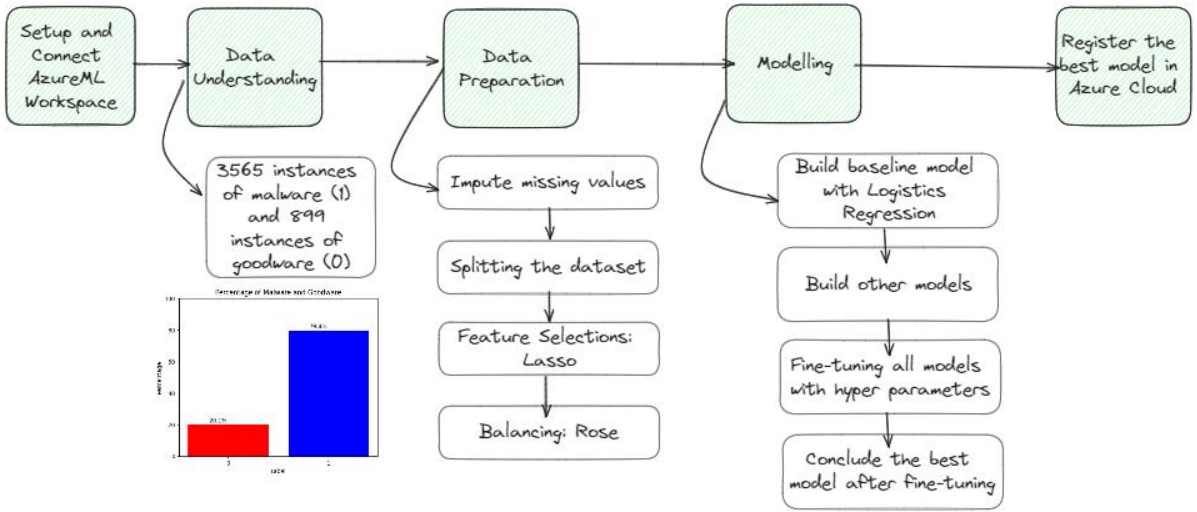
Methodology and Workflow

- The methodology applied for this dataset is based on the CRISP-DM Model.
- In the data understanding step, one row with a missing value is identified. This row is subsequently deleted in the data preparation step
- The dataset then undergoes feature selection using the Lasso method and balancing with the ROSE method due to the imbalance in the dataset (3565 instances of malware (1) and 899 instances of goodware (0)).



Primary Evaluation

- In total, 10 models are built to predict the dataset. Here are the final results in terms of Accuracy and AUC indicators.

	Accuracy	AUC
Logistic Regression	0.9697	0.9582
SVM	0.9944	0.9944
MLPClassifier	0.9922	0.9998
AdaBoostClassifier	0.9877	0.9784
KNeighborsClassifier	0.9821	0.9964
RandomForestClassifier	0.9966	0.9958
DecisionTreeClassifier	0.9922	0.9860
GaussianProcessClassifier	0.9944	0.9995
HistGradientBoostingClassifier	0.9933	0.9999
GradientBoostingClassifier	0.9933	0.9999
GaussianNB	0.9698	0.9787

baseline

best model

- As can be seen, the baseline model does not perform very well compared to the other models.
- All the models achieve notably high results, raising concerns that overfitting might occur during modeling.
- In a preliminary evaluation, the Random Forest model emerges as the best-performing model with an Accuracy score of 0.9966 and an AUC score of 1.0.
- Subsequently, fine-tuning will be performed to assess whether the models can achieve even higher scores.

Final Evaluation

- After fine-tuning, here are the new table results:

	Accuracy	AUC
Logistic Regression	0.9899	0.9909
SVM	0.9955	0.9951
MLPClassifier	0.9944	0.9996
AdaBoostClassifier	0.9843	0.9889
KNeighborsClassifier	0.9855	0.9942
RandomForestClassifier	0.9966	0.9958
DecisionTreeClassifier	0.9933	0.9888
GaussianProcessClassifier	0.9944	0.9995
HistGradientBoostingClassifier	0.9922	0.9999
GradientBoostingClassifier	0.9933	0.9995
GaussianNB	0.9698	0.9787

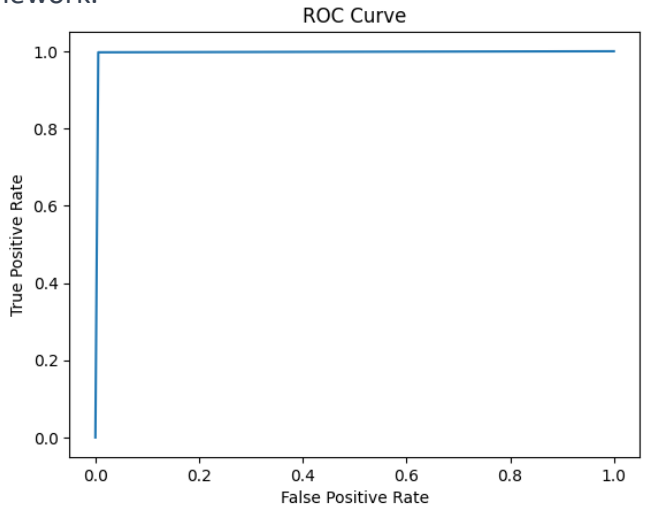
baseline

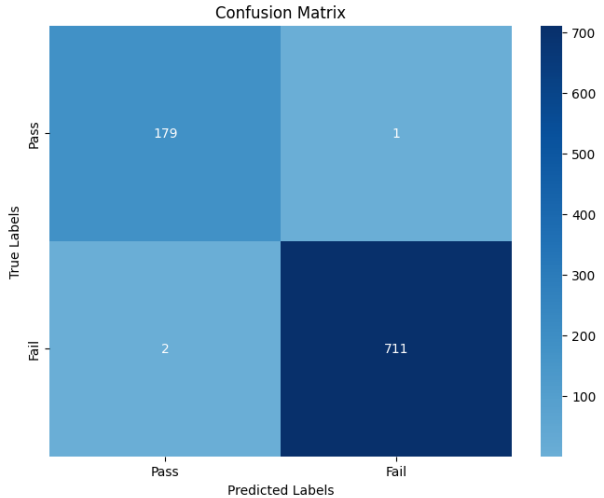
best model

- Most of the models yield higher results after fine-tuning.
- Although the results for Random Forest remain the same, it is still considered the best model among all.
- In conclusion, Random Forest will be chosen as the optimal model for detecting malware for the company.

About the best model – Random Forest

- The model excels in malware and goodware detection with an exceptional accuracy of 99.66%.
- The confusion matrix indicates only 3 misclassifications out of 893 instances, showcasing robust performance in identifying both malware and goodware.
- Furthermore, the AUC Score of 99.58% underscores the model's excellent discrimination ability. In summary, the model's outstanding performance makes it a highly reliable choice for deployment in the company's security framework.





Reference

- https://www.kaggle.com/code/joebeachcapital/malware-detection-data-import-eda-starter
- https://chat.openai.com/share/86bd6096-7b2d-4fed-b1d6-397c56dd2eed
- https://chat.openai.com/share/183e8910-b632-474a-9d98-21ac29ea9697
- https://chat.openai.com/share/1e5dc19e-899a-4c3d-a3b6-fbc956d65465