# Comparing Machine Learning Models for Malware Detection

John Williams

# Introduction - Malware

- Ransomware and cybersecurity breaches are a growing threat to businesses.
- There is a growing need to detect and block malware as quickly as possible while minimizing false positives.
- Security companies are using machine learning to evaluate files and processes to determine if the files are malicious or not.

# Introduction - Value

- Classifying malware as malicious or not is valuable to organizations because of cybersecurity threats.
- Coding value is found in comparing the algorithms. This allows an organization to pick an algorithm that completes the classification task.

# Literature Review – Survey Papers

- Academic papers surveying malware detection with machine learning were valuable.
- The survey papers aimed to explain the basics of malware and detection. Academic papers presenting machine learning detection were shared.
- These papers provided information on how to implement malware detection using machine learning models.

# Literature Review – ML Models

- Support Vector Model and Random Forest algorithms were researched during the literature review.
- Academic papers presented good results, >95% detection accuracy, using SVM and RF.
- The data imported in the models varied. PE Header data was researched and found to be a good dataset for detection.

# Methodology – Planning

- Perform data processing and feature extraction.
- Write python code to run the machine learning models.
- Measure speed and accuracy of the models.

# Methodology – Algorithms

- Support Vector Model – algorithm that splits linear or non-linear data points across n-dimensional planes. Good for high dimensional analysis. High compute requirements with large datasets, extremely slow.

- Random Forest – algorithm that uses multiple decision trees and averages the trees. This prevents overfitting. Extremely fast compared to SVM.

# Implementation – Programming

- Use Python modules, Pandas and Numpy, to extract and process data and features.
- Use Sklearn modules to setup the learning models.
- Use the Python time module to measure speed of execution.
- The code runs in Google Colab. Allows for a standardization for hardware.

# Implementation - Dataset

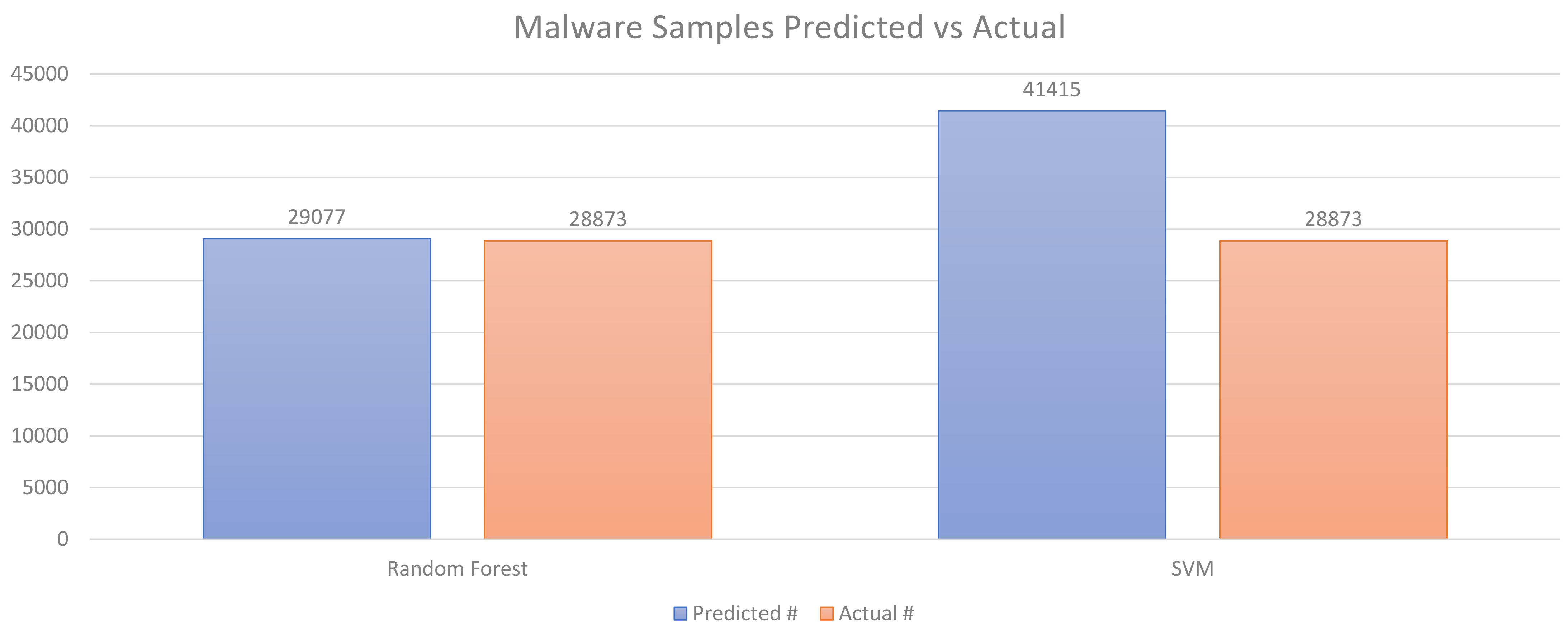
- Data will be standardized.
- Perform data cleaning, remove any null values.
- Use the train_test_split Sklearn module to prepare the dataset.

# Experimental Setup – Metrics

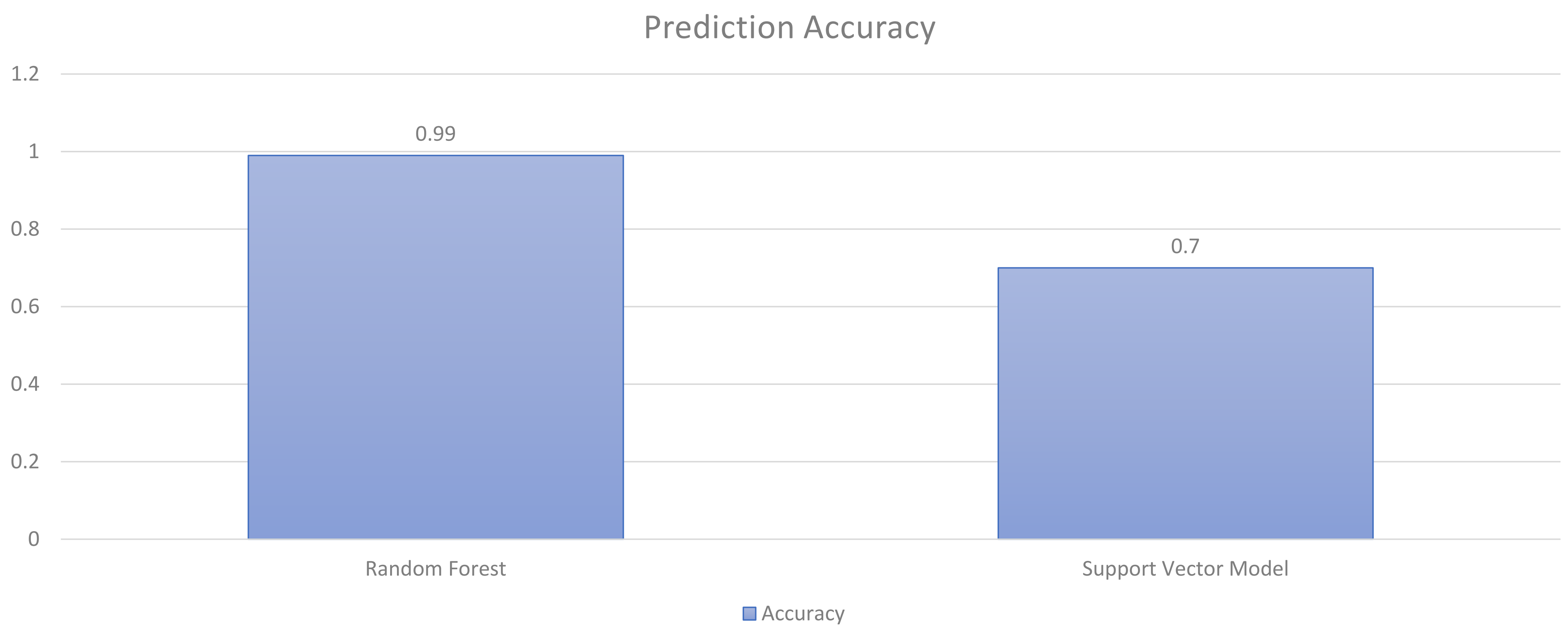
- Use mean absolute error, and execution time to determine which model performs better.
- Validation dataset size is 10% of the dataset.

# Demo

# Initial Results Analysis– Prediction Count Comparison



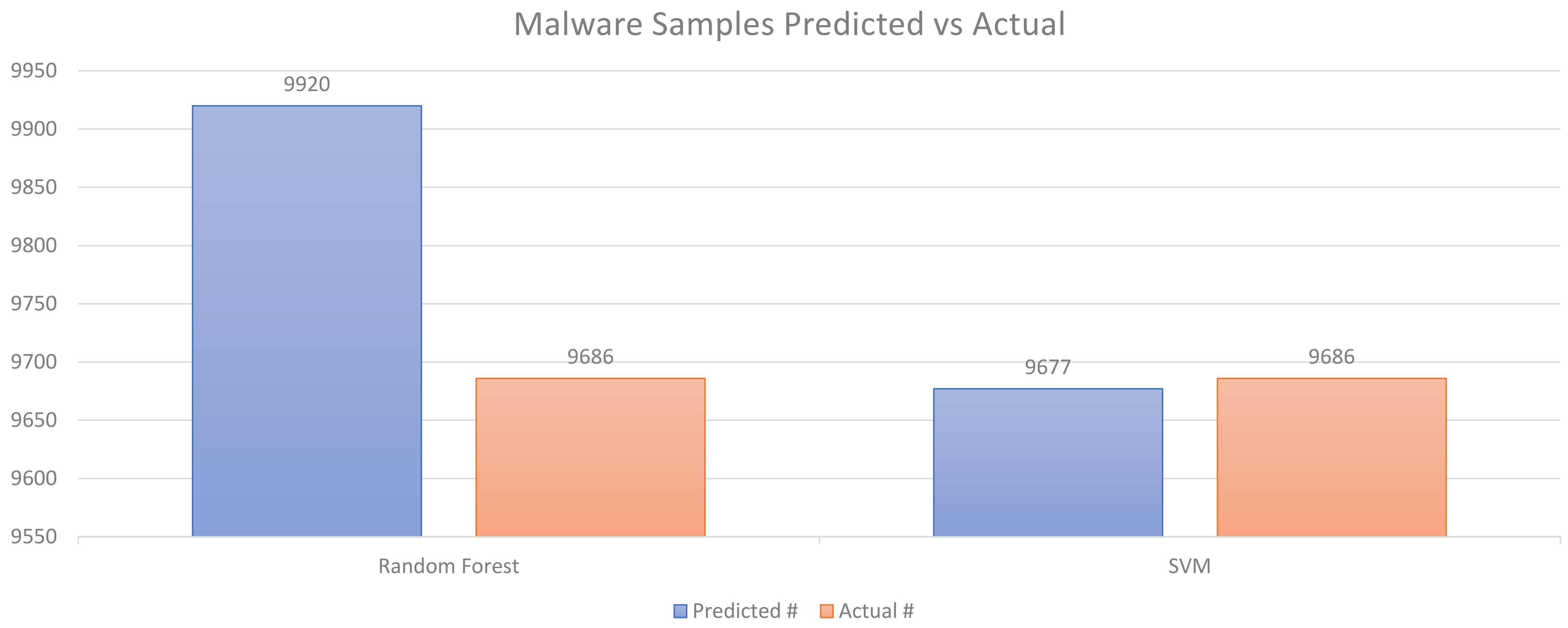
# Initial Results Analysis - Accuracy



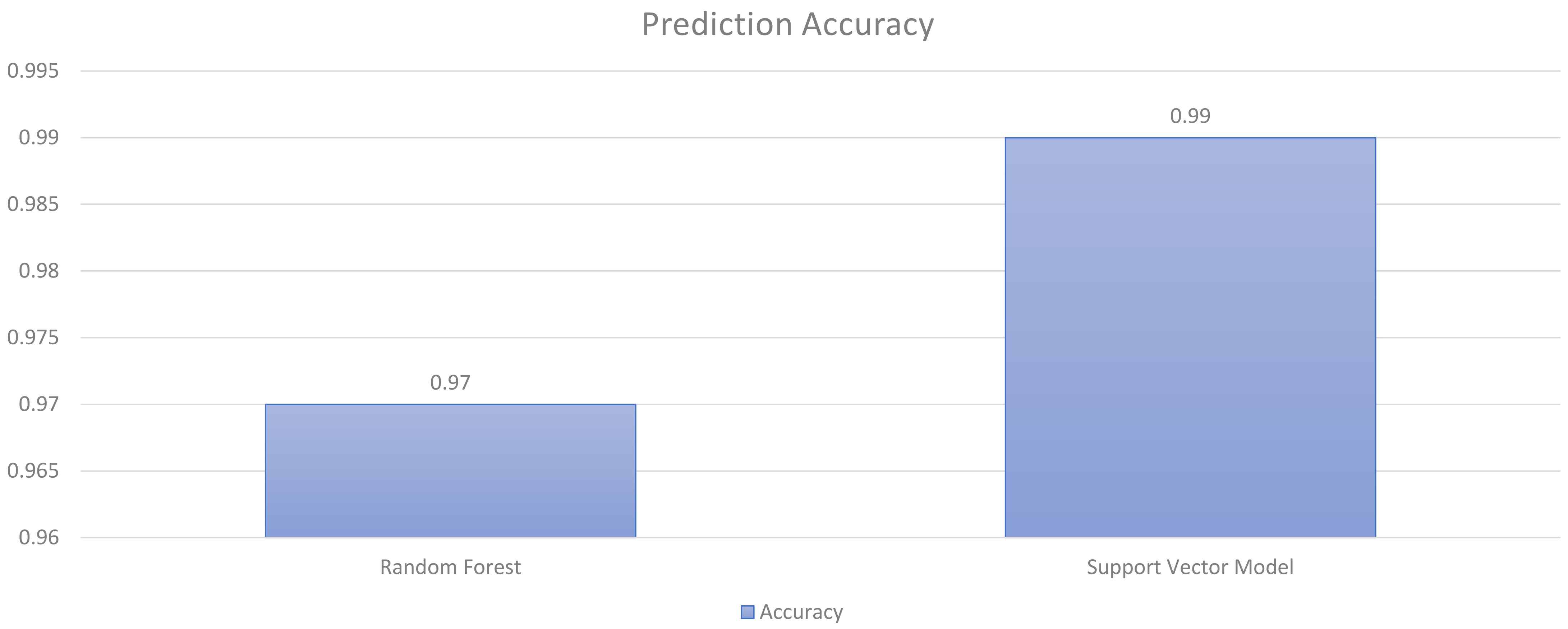
# Testing Modifications

- Changed the test sample size from 25% to 10% of the dataset.
- Scaled the dataset using Sklearn module preprocessing.standardscaler.

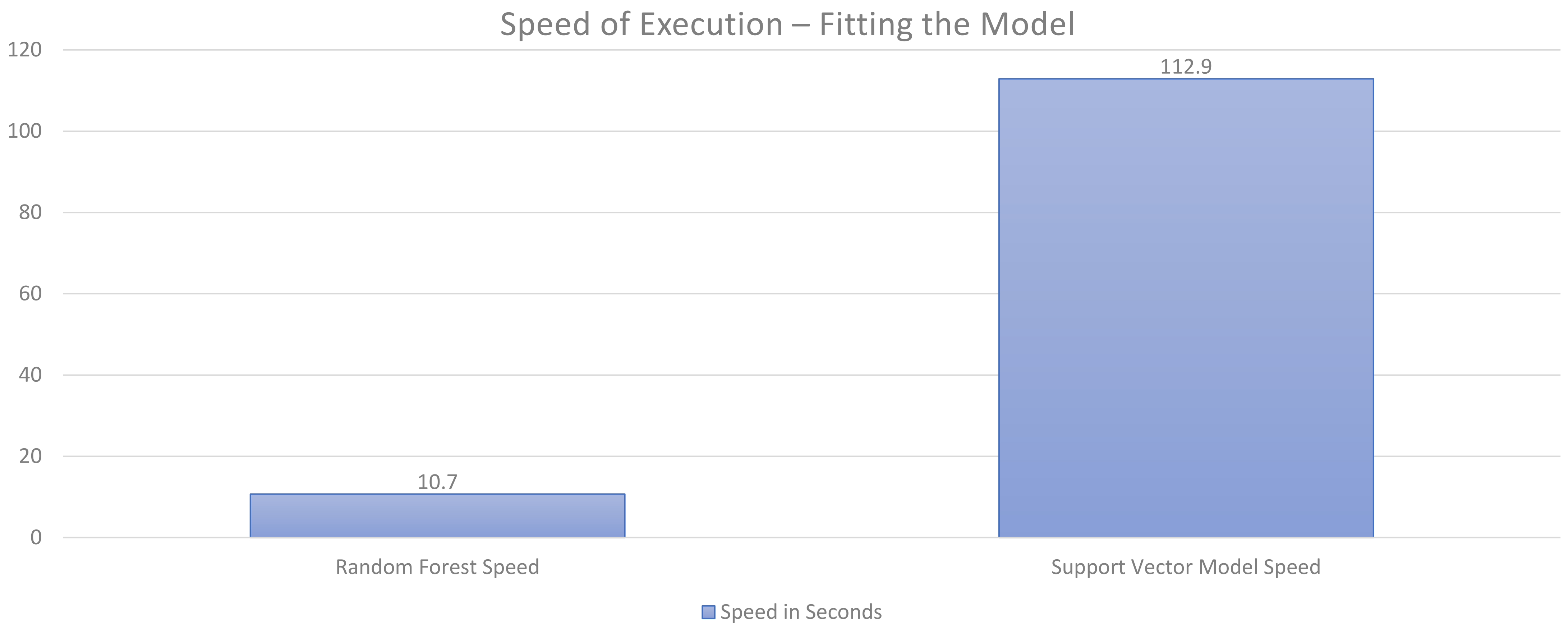
# Final Results Analysis - Prediction Count Comparison



# Final Results Analysis - Accuracy



# Final Results Analysis - Speed



# Conclusion

- Initial results showed that Random Forest was more accurate.
- After scaling the dataset, SVM had better accuracy.

# References


- Jagsir Singh, Jaswinder Singh, A survey on machine learning-based malware detection in executable files, Journal of Systems Architecture, Volume 112, 2021, 101861, ISSN 1383-7621, https://doi.org/10.1016/j.sysarc.2020.101861.Analyzing Machine Learning Approaches for Online Malware Detection in Cloud
- Daniel Gibert, Carles Mateu, Jordi Planes, The rise of machine learning for detection and classification of malware: Research developments, trends and challenges, Journal of Network and Computer Applications, Volume 153, 2020, 102526, ISSN 1084-8045, https://doi.org/10.1016/j.jnca.2019.102526.
- Damaševičius R, Venčkauskas A, Toldinas J, Grigaliūnas Š. Ensemble-Based Classification Using Neural Networks and Machine Learning Models for Windows PE Malware Detection. Electronics. 2021; 10(4):485. https://doi.org/10.3390/electronics10040485
- Jeffrey C Kimmell, Mahmoud Abdelsalam, Maanak Gupta, Analyzing Machine Learning Approaches for Online Malware Detection in Cloud, 2021, https://arxiv.org/pdf/2105.09268.pdf
- Edward Raff, Jared Sylvester, Charles Nicholas, Learning the PE Header, Malware Detection with Minimal Domain Knowledge, 2017, https://arxiv.org/ftp/arxiv/papers/1709/1709.01471.pdf