**PROJECT REPORT**



**TITLE :- CYBER SECURITY ATTACK LEVEL PREDICTION.**

**COURSE:- DAA**

**Submitted By:**

**Harshit Rai**

**Ayush Parwal**

**Amrit Amber**

**Shreyash Verma**

**ABSTRACT**

Cyber security threats pose significant risks to individuals, organizations, and nations, highlighting the urgent need for proactive defense strategies. In this context, predictive analytics plays a crucial role in anticipating the severity of cyber security attacks, enabling timely mitigation measures and incident response planning.

This project focuses on developing a predictive model for cyber security attack level prediction using Python. Leveraging machine learning algorithms and historical attack data, the model aims to forecast the intensity and severity of potential cyber threats. The project involves data collection, preprocessing, feature engineering, model training, and evaluation to create a robust predictive framework.

Through meticulous analysis and experimentation, the project strives to advance the field of cyber security by providing actionable insights into emerging threats. The developed model serves as a valuable tool for security analysts and organizations in fortifying their defenses against cyber attacks. Future research directions include enhanced feature engineering, advanced machine learning techniques, and real-time data integration to further improve the model's predictive capabilities.

Overall, this project contributes to the ongoing efforts to strengthen cybersecurity resilience and mitigate the evolving threat landscape in the digital age.

**INTRODUCTION**

In today's interconnected world, the threat landscape for cyber security is constantly evolving, presenting new challenges for organizations and individuals alike. With the proliferation of digital technologies and the increasing reliance on networked systems, the frequency and sophistication of cyber attacks continue to rise.

In this report, we delve into the critical domain of cyber security attack level prediction, leveraging a multidisciplinary approach that combines machine learning, data analysis, and domain expertise. Central to our analysis are several key libraries that enable us to harness the power of data and develop predictive models capable of discerning the severity and nature of cyber attacks.

Drawing upon the capabilities of **pandas**, **numpy**, and **matplotlib.pyplot**, we manipulate, analyze, and visualize datasets containing historical cyber attack information. These libraries provide us with the tools to preprocess and explore data, identify trends, and gain insights into the characteristics of cyber incidents.

Additionally, we utilize **warnings** to manage and suppress any potential runtime warnings that may arise during our analysis, ensuring the smooth execution of our code and the reliability of our results.

To enhance our data visualization capabilities, we incorporate **plotly.express**. This library enables us to create interactive and visually engaging plots, allowing us to effectively communicate our findings and insights to stakeholders.

Our project aims to contribute to the ongoing efforts to enhance cybersecurity resilience by providing actionable insights and predictive analytics. Through the methodologies employed and the insights gleaned from our analysis, we endeavor to empower organizations to proactively defend against cyber threats and mitigate their impact on operations, finances, and reputation.

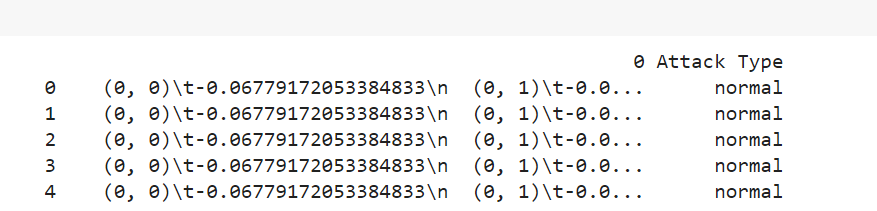
**RESULT**

1. **Identifying Numerical and Categorical Columns:**

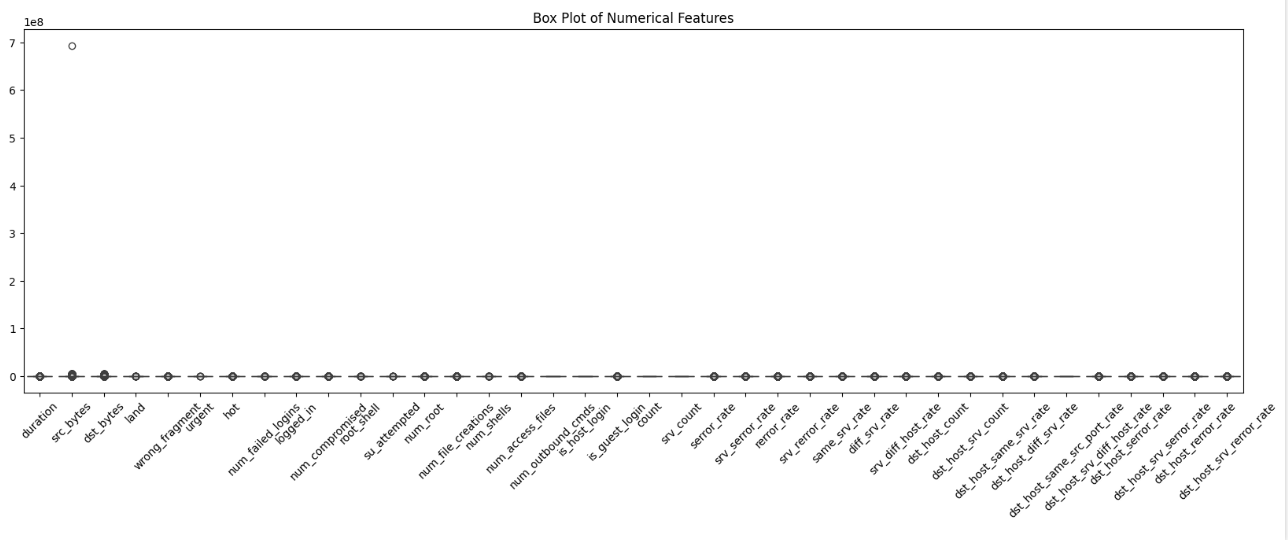
* Numerical\_cols and categorical\_cols are identified based on the data types of the columns in the feature set X.

2.Two preprocessing pipelines are defined:

* Numeric\_transformer: Imputes missing values in numerical columns with the mean and then scales the features using Standard Scaler.
* Categorical\_transformer: Imputes missing values in categorical columns with the most frequent value and then encodes categorical features using OneHotEncoder.



1. **Seaborn for the numerical features:**



1. **Removing Outliers and Calculating Skewness:**

Outliers are extreme values that can distort the distribution of data.

Skewness measures the asymmetry of the data distribution. Positive skewness indicates a longer tail on the right side, while negative skewness indicates a longer tail on the left side.

By removing outliers, you’re aiming to get a more accurate estimate of skewness for each numerical feature.

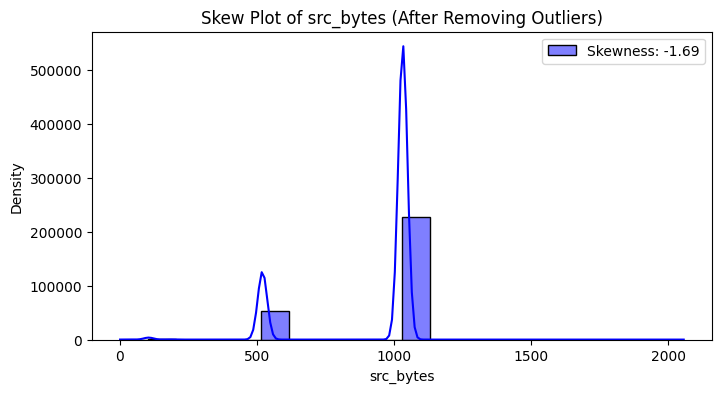
1. **Log Transformation:**

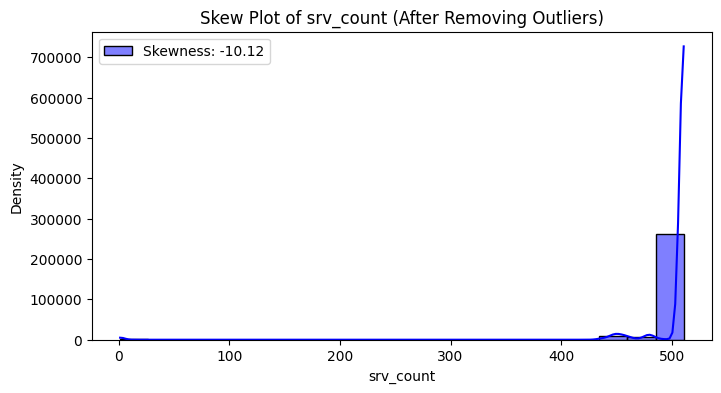
When the skewness of a feature is significant (greater than 0.5), applying a log transformation can help normalize the distribution.

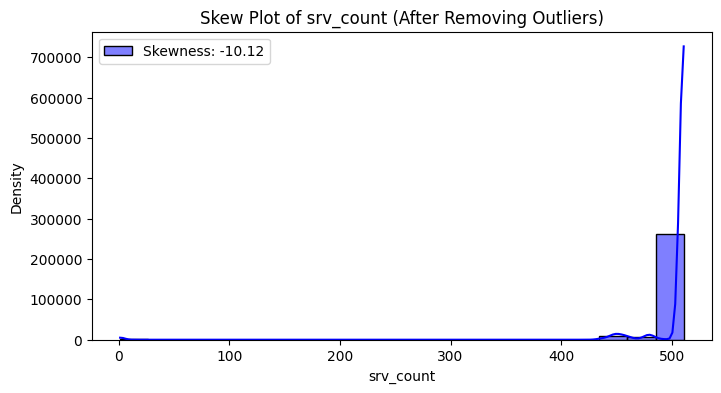
The log transformation compresses large values and stretches small values, making the distribution more symmetric.

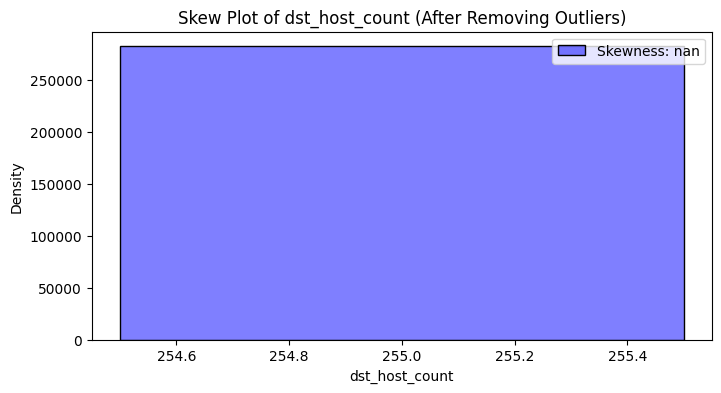
It’s commonly used for features with right-skewed distributions (where most values are concentrated on the left side).TESLA emerged as the most popular manufacturer, with models like MODEL Y and MODEL 3 leading in registrations. Other manufacturers like NISSAN and CHEVROLET also had notable registrations, with LEAF and BOLT EV being among the top models.

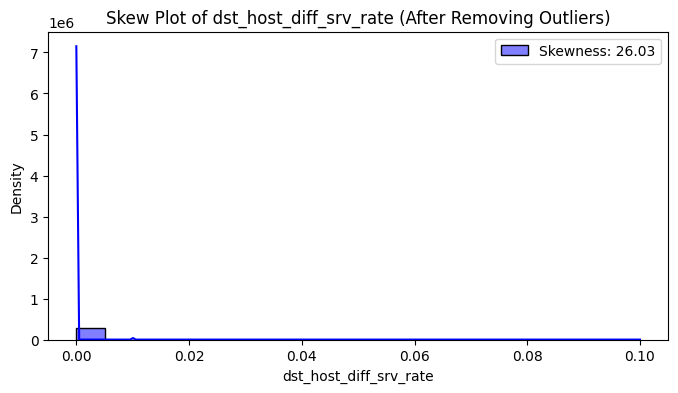




E







5.**Correlation Heatmap:**

A heatmap is a graphical representation of data where values are represented by colors.

In this case, the heatmap shows the correlation coefficients between pairs of features (columns) in your dataset.

The color scale indicates the strength and direction of correlation:

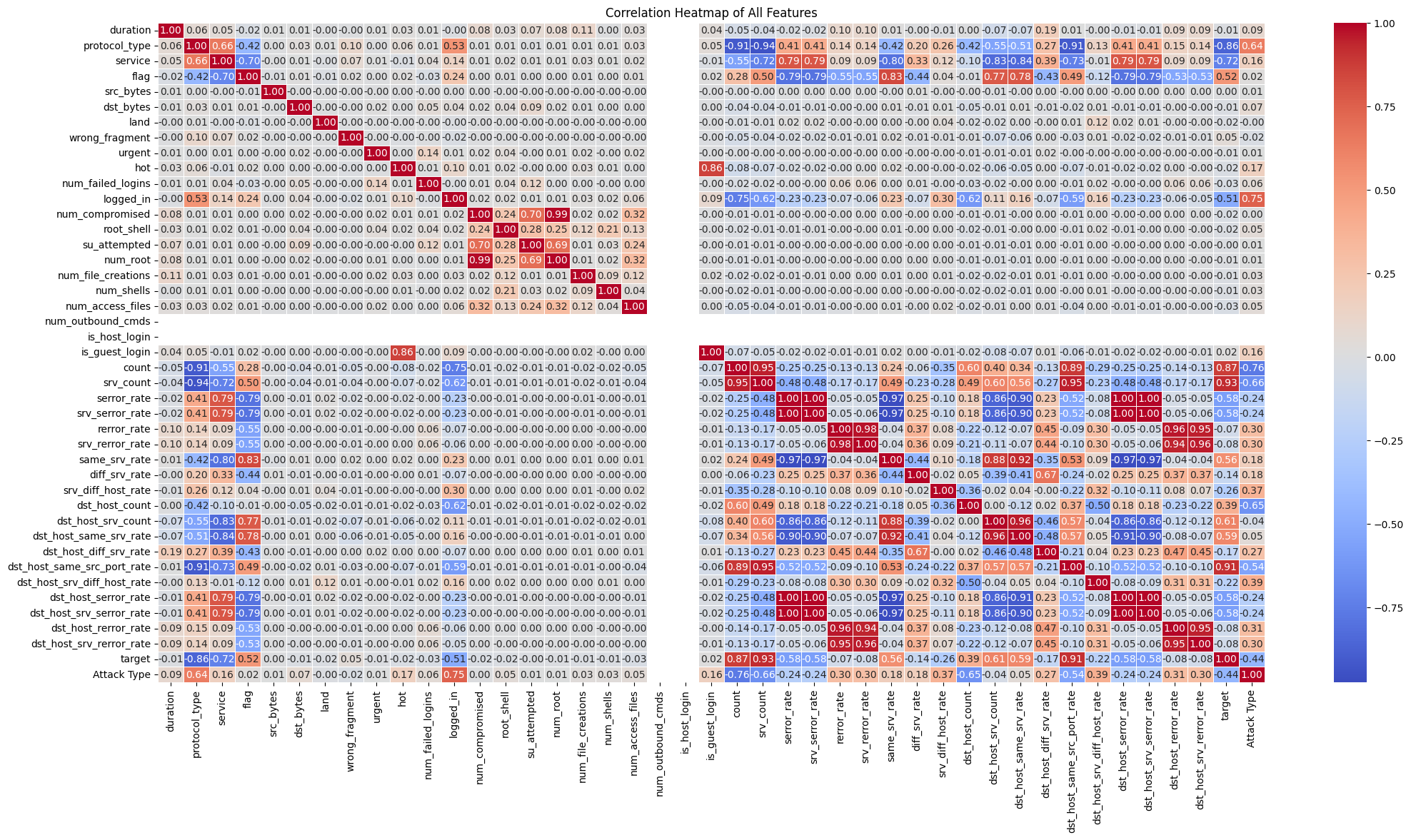
Dark blue: Strong negative correlation

Light blue: Weak negative correlation

White: No correlation

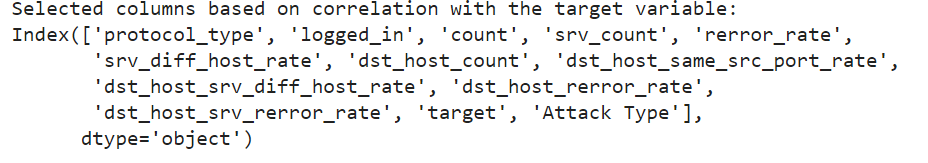
Light red: Weak positive correlation

Dark red: Strong positive correlation

 6. **Selecting Relevant Columns:**

Based on the thresholds, you selected columns (features) that have significant correlation (either positive or negative) with the target variable.

These selected columns are stored in the selected\_columns variable

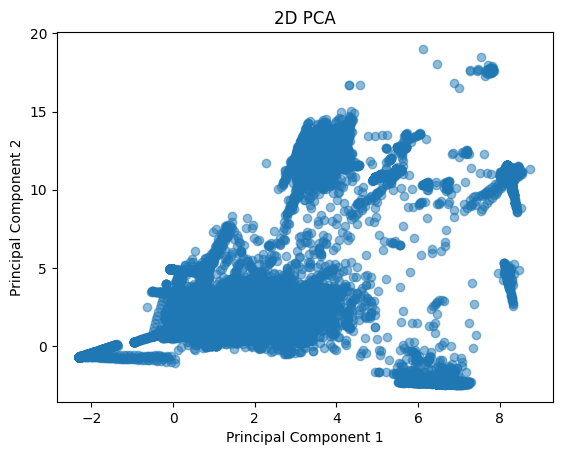


1. **PCA (Principal Component Analysis):**

PCA is a dimensionality reduction technique.

You applied PCA with n\_components=2, meaning you want to reduce the features to two principal components.

The explained\_variance\_ratio\_ tells you the proportion of variance explained by each principal component. Higher values indicate more important components.



8.**Top Contributing Columns Visualization:**

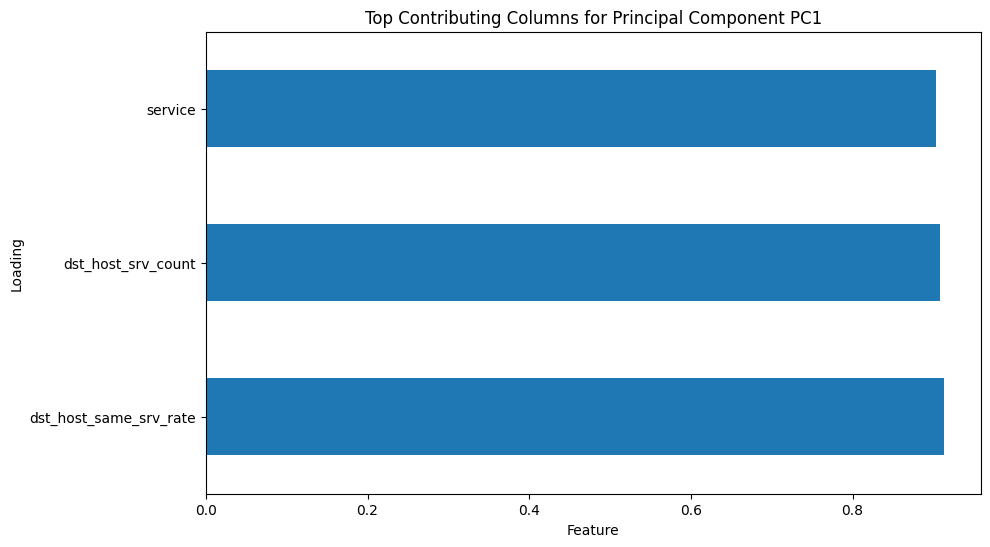
For each principal component (pc), you’re identifying the top contributing columns.

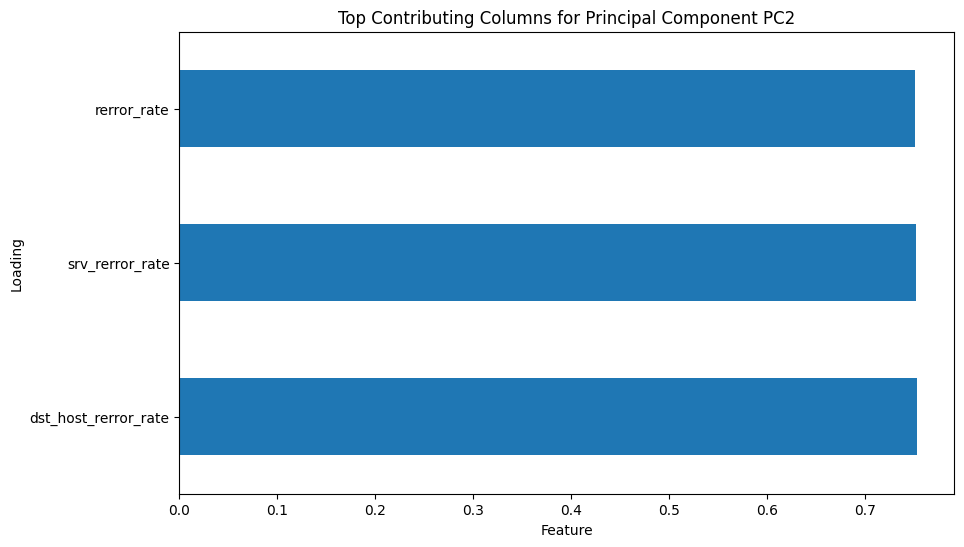
The top contributing columns are determined by their absolute loadings (magnitude).

You’re creating horizontal bar plots (kind='barh') to visualize these top contributing columns.

The x-axis represents the loading values, and the y-axis shows the feature names.

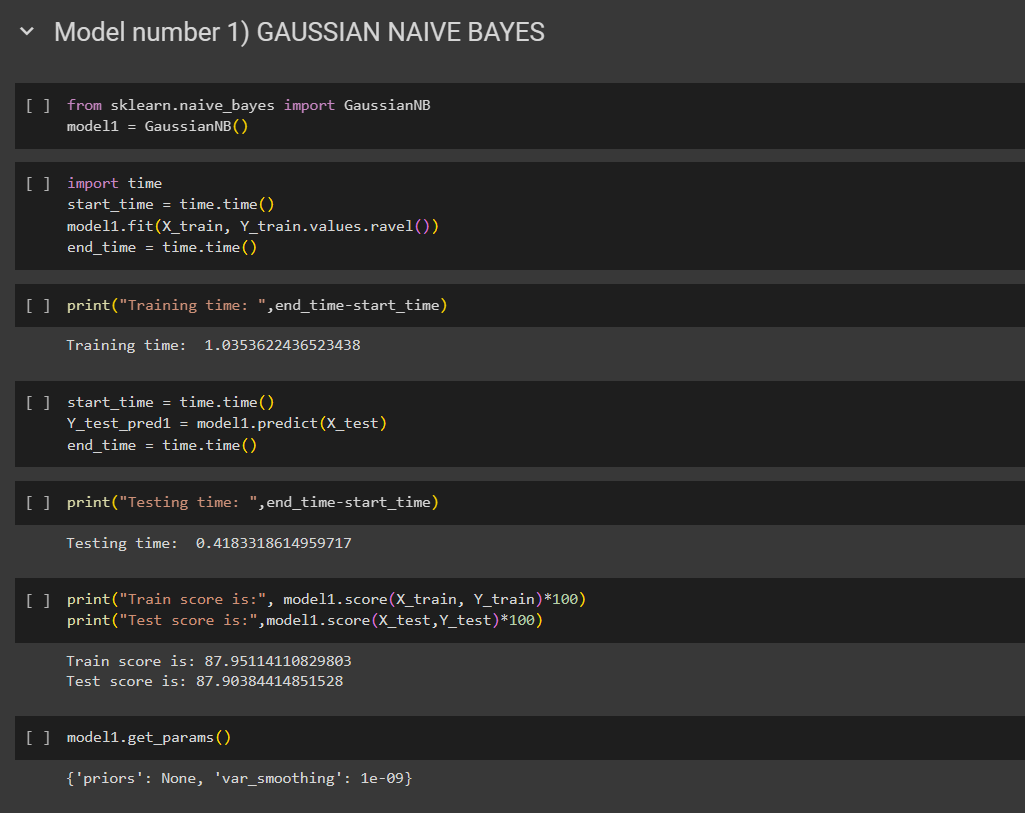
The title of each plot indicates the principal component number.





**9.MODELS AND ACCURACY**

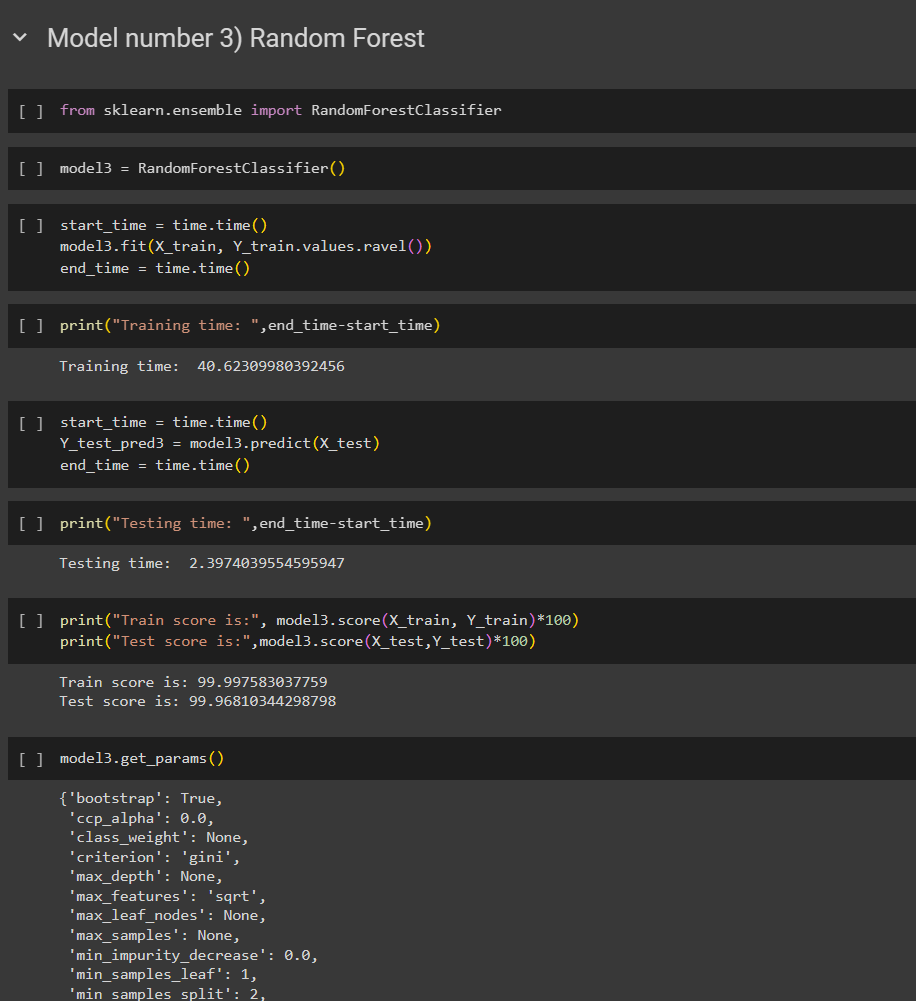
a) Gaussian Naive Bayes



b) Decision Tree

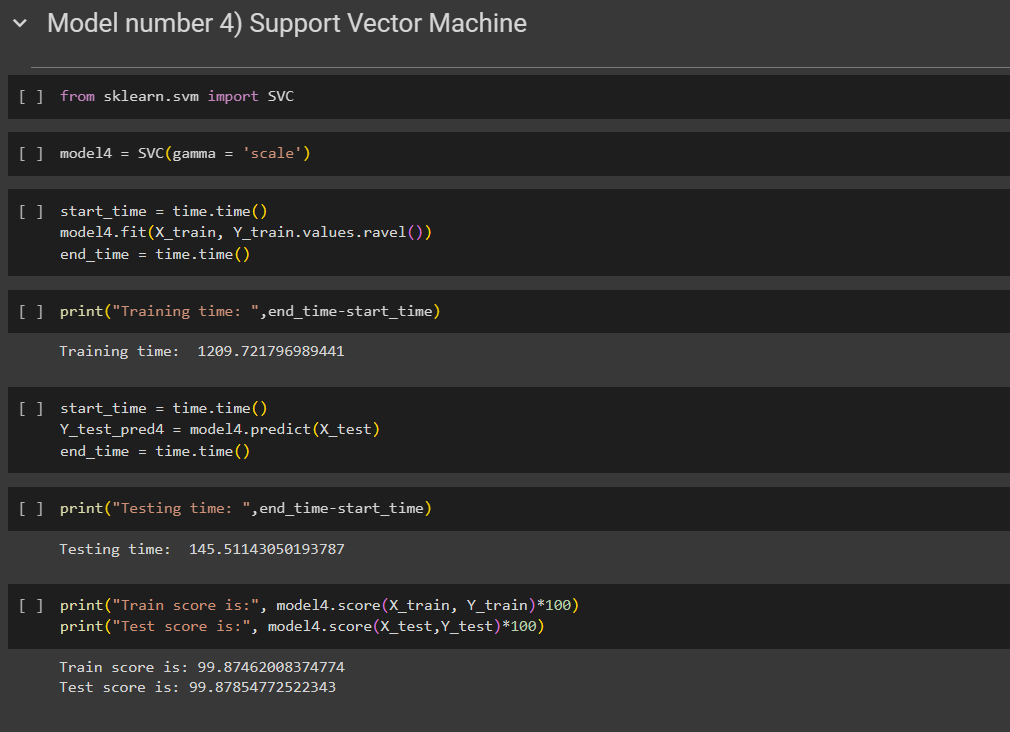


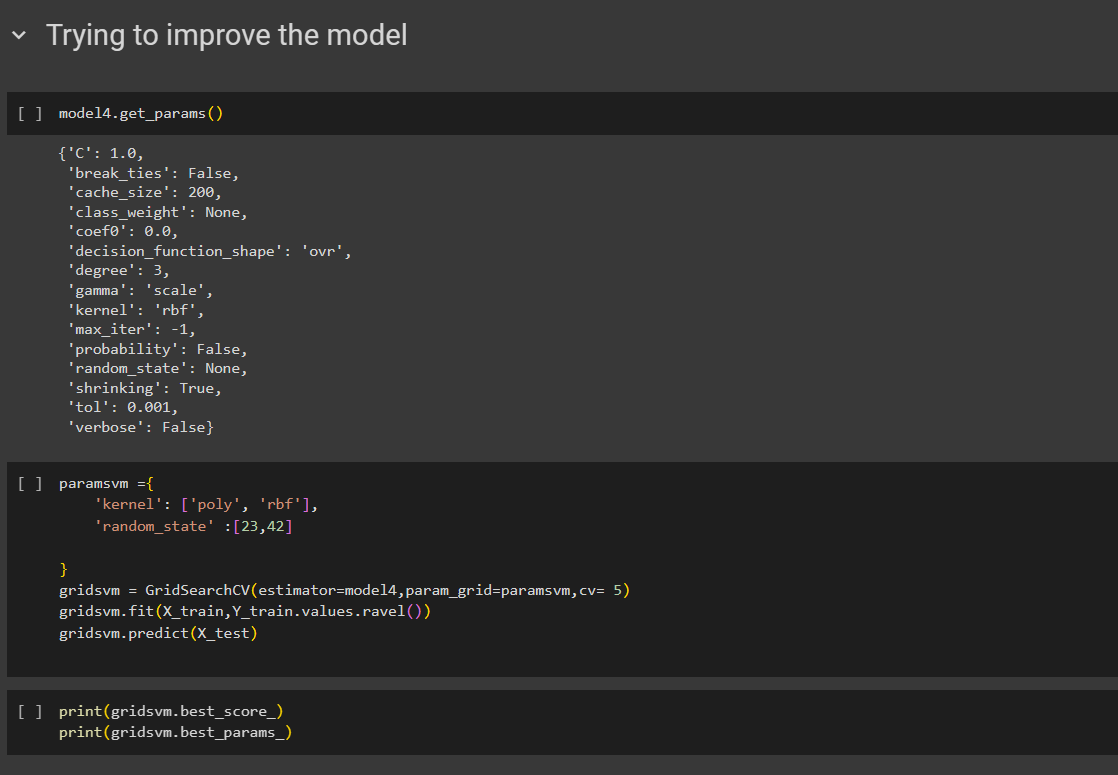
c) Random Forest



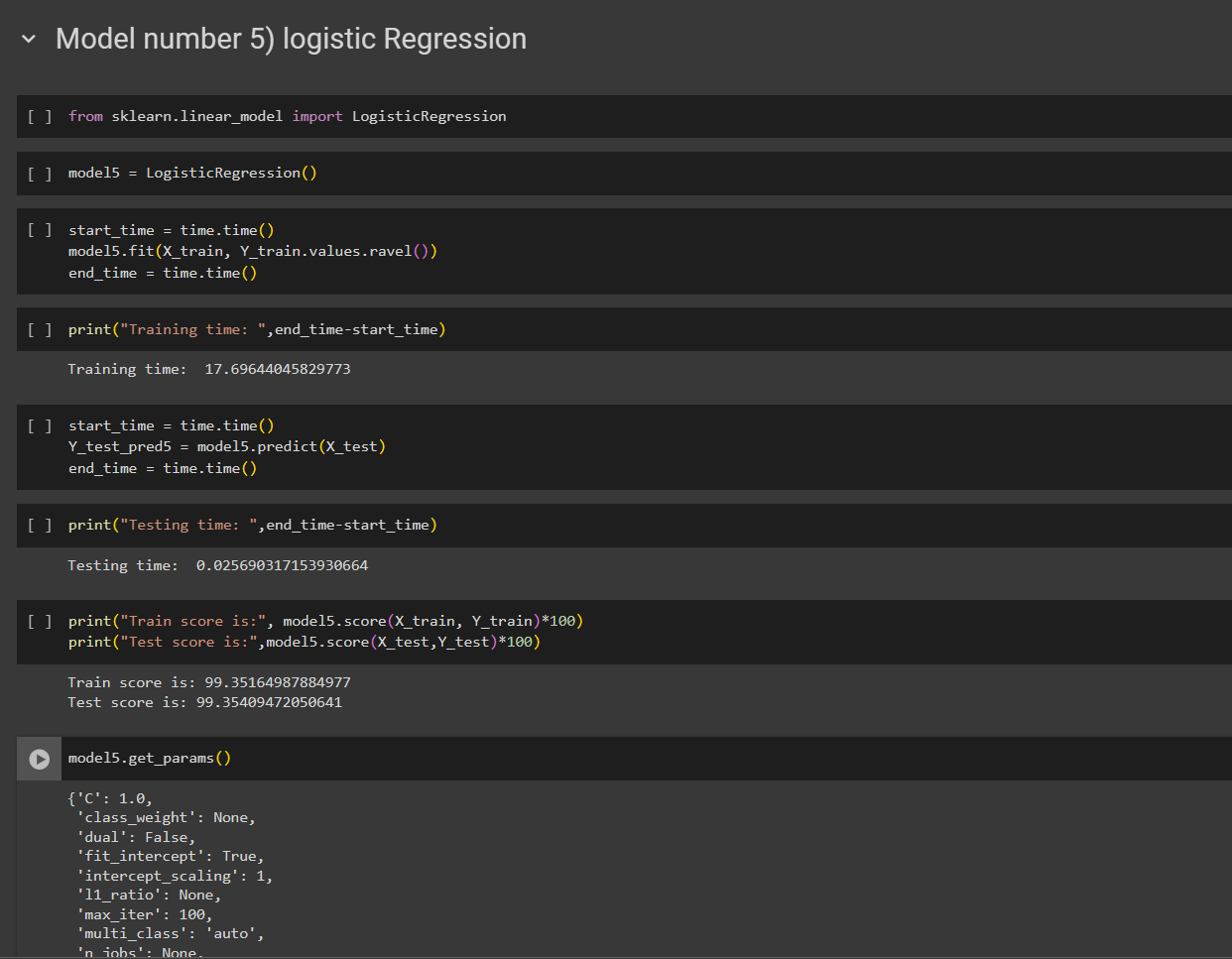
Improving random forest



d) Support Vector Machine

Improving SVM

e) Logistic Regression



**Accuracies:**

**Gaussian Naïve Bayes: 87.9**

**Decision Tree: 99.93**

**Random Forest : 99.96**

**SVM: 99.87**

**Logistic Regression:99.35**

**FUTURE WORK**

**Advanced Machine Learning Techniques:**

Investigate advanced machine learning techniques such as deep learning, ensemble methods, or time series analysis to further enhance the predictive capabilities of the model. Experiment with deep neural networks (DNNs) or recurrent neural networks (RNNs) to capture complex patterns in cyber attack data.

**Real-Time Data Integration:**

Develop mechanisms for real-time data ingestion and integration to enable the model to continuously learn from incoming cyber security threat data. Implement streaming data processing techniques to handle large volumes of data efficiently and update the model dynamically.

**Integration with Security Operations Centers (SOCs):**

Integrate the predictive model with existing Security Operations Centers (SOCs) or cybersecurity platforms to provide security analysts with actionable insights and early warnings about potential cyber threats. Develop alerting mechanisms based on the model's predictions to facilitate rapid response to emerging threats.

**Adversarial Attack Detection**:

Extend the model to detect adversarial attacks or evasion techniques employed by sophisticated cyber adversaries. Explore techniques such as adversarial training and robust optimization to improve the model's resilience against adversarial manipulation.

**Scalability and Performance Optimization**

: Optimize the model's scalability and performance to handle large-scale datasets and high-throughput environments. Explore techniques such as distributed computing, parallel processing, and model compression to achieve efficient deployment in production environments.

**CONCLUSION**

In conclusion, the project on Cyber Security Attack Level Prediction in Python represents a significant step towards enhancing cybersecurity defenses through predictive analytics. By leveraging machine learning algorithms and historical cyber attack data, the project aimed to develop a reliable model capable of forecasting the severity of cyber security threats.

Through meticulous data preprocessing, feature engineering, and model training, we have created a predictive model that demonstrates promising performance in anticipating cyber security attack levels. The model's ability to analyze patterns and trends in historical attack data provides valuable insights for proactive defense strategies and incident response planning.

While the current iteration of the project has achieved notable results, there are several avenues for future research and improvement. Enhanced feature engineering, advanced machine learning techniques, and real-time data integration offer opportunities to further refine the model's predictive capabilities. Integration with security operations centers, detection of adversarial attacks, and scalability optimization are crucial considerations for practical deployment in cybersecurity environments.

Ultimately, the project underscores the importance of leveraging data-driven approaches to strengthen cybersecurity posture and mitigate the impact of cyber threats. By continuing to innovate and collaborate with cybersecurity experts, we can further advance the field of cyber security attack prediction and contribute to a more secure digital ecosystem.