Regression and Classification Models

Student 1 Name: مريم ابراهيم رشدي ابراهيم

ID: 20191700623

Student 2 Name:محمد اسامة عبدالراضي عفيفي

ID: 20191700501

Student 3 Name:عمرو عزت حنفي محمد

ID: 20191700428

Student 4 Name:مريم عبد الرحيم وزيري

ID: 20191700629

Student 5 Name: مريم محمود عبد القادر محمد سويلم

ID:20191700635

Regression Report

Analysis:

- The data consists of
  - - multiple features for a player that affect his value in the game, such as his club's information (jersey number, position, etc.).
  - his national team's information (rating, position, etc.).
  - and normal required information about the player such as (age, positions, nationality, etc.).
- The first three columns in the file (id, name, full name) were dropped early on since they have no effect on the value.
- The fourth column (birthdate) was dropped at the start; due to the player's age existing in the file, which means it wouldn't have affected the player's value.
- Analysis is applied by dropping the features which have more than 10% missing values.
- Rows with missing values are then dropped which are approximately 300 rows (which is considered almost 2% of the whole data).

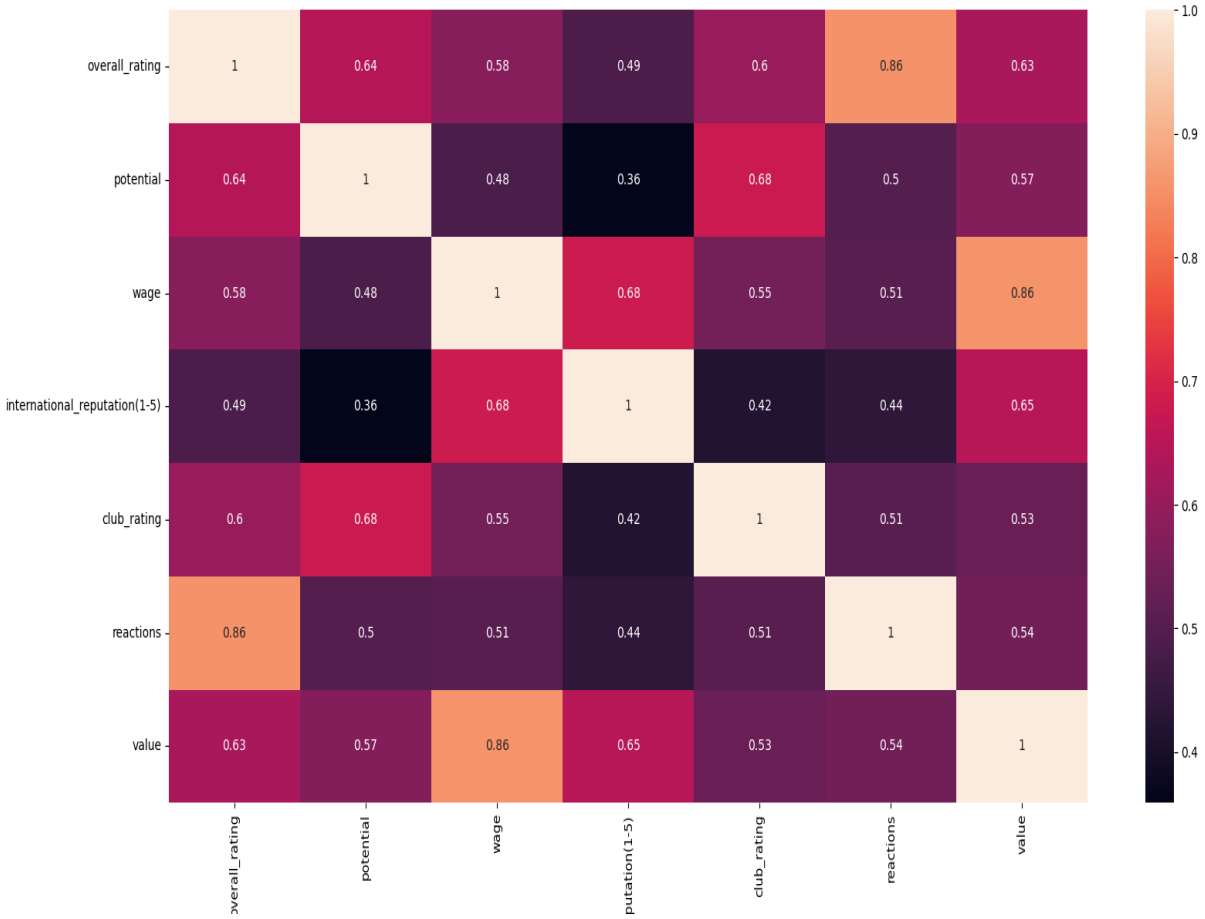
Preprocessing:

- Label encoding was applied to multiple features such as: nationality, preferred foot, body_type, club_team, club_position through a built-in function.
- Work_rate: We applied ordinal encoding to the column by manually mapping the values such that: High – 3, Medium – 2, Low – 1. The values are then summed since the players have two value per row.

- Positions: We applied One Hot Encoding by dividing each distinct values to columns.
- **Feature Selection:**

Feature Selection was applied depending on Correlation between features and value, we chose the top features depending on which of them has Correlation greater than 0.5.

The top features were (Overall rating, wage, potential, international reputation [1-5], club rating, reactions)



- ***Feature Scaling*** was applied to the columns of the top features to have their values in range of [0,1].

Regression:

- Polynomial regression and multi variable linear regression were applied on the data.
- The size of the test data was 20% while the training data was 80% as it was the most appropriate percentage to get best results.
- We changed the value of Random_state to (120) as it gave the least MSE with the two mentioned regression techniques.
- With Polynomial Regression we chose its degree = 3, it gave the best results, and any higher or lowers degrees gave us higher MSE and lower accuracy.
- The difference between each model is that Polynomial regression gave the best results since it had the least MSE and the highest R2 accuracy score (0.94) but it had higher training time. While with Linear Regression had a higher MSE and had an R2 accuracy of (0.79) with lower training time.

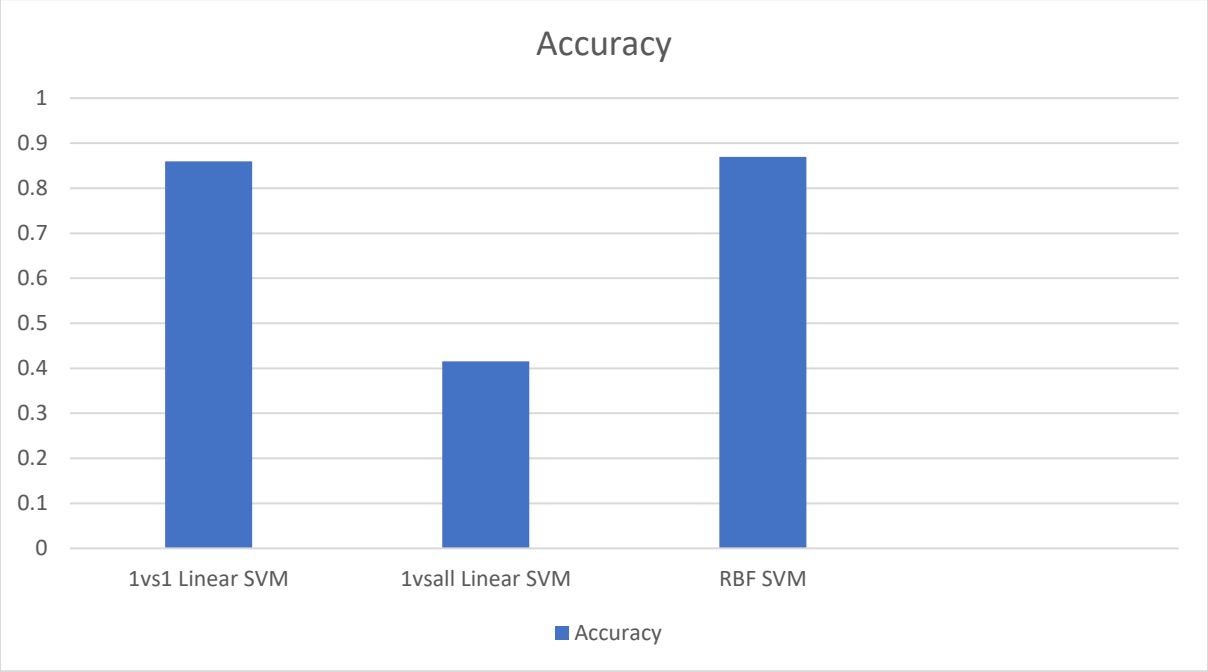
```
Mean Square Error of linear regression: 6766713461897.042
Model Accuracy with Linear Regression  0.7944535259580948
Time in linear regression =  0.0029239654541015625
Mean using Polynomial Regression:
Mean Square Error 1853189182192.0554
Model Accuracy with Polynomial Regression  0.9437073101621494
Time in polynomial =  0.04880046844482422
```

Conclusion:

- We thought it would be better to start by preprocessing all the features in the data, but later we realized it wasn't the best idea since many of those features wouldn't be used later the code, so we decided to start by analyzing the data to drop some features and work on preprocessing the remaining ones later.
- During the analysis phase, we chose to drop columns with 10% missing values at the start, then drop rows with any missing values which ended up being over (3000) rows which is 20% precent of the data, so we lowered the percentage of the missing value in the columns to 10%.

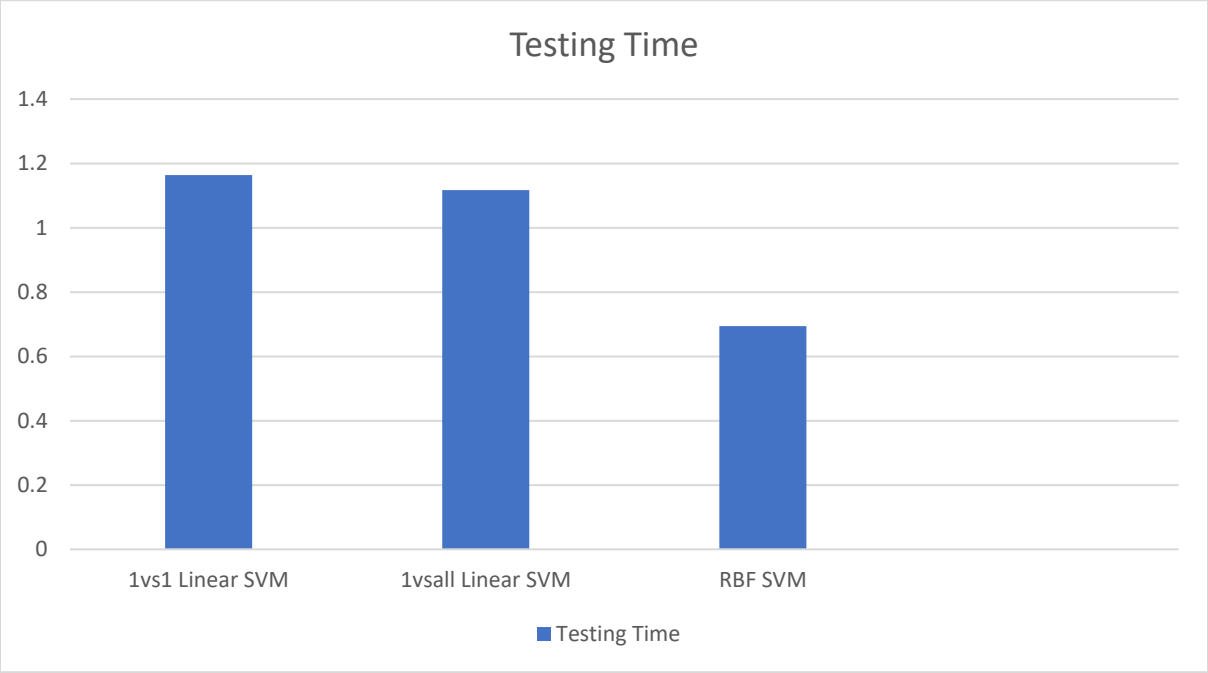
Classification Report

Classification Accuracy:



Total Training Time:

Total Testing Time:



Feature Selection:

The Player Value was changed and had to be label encoded but the feature selection process wasn't changed(using Correlation). The regression model had less number of top features than the classification model despite having the same correlation value.

New features were added in the top features such as (vision, composure, short_passing)

Hyperparameter Tuning:

Using the RBF SVC model we managed to change the value of the two hyperparameter value C and Gamma. While increasing C's value(c=90),

the model tries to minimize the number of misclassified examples -due to high penalty- which results in better model performance and higher accuracy. While decreasing C's value(c=0.00001), the model results in lower accuracy and performs underfitting.

While increasing Gamma's value (gamma = 9), the model performs better and gives higher accuracy. While decreasing Gamma's value (gamma = 0.0001), the model performs badly and gives off lower accuracy close to underfitting.

Conclusion:

We first thought of using (One Vs. One) classifier since it gives higher accuracy than (One Vs. All) classifier and it was proved right when the accuracy of the chosen classifier was higher than the later one.

Since we had a little number of features and a high number of training examples we decided to use polynomial SVC model in order to get higher accuracy results which was proved by having the poly SVC kernel function giving us the highest accuracy in the classification model.