

**RANSOMWARE DETECTION AND RECOVERY PLANNING**

(Team Members: Harshith Martha- TZ18899, Preet Patel- CB69612, Sai Sri Harsha Kumbam - EF41093, Ramya Jyotsna Neelakantrao - WI30334, Jaswanth Ram Nagabhyrava - NW46934)

University of Maryland, Baltimore County

IS 734: Data Analytics for Cybersecurity

Faisal Quader, PhD

TABLE OF CONTENTS:

[1. ABSTRACT 2](#_Toc166179641)

[2. INTRODUCTION 2](#_Toc166179642)

[3. OBJECTIVES 3](#_Toc166179643)

[4. BACKGROUND WORK 3](#_Toc166179644)

[5. METHODOLOGY 5](#_Toc166179645)

[5.1 DATA COLLECTION AND SOURCE 5](#_Toc166179646)

[5.2 DATASET OVERVIEW 5](#_Toc166179647)

[5.3 DATA PRE-PROCESSING 6](#_Toc166179648)

[5.4 FEATURES ENGINEERING 7](#_Toc166179649)

[6. MODEL TRAINING AND EVALUATION 8](#_Toc166179650)

[6.1 DATA SPLITTING 8](#_Toc166179651)

[6.2 MODELS BUILT 8](#_Toc166179652)

[7. RESULTS AND VISUALIZATIONS 9](#_Toc166179653)

[7.1 RESULTS 9](#_Toc166179654)

[7.2 VISUALIZATIONS 10](#_Toc166179655)

[8. CHALLENGES FACED: 12](#_Toc166179656)

[9. CONCLUSION 13](#_Toc166179657)

[10. FUTURE WORK 13](#_Toc166179658)

[11. REFERENCES 13](#_Toc166179659)

# ABSTRACT

Ransomware is a persistent and complex threat to cybersecurity, necessitating advanced detection and response strategies beyond standard safeguards. In this study, we utilize cutting-edge machine learning techniques to create a reliable prediction model for the early detection of ransomware. We collect a large dataset and use various machine learning algorithms to find minute patterns that point to ransomware activity, like file access and modification activities. We show that our models are effective in improving the predictive power of cybersecurity defenses through stringent evaluation procedures. In addition, our project investigates recovery planning techniques designed to lessen the effects of ransomware attacks, protecting operational resilience and business continuity by reducing downtime and data corruption. Through the use of data analytics to clarify the technological aspects of ransomware detection, our research advances knowledge regarding cybersecurity procedures. Furthermore, our results highlight how machine learning integration can be used to strengthen digital infrastructures against new cyber threats.

# INTRODUCTION

Cybersecurity on an individual and organizational level has been seriously threatened by the rise in ransomware attacks in recent years. Ransomware, a kind of malicious software that encrypts a target's data and requests a ransom to unlock it, has developed into one of the most significant types of online crime. These assaults damage critical data and cause significant financial losses in addition to disrupting operations. Even while they are still important, standard cybersecurity solutions are becoming less and less effective against the dynamic and complex threats posed by ransomware.

The quick adaptation and evolution of these harmful programs highlights the need for more robust protection against ransomware. Because ransomware operators are always improving their techniques to take advantage of software and human behavior flaws, cybersecurity systems must be able to anticipate and proactively counter possible threats. This project was started in response to this mounting difficulty with the goal of improving ransomware recovery and detection techniques by utilizing machine learning.

Our goals are twofold: first, we want to create a predictive model that makes use of machine learning algorithms to effectively identify early signs of ransomware activity; and second, we want to investigate and put recovery planning techniques into place that lessen the impact of such attacks. This research uses a variety of machine learning approaches, including Random Forest, Logistic Regression, and Support Vector Machines, to accomplish these aims by utilizing a dataset of 62,000 occurrences, which includes both benign and malicious files. With the help of the prediction model, cybersecurity professionals will be able to avert possible ransomware attacks with more precision and with more time to spare.

To sum up, this introduction lays the groundwork for a discussion of the specific techniques, outcomes, and project consequences while highlighting the critical role that advanced data analytics can play in the battle against ransomware.

# OBJECTIVES

The principal aim of this project is to improve cybersecurity defenses against ransomware threats by utilizing machine learning techniques to develop a predictive model. Using sophisticated algorithms like Random Forest, Logistic Regression, and Support Vector Machines, the goal is to find early warning signs of ransomware activity so that preventive mitigation measures can be put in place before serious damage is done.

Establishing effective recovery plans is crucial to restoring systems quickly, minimizing downtime, and minimizing financial losses after a ransomware attack, in addition to early detection. The recovery plans shall be carefully designed to cover a variety of scenarios with an emphasis on prompt response and restoration to guarantee operational resilience and business continuity.

In addition, this project uses insights from a large dataset with 62,000 occurrences to improve the responsiveness and adaptability of cybersecurity defenses. The project aims to improve the accuracy and consistency of ransomware detection by refining and optimizing detection methods through rigorous analysis and experimentation. The research aims to provide organizations with the tools and strategies needed to effectively combat evolving ransomware threats by advancing the development of adaptive cybersecurity defenses.

In conclusion, this project's goals include creating a predictive model for early ransomware detection as well as putting strong recovery plans into place and strengthening adaptive cybersecurity defenses. By using these diverse strategies, the project aims to increase cybersecurity resilience and lessen ransomware's effects on individuals and organizations alike.

# BACKGROUND WORK

Ransomware has emerged as a significant cybersecurity threat, prompting extensive research into efficient detection and mitigation strategies. This section outlines key studies and methodologies that have influenced and shaped the current project's focus on ransomware detection and recovery planning.

The dataset used in this project was sourced from Kaggle, a popular platform for data science and machine learning. The dataset comprises approximately 62,000 instances and encompasses a diverse array of attributes pertinent to ransomware detection, including MD5 hashes, file names, machine architecture, debug sizes, and resource sizes. This dataset provides a robust foundation for training and evaluating machine learning models.

Siddiqui et al. (2008) were the pioneers in utilizing static binary feature extraction to identify ransomware. Their approach included analyzing file attributes, such as file size, entropy, and import tables, which are indicative of ransomware behavior. This foundational work provided a framework for identifying malicious files based on their static features, which has helped shape subsequent research in this domain.

Mohaisen and Alrawi (2015) extended Siddiqui's research by utilizing machine learning algorithms to examine supplementary file characteristics, including dynamic link library (DLL) imports and function calls. Their research has demonstrated the efficacy of utilizing machine learning to analyze intricate file behaviors, resulting in enhanced precision in the detection of ransomware. This study significantly influenced the feature engineering and model selection processes of the current project.

Breiman (2001) introduced the Random Forest algorithm, which is an ensemble learning method that combines multiple decision trees to enhance classification accuracy and reduce overfitting. This algorithm was a suitable choice for this project's ransomware detection model due to its ability to handle a variety of features and its robustness against noise and overfitting.

Cortes and Vapnik (1995) devised Support Vector Machines, a potent classification technique that is particularly effective in high-dimensional spaces. The ability of SVM to establish a distinct distinction between classes renders it an ideal candidate for distinguishing between benign and malicious files within the dataset.

Logistic Regression, a fundamental machine learning algorithm, provides probabilistic outputs and is extensively utilized for binary classification tasks. Its simplicity and interpretability make it an ideal benchmark for comparing the performance of more complex models.

Techniques such as data normalization, outlier detection, and feature creation were informed by the extensive methodologies outlined in Han, Kamber, and Pei (2011), and Friedman, Hastie, and Tibshirani (2001) These steps ensure that the data are clean, balanced, and optimized for model training.

The use of metrics such as accuracy, precision, recall, F1 score, and AUC-ROC, along with visualization techniques such as histograms, scatter plots, and confusion matrices, was guided by best practices in data science and machine learning, as described by Provost and Fa These metrics and visualizations hold significant importance in evaluating model performance and comprehending data distributions.

# METHODOLOGY

## 5.1 DATA COLLECTION AND SOURCE

This dataset was acquired using the open-source website Kaggle, a well-known center for data science and machine learning that offers a collaborative environment for experts and beginners to research and collaborate on difficult data problems.The present example's data set was selected with care to satisfy the objectives of the model-building and analysis procedures, which were based on reports of ransomware attacks. It consists of anonymous data that has been recorded from numerous ransomware outbreaks, along with related data and DLL characteristics.

## 5.2 DATASET OVERVIEW

The selected dataset provides a comprehensive overview of all the related attributes and instances that might be necessary for further processes. With the help of the dataset, it was possible to build machine learning models and predict whether the file was malicious or benign.

The dataset consists of 62485 instances and 18 attributes. The attributes are mentioned as below:

1. FileName: The name of the file.
2. md5Hash: MD5 hash of the file, which is a unique identifier for the contents of the file.
3. Machine: An identifier for the system architecture.
4. DebugSize: Information related to the debugging data.
5. DebugRVA: Information related to the debugging data.
6. MajorImageVersion: Version information of the image.
7. MajorOSVersion: Version information of the operating system.
8. ExportRVA: Information about the export table.
9. ExportSize: Information about the export table.
10. IatVRA : Import Address Table Virtual Relative Address.
11. MajorLinkerVersion : Version of the linker used.
12. MinorLinkerVersion : Version of the linker used.
13. NumberOfSections : Number of sections in the file.
14. SizeOfStackReserve: Size of stack reserve.
15. DllCharacteristics: DLL characteristics flags.
16. ResourceSize: Size of resources in the file.
17. BitcoinAddresses: Indicates the presence of Bitcoin addresses in the file (potentially relevant for ransomware analysis).
18. Benign: A label indicating whether the file is benign (1) or malicious (0).

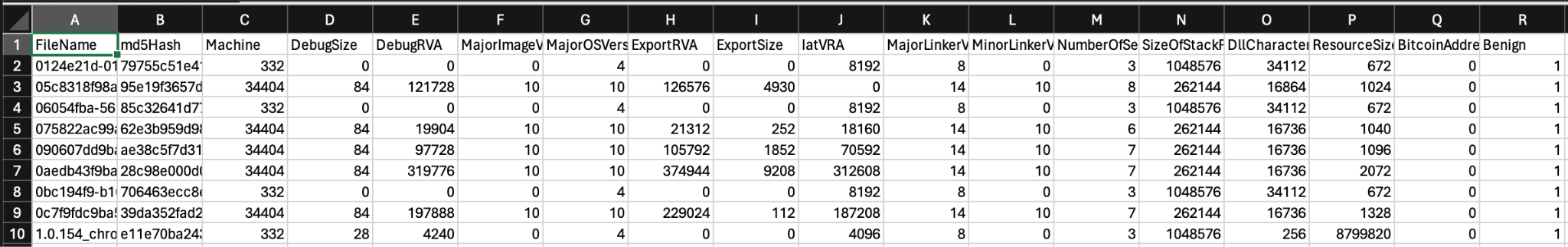


Figure:1 Sample Data



Figure 2: No. of Malware and Benign Files in raw data.

## 5.3 DATA PRE-PROCESSING

In order to guarantee the robustness and dependability of our machine learning models, the dataset was first processed. To avoid biases in the model training process, we started by locating and eliminating duplicate entries. When missing values were statistically significant, they were carefully managed using either imputation or exclusion in situations where data integrity might be jeopardized.

Using traditional scaling procedures, all numerical features in the dataset were standardized to have a mean of zero and a standard deviation of one. By normalizing the data, it is ensured that no single feature's scale will have an undue influence on the model. In order to make comparisons across variables easier, min-max scaling was also used for some features to move the data into a 0–1 range.

By utilizing the Interquartile Range (IQR) approach to address outliers, further data integrity was strengthened. This method assisted in lessening the negative effects of extreme values on the analysis. Data reduction techniques were utilized in conjunction with these stages to narrow down the analysis to the most trustworthy and relevant attributes.

With the help of these preparatory stages, the dataset was guaranteed to be fully prepared, providing a strong basis for further analytical methods, such as machine learning modeling, which extract useful insights and improve prediction accuracy.

## 5.4 FEATURES ENGINEERING

In the machine learning pipeline, feature engineering is a critical stage where unstructured data is converted into useful features to enhance interpretability and model performance. Several feature engineering techniques were used in this project to preprocess the dataset and extract pertinent data for ransomware detection.

We optimized the dataset for machine learning model training in our project by performing a thorough feature engineering process. In order to maintain consistency and stop any one feature from controlling the learning process, we started by normalizing and standardizing numerical features. Utilizing the StandardScaler from sklearn.preprocessing, we set the mean and standard deviation of numerical features to zero and one, respectively.

Then, in order to order the most informative attributes for model training, we used feature selection techniques. Based on the ANOVA F-test, we chose the top k features using the SelectKBest function from sklearn.feature\_selection. By identifying the characteristics that had the strongest correlation with the target variable (malicious or benign files), we were able to decrease the dimensionality of the model and increase its discriminatory power.

We also dealt with dataset outliers, which have the potential to distort model predictions. We separated outliers that differed noticeably from the rest of the data by determining z-scores for numerical features and establishing an outlier detection threshold. This step reduced the impact of anomalous data points, ensuring the model's robustness and dependability.

Using seaborn and matplotlib, we created a heatmap visualization of the correlation matrix to better understand feature relationships and multicollinearity. During model training and feature selection, this visualization technique helped to uncover important feature correlations and make well-informed decisions.

We improved the dataset for machine learning model training through thorough feature engineering, which improved the predictive model's interpretability, robustness, and accuracy for ransomware detection. These preprocessing procedures improved cybersecurity defenses at the individual and organizational levels by laying the foundation for a dependable defense against ransomware attacks.

# MODEL TRAINING AND EVALUATION

For the purpose of creating a reliable predictive system for ransomware detection, the model training and evaluation step was crucial. We used a number of machine learning methods, such as Support Vector Machines (SVM), Random Forest, and Logistic Regression, and Gradient Boost all of which were selected for their special abilities to handle classification difficulties and find patterns in intricate datasets.

## 6.1 DATA SPLITTING

To evaluate the models' performance, the dataset was divided into training and testing sets prior to starting the model training process. Two subsets of the dataset were created: one for training the models (80%) and another for testing their performance (20%). This permits an independent assessment of the models' performance on unobserved data and guarantees that they are trained on an adequate volume of data.

## 6.2 MODELS BUILT

1. *Logistic regression:* Based on the features of the dataset, the logistic regression model was trained to forecast ransomware activity. Predictions were made on the testing data following initialization and training with the LogisticRegression class from sklearn.linear\_model. This produced an accuracy score and a classification report with specific metrics like precision, recall, and F1-score for both benign and malicious classifications. The removal of outliers significantly improved the model's performance, demonstrating the importance of data preprocessing in improving model accuracy.
2. *Random Forest:* Based on the features of the dataset, the Random Forest classifier was used to forecast ransomware activity. After initializing and training on the training data, the model was trained using the RandomForestClassifier from sklearn.ensemble. Following that, predictions were made using the testing data, and the accuracy score, classification report, and confusion matrix were used to assess the model's performance. Furthermore, the model's capacity to differentiate between benign and malicious classifications was evaluated by computing the area under the curve (AUC) score and the receiver operating characteristic (ROC) curve.
3. *Support Vector Machine:* Based on the features of the dataset, the Support Vector Machine (SVM) classifier was used to forecast ransomware activity. The model was trained on the training set of data after being initialized with probability estimation enabled using the SVC class from sklearn.svm. Following that, predictions were made using the testing data, and the accuracy score, classification report, and confusion matrix were used to assess the model's performance. Furthermore, the model's capacity to differentiate between benign and malicious classifications was evaluated by computing the area under the curve (AUC) score and the receiver operating characteristic (ROC) curve.
4. *Gradient Boost:* The Gradient Boosting classifier was utilized to predict ransomware activity based on the dataset features. Using the GradientBoostingClassifier from sklearn.ensemble, the model was initialized and trained on the training data. Predictions were then made on the testing data, and the model's performance was evaluated using accuracy score, classification report, and confusion matrix. Moreover, the model's ability to distinguish between benign and malicious classifications was assessed through the calculation of the receiver operating characteristic (ROC) curve and the area under the curve (AUC) score.

All models performed well, reliably predicting both benign and malicious files with high accuracy. This demonstrates how well they detect ransomware and offer strong cybersecurity protections.

# RESULTS AND VISUALIZATIONS

We were able to evaluate model performance, comprehend data trends, and effectively convey conclusions thanks to the invaluable assistance of statistical methods and visualizations in our project. Through feature engineering and model evaluation, these tools guided us as we gained insights into the behavior of the models and the structure of the dataset.

## 7.1 RESULTS

The performance of each model in predicting ransomware activity was evaluated using several key metrics, including accuracy, precision, recall, F1-score, and area under the curve (AUC) for the receiver operating characteristic (ROC) curve. The results are summarized in the table below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-score** | **AUC** |
| Logistic Regression | 0.9257 | 0.9533 | 0.8686 | 0.9089 | 0.9736 |
| Random Forest | 0.9978 | 0.9977 | 0.9970 | 0.9974 | 0.9994 |
| SVM | 0.9788 | 0.9775 | 0.9728 | 0.9751 | 0.9898 |
| Gradient Boosting | 0.9929 | 0.9954 | 0.9879 | 0.9916 | 0.9988 |

Table1: Model Results

The evaluation results show how well the models work at correctly identifying files that are malicious or benign. They demonstrate strong performance in differentiating between the two classes, with accuracy scores above 92% for all models. Strong recall, precision, and F1-score values further confirm their ability to accurately detect ransomware instances while reducing false positives and negatives. The models' persistently high area under the curve (AUC) values highlight their capacity to distinguish between malicious and benign files, demonstrating their dependability in the identification of ransomware. These findings demonstrate how machine learning techniques can significantly improve cybersecurity defenses by offering precise and trustworthy malware presence predictions.

## 7.2 VISUALIZATIONS

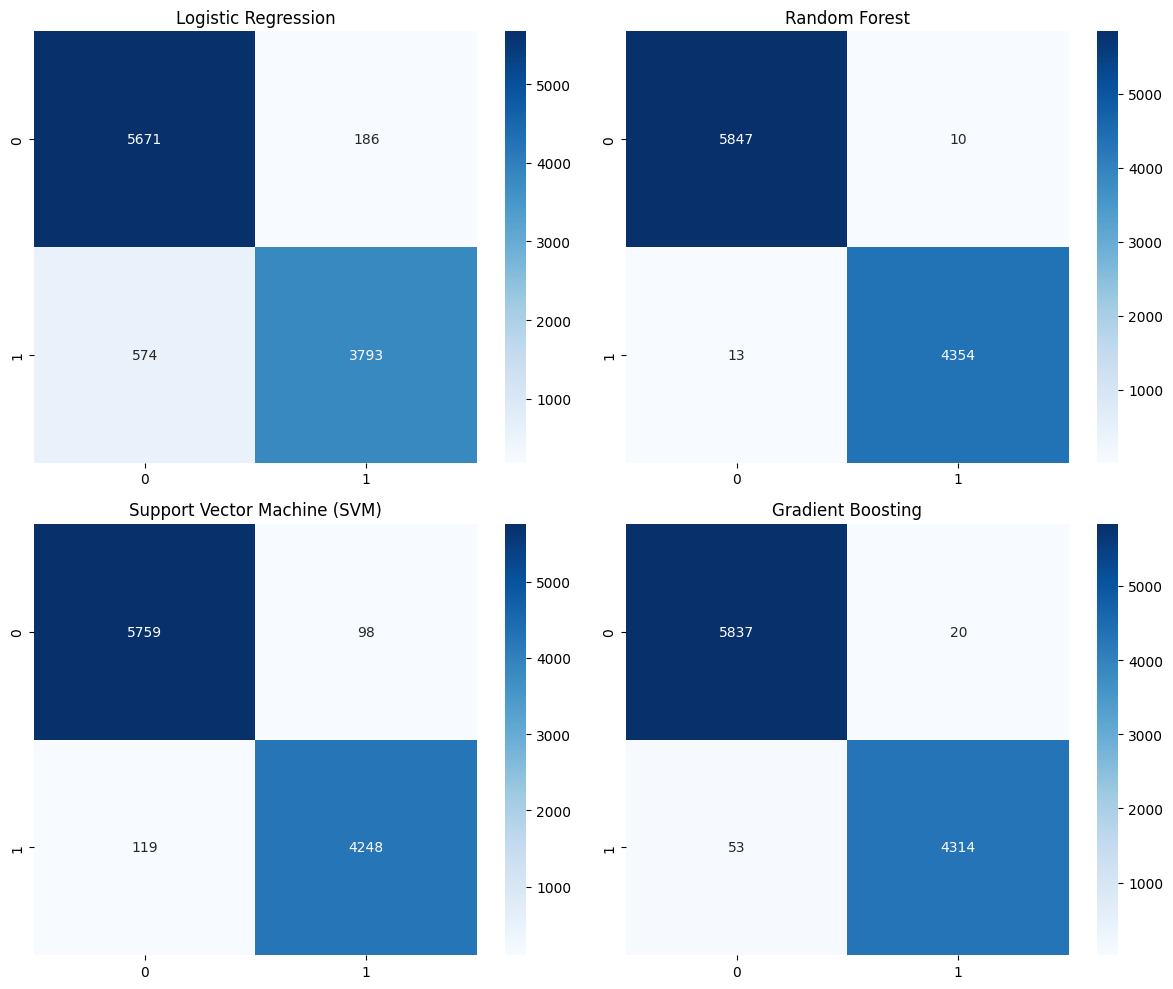
****

Figure 3: Confusion Matrix of all the Models

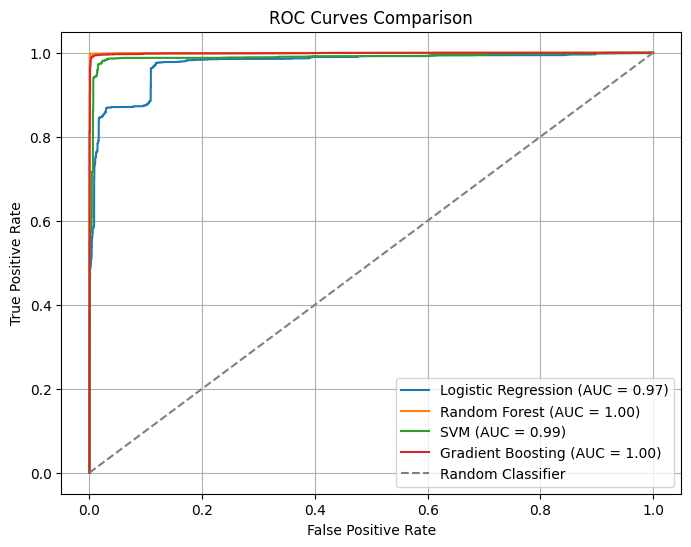


Figure 4: ROC curves of the Models

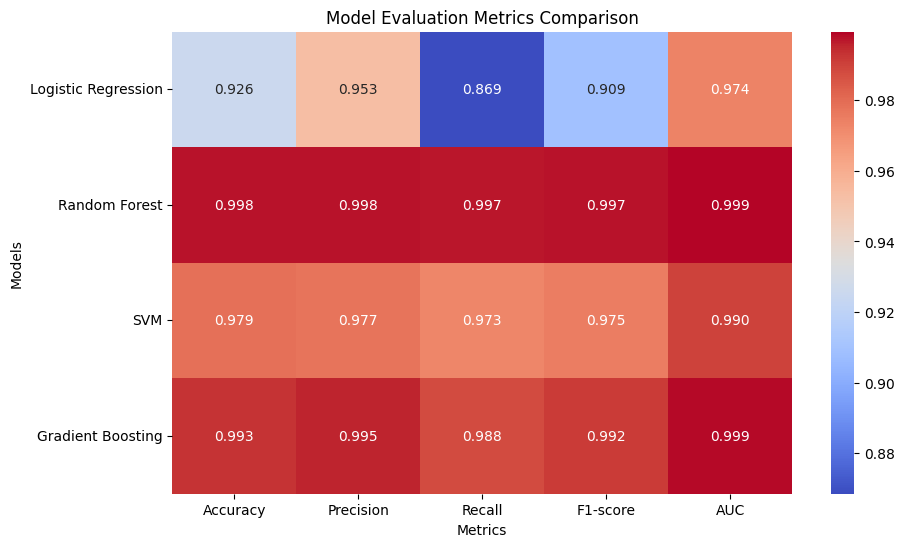


Figure 5: Comparison of the Models

In general, statistical methods and visualizations were essential to our study since they gave us the means to investigate, comprehend, and improve our models. Our cybersecurity solution was strengthened by these techniques, which guaranteed the accuracy and dependability of our predictive models for ransomware detection.

# CHALLENGES FACED:

During the course of the project, several challenges were encountered and addressed. Initially, procuring a suitable dataset proved to be challenging, thereby hindering the efficacy of evaluating model performance until a comprehensive dataset was obtained from Kaggle. The Random Forest model initially performed poorly, requiring extensive parameter tuning and feature importance analysis to improve its accuracy. Another obstacle was the complexity of feature engineering, as unfamiliarity with the dataset's attributes initially complicated the process, requiring thorough exploratory data analysis and statistical tests to identify significant features. It was challenging to handle the imbalance between benign and malicious files in the dataset, as this led to model bias towards the majority class. This was mitigated through resampling techniques and the use of metrics like the F1 score and AUC-ROC. Various machine learning models required extensive tuning and validation, which increased the complexity and duration of development. It was critical to ensure scalability and generalization of the models in order to prevent overfitting and enhance performance on unseen data. This was achieved through cross-validation and testing on separate validation sets. Furthermore, the implementation of efficient recovery planning strategies necessitated meticulous planning and testing to minimize downtime and data loss, with a particular emphasis on automated backup and restoration procedures.

Furthermore, the complexity of data preprocessing, including normalization, handling missing values, and outlier detection, required the establishment of a systematic preprocessing pipeline to streamline data preparation and maintain consistency across model iterations. These challenges collectively underscored the necessity for robust methodologies and continuous optimization to develop efficient ransomware detection and recovery solutions.

# CONCLUSION

Robust machine learning models that can accurately detect ransomware have been developed by the project. By using an extensive dataset, these models are able to recognize early warning signs of ransomware activity, which improves cybersecurity defenses and predictive capabilities. Machine learning has a great deal of promise to prevent detect and mitigate ransomware threats when integrated into cybersecurity frameworks. The project ensures comprehensive management of ransomware incidents by concentrating on both detection and recovery planning, thereby mitigating potential downtime and financial losses. Furthermore, the study offers insightful information about the use of machine learning and advanced data analytics in cybersecurity, particularly in the identification and mitigation of ransomware threats. Emphasizing the significance of continuous model evaluation and feature engineering, the study underscores the importance of sustaining and enhancing model performance over time. Overall, the project's findings highlight the crucial role of machine learning in fortifying cybersecurity measures against evolving ransomware threats.

# FUTURE WORK

1. Exploration of advanced data mining techniques such as clustering, anomaly detection, and deep learning to enhance granularity and precision in data analysis and classification for improved ransomware detection.
2. Integration of real-world ransomware and benign file records to validate and refine the model's effectiveness across diverse and dynamic environments, minimizing the risk of overfitting to test data.
3. Enhancement of recovery planning strategies by developing comprehensive and automated procedures, including automated backup and restore processes, incident response automation, and integration with broader cybersecurity frameworks, to minimize system downtime and financial consequences following ransomware attacks.

# REFERENCES

*Ransomware detection data set*. (2022, July 31). Kaggle. <https://www.kaggle.com/datasets/amdj3dax/ransomware-detection-data-set>

Siddiqui, M. A., Wang, M. C., & Lee, J. (2008). *A survey of data mining techniques for malware detection using file features*. <https://doi.org/10.1145/1593105.1593239>

Mohaisen, A., Alrawi, O., & Mohaisen, M. (2015). AMAL: High-fidelity, behavior-based automated malware analysis and classification. *Computers & Security*, *52*, 251–266. <https://doi.org/10.1016/j.cose.2015.04.001>

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. <https://doi.org/10.1007/bf00994018>

Hastie, T., Tibshirani, R., & Friedman, J. H. (2001). The elements of statistical learning. In *Springer series in statistics*. <https://doi.org/10.1007/978-0-387-21606-5>

*Data science for business*. (n.d.). Google Books. <https://books.google.com/books?hl=en&lr=&id=EZAtAAAAQBAJ&oi=fnd&pg=PP1&dq=Provost,+F.,+%26+Fawcett,+T.+(2013).+&ots=ymWMQu6QA-&sig=zmXVNhj87XjXEahMDzCSwdfLgBk#v=onepage&q=Provost%2C%20F.%2C%20%26%20Fawcett%2C%20T.%20(2013).&f=false>

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection. *ACM Computing Surveys*, *41*(3), 1–58. <https://doi.org/10.1145/1541880.1541882>