Evaluation Report

Contents

[Introduction 2](#_Toc82979079)

[Research setup 2](#_Toc82979080)

[Dataset 2](#_Toc82979081)

[Experiments 3](#_Toc82979082)

[Analysis 1: Prediction model 4](#_Toc82979083)

[AUC Score 4](#_Toc82979084)

[F1-score 5](#_Toc82979085)

[Analysis 2: Feature selection method 6](#_Toc82979086)

[Analysis 3: Number of features reduced 6](#_Toc82979087)

[Conclusion 6](#_Toc82979088)

[Appendix 6](#_Toc82979089)

[References 7](#_Toc82979090)

# Introduction

This report covers our research analysis for our application’s performance based on the various configurations within our application. Our objective is to find the combination which displays the best results, particularly when handling imbalanced datasets. Additionally, we believe the results gained will provide insights towards this field of research.

# Research setup

## Dataset

During our research, we analysed the datasets from NASA and promise repository. Through our analysis, we studied the properties of each dataset and tabulate the information as followed:

**Table 1: Properties of datasets from NASA (Shepherd et al., 2014) repository**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Dataset** | **Procedural Metric** | **Object oriented Metric** | **Misc.**  **Metric** | **Metric count** | **Defective** | **Not defective** | **Module count** | **Degree of imbalance** |
| CM1.arff | 25 | 11 | 2 | 38 | 12.84% | 87.16% | 327 | Moderate |
| JM1.arff | 17 | 4 | 0 | 21 | 20.88% | 79.12% | 7720 | Mild |
| KC1.arff | 17 | 4 | 0 | 21 | 25.3% | 74.7% | 1162 | Mild |
| KC3.arff | 25 | 13 | 2 | 40 | 18.56% | 81.44% | 194 | Moderate |
| KC4.arff  (Raw) | 24 | 14 | 2 | 40 | 48.8% | 51.2% | 125 | Normal |
| MC1.arff | 24 | 13 | 2 | 39 | 1.84% | 98.16% | 1952 | High |
| MC2.arff | 24 | 13 | 2 | 39 | 35.48% | 64.52% | 124 | Mild |
| MW1.arff | 24 | 11 | 2 | 37 | 10% | 90% | 250 | Moderate |
| PC1.arff | 24 | 11 | 2 | 37 | 8.1% | 91.9% | 679 | Moderate |
| PC2.arff | 23 | 11 | 2 | 36 | 2.22% | 97.78% | 722 | High |
| PC3.arff | 24 | 11 | 2 | 37 | 12.35% | 87.65% | 1053 | Moderate |
| PC4.arff | 24 | 11 | 2 | 37 | 13.86% | 86.14% | 1270 | Moderate |
| PC5.arff | 23 | 13 | 2 | 38 | 27.04% | 72.96% | 1694 | Mild |

**Table 2: Properties of datasets from Promise (Menzies, 2004) repository**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Dataset** | **Procedural Metric** | **Object oriented Metric** | **Misc.**  **Metric** | **Metric count** | **Defective** | **Not defective** | **Module count** | **Degree of imbalance** |
| kc1.arff | 17 | 4 | 0 | 21 | 15.46% | 84.54% | 2109 | Moderate |
| cm1.arff | 17 | 4 | 0 | 21 | 9.84% | 90.16% | 498 | Moderate |
| kc2.arff | 17 | 4 | 0 | 21 | 20.5% | 79.5% | 522 | Mild |
| jm1.arff | 17 | 4 | 0 | 21 | 19.35% | 80.65% | 10885 | Moderate |
| pc1.arff | 17 | 4 | 0 | 21 | 19.35% | 80.65% | 1109 | Moderate |

These datasets will be used throughout this research. Furthermore, these datasets were also used for our second evaluation report which focuses on comparing our results with other research papers.

For the following experiments, our focus would mainly be on datasets which are highly imbalanced. As such, our experiments were conducted using datasets that were categorized to have a moderate or high degree of imbalance. The selected datasets are as followed:

* Moderate: CM1, KC3, MW1, PC1, PC3, PC4
* High: MC1, PC2

## Experiments

Our devised experiments were used to study the performance of our algorithm based on the following factors:

1. Degree of imbalance
2. The prediction models selected
3. The feature selection used
4. The number of features reduced

The application we devised allows us to perform experiments with variations available on these factors mentioned. In addition, our program allows us to test and evaluate the performance of our built models using 4 evaluation metrics.

From the results gained, we will be able to determine the potency of which each factor influences the performance of our fault detection method. Furthermore, we will be able to identify the optimal values and selections which allows will the best results to be achieved.

# Analysis 1: Prediction model

There are several prediction models included within our program, each having different levels of proficiency for handling imbalanced datasets. To compare the performance of each model, we used our program with all the models selected to obtain the performance result in chart and tabular form. An additional note, no feature selection methods were used as this analysis focuses solely on the effectiveness of each model.

## AUC Score

**Table 3: AUC results for the first analysis**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model name | Complement Naive Bayes | Decision Tree | Logistic regression | Multi-Layer Perceptron | Naive Bayes | Random Forest | Rotation Forest | Voting |
| CM1.arff | 0.559 | 0.669 | 0.705 | 0.481 | 0.697 | 0.798 | 0.723 | 0.745 |
| KC3.arff | 0.415 | 0.566 | 0.501 | 0.424 | 0.506 | 0.545 | 0.533 | 0.528 |
| MC1.arff | 0.602 | 0.93 | 0.909 | 0.61 | 0.82 | 0.968 | 0.933 | 0.926 |
| MW1.arff | 0.475 | 0.455 | 0.508 | 0.378 | 0.461 | 0.636 | 0.6 | 0.566 |
| PC1.arff | 0.611 | 0.818 | 0.75 | 0.714 | 0.796 | 0.864 | 0.769 | 0.841 |
| PC2.arff | 0.646 | 0.665 | 0.92 | 0.479 | 0.876 | 0.788 | 0.717 | 0.878 |
| PC3.arff | 0.612 | 0.739 | 0.831 | 0.563 | 0.766 | 0.845 | 0.788 | 0.836 |
| PC4.arff | 0.672 | 0.893 | 0.831 | 0.646 | 0.783 | 0.909 | 0.829 | 0.894 |

Chart, bar chart

Description automatically generated**Fig. 1 Bar chart which displays the AUC evaluation scores for the first analysis**

As shown, most models performed outstandingly well in this field. There are various

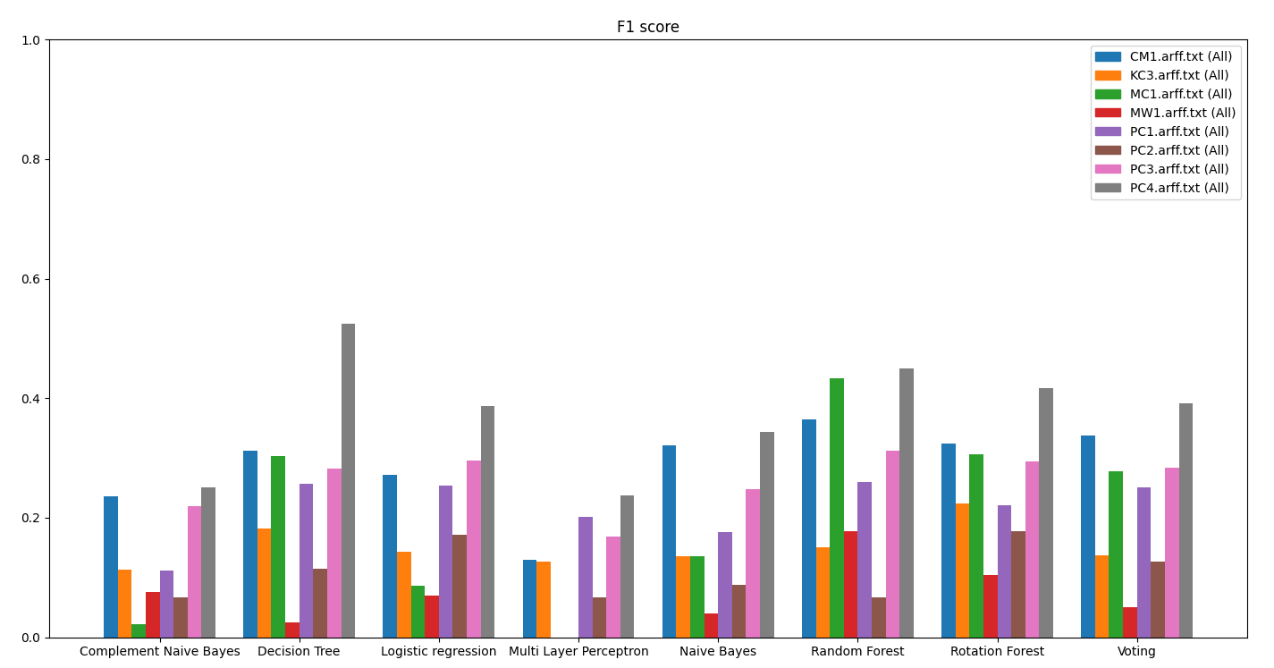
In literature, Naïve Bayes and Logistic Regression are known to be proficient for handling imbalanced dataset. So, the scores for these two predictions were expected to perform greater than other base predictors. If observe from Fig 1, these two base models shown to achieve scores which outperform others, these scores fall between acceptable and excellent range. If we look at the datasets with the highest degree of imbalanced, MC1 and PC2, both models show outstanding performance. Overall, the Logistic regression model shown to have results for handling imbalanced datasets.

For the ensemble predictors, these models overall perform greater than most base prediction models. On this field, the voting and random forest models shown to have the best performance.

## F1-score

**Table 4: F1-score results for the first analysis**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model name | Complement Naive Bayes | Decision Tree | Logistic regression | Multi-Layer Perceptron | Naive Bayes | Random Forest | Rotation Forest | Voting |
| CM1.arff | 0.236 | 0.312 | 0.272 | 0.13 | 0.321 | 0.365 | 0.324 | 0.337 |
| KC3.arff | 0.113 | 0.182 | 0.143 | 0.126 | 0.135 | 0.15 | 0.224 | 0.137 |
| MC1.arff | 0.021 | 0.303 | 0.086 | 0 | 0.136 | 0.433 | 0.306 | 0.277 |
| MW1.arff | 0.076 | 0.025 | 0.069 | 0 | 0.04 | 0.178 | 0.104 | 0.05 |
| PC1.arff | 0.111 | 0.257 | 0.253 | 0.201 | 0.176 | 0.259 | 0.22 | 0.25 |
| PC2.arff | 0.067 | 0.114 | 0.171 | 0.067 | 0.087 | 0.067 | 0.178 | 0.126 |
| PC3.arff | 0.219 | 0.282 | 0.296 | 0.168 | 0.247 | 0.312 | 0.294 | 0.283 |
| PC4.arff | 0.251 | 0.525 | 0.387 | 0.237 | 0.343 | 0.449 | 0.416 | 0.392 |



**Fig. 1 Bar chart which displays the F1-score evaluation scores for the first analysis**

For F1-score, none of models were able to achieve outstanding results, which is a common result for this field of study. For the MW1 dataset in particular, every model shown to perform poorly as observed in Table 4.

An interesting observation was that the best scores were achieved mainly by the tree-based models. The decision tree achieved promising results and shows to have the best results when compared with the other base prediction models. The Rotation Forest had the best performance, achieving remarkable scores and shown good consistency.

One thing to highlight would be the performance of the Logistic regression and Voting models. While the scores achieved by these models were not greater than the tree-based models, they are above average and show consistency.

# Analysis 2: Feature selection method

For this analysis, we introduced the feature selection methods to our experiment which reduces the metrics used when fitting the data to the models. The idea behind the feature selection methods is that not all metrics are good indicators for faultiness of a software. These algorithms are used to reduce the metrics so that only the useful metrics remain which will improve the performance of our program.

# Analysis 3: Number of features reduced

# Conclusion

From the results gained, our proposed method towards handling imbalanced dataset can be described as the following:

# Appendix

# References

Menzies, T. (2004). *Promise Software Engineering Repository.* Retrieved May 15, 2021 from

<http://promise.site.uottawa.ca/SERepository/datasets-page.html>

Shepherd et al. (2014). *NASADefectDataset.* Retrieved May 15, 2021 from

<https://github.com/klainfo/NASADefectDataset/tree/master/CleanedData/MDP/D>''