MAJOR PROJECT EXTERNAL VIVA

PRESENTED BY

Nikitha Maramraju
1602-19-737-145

Guggilla Shri Pallavi
1602-19-737-171

Team - 8

Guide: D.R.L. Prasanna (Assistant Professor)

Contents

TITLE

A Comparative Investigation on the Application of Machine Learning for Ransomware Detection

ABSTRACT

- The rise of ransomware attacks in recent years has highlighted the need for effective methods to detect and prevent such attacks.
- Ransomware is a type of malicious software that encrypts files on a computer system and demands a ransom in exchange for the decryption key.
- Ransomware attacks can cause significant financial losses and data breaches.
- Our proposal is to design a Ransomware detection system based on machine learning.

PROBLEM STATEMENT

Ransomware is malicious software that encrypts a user's data and demands a ransom payment to access the data again. To overcome this, machine learning algorithms can be used to detect ransomware and protect users from its effects.

- The objective of this project is to develop a machine learning-based approach to detect ransomware.
- The proposed solution should be able to accurately detect ransomware with high precision and recall.
- The project aims to contribute towards the development of effective methods to detect and prevent ransomware attacks.

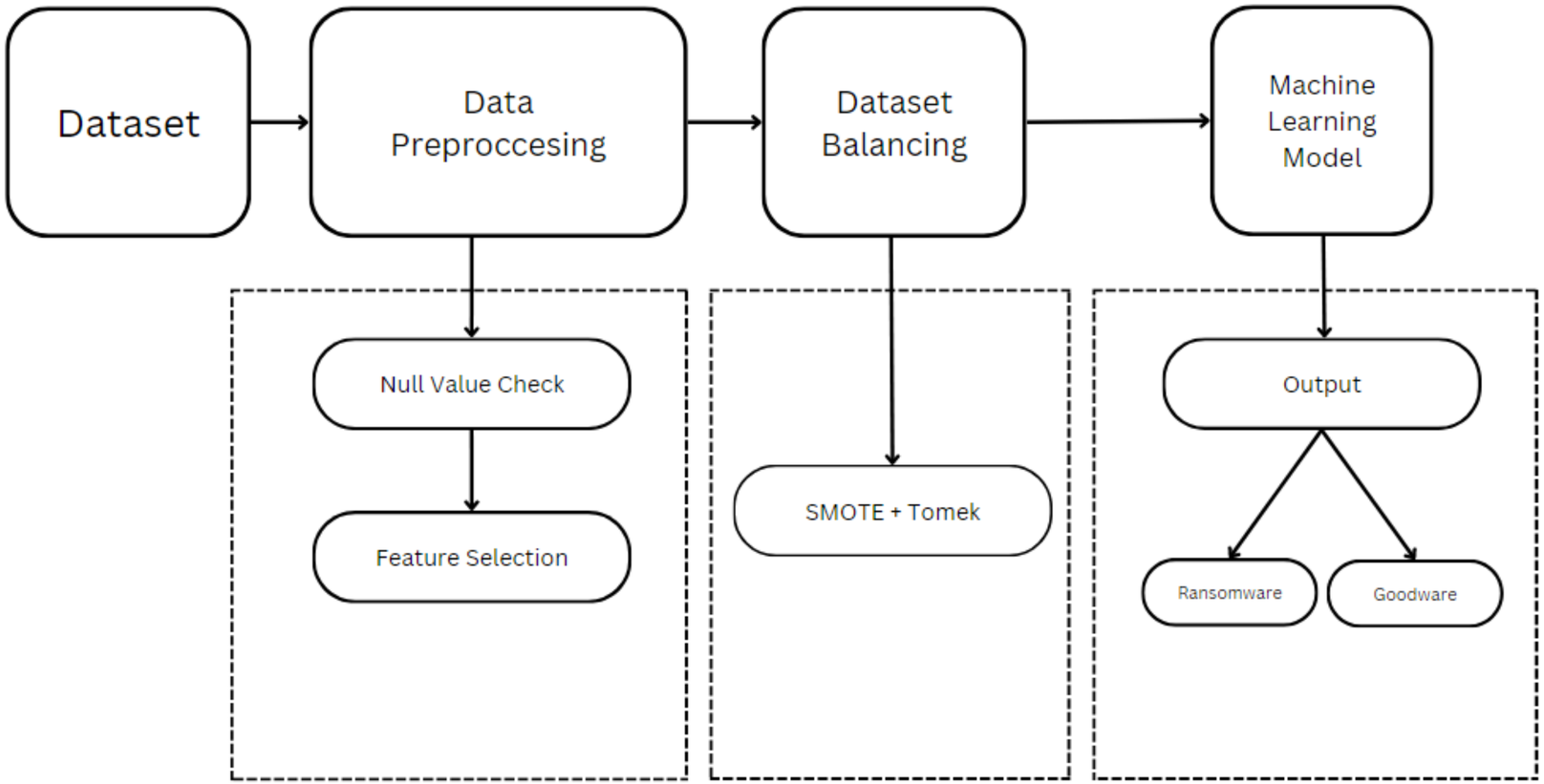
LITERATURE SURVEY

- Detecting Ransomware with Machine Learning Techniques by F. T. A. Alhawawreh et al. (IEEE Access, 2018)
- Ransomware detection using machine learning algorithms: A review by M. S. Mahmud et al. (Computers & Security, 2021)
- Feature selection based on VIF and mutual information for classification of cancer data by V. Gayathri and V. Saravanan from ScienceDirect. This paper proposes a feature selection method based on VIF and mutual information for classifying cancer data using machine learning.
- A Comparative Study on the Effectiveness of SMOTE and Tomek Links in Handling Data Imbalance" by R. R. Nair and P. R. Nair.

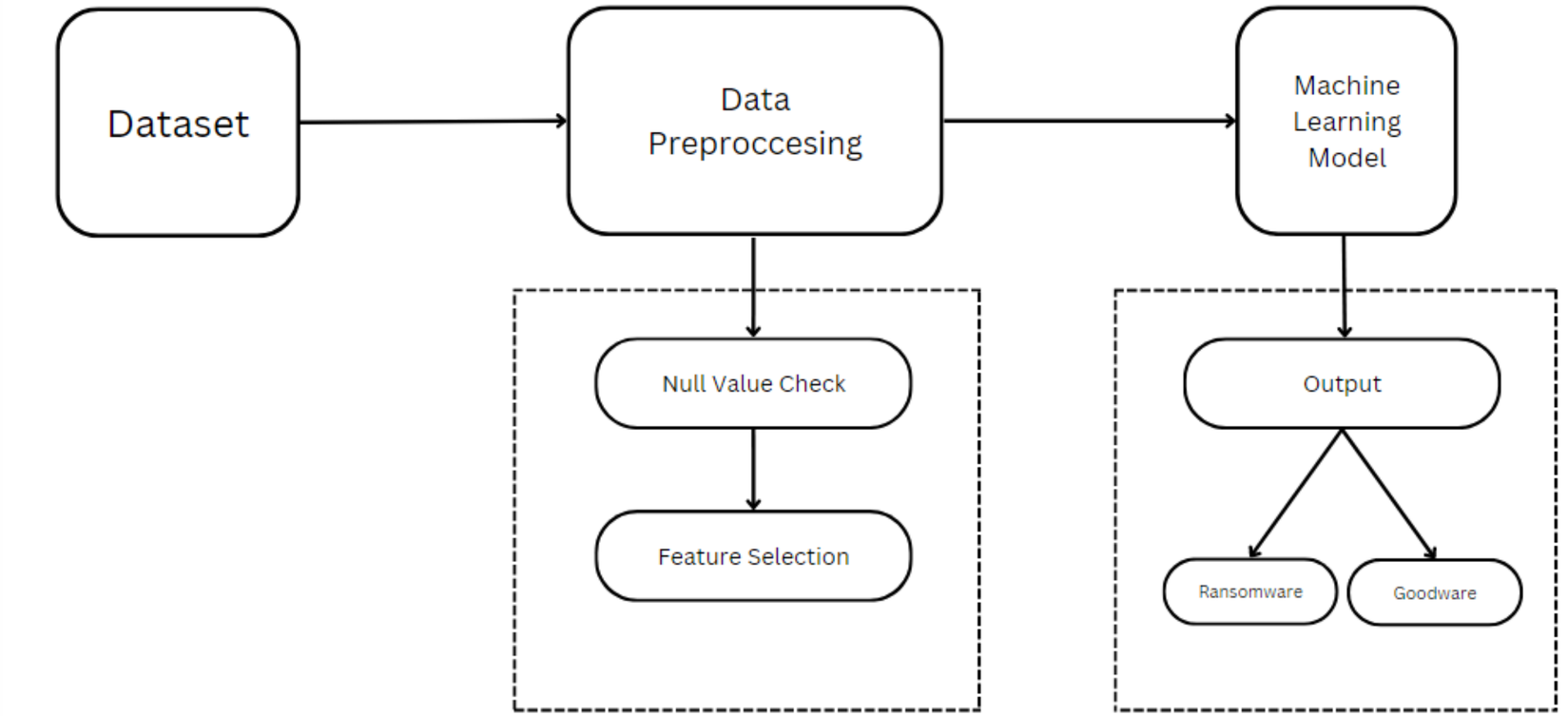
PROPOSED METHODOLOGY

- Collecting Dataset
- Data Preprocessing
- Feature Selection
- Applying a Machine Learning Model after Data Balancing
- Applying a Machine Learning Model without Data Balancing

PROPOSED METHODOLOGY



PROPOSED METHODOLOGY



DATASET

	A	B	C	D	E	F	G	H	I	J	K
1	**Name**	**md5**	**Machine**	**SizeOfOptionalHeader**	**Characteristics**	**MajorLinkerVersion**	**MinorLinkerVersion**	**SizeOfCode**	**SizeOfInitializedData**	**SizeOfUninitializedData**	**AddressO**
2	memtest.exe	631ea355665f28d4707448e442fbf5b8	332	224	258	9	0	361984	115712	0	613
3	ose.exe	9d10f99a6712e28f8acd5641e3a7ea6b	332	224	3330	9	0	130560	19968	0	8177
4	setup.exe	4d92f518527353c0db88a70fddcfd390	332	224	3330	9	0	517120	621568	0	35089
5	DW20.EXE	a41e524f8d45f0074fd07805ff0c9b12	332	224	258	9	0	585728	369152	0	45125
6	dwtrig20.exe	c87e561258f2f8650cef999bf643a731	332	224	258	9	0	294912	247296	0	21738
7	airappinstaller.exe	e6e5a0ab3b1a27127c5c4a29b237d823	332	224	258	9	0	512	46592	0	448
8	AcroBroker.exe	dd7d901720f71e7e4f5fb13ec973d8e9	332	224	290	9	0	222720	67072	0	21933
9	AcroRd32.exe	540c61844ccd78c121c3ef48f3a34f0e	332	224	290	9	0	823808	650240	0	58766
10	AcroRd32Info.exe	9afe3c62668f55b8433cde602258236e	332	224	290	9	0	4096	7168	0	675
11	AcroTextExtractor.exe	ba621a96e44f6558c08cf25b40cb1bd4	332	224	290	9	0	29696	12800	0	2705
12	AdobeCollabSync.exe	bf0a35c0efcaf650550b9e346dfcbd33	332	224	290	9	0	917504	316928	0	83380
13	Eula.exe	1556a34d117a80bdc85a66d8ea4fbcf2	332	224	290	9	0	53248	34816	0	5360
14	LogTransport2.exe	c4005b63df77068bce158ac8ef7c522b	332	224	258	9	0	206848	102400	0	11015
15	reader_sl.exe	e595f220ed529885d8bc0ef42e455e4d	332	224	259	9	0	14848	14336	0	1652
16	AcrobatUpdater.exe	0e9dee95fdf47d6195da804a0deeda5b	332	224	258	9	0	178688	134144	0	7808
17	AdobeARM.exe	47c1de0a890613ffcff1d67648eedf90	332	224	258	9	0	413184	518144	0	16019
18	armsvc.exe	11a52cf7b265631deeb24c6149309eff	332	224	258	9	0	37376	20992	0	3098
19	ReaderUpdater.exe	5ed9b78b308d302c702d44f4505b3f46	332	224	258	9	0	178688	134144	0	7808

DATASET

- This dataset contains details of various files of different extensions such as .exe, .dll, etc.
- Rows - 138047
- Columns - 57
  - **Name -** The name of the file
  - **md5 -** The MD5 hash value of the file
  - **Machine -** The type of machine architecture for which the file was created
  - **SizeOfOptionalHeader -** contains information such as the size of the code and data sections
  - **Characteristics -** The flags that specify the characteristics of the file, such as whether it is a DLL (dynamic-link library), whether it can be run on a 64-bit machine
  - **MajorLinkerVersion -** The major version number of the linker used to create the file
  - **MinorLinkerVersion -** The minor version number of the linker used to create the file
  - **SizeOfCode -** The size of the code section
  - **legitimate -** A binary flag indicating whether the PE file is legitimate (1) or ransomware (0)

EVALUATION METRICS

- **False Negative Ratio (FNR)**
- **False Positive Ratio (FPR)**
- **True Negative Ratio (TNR)**
- **True Positive Ratio (TPR)**
- **F-Measure (the HM of precision and recall)**
- **Accuracy**

$$TRP = \frac{TP}{TP + FN}, \quad FRP = \frac{FP}{FP + TN},$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}, \; Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$\text{F} - \text{Measure} = 2 * \frac{(Precision * Recall)}{Precision + Recall}$$

True Positive (TP): the number of ransomware that is correctly predicted as ransomware.

True Negative (TN): the number of goodware files that are correctly classified as goodware.

False Positive (FP): number of goodware files misclassified as ransomware.

False Negative (FN): number of ransomware which is misclassified as goodware.

RESULTS

Model	ACC	TPR	FPR
Linear SVC	0.968	0.932	0.015
BernoulliNB	0.940	0.822	0.008
Random Forest	0.994	0.993	0.004

Without Data Balancing

RESULTS

Model	ACC	TPR	FPR
Linear SVC	0.967	0.957	0.028
BernoulliNB	0.945	0.847	0.010
Random Forest	0.995	0.996	0.005

With Data Balancing

CONCLUSION

- We proposed a comparative analysis between various classification models. We used feature selection and data balancing techniques and observed their effects on different classification models, for the detection of ransomware.
- We employed the Variance inflation factor technique for feature selection and eliminated the unnecessary features. The major advantage it played was that they reduced the training complexity classification algorithms successfully.
- Next, we applied the resampling technique, SMOTE and Tomek Links, to rebalance the imbalanced datasets and compared the accuracies without balancing and with balancing on different machine learning models such as Linear SVC, Bernoulli Naive Bayes and Random Forest.
- From the experiments we conducted, we have found that the Random Forest Classifer gave better results compared to Linear SVC, Bernoulli Naive Bayes.

FUTURE SCOPE

In the future, we recommend exploring additional oversampling and undersampling techniques to further enhance the performance of the classification algorithms. Additionally, we emphasize the importance of developing strategies for eradicating ransomware after its detection, as this remains a critical aspect in mitigating the impact of such malicious attacks. By considering both detection and eradication in future research, we can work towards more robust and comprehensive ransomware defense mechanisms.

Thank You