AGENDA

- Vinci Tool Integration: Integrate Vinci Tool to enhance software security through advanced vulner detection, transcending the capabilities of traditional static code analyzers.
- Threat Level Evaluation: Utilize Vinci Tool's feature to assess threat levels of identified vulnerabilities for comprehensive security understanding.
- Enhancing Static Code Analysis: Transform static code analysis by identifying deeper vulnerabilities, and addressing a critical security gap.
- Introduction to Vinci Tool's final integration - ML4Cyber Dashboard.
- Demonstration of the dashboard's capabilities in real-time threat detection and analysis.
- Discussing the dashboard’s impact on transforming reactive cybersecurity strategies into proactive solutions.
- Early Cyberattack Prevention: Emphasize early detection and prevention of cyberattacks through Vinci Tool's application in real-world scenarios, showcasing the practical benefits of machine learning in cybersecurity.

INTRODUCTION

- The cybersecurity domain is facing an ever-increasing urgency to evolve and adapt due to the escalating and evolving nature of cyber threats. This urgency underscores the need for innovative solutions to stay ahead of potential security breaches.

-Vinci Tool, unlike conventional static code analyzers, utilizes machine learning to predict and assess anomalies in software code, offering a more proactive approach to software security.

- Integrating advanced machine learning with traditional security measures revolutionizes software security, enables proactive approach to identifying and mitigating potential threats, thus enhancing overall cybersecurity posture.

- The goal is to enhance software security with Vinci Tool, revolutionizing against emerging threats by identifying and evaluating vulnerabilities, safeguarding digital assets amidst growing cyber threats.

- The advent of Vinci's ML4Cyber Dashboard ushers in a transformative approach to software security. By converging machine learning precision with user-friendly analytics, we empower organizations to not only predict threats but also to prevent them effectively, ensuring a resilient cybersecurity posture.

BUSINESS PROBLEM

The business question explores how Vinci Tool's integration can transition cybersecurity from Our mission transcends traditional cybersecurity methods.

The integration of the ML4Cyber Dashboard enables a shift from reactive defense to proactive protection.

This strategic tool leverages predictive analytics to safeguard digital infrastructure, ensuring the continuity of business operations and elevating risk management to a strategic function within the enterprise

DATA PREPARATION

Throughout the project, we've had the privilege of receiving sample data files from our generous sponsor.

These data files are comprehensive reports from widely available Static Code Analyzer tools

Our strategic choice of using ESLint aligns with our goal of enhancing software security.

The insights from ESLint reports form a strong foundation for identifying coding flaws and security weaknesses

EXPLORATORY DATA ANALYSIS

DATASET OVERVIEW

	ID	Severity	CWE	Type	Tool	Location	Path	FileName	Line	Status
0	66621	High	22.0	Path Traversal	ESLint	Atom x64\resources\app\apm\lib\link.js	Atom x64\resources\app\apm\lib	NaN	52	Escalated
1	66622	Medium	665.0	Initialization	ESLint	Atom x64\resources\app\apm\lib\link.js	Atom x64\resources\app\apm\lib	NaN	4	False Positive
2	66623	Medium	665.0	Initialization	ESLint	Atom x64\resources\app\apm\lib\link.js	Atom x64\resources\app\apm\lib	NaN	8	False Positive
3	66624	High	22.0	Path Traversal	ESLint	Atom x64\resources\app\apm\lib\link.js	Atom x64\resources\app\apm\lib	NaN	35	Escalated
4	66626	Medium	665.0	Initialization	ESLint	Atom x64\resources\app\apm\lib\link.js	Atom x64\resources\app\apm\lib	NaN	6	False Positive
...	...	...	...	...	...	...	...	...	...	...
702	89569	Medium	665.0	Initialization	ESLint	Atom x64\resources\app\apm\script\download-nod...	Atom x64\resources\app\apm\script	NaN	9	False Positive
703	89570	Medium	665.0	Initialization	ESLint	Atom x64\resources\app\apm\script\download-nod...	Atom x64\resources\app\apm\script	NaN	72	False Positive
704	89571	High	22.0	Path Traversal	ESLint	Atom x64\resources\app\apm\script\download-nod...	Atom x64\resources\app\apm\script	NaN	98	Escalated
705	89572	Medium	665.0	Initialization	ESLint	Atom x64\resources\app\apm\script\download-nod...	Atom x64\resources\app\apm\script	NaN	87	False Positive
706	89573	Medium	665.0	Initialization	ESLint	Atom x64\resources\app\apm\script\download-nod...	Atom x64\resources\app\apm\script	NaN	1	False Positive

707 rows × 10 columns

We initiated our data analysis by using the "Atom.csv" file and Python.

707 records in our labeled dataset.

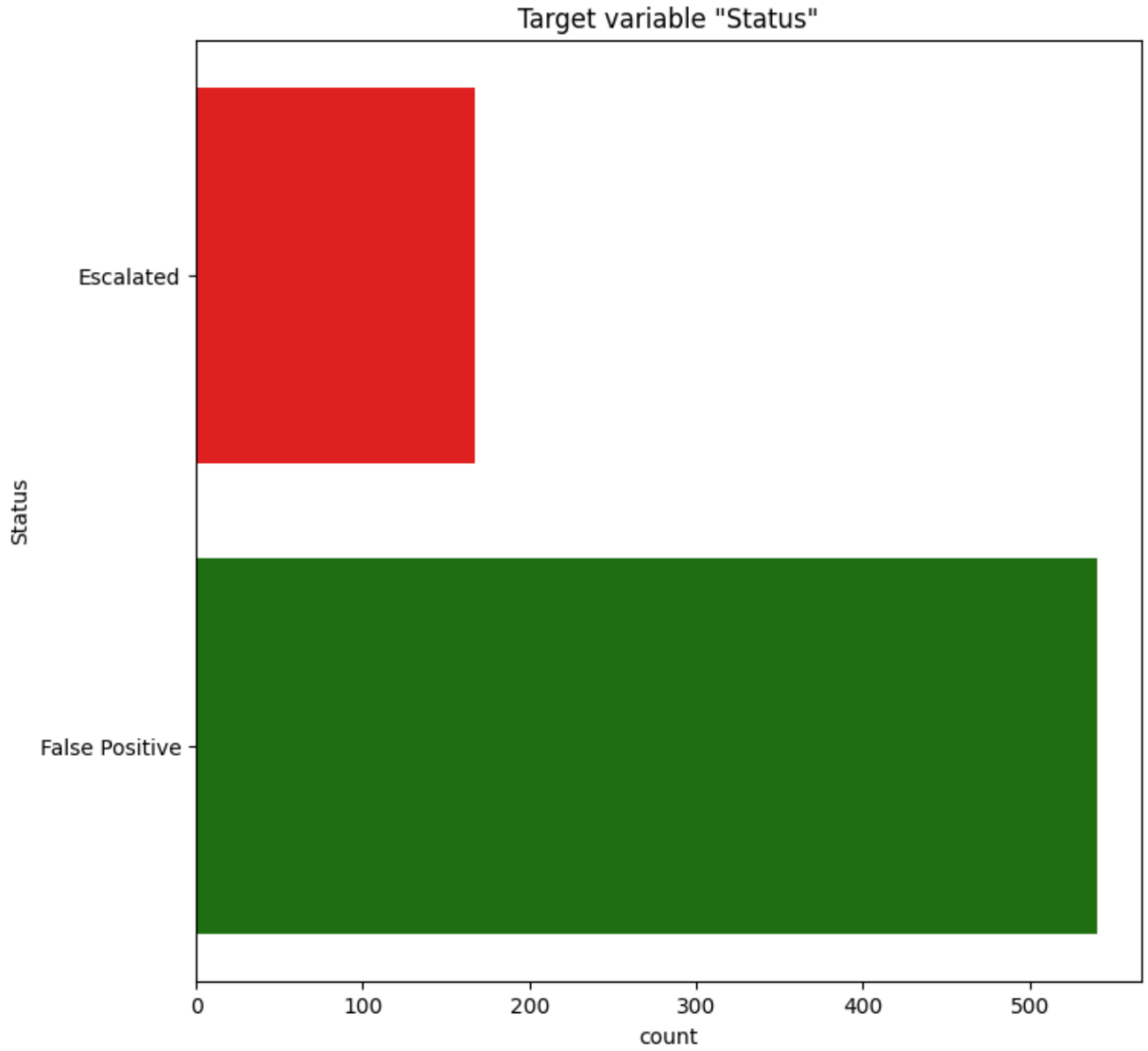
Target variable: "Status" (categorized as "escalated" or "false positive").

"ID" for record identification, "Severity" for vulnerability severity.

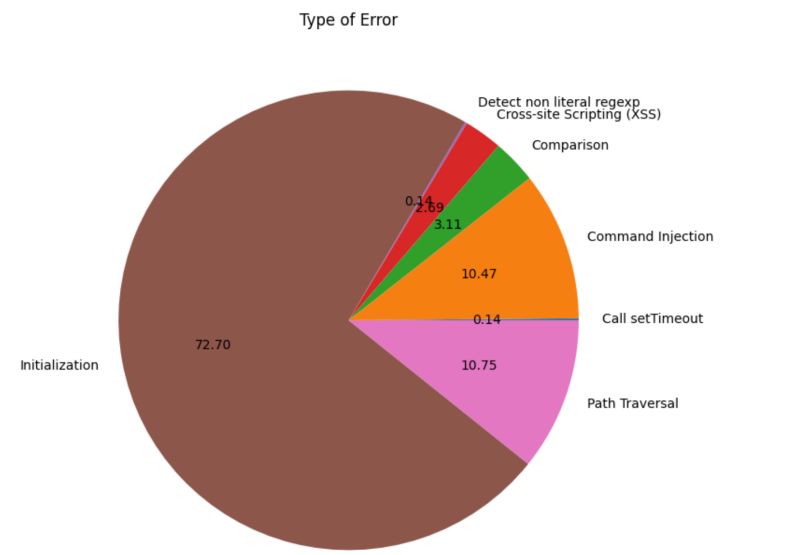
"CWE" assigns a numerical score based on vulnerability type.

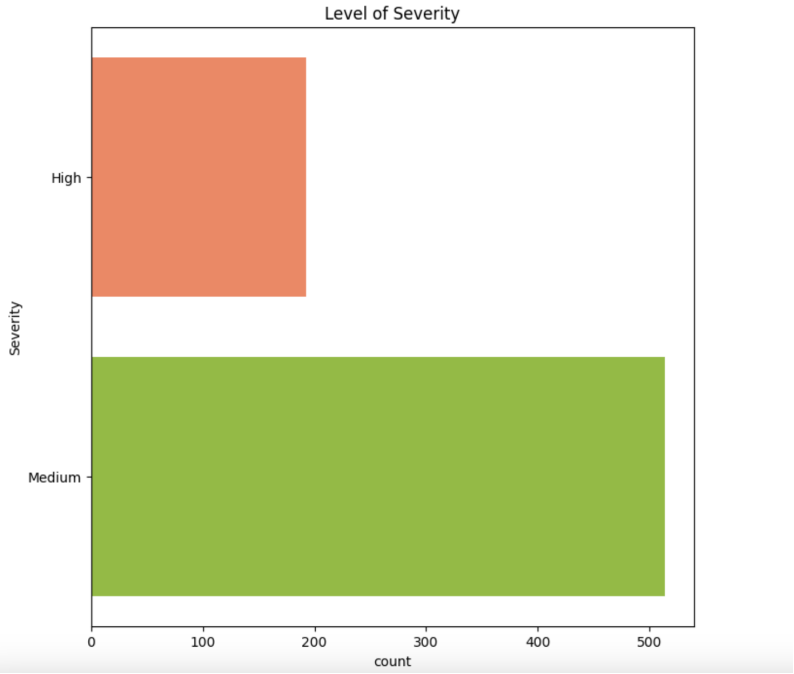
"Tool," "Location," and "Path" provide context and location in the code.

"Line" specifies the line number in the code



- The "Status" column is our target variable, classifying vulnerabilities as "escalated" or "false positive."
- In the image, we observe a distribution of our "Status" target variable.
- "False Positive" instances are more prevalent than "Escalated" ones.





- In Image 1, we present a pie chart depicting the types of errors in the code.
- Initialization errors stand out as the most common.
- Initialization errors often result from issues during software development and are a key focus area for security analysis.
- Addressing these vulnerabilities can significantly enhance software robustness and security
- Image 2 illustrates the distribution of vulnerability "Severity."
- "Medium" severity vulnerabilities outnumber "High" severity ones

MODEL IMPLEMEMENTATION

- One-hot encoding for categorical variables to convert unique variable values into equivalent 0s and 1s
- Train-test split of 10-90 was performed to train the model
- Used suitable algorithm for binary classification - Logistic regression and ensemble modelling technique
- Model performance was analyzed based on the evaluation metric parameters - Accuracy, precision, recall,

  and F-1 score

LOGISTIC REGRESSION

- The confusion matrix visually represents how well the model predictions align with true values
- The loss factor of 7.44% indicates the misclassification of vulnerabilities
- According to the classification report, Model accuracy – **92.13%**

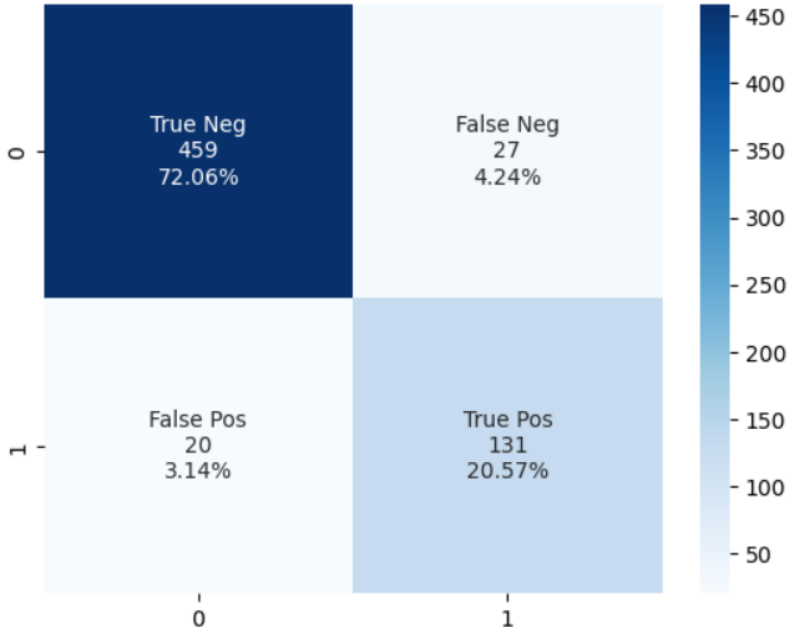


Image : Confusion Matrix for logistic regression

X-axis –Actual values,
Y-axis – Predicted values

DECISION TREE CLASSIFIER

- Constructed as tree-like structure, where each node represents a feature, a branch denotes the decision
- It follows the pattern of node splitting and recursive partitioning by breaking down the problem into series of questions
- Leaf node indicates the predicted outcome on the provided inputs
- According to classification report – Model accuracy **95.91%**
- A lower Gini index of 4 indicates the less likelihood of misidentification of the vulnerability

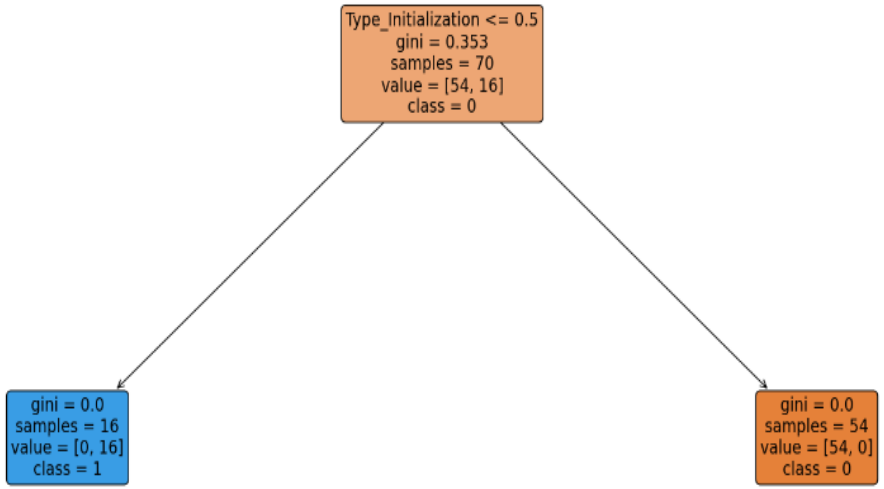


Image: Decision tree classifier

GRADIENT BOOST MODEL

- According to classification report – Model accuracy **95.91%**

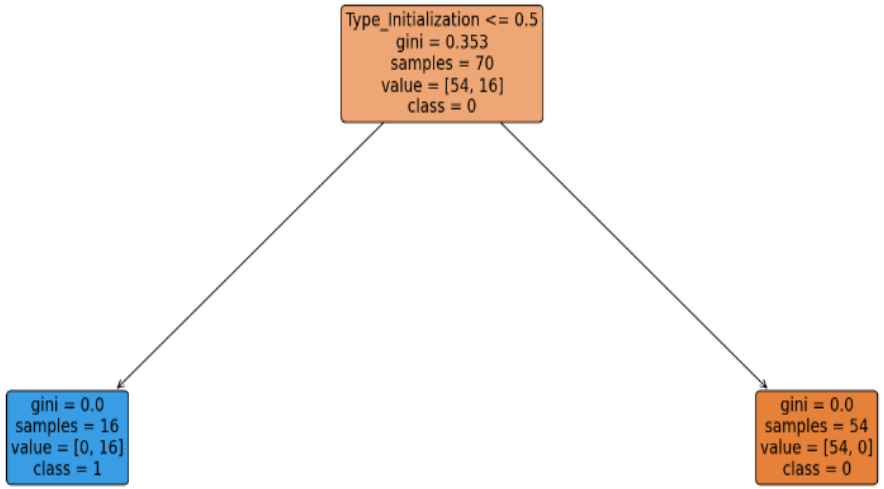


Image: Decision tree classifier

CLASSIFICATION REPORT

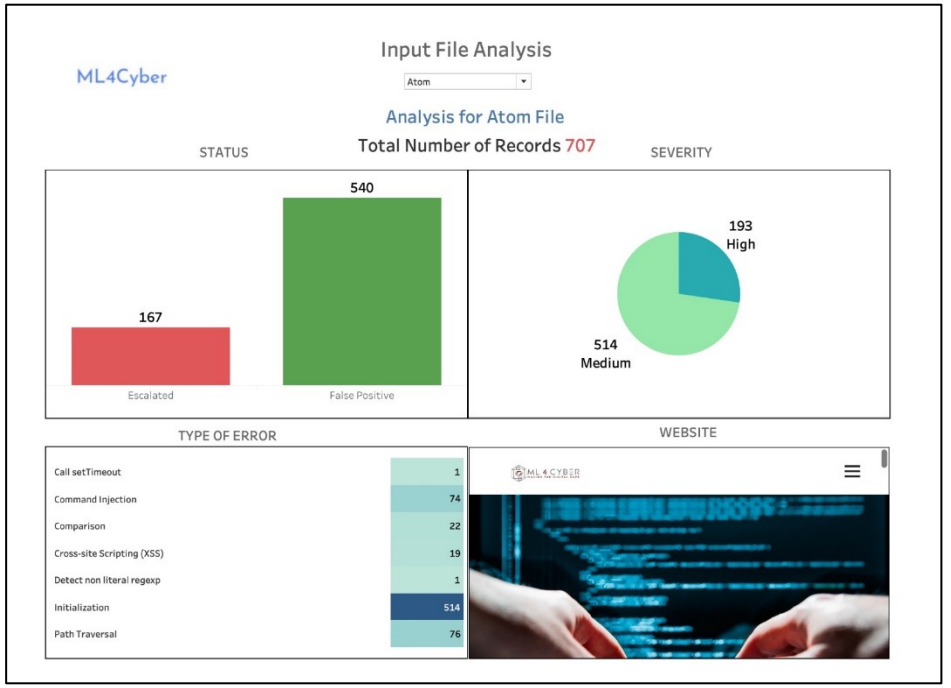
- Precision: from all the classes we have predicted as positive, how many are actually positive?
- Recall: from all the positive classes, how many we predicted correctly.
- F-measure: It is used to compare scenarios where the occurrence of low precision and high recall and vice versa is observed

Model	Logistic		Decision Tree	
Accuracy	92.62%		95.91%	
	0	1	0	1
Precision	0.96	0.83	1.00	0.85
Recall	0.94	0.87	0.95	1
F-1 Score	0.95	0.85	0.97	0.92

ML4Cyber Dashboard

Our project's journey has led to the creation of the ML4Cyber Dashboard, a revolutionary tool designed to transform the landscape of software security analytics. It embodies our commitment to integrating machine learning with cybersecurity, providing a robust solution for predicting and prioritizing software vulnerabilities. The features that highlight this dashboard are –

- Real-time Analytics: Instant visualization of security status, enabling proactive threat management.
- User-Centric Design: A dashboard interface that is both intuitive and informative, facilitating a seamless user experience for security professionals.
- Advanced Threat Categorization: Utilizes our predictive models to classify threats by severity, reducing response times and focusing efforts where they are needed most.



Future Steps and implementation

- Leverage the capabilities of the ML4Cyber Dashboard to automate vulnerability analysis further, incorporating real-time data feeds and predictive analytics.
- Iterate on our current models to enhance accuracy, reducing false positives and negatives, by integrating the latest advancements in AI and ML research.
- Introduce adaptive learning mechanisms to continuously improve the model's performance as new data is processed and threats evolve.
- Develop a phased implementation roadmap for integrating the ML4Cyber Dashboard across different organizational units, ensuring a smooth adoption curve.
- Initiate a pilot program to gather user feedback on dashboard usability and model effectiveness in real-world scenarios

Future Steps and Recommendations

- Refine the ML4Cyber Dashboard's algorithm to dynamically adjust threat levels based on evolving patterns, ensuring the model remains robust against novel vulnerabilities."
- Incorporate an anomaly detection feature that learns from historical trends, aiding in the early identification of potential zero-day exploits.
- Enhance the dashboard's interactivity, allowing users to simulate potential security scenarios and visualize the impact of hypothetical threats on their systems.
- Expand the model's capabilities to include prescriptive analytics, offering actionable steps for threat mitigation and risk reduction."
- Develop advanced training modules within the dashboard to help users understand the predictive model's insights and effectively respond to alerts.
- Implement a modular update system for the dashboard, allowing seamless integration of new machine learning models and data sources without disrupting the user experience

CONCLUSION

- Our findings clearly demonstrate that Decision Tree models surpass Logistic Regression in predicting vulnerabilities with higher precision. This validates our approach, leveraging the inherent strengths of tree-based models for complex pattern recognition in cybersecurity data.
- Through advanced feature engineering, we've significantly minimized loss factors, unlocking deeper insights and strengthening the predictive framework of our ML tools.
- We advocate for integrating gradient boosting techniques, which have been shown to bolster the predictive accuracy and robustness of our vulnerability prediction models, as evidenced by the ML4Cyber Dashboard's performance.
- Investigate the application of advanced boosting algorithms such as CatBoost and LightGBM to further enhance model precision.
- Evaluate the impact of different boosting approaches on our dataset to identify the most effective method for our specific use case.

FUTURE RESEARCH & SCOPE

Our models have set a benchmark in vulnerability prediction, with Logistic Regression achieving a 92.62% accuracy rate and the Decision Tree model further elevating this to 95.91%. These results underscore the robustness of our ML approach in pinpointing software vulnerabilities.

Key takeaways from our analysis include:

- The Decision Tree model demonstrates superior performance across key metrics, including accuracy, precision, recall, and the F-1 score."
- "Feature analysis revealed that 'Type Initialization' is critical for model accuracy, guiding our focus for future model enhancements.

Future strategies and products:

- Augment the ML4Cyber Dashboard to further streamline the vulnerability analysis process, integrating real-time threat intelligence for quicker, more accurate assessments."
- "Refine our models to not only differentiate between true vulnerabilities and false positives but also to predict the potential impact and suggest mitigative actions."
- "Adopt cutting-edge AI and ML advancements to evolve our tool's predictive power, focusing on continuous learning capabilities to adapt to new threat landscapes.