Software Quality Evaluation Project

Vladi Valsecchi 730030

Objective

- Perform the analysis of the ant-1.7 dataset in order to assess whether there are any problems / limitations in the use of the data for the recognition of software faultiness.

Used techniques

- Linear regression, with one indipendent variable and a dependent variable;
- Logistic regression, using three independent variables to classify data into two categories (bugged, non-bugged)

Dataset

Contains 745 elements and their mesurements.

The file contains 24 fields like:

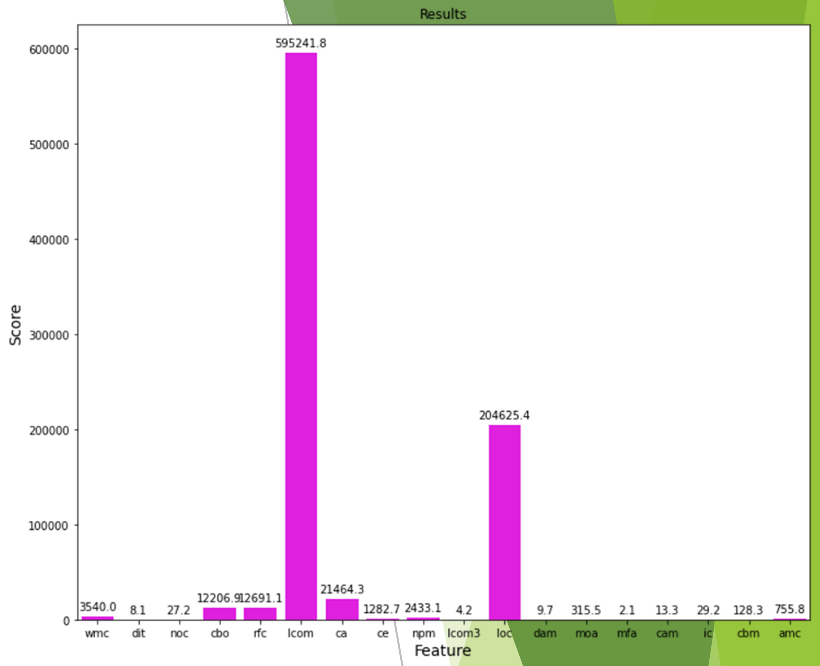
- Wmc
- Rfc
- Loc
- Lcom
- …

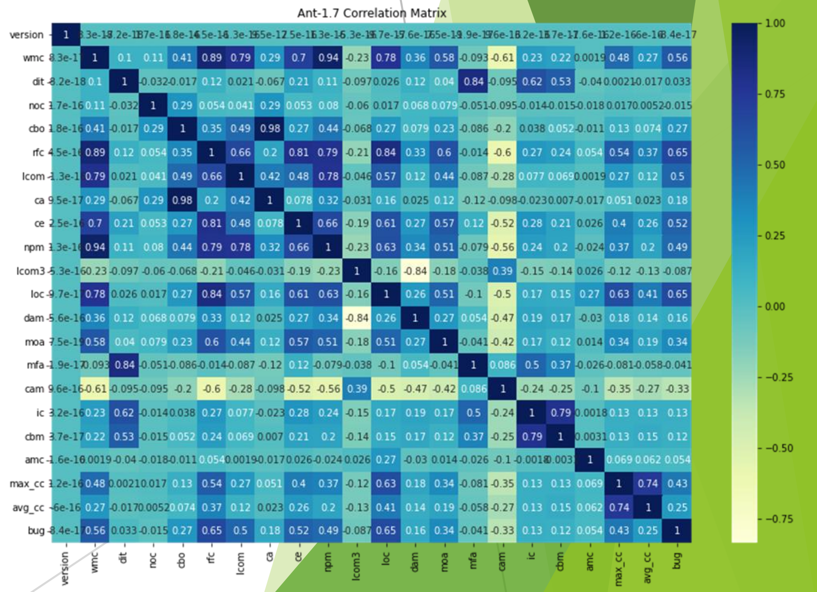
Feature selection

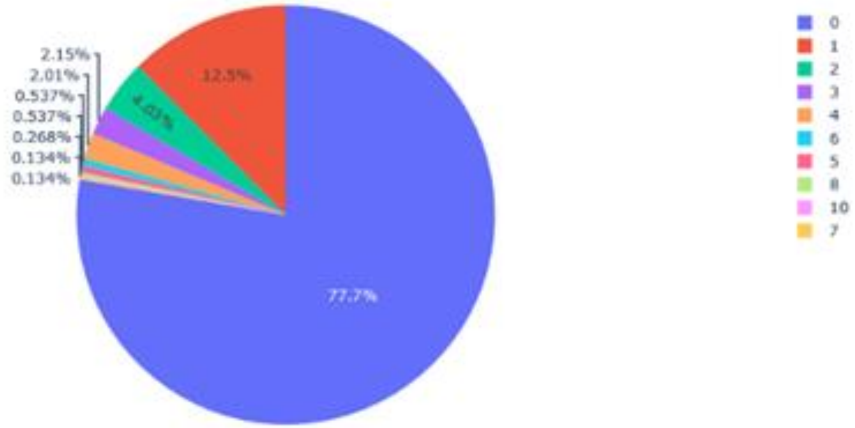
Used 2 approaches, correlation matrix and SelectKbest.

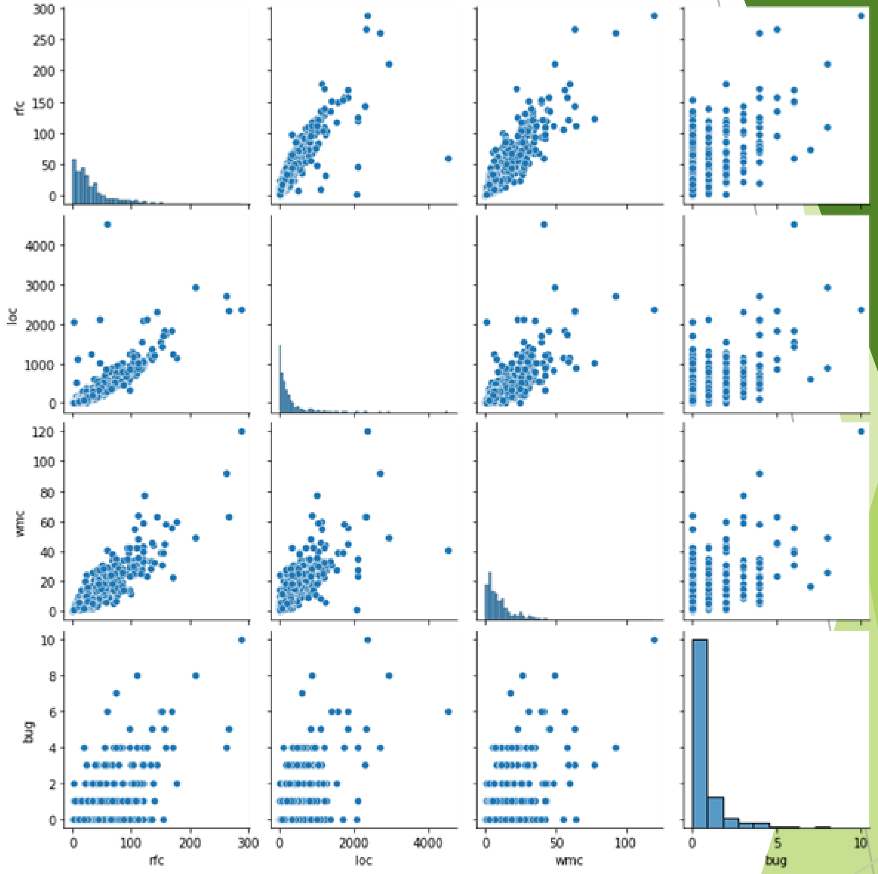
The features were selected through correlation matrix approach after a quick test to evaluate the R2 coefficient.Selected features are:

- Rfc
- Lcom
- Wmc



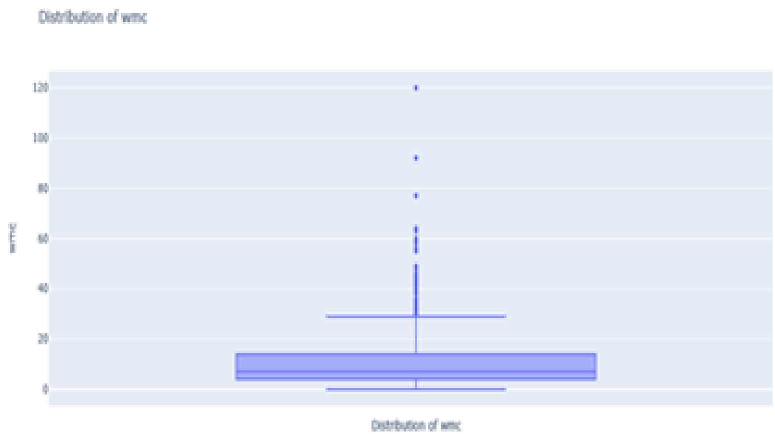


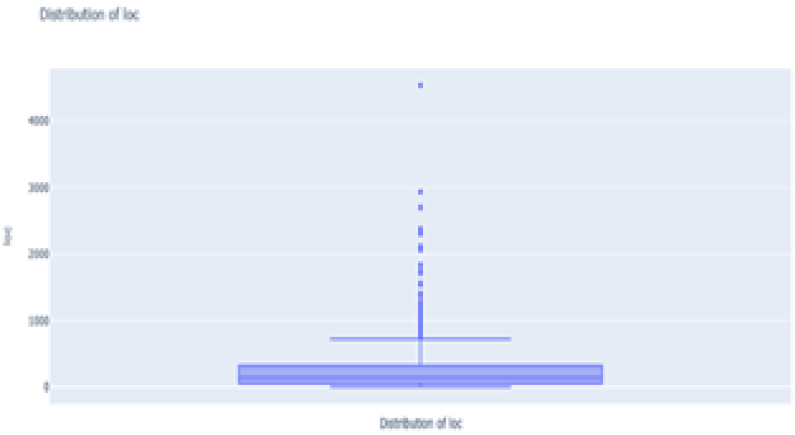


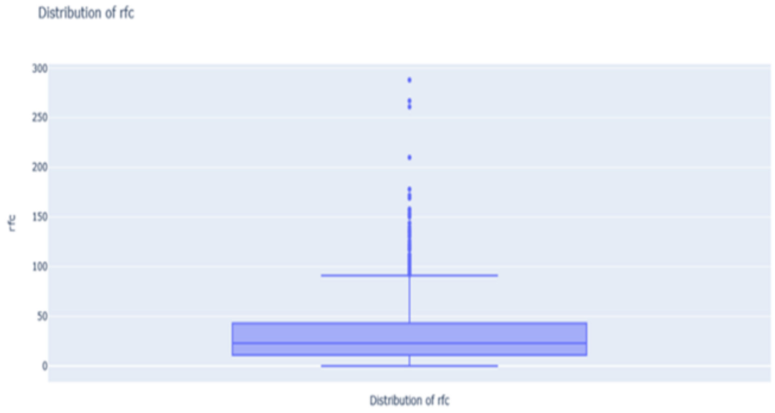


Data visualization

Data visualization 2







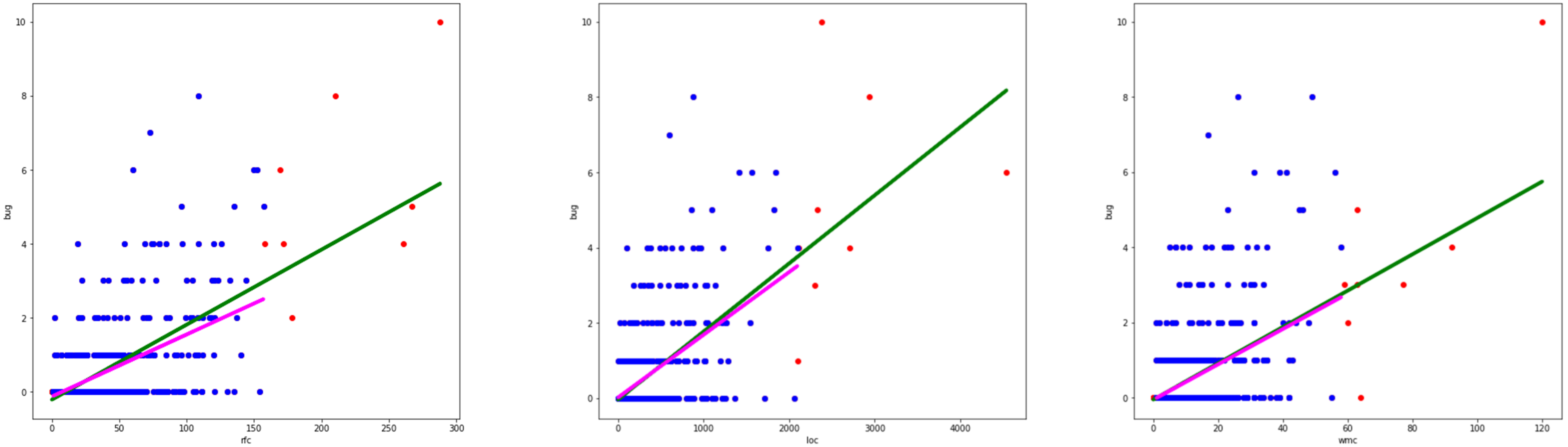
Descriptive statistics

# of bugs	Total records
0	579
1	93
2	30
3	16
4	15
6	4
5	4
8	2
10	1
7	1

The minimum number of rfc : 0
The maximum number of rfc : 288
The average number of rfc : 34.36241610738255
The value of the standard deviation of the feature rfc : 36.02497169398523

The minimum number of loc : 0
The maximum number of loc : 4541
The average number of loc : 280.07114093959734
The value of the standard deviation of the feature loc : 411.87207539635864

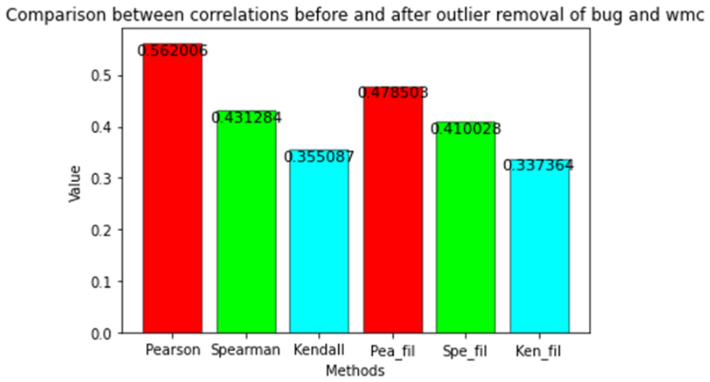
The minimum number of wmc : 0
The maximum number of wmc : 120
The average number of wmc : 11.071140939597315
The value of the standard deviation of the feature wmc : 11.97596324330988

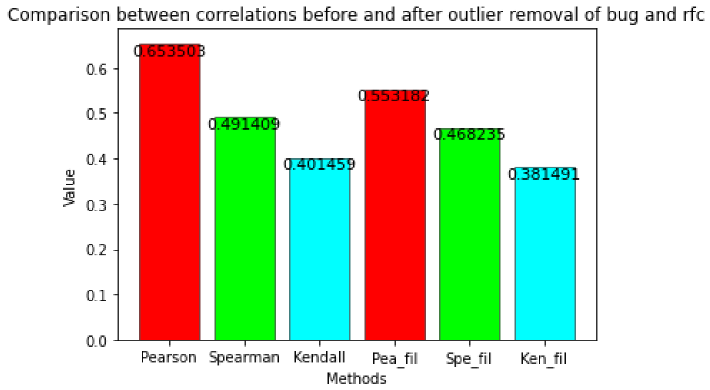


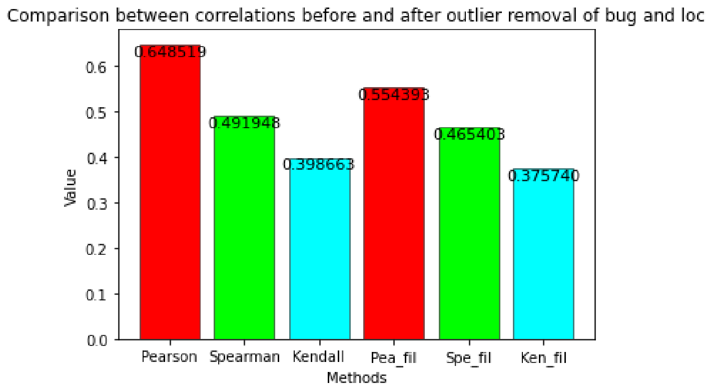
Linear regression

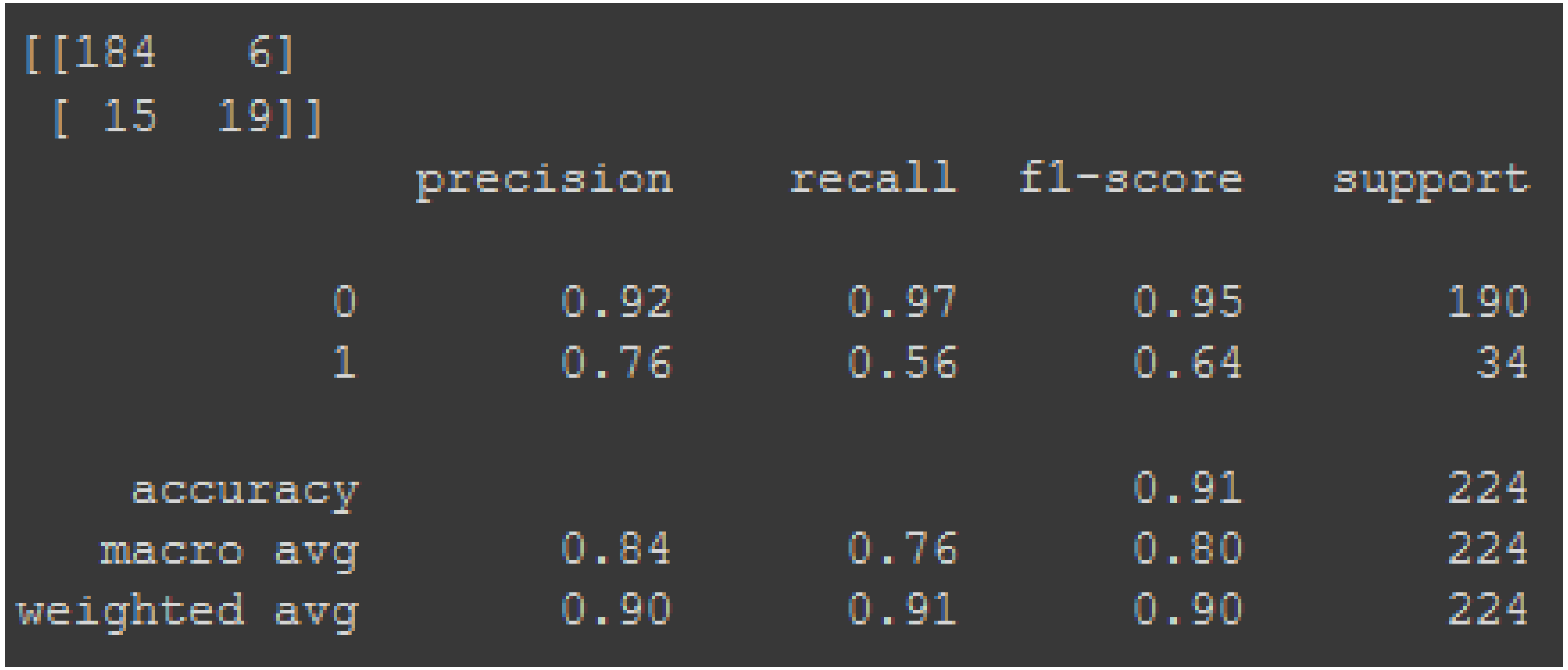
Rfc, loc, wmc

Correlation before and after outlier removal


Comparison between correlations before and after outlier removal of bug and wmc


Comparison between correlations before and after outlier removal of bug and rfc


Comparison between correlations before and after outlier removal of bug and loc



```
[[184   6]
 [ 15  19]]
              precision    recall  f1-score   support

           0       0.92      0.97      0.95       190
           1       0.76      0.56      0.64        34

    accuracy                           0.91       224
   macro avg       0.84      0.76      0.80       224
weighted avg       0.90      0.91      0.90       224
```

Logistic regression

Result description

- Linear regression: The model works well enough for predicting bug values. Rfc is the feature that performs better.
- Logistic regression: Despite the high accuracy, the data provided are not the best for training the model as they are unbalanced, containing 77% of non-bugged elements and only 23% of elements with at least one bug. Providing more varied data could train the model better.