Assignment 1 - Reproducibility

Maria Rafaela Alves Abrunhosa 107658

Mining Large Scale Datasets, University of Aveiro, March 2025

Abstract: Reproducing a study is a common practice for most researchers, as it can help to confirm research findings, gain inspiration from the work of others, and avoid reinventing the wheel [3]. In this paper, our aim is to reproduce an application of a data mining process that has been published in a scientific venue [1]. In this work, we will focus on replicating the methodology by replicating the predictive modelling of the given study. We train five different classifiers using cross-validation, as in the previous study, and compare the results, using accuracy as the metric for performance evaluation.

Keywords: Reproducibility, predictive modelling, cross-validation, random forest, extra tree, ada boost, xgboost, gradient boosting, accuracy

I. Introduction

This project aims to reproduce an application of a data mining process published in a scientific venue [1]. We focus on the reproduction of predictive modelling as part of the data mining process. The published paper used as a guide for this data mining process reproduction presents two malware feature datasets on two different platforms to support the validation of the effectiveness of a malware detection method.

The article consists of three contributions, starting with the development of two structures for creating malware datasets for Windows and Android, extracting features from malware analysis reports to create two datasets of malware features, and an evaluation of the usefulness of the datasets. However, in this work we will only attempt to reproduce the third point, the evaluation of the usefulness of the datasets in a supervised framework, as we have no way of replicating steps one and two of the article.

A. Reproducibility

Reproducibility is the ability of a researcher to replicate the results of a previous study, using the same materials that the original researcher used, in an attempt to obtain the same results as the previous study, with the same data and tools [2].

Reproducibility is generally difficult to achieve, especially in ML and DL, because of the many sources of nondeterminism.

In machine learning, the performance of algorithms depends on the hyperparameters, and these algorithms often show a high sensitivity to the setting of the hyperparameters. To reproduce the results of another study, we need a spacing of the range of hyperparameters considered, as well as the method used to select the best hyperparameters.

We also need a clear definition of the statistics used to report the results, a description of the results as well as the number of evaluation runs and other performance evaluation metrics specifications [3].

II. Replicability Process

A. Dataset Description and Pre-processing

Firstly, it is important to note that the previous study has two datasets of malware characteristics, TUMALWD and TUANDROMD respectively. However, for the purposes of this reproduction, we will only use the TUANDROMD dataset as it was the only one provided and we couldn't find the other or anything identical.

1) Dataset TUANDROMD characteristics: In order to study if our dataset is similar to the one used in the previous study, we can check some of the dataset characteristics, following some points presented in the previous study report.

According to the study report we should have 72 labels, 71 representing the malware family and the last one representing the goodware. However, as we can see from our code and results, our dataset only has two classes, the malware and the goodware, there is no distinction between the different types of malware. So we now have a binary class problem instead of a multiclass problem, which means that the results and reproducibility are no longer possible.

In the previous study, they have a total of 25,553 instances, while we only have 4465. In terms of features we have 241 and the last column is the target value, the label.

Regarding the balance between classes, our dataset is also very unbalanced as the one used in the guidance paper. We have 79.9% of the data classified as malware and 20.1% as goodware.

In terms of recency, we can not say if the data is recent, and unlike the study, the dataset is not updated with this new data.

We can say that the features make it possible to distinguish between malware and goodware applications in terms of relevance. In addition, we can examine the top 15 features in our dataset and compare them with those in the study provided.

2) Top ranked features for TUANDROMD: To check the top features of our dataset, we use a Random Forest Classifier to extract the most important features and compare our results with the study results. We chose this classifier to test the extraction of the most important features, since we had to choose something and the previous study does not have any information about its extraction of the most important features, which means that we are likely to get different top features.

We can see in **TABLE I** and **TABLE II** from the notebook that we have some top features in common with the previous study and some very important features reported in the paper are not the very important features we have extracted. For example, the top 1 feature in the previous study is *SEND_MSG*, but in our results this feature is ranked 23rd and does not even appear in the top 15 features. This means that our datasets have different importance for the same features, which means that the predictions can be different when training, evaluating and validating the models.

B. Performance evaluation and validation

As mentioned above, the dataset provided is for binary classification, while the dataset in the study is multiclass. Therefore, we can assume in advance that the results of our models will not be the same as the results of the study. Furthermore, the importance of the features is different between the study dataset and our dataset, which may lead to different results. From the start, this work cannot be replicated with the small amount of information and differences we have.

1) Cross-validation: Cross validation is a technique used in machine learning to evaluate the performance of a model on unseen data [5]. The results of each training and validation are stored and output, as well as an average result of all the validations.

2) Machine Learning Models: To evaluate performance, we used the same five classifiers as in the previous study, and for each classifier, we used a K-fold cross-validation as explained earlier, with k = 10.

3) Parameters: Regarding the parameters, the provided study does not give any information on which parameters are used and with which values. For this reason, we used the default values as much as possible, assuming that these models were created with default values.

We chose the number of estimators as 100 only because many of the models may stop before reaching this value if there are some early stopping criteria, and also chose the random state value as 42. This number works as a convention of random state because it ensures reproducibility - the same random numbers are generated each time the code is run - and consistency.

4) Metrics: The only metric used to evaluate performance in the previous study was the test accuracy. For this reason, it is the only metric we have evaluated and used to compare the results.

It is important to note that even though this metric is useful to compare our results with the results of the study, this metric is not good for evaluating the performance of the models because, as explained before, both datasets are very unbalanced, so accuracy is not a good metric for these cases.

5) Conclusions: Looking at **TABLE III** and **TABLE IV**, we can see that the results are not the same, although they are not very different. We have some higher results for the first three classifiers than the results of the study, but lower results for the last two classifiers. These differences in the results could be due to the different datasets, to the change between a multiclass classification and a binary classification (our case) or even to the different parameters used in the classifiers.

We can now conclude that the study modelling process is not reproducible for many factors and reasons, but mainly because we do not have the information needed to reproduce the different steps in this process.

C. Performance evaluation and validation using only the top 15 features

As an experiment, we decided to train the models on the most relevant features rather than all of them. We used the 15 most important features resulting from the random forest explained above. This test was done to see if the results were closer to those obtained in the study presented. We also added some random parameters instead of all the standard ones.

Looking at **TABLE V**, we can see that even if a classifier has the same accuracy in both studies, the results are not the same. These differences in the results could be due to the same reasons mentioned before. Despite this approximation of the values, we cannot conclude that this study is reproducible, as the same results could be a matter of luck.

III. Discussion and Conclusion

A. Evaluation of Replication Success or Failure

We could not replicate the methodology or the results. The only thing we did exactly as in the previous study was to use the value k=10 in the cross-validation and to use the same five classifiers as in the performance evaluation. None of this is relevant or sufficient to replicate the predictive modelling.

As mentioned above, our dataset is very different from the one used in the previous study. Comparing only the simplest characteristics, we realise that it is already in the part of the dataset where this replication is not possible. We have even moved from a multiclass problem classification to a binary class problem classification.

Furthermore, the hyperparameter supply, which is very important, does not exist. We have no information about the parameters used to train the models. Since we know that the performance of the algorithms depends on the hyperparameters and since we do not have any indication of which parameters were used, we can never reproduce the predictive modelling with the same factors as the study provided.

B. Proposed Solutions

There are some aspects of the publication that could make the replication process easier. For example, they could publish the dataset they created and add a link to the references so that we can access the original datasets used in the study.

Also, for each classifier, they could write a small table explaining which parameters were used in each classifier and even justify their choices.

C. Conclusion

At the end of this study, we have already seen that our main objective has not been achieved, as we have not been able to reproduce the results of the previous study. Even with the cross-validation and training of the same five classifiers used in the previous study, we could not get the same results.

In order to reproduce the previous study, we would need some more information, for example, the dataset used should be provided, because for different datasets the results are obviously different, and the parameters used to train the models should also be provided. These two pieces of information are the key to a reproducible paper, as they are the most meaningful parts of a predictive modelling process.

We also understand the importance of reproducibility and transparency when publishing in a scientific venue. It is important for the machine learning based scientific community to be transparent so that we can reproduce some of the published work. Without this, we cannot prove that a new technique we have implemented is an improvement if we cannot get the same accuracy and results for the original research. If no one gets the same results as us, it is likely that we are doing something wrong; by sharing all the data and tools, we can discuss the results with more researchers. Furthermore, there may be cases where if we cannot produce the same results, the same good results may be due to randomness, and non-reproducible models are less reliable since we cannot prove that these results can be achieved, which means they can be inferred for some anomaly.

REFERENCES

[1] IEEE. (2025). Document 9312053. Available at https://ieeexplore.ieee.org/document/9312053. Accessed on February 22, 2025.

[2] Trifan, A. (2025). MDLE Aula 02. Available at https://uapt33090-my.sharepoint.com/:b:/r/personal/alina_trifan_ua_pt/Documents/MDLE/aula%2002/MLD_02.pdf?csf=1&web=1&e=0AwwEE. Accessed on February 22, 2025.

[3] Carnegie Mellon University. (2020). The importance of reproducibility in machine learning. Available at https://blog.ml.cmu.edu/2020/08/31/5-reproducibility/. Accessed on February 25, 2025.