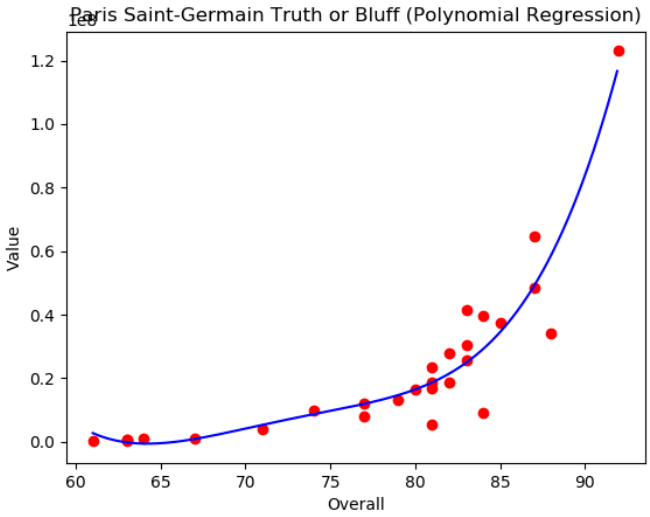
Attachment 16: Polynomial Regression Estimation of Manchester City Team Data (143.5 million Euro)



Attachment 17: Linear Regression Figure of Paris Saint-Germain

```
[56171139.69198197]
```

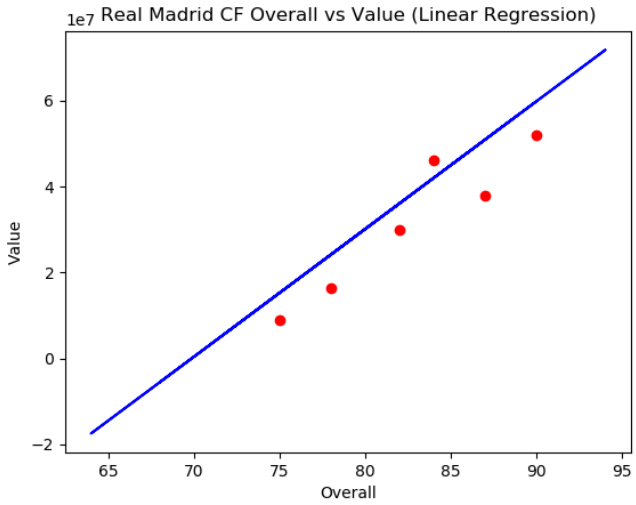
Attachment 18: Linear Regression Estimation of Paris Saint-Germain Team Data (56.1 million Euro)



Attachment 19: Polynomial Regression Figure of Paris Saint-Germain

```
[1.65550056e+08]
```

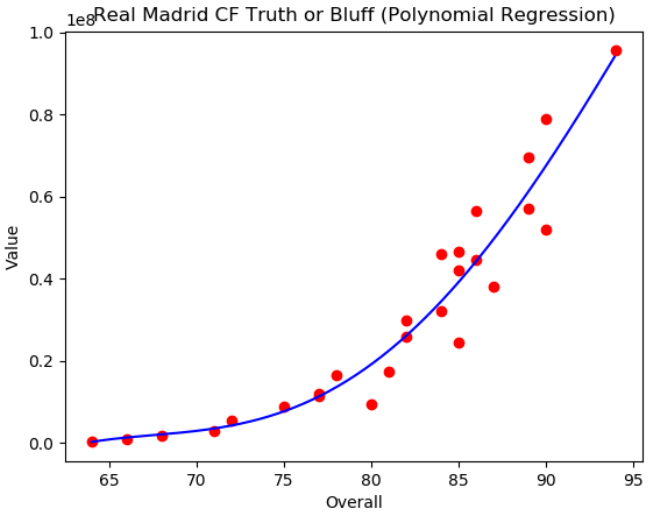
Attachment 20: Polynomial Regression Estimation of Paris Saint-Germain Team Data (165.5 million Euro)



Attachment 21: Linear Regression Figure of Real Madrid

```
[71757790.34772834]
```

Attachment 22: Linear Regression Estimation of Real Madrid Team Data (71.7 millon Euro)



Attachment 23: Polynomial Regression Figure of Real Madrid

```
[94647638.57109165]
```

Attachment 24: Polynomial Regression Estimation of Real Madrid Team Data (94.6 millon Euro)

BIBLIOGRAPHY

[1] Yıldız, Mümin - Modeling Customer Loyalty in Telecommunication Sector with Data Mining Techniques – 2017

[2] Leonardo Cotta, Pedro O.S. Vaz de Melo, Fabrício Benevenuto, Antonio A.F. Loureiro – Using Soccer Video Game Data for Soccer Analytics

[3] Result Prediction by Mining Replays in Dota 2, Filip Johansson, Jesper Wikström – 2015

[4] Game Data Mining: Clustering and Visualization of Online Game Data in Cyber-Physical Worlds,, Peter Braun, Alfredo Cuzzocrea, Timothy D. Keding, Carson K. Leung, Adam G.M. Padzor, Dell Sayson (2017)

[5] Spiers, N., Manktelow, B. Ve Hewitt, M. J.(2009), Practical Statistics Using SPSS, National Institude for Health Research NHS, England.

[6] http://scikit-learn.org/stable/modules/linear_model.html

[7] http://scikit-learn.org/stable/modules/linear_model.html#polynomial-regression-extending-linear-models-with-basis-functions

[8] http://scikit-learn.org/stable/modules/linear_model.html#lasso

[9] http://scikit-learn.org/stable/modules/svm.html#regression

[10] python.org

[11] https://spectrum.ieee.org/computing/software/the-2017-top-programming-languages

[12] Python for Data Analysis, 2nd Edition Data Wrangling with Pandas, NumPy, and IPython, William McKinney, 2017

RESUME

Name Surname : Selchuk HADZHAAHMED
Date of Birth : 04.08.1995
Place of Birth : Bakırköy/İSTANBUL
Highschool : 2009 – 2013 Bursa Turhan Tayan Anatolian Highschool
Place of Intern : Yapı and Kredi Bank Inc. Information Technologies Department – Gebze – (2016/4 weeks)
Yapı and Kredi Bank Inc. Information Technologies Department – Gebze – (2017/4 weeks)
Gamelab İstanbul Game Development – İstanbul – (2017/20 work days)
Encode IT Management and Consultancy – İstanbul – (2018 - 2020)