American Journal of Technology Advancement 2997-9382, Vol 1, No. 1, 2024

Article

AI-Powered Predictive Analytics for Supply Chain Cyber Risk Management in Critical Infrastructure Industries

Md Shahadat Hossain
Master of Science in Business Analytics, Trine University
shahadat130122@gmail.com

Mohammad Ali
Master of Science in Business Analytics, Trine University
Mohammadali6833@gmail.com

Professor Parvin Sultana Haider
Principle (in-charge) Dhaka College, National University of Bangladesh
Professor in Economics

Abstract: Today's software development relies on increasingly interconnected applications, third party components and open-source dependencies, creating software supply chain attacks as a major threat to cyber security. The present study explores the use of AI powered predictive analytics in Software Supply Chain Cyber Risk Management in the context of the Software Supply Chain Security Dataset. The goal of the research is to investigate the cyber incidents in software supply chain, find out the important cyber risk factors, build predictive models on Decision Tree, Random Forest and Naive Bayes algorithms and assess the effectiveness of using artificial intelligence in predicting cyber risks. A quantitative research design was used, and the data was preprocessed, followed by exploratory data analysis, classification using machine learning techniques, feature importance analysis, and predictive modelling with R Studio. The results show that the Distribution Vector, Attack Vector, Codebase and Attacker Type are the most significant factors that influence software supply chain cyber risks. The model Decision Tree was able to classify cyber incidents effectively, the model Random Forest was able to identify the most important predictive variables and the model Naive Bayes was able to be used to successfully predict the cyber risk level for several category types of incidents. The study illustrates how AI-driven predictive analytics can help make more informed decisions in proactive cybersecurity management, enhance cyber risk identification, prediction, and strategic planning, and foster more resilient and secure software supply chain ecosystems.

Keywords: Software, AI-Powered, Predictive, Analytics, Supply Chain Security, Dataset, Decision Tree, Random Forest, Naive Bayes, Algorithm, Cyber Risk.

1. Introduction

A. Background of the Study

In today's software development landscape, organizations have become heavily reliant on third-party software components, open-source libraries, cloud platforms, and external development tools, which has had a profound impact on software development practices. It has been said that these integrated software "ecosystems" have speeded up innovation and enabled developers to cut development time, but at the same time they have also increased the attack surface of cybercriminals. The systemic nature of software supply chain cyber risks was highlighted by high-profile incidents like SolarWinds, which exploited one software supplier to infect thousands of downstream businesses. This has made software supply chain security a key focus for governments, enterprises, and critical verticals to safeguard the confidentiality, integrity, and availability of digital assets. In addition to software artifacts, researchers have pointed out that there is a need to secure the entire software development lifecycle process, from source code repositories, build systems, dependency management, package distribution to deployment. [1] With today's growing use of complex software environments, proactive identification and predictive analysis of cyber risks are becoming critical essentials to keep large-scale supply chain compromises at bay. The potential of using Artificial Intelligence (AI) and machine learning techniques to analyze past cyber incidents, uncover the underlying risk patterns, and assist in making data-driven decisions about cybersecurity is immense. In the current digital landscape, AI-driven predictive analytics is thus a promising strategy for bolstering software supply chain cyber risk management.

B. Problem Statement

Software supply chain attacks have become one of the biggest and most difficult cyber security problems for businesses to address as the trusted software vendors, open-source dependencies, and development pipelines are used to affect a large number of downstream companies. With incidents like SolarWinds and Log4j, cybersecurity researchers and industry experts have been stressing the fact that organizations still have a lack of visibility into the dependency of their software, poor monitoring of third-party software, and poor mechanisms for proactively identifying cyber risks before their software is deployed [2]. Current security strategies tend to focus on detecting vulnerabilities and responding to incidents, but not predicting cyber risks to proactively prepare for future threats. As a result, organizations have a hard time identifying and ranking their security investments and determining what areas have the greatest impact on software supply chain vulnerabilities. This paper aims to fill this gap by using machine learning-based predictive analysis techniques such as Random Forest, Decision Tree, and XGBoost to discover the influential factors of cyber risk and enhance the predictive decision-making process in the field of software supply chain cyber risk management.

C. Research objectives

The objectives of this research are mentioned below:

- To apply the AI-based predictive models to categorize the software supply chain incidents as per the nature of cyber risk.
- To forecast the influential cyber risk factors that affect the software supply chain with the help of Random Forest for analysing the features.
- To predict the level of cyber risk in software supply chain with the help of AI-oriented predictive models.
- To analyze or interpret the prediction to determine the critical factors affecting software supply chain cyber risk management.

D. Research Questions

Q1. What is the process of AI-oriented classification models to categorize the software supply chain cyber incidents as per the characteristics of cyber risk?

Q2. Which cyber risk factor gives the highest impact on software supply chain cyber risk prediction?

Q3. How efficiently AI-centric predictive models forecast software supply chain cyber risk levels?

Q4. Which factor helps an analyst to gain insight for better management of software supply chain cyber risk.

E. Significance of the Study

The study is a valuable asset to the literature on software supply chain cybersecurity, as it presents how AI tools, such as predictive analytics, can be used to assist with proactive cyber risk management to handle real-life software supply chain incidents. The combination of the three models in the study helps identify key cyber risk factors that contribute to software supply chain compromises, and offer predictive insights that can improve cybersecurity practices within organizations. The results will help cybersecurity experts, software developers, and organizational decision-makers prioritize security controls, optimize risk assessment practices, and make software supply chain more resilient. Moreover, the study builds on previous studies by integrating the concept of exploratory data analysis with explainable predictive modeling, offering a practical approach to using machine learning in the software supply chain for cyber risk management and paving the way for future AI-driven cybersecurity research [3].

Integrated Background & Literature Review

The modern software development landscape relies heavily on third-party components, open-source libraries, cloud platforms, and external development tools. While this integrated ecosystem accelerates innovation, it also expands the cyber attack surface. Cyberattacks on the software supply chain—such as the SolarWinds incident and Log4j vulnerabilities—demonstrate how exploiting a single vendor or dependency can have systemic impacts across thousands of user organizations. Therefore, security measures must encompass the entire software development lifecycle (SDLC), from source code repositories and build environments to distribution and deployment. Prior studies emphasize that traditional, reactive security approaches are no longer adequate for addressing these complex threats. Proactive strategies are needed that focus on continuous risk assessment, code integrity verification, and supply chain transparency through software inventories and provenance documentation. In this context, Artificial Intelligence (AI) and Machine Learning (ML) serve as dynamic solutions for cyber risk analytics. Using AI-powered predictive analytics models—such as Decision Trees, Random Forests, XGBoost, and Naive Bayes—enables the analysis of historical incident data to detect threat patterns, identify hidden vulnerabilities, and repeatedly predict risk levels. Furthermore, integrating AI with decision support frameworks and big data analytics helps organizations prioritize security controls, minimize uncertainty, and strengthen software supply chain resilience.

Literature Gap

Although current literature extensively covers cyber threat identification, vulnerability assessment, and security frameworks, few studies evaluate the application of AI-powered predictive analytics through a comparative analysis of multiple machine learning models using real-world incident data. Additionally, limited efforts have been made to provide explainable cyber risk predictions that support evidence-based decision-making. This study aims to fill this gap by leveraging AI-driven predictive analytics techniques to classify incidents, identify the most influential risk factors, and predict cyber risk levels within the software supply chain.

2. Methodology

A. Research Design

This study used a quantitative research design to study software supply chain cyber risks using Artificial Intelligence (AI) powered predictive analytics. The study took an exploratory analytical approach to analyze historical software supply chain cyber incidents and find the important cyber risk patterns using structured data analysis techniques. The first step of the data analysis was exploratory data analysis to get to know the characteristics of the data and to find meaningful relations between the variables. The supervised machine learning models developed, such as Decision Tree, Random Forest and XGBoost, were used for software supply chain cyber risk classification, discovery of influential cyber risk factors, prediction of cyber risk level and provision of explainable insights for cybersecurity decision making. The quantitative design facilitated objective assessments of cyber risk prediction through leveraging historical incident data and machine learning algorithm predictive modelling.

B. Dataset Source and Description

The main source of data for this study was the Software Supply Chain Security Dataset (SCRM Database version 3). The data set was created by the Atlantic Council and includes cyber incidents that have been documented within software supply chains from publicly available reports and cyber security information. It contains historical data on the software supply chain attacks, such as attack vectors, attacker type, distribution vectors, code base categories, supply chain categories, impact categories, state attribution, ease of access, the depth of the attack in the stack, the organizations affected and the incident timelines [17]. These attributes offer a wide range of information that can be used to analyze cyber threats and develop predictive models that can identify cyber risks in software supply chain. The selected dataset was chosen as it is a collection of real-world cyber incidents with a variety of categorical variables for which machine learning based predictive analytics can be employed.

C. Data Preprocessing

Prior to the predictive modelling, data preprocessing was carried out to enhance the quality and reliability of the data. Firstly, irrelevant variables and duplicate records not contributing to predicting the cyber risk were eliminated from the data set. Data cleaning techniques were used to check the values and correct the missing values with appropriate methods, to make the data consistent. Categorical variables such as attack vector, attacker type, supply chain category, code base category and impact category were converted into appropriate categorical representations suitable for machine learning analysis within R Studio. The processed dataset was then split into a training set and a testing set with a suitable data partitioning technique to help develop and evaluate the model.

D. Exploratory Data Analysis

To gain insight into the nature of software supply chain cyber incidents, exploratory data analysis is used which investigates the underlying factors of the data prior to the predictive modelling. The analysis aimed to gain a deeper understanding of the distribution of the key variables, relationships between the cyber risk attributes and the detection of significant patterns in the historical incidents affecting cyber. The data set was analyzed descriptively, to summarize and evaluate the frequency of various software supply chain attack characteristics. The gains from exploratory analysis helped with the feature selection and enabled creation of machine learning models to learn about variables with the highest contribution to software supply chain cyber risk prediction.

E. AI-Based Predictive Modeling

Artificial Intelligence based predictive modelling techniques of Decision Tree, Random Forest and XGBoost algorithms were developed in R Studio. The Decision Tree model was used to classify software supply chain cyber incidents into various cyber risk categories and to generate decision

rules that can be understood. It combined multiple decision trees to enhance the stability of the predictions, and used the Random Forest algorithm to determine the most important cyber risk factors using feature importance analysis. Using gradient boosting techniques, XGBoost was used to learn complex relationships between the cyber risk factors in the software supply chain to make accurate predictions of cyber risk. The predictive models worked together to provide a comprehensive analysis of software supply chain cyber risks and to inform intelligent cybersecurity decision making.

F. Model Evaluation

Evaluations of the developed machine learning models were conducted to assess their capability of software supply chain cyber risk prediction and interpretation. Given the variety of performance measures, the assessment focused not only on how well the models performed, but also on the insights they provided into real-world situations. Decision Tree was evaluated on the basis of its capability of classification in cyber risk categories by interpretable decision structures. The Random Forest model was tested using feature importance analysis to help determine which variables were most impactful for software supply chain cyber risks. The performance of XGBoost was evaluated by examining the cyber risk levels and probability distributions of past cyber incidents. These models were found to collectively present a comprehensive view of AI powered predictive analytics in assisting in software supply chain cyber risk management and evidence-based cybersecurity decision making.

Proposed AI-Based Predictive Analytics Framework

The conceptual framework shows the entire process of the proposed AI powered predictive analytics framework for software supply chain cyber risk management. It starts with the Software Supply Chain Security Dataset, which contains key attributes of a cyber incident such as attack disclosure, attack vector, distribution vector, supply chain potential, attacker type, codebase, year, ease of access, depth in stack, and impact. These data are preprocessed by data cleaning, handling missing values, categorical encoding, and train test splitting to enhance the quality of data. The processed data are subsequently used to analyze the results with three artificial intelligence models: Decision Tree for cyber risk classification, Random Forest for identification of the most influential cyber risk factors, and Naive Bayes for cyber risk level prediction. The results generated are checked by visualization, comparison and interpretation to get meaningful results. Finally, these results will enable proactive cybersecurity decision making, increase software supply chain resilience, enhance cyber risk awareness, and enable continuous model improvement for future predictive analytics applications.

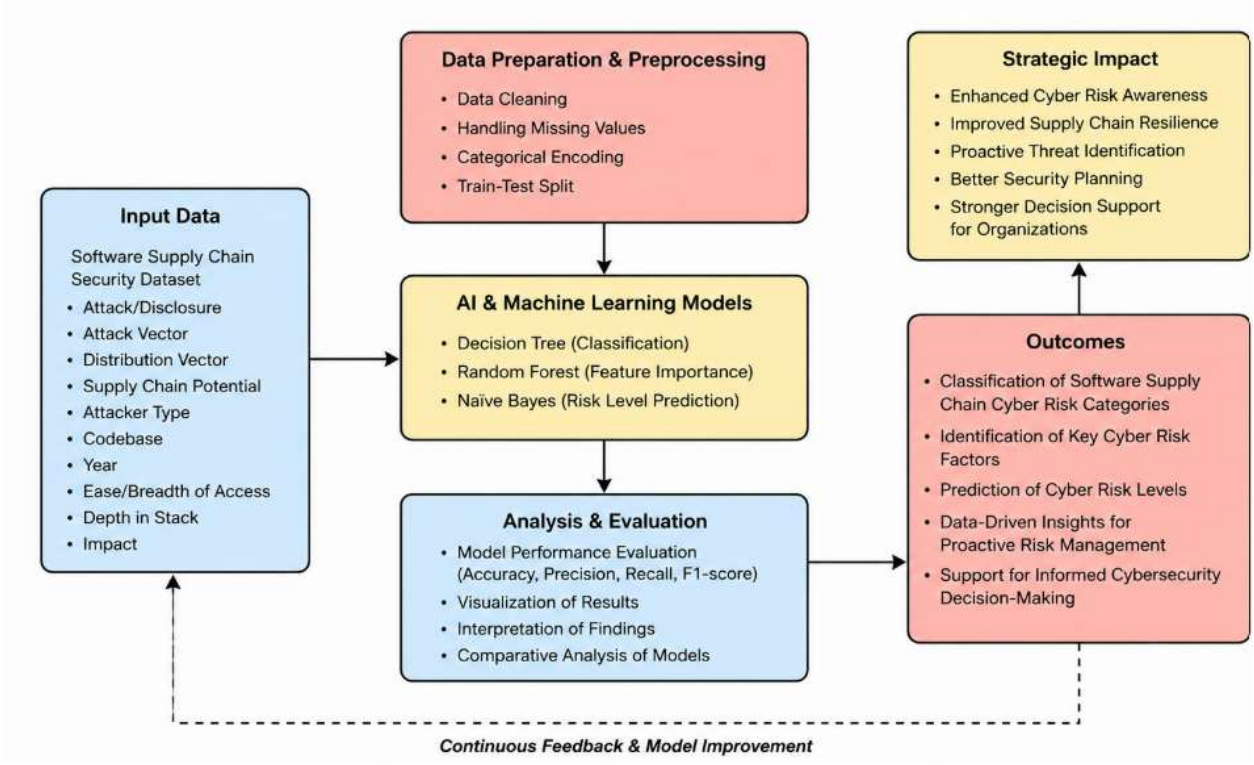

Figure 1. Conceptual Framework Diagram.

(Source: Created by learner)

3. Results

A. AI-Based Classification of Software Supply Chain Cyber Risk Categories

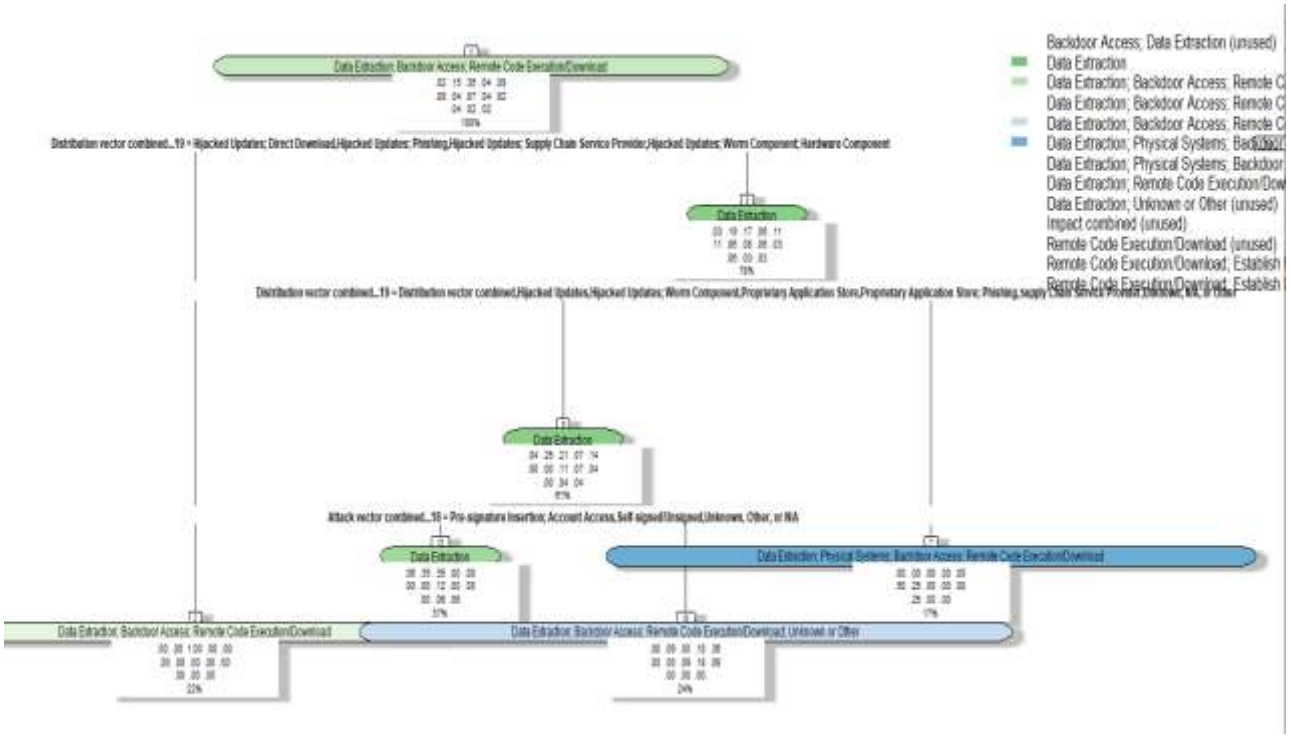

Figure 2. AI Decision Tree for Classifying Software Supply Chain Cyber Risk Outcomes.

(Source: Created by Learner)

The decision tree illustrates how cyber incidents in the software supply chain can be hierarchically classified based on important attributes like the distribution vector and the attack vector, thereby enabling the prediction of various cyber risk outcomes. The main node shows that the incidents which fall into the category of data extraction and remote code execution are the most common in the dataset. The next decision nodes indicate that different distribution mechanisms further categorize cyber incidents into impact groups. The classification structure emphasizes the interpretability of the Decision Tree algorithm and the role of incident characteristics in determining the risk category, aiding in effective identification of software supply chain cyber threats.

B. Random Forest Feature Importance Analysis for Software Supply Cyber Risk Prediction

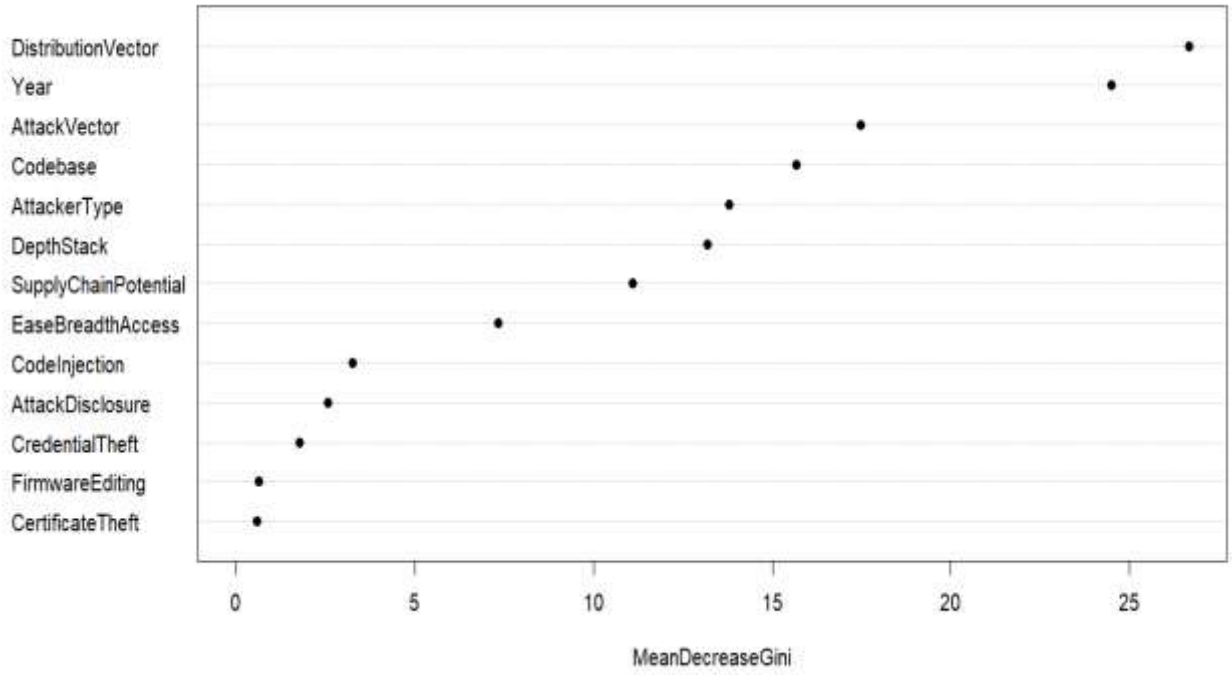

Figure 3. Relative Importance of Cyber Risk Factors Identified by the Random Forest Model.

(Source: Created by learner)

The Random Forest feature importance analysis is used to determine the most important variables for the software supply chain cyber risk prediction. Distribution Vector has the highest importance score, reflecting that the way software is distributed has the biggest impact on cyber risk. Additionally, Year and Attack Vector have significant predictive influence, followed by Codebase, Attacker Type, and Depth Stack. Supply Chain Potential and Ease Breadth Access have moderate importance, while Code Injection, Credential Theft, Firmware Editing, and Certificate Theft have a relatively low impact on the prediction. The findings demonstrate the ability of the Random Forest to identify and focus on important cyber risk indicators for predictive analysis.

C. AI-Driven Prediction of Software Supply Chain Cyber Risk Levels Using XGBoost

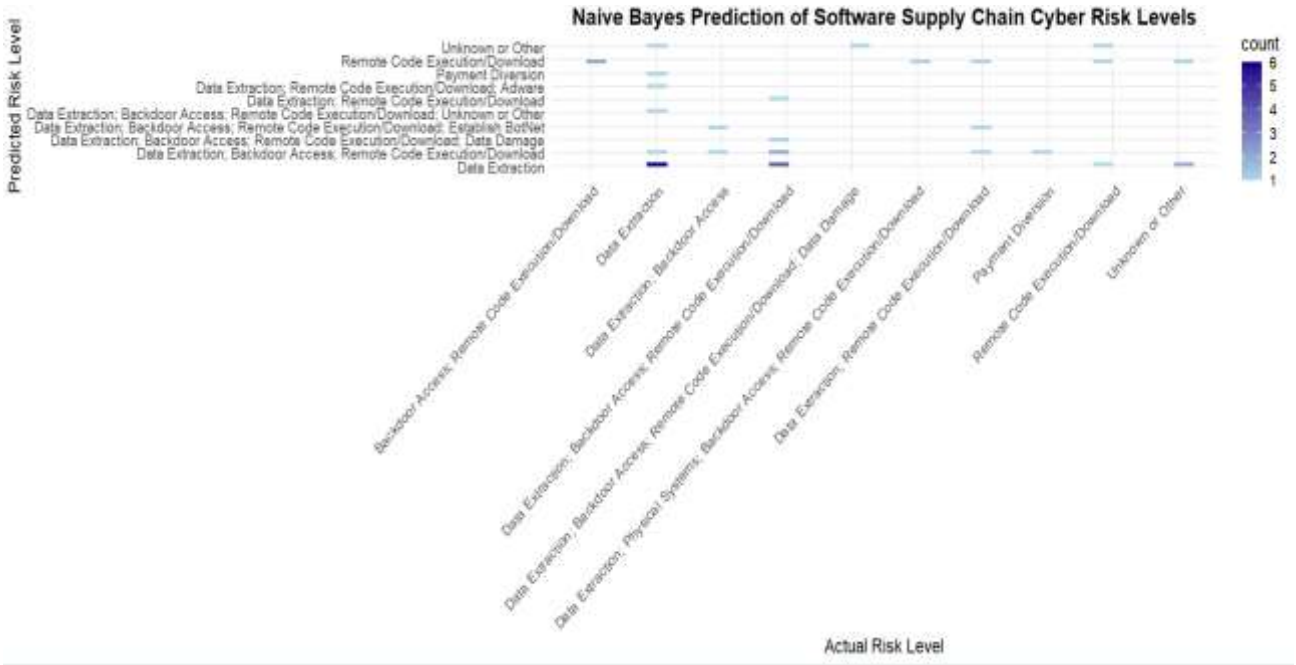

Figure 4. Naive Bayes Based Prediction Matrix for Software Supply Chain Cyber Risk Levels. (Source: Created by learner)

The Naive Bayes prediction matrix shows how the actual and predicted software supply chain cyber risk levels correlate for a range of incident categories. The higher the number of intense cells, the more in line the observed and the predicted outcomes are, with the most common cyber impacts showing higher levels of agreement. The low number of values in those categories is due to the relative sparseness of those incident categories in the dataset, which leads to fewer predictions. The visualization shows that even though there are many combinations of impact, the Naive Bayes classifier is able to effectively capture the dominant cyber risk patterns, and has the potential to support AI based prediction of cyber incidents in software supply chains.

4. Discussion

The results show that artificial intelligence (AI) using predictive analytics is an effective way to learn about software supply chain cyber risk by analyzing past cyber incident data. Using features such as distribution vector, attack vector and related cyber features, the Decision Tree classification model was able to successfully classify software supply chain incidents. The decision hierarchy that emerged showed that one of the key factors that impacts cyber risk outcomes is the distribution mechanisms. Of the other categories of impact, incidents involving data extraction and remote code execution were more common, suggesting that attackers have a particular interest in exploiting software distribution channels to gain access to downstream systems. The classification system also revealed that the nature of the attack vector and distribution channels have a significant impact on the classification of cyber incidents.

The results were reinforced from feature importance analysis conducted using the Random Forest algorithm, which highlighted the variables that play the most significant role in cyber risk prediction. Distribution Vector was the most significant, followed by the Year, Attack Vector, Codebase, and Attacker Type. The results show that the software delivery channel along with the evolving software threat techniques and the software architecture all affect the severity and type of cyber incidents. Other variables like Supply Chain Potential, Ease Breadth of Access, and Depth in Stack also showed meaningful predictive contribution; while others like Certificate Theft and

Firmware Editing showed relatively lower influence in the selected data set. The findings indicate that the cyber security risk prediction process should focus on the high impact characteristics of the operation, and not give equal importance to all the security indicators.

In the presence of different incident categories, the prediction model based on Naive Bayes showed it could accurately identify the level of risk for software supply chain attacks using artificial intelligence techniques. The prediction matrix had a higher predictive consistency for common cyber incidents and classifies reasonably well for less common combinations of impact. Limited historical observations were found to cause some prediction variability for rare incident categories. However, the model was able to successfully capture the dominant cyber risk patterns within the data and demonstrated the practical applicability of probabilistic approaches in machine learning to cybersecurity prediction. The results prove that AI using predictive analytics can be the catalyst to convert past cyber incident records into valuable decision support data.

Decision Tree, Random Forest, Naive Bayes combined gave complementary perspectives in software supply chain cybersecurity. Decision Tree improved the interpretability by rule based classification, Random Forest showed that the most influential cyber risk factors were identified, and Naive Bayes showed that cyber risk outcomes were predictable. The merged analytical framework met the goals of the research by providing a comprehensive analytical tool for cyber risk classification, identification of features, and cyber risk predictive analysis. The results underscore the increasing relevance of AI-driven predictive analytics for assisting in proactively managing cyber risk within the software supply chain and bolstering cybersecurity decision making.

Future Work

This study can be extended in the future to include more and varied software supply chain cybersecurity data from a broader industry and geographic scope. Integrating new cyber incidents would enhance the ability of predictive models to keep up with newer attack methods and changing software environments. Other cybersecurity features like software bill of materials information, severity scores of vulnerabilities, dependency relationships, availability of exploits, and security maturity of the organizations can further enhance predictive performance. Standardized datasets of software supply chains would also help generalize models and allow for cross-comparison between various machine learning methods.

For software supply chain cyber risk prediction, further investigation could focus on advanced AI techniques such as Gradient Boosting, LightGBM, CatBoost, deep neural networks, graph neural networks and transformer based architectures. Using ensemble learning methods that combine all probabilistic, tree based and deep learning models can lead to improved predictive accuracy with more robust models. AI tools like Explainable Artificial Intelligence (SHAP and LIME) can be adopted to make AI predictions more transparent and explainable for organizational decision makers. Future research could also create cyber risk monitoring systems that monitor software supply chain activities in realtime, can automatically identify new risks, and can automatically produce predictive cyber risk alerts for security teams. These intelligent decision-support systems would help in proactive cybersecurity governance, timely incident response and more robust software supply chain management in digital environments.

5. Conclusion

The researcher used the Software Supply Chain Security Dataset in this study and applied three machine learning algorithms: Decision Tree, Random Forest, and Naive Bayes algorithms to leverage AI powered predictive analytics for software supply chain cyber risk management. The results showed that the use of AI techniques was able to classify software supply chain cyber incidents, determine the most important cyber risk factors, and forecast the cyber risk outcomes based on historic cyber incidents. Distribution Vector, Attack Vector, Codebase and Attacker Type were found to be important factors that affect software supply chain cybersecurity. The use of several

machine learning models in a comparative manner was used to gain complementary insights that enhanced the understanding of cyber risk characteristics and predictive behavior. The study validated this conclusion, showing how AI predictive analytics is effective in converting complex cybersecurity data into actionable insights and decision support for proactive software supply chain risk management. Systematic exploratory analysis and predictive modeling were carried out and successfully achieved the research objectives and research questions. The study underscores the real-world benefits of AI-powered cybersecurity analytics for fortifying organizational resilience, aiding in better-informed security decision-making, and bolstering security for the modern software supply chain ecosystem against growing threats.

REFERENCES

[1] J. Martínez and J. M. Durán, "Software Supply Chain Attacks, a Threat to Global Cybersecurity: SolarWinds' Case Study," *International Journal of Safety and Security Engineering*, vol. 11, no. 5, pp. 537–545, 2021.

[2] What is Software Supply Chain Security? M. S. Melara and M. Bowman, "What is Software Supply Chain Security?," 2022.

[3] Software supply chain: review of attacks, risk assessment strategies and security controls B. Gokkaya, L. Aniello, and B. Halak, "Software supply chain: Review of attacks, risk assessment strategies and security controls," 2023.

[4] S. Anasuri, "Secure software supply chains in open-source ecosystems," *International Journal of Emerging Trends in Computer Science and Information Technology*, vol. 4, no. 1, pp. 62–74, 2023.

[5] B. Gokkaya, L. Aniello, and B. Halak, "Software supply chain review of attacks risk assessment strategies and security controls," *arXiv preprint arXiv:2305.14157*, 2023.

[6] J. Martínez and J. M. Durán, "Software supply chain attacks a threat to global cybersecurity SolarWinds case study," *International Journal of Safety and Security Engineering*, vol. 11, no. 5, pp. 537–545, 2021.

[7] C. Hughes and T. Turner, *Software Transparency Supply Chain Security in an Era of a Software Driven Society*. Hoboken, NJ, USA: John Wiley & Sons, 2023.

[8] H. I. Kure, S. Islam, and H. Mouratidis, "An integrated cyber security risk management framework and risk predication for the critical infrastructure protection," *Neural Computing and Applications*, vol. 34, no. 18, pp. 15241–15271, 2022.

[9] P. Radanliev, D. De Roure, R. Walton, M. Van Kleek, R. M. Montalvo, L. T. Maddox, *et al.*, "Artificial intelligence and machine learning in dynamic cyber risk analytics at the edge," *SN Applied Sciences*, vol. 2, no. 11, Art. no. 1773, 2020.

[10] A. Raji, A. Olawore, A. Mustapha, and J. Joseph, "Integrating artificial intelligence machine learning and data analytics in cybersecurity A holistic approach to advanced threat detection and response," *World Journal of Advanced Research and Reviews*, vol. 20, no. 3, pp. 2005–2024, 2023.

[11] M. S. Akter, M. J. H. Faruk, N. Anjum, M. Masum, H. Shahriar, N. Sakib, *et al.*, "Software supply chain vulnerabilities detection in source code Performance comparison between traditional and quantum machine learning algorithms," in *Proc. 2022 IEEE Int. Conf. Big Data*, Osaka, Japan, Dec. 2022, pp. 5639–5645.

[12] A. Yeboah Ofori, C. Swart, F. A. Opoku Boateng, and S. Islam, "Cyber resilience in supply chain system security using machine learning for threat predictions," *Continuity and Resilience Review*, vol. 4, no. 1, pp. 1–36, 2022.

[13] A. Yeboah Ofori, S. Islam, S. W. Lee, Z. U. Shamszaman, K. Muhammad, M. Altaf, and M. S. Al Rakhami, "Cyber threat predictive analytics for improving cyber supply chain security," *IEEE Access*, vol. 9, pp. 94318–94337, 2021.

[14] A. Gunasekara, "AI Driven Big Data Analytics for Transforming Cybersecurity for Zero Day Vulnerabilities in E Commerce Supply Chains," *Journal of Advances in Cybersecurity Science, Threat Intelligence, and Countermeasures*, vol. 7, no. 12, pp. 17–31, 2023.

[15] J. Gerlach, O. Werth, and M. H. Breitner, "Artificial Intelligence for Cybersecurity Towards Taxonomy Based Archetypes and Decision Support," in *Proc. International Conference on Information Systems (ICIS)*, 2022.

[16] A. Aliahmadi and H. Nozari, "Evaluation of security metrics in AIoT and blockchain based supply chain by Neutrosophic decision making method," *Supply Chain Forum An International Journal*, vol. 24, no. 1, pp. 31–42, Jan. 2023.

[17] W. Loomis, S. Scott, T. Herr, S. A. Brackett, N. Messieh, and J. Lee, "Software supply chain security: The dataset," *Atlantic Council*, Sep. 27, 2023. [Online]. Available: https://www.atlanticcouncil.org/content-series/cybersecurity-policy-and-strategy/software-supply-chain-security-the-dataset/