|  |  |  |
| --- | --- | --- |
| Section | ID | Name |
| 3 | 20191700376 | عبدالعزيز جمال على محمد |
| 3 | 20191700368 | عبدالرحمن عمرو محمد |
| 3 | 20191700365 | عبدالرحمن علاء حموده |

**1-Load data:**

* Read data.
* View data information.

**2-Exploratory data analysis (EDA):**

* View missing values data.
* View Histogram of target column.
* Explore the correlation of continuous features.

A screenshot of a graph

Description automatically generated with low confidence**missing values percentage**

**Histogram of target column**A picture containing diagram, line, screenshot

Description automatically generated

![A picture containing text, screenshot, diagram, line

Description automatically generated]()**correlation of continuous features**

**Preprocessing:**

* remove columns that does not affect on classification.
* Select categorical and numerical columns.
* remove columns that have null values more than 300.
* Handel encoded categorical columns by converting to factor.
* Transform all categorical columns to factors.
* Fill null values in numerical columns with mean.
* Fill null values in factor columns with mode.
* Get correlation of continuous columns with target column.
* Select features with correlation >|0.2|
* Scale continuous data.
* Fit encoders to each of factor columns and apply encoders.

**Test Preprocessing:**

* First select only the features that model trained on.
* Apply the same preprocessing as for training data.
* Handel unseen values in encoding with (zero).

**Modelling:**

* Models used: -
* SVM
* Random Forest
* generalized linear models(glm).
* Visualization

A picture containing text, screenshot, diagram, line

Description automatically generatedSVM features importance

A picture containing text, line, diagram, screenshot

Description automatically generatedSVM

A picture containing text, line, diagram, screenshot

Description automatically generatedRandom Forest

GLM A picture containing text, line, diagram, screenshot

Description automatically generated

* **Prediction**

|  |  |  |
| --- | --- | --- |
| Model Name | MSE | Score in Kaggle |
| SVM | 257347170.837942 | **0.13573** |
| Random Forest | 136193706.447105 | 0.14583 |
| GLM | 960924740.520531 | 0.17299 |

A graph with red and green squares

Description automatically generated with low confidenceMSE Histogram

A picture containing text, screenshot, font, number

Description automatically generatedKaggle Scores

* **Conclusion:**

SVM model perform better than other models

As test score is better.