Table of Contents

Subjects 1

preprocessing techniques 1.1

[dataset analyzing 1.2](#_Toc90367860)

[regression techniques 1.3](#_Toc90367860)

[differences between models and some details 1.4](#_Toc90367860)

[features used or discarded 1.5](#_Toc90367860)

training, testing and validation sets sizes 1.6

screenshot 1.7

Conclusion2

**Subjects**

* **Preprocessing techniques and analyzing**

1. Handling the Null/Missing Values: we replaced the missing values of (value and wage) columns with the mean of each one

### Encoding the Categorical Variables: encoded string values such as (nationality, work rate and club team ) date values (contract\_end\_year, birth\_date, club\_join\_date)

1. Dropping ineffective columns: -as id, name, birth date, positions (we split positions to 4 columns so we don’t need it any more)
2. Data Transformation: at the positions rating columns we added the bonus value to the rating so we use it as numerical data ('LS','ST','RS','LW','LF','CF','RF','RW','LAM','CAM','RAM','LM','LCM','CM','RCM','RM','LWB','LDM','CDM','RDM','RWB','LB','LCB','CB','RCB','RB')
3. Scaling data:

We performed data scaling from 1 to 100 so that each feature contributes approximately proportionately to the final distance.

* **Regression techniques**

1. Polynomial regarrison (degree = 3)
2. Linear regarrison

* **Differences between models and some details**

|  |  |
| --- | --- |
| **Polynomial regarrison (degree = 3)** | **Linear regarrison** |
| Polynomial regarrison makes a higher degree equation to cover most of the data with respect to the variance bias and edges of the data and try | linear regarrison is just a linear equation that cover the mean of the data not the data itself most of times |
| Mean Square Error = 136597209328.58661  accuracy = 98.83076997234913% | Mean Square Error = 363483567620.4137  accuracy = 98.83260821551536% |
| training time = 0.36354589462280273 | training time = 0.008978128433227539 |

* **Features used or discarded**

1. Features used: overall\_rating, potential, wage, international\_reputation (1-5), release\_clause\_euro, club\_rating, ball\_control, LS, LF, LAM, LM, LCM, RM.
2. Features discarded: reactions, composure, CM, RCM, ST, RS, CF, RF, CAM, RAM.

* **Training, testing and validation sets sizes**

training is 80%

testing is 20%

* **Screenshot**

**A screenshot of a computer

Description automatically generated with medium confidence**

**Conclusion:**

We concluded that polynomial regression model is better than the linear regression model as it has smaller Mean Square Error, but the linear model is faster at learning time than polynomial model