JOURNAL OF STATISTICAL MODELING & ANALYTICS (JOSMA)
(ISSN: 2180-3102)
UNIVERSITI MALAYA

Developing a New Feature for Vulnerability Risk Scoring Model for Enhanced Cybersecurity

Lim Joey[1] and Noryanti Muhammad[1]

[1]*Centre for Mathematical Sciences, Universiti Malaysia Pahang Al-Sultan Abdullah, Lebuh Persiaran Tun Khalil Yaakob, 26300 Gambang, Pahang, Malaysia.*

Corresponding author: noryanti@umpsa.edu.my

Received 25 Nov 2024
Accepted 10 Dec 2024
Published

RESEARCH ARTICLE

Abstract

As organisations increasingly rely on technology, the risk of cyber threats to data integrity and security grows significantly. Traditional vulnerability risk-scoring models may not adequately address the rapidly evolving nature of cyber threats, necessitating the development of more adaptable and context-specific models. This research aims to achieve two primary objectives: developing a flexible risk-scoring model that can be customised for different industries, companies, or situations, and developing a new feature to accurately reflect the risk score derived from the current dataset. The study employs correlation analysis and machine learning-based regression modelling, utilising appropriate evaluation metrics to assess model performance. Results indicate that the K-Nearest Neighbors regression model performs particularly well using Regression evaluation metrics. The newly developed risk score feature demonstrated its effectiveness in aligning with cybersecurity priorities. It successfully provided a systematic and interpretable metric for evaluating vulnerability risks, with higher scores corresponding to critical vulnerabilities requiring immediate attention. The research provides a scalable and adaptable framework for developing tailored risk-scoring models, improving the effectiveness of cybersecurity strategies in diverse contexts and datasets.

Keywords: Correlation-Based Weighting, Regression Modelling, Risk-scoring model, Dynamic Threat Landscape

1. Introduction

Malaysia has rapidly embraced the digital age, aligning with the global trend. Virtually every individual or entity is now linked to the online sphere. Figure 1 depicts the trajectory of internet user numbers in Malaysia from 2013, encompassing forecasts up to 2028. According to Simon Kemp (2023), approximately 33.03 million internet users were recorded in Malaysia from the beginning of 2023. As of January 2023, with Malaysia's population standing at 34.13 million, a staggering 96.78% of Malaysians were actively engaged as internet users. By using Figure 1 as a reference, the number of internet users in Malaysia witnessed growth between 2013 and 2023, with a projected ongoing increase to reach 35.33 million by 2028. Malaysia's digital user base is expected to grow rapidly, with a

significant increase in the number of people engaging with digital platforms shortly (Simon Kemp, 2023).

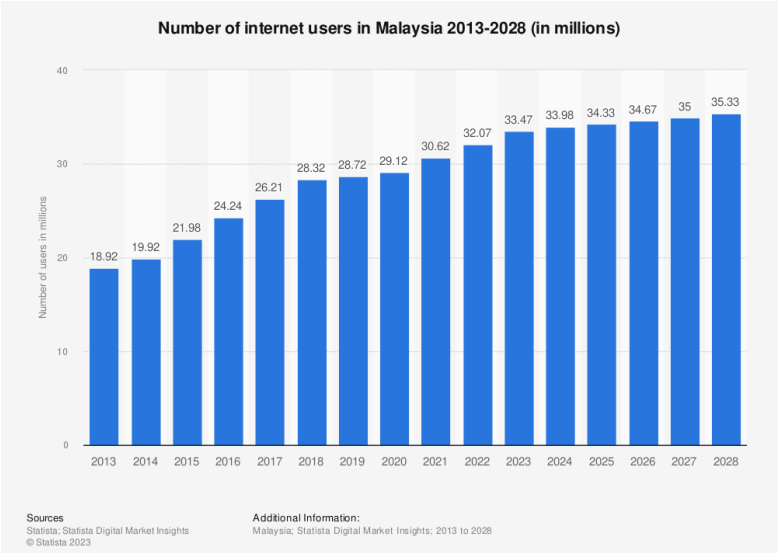

Figure 1. Number of Internet users in Malaysia 2013-2028 (in millions)
Source: Simon Kemp (2023)

In an era where businesses extensively utilise interconnected technologies like mobile devices, social media, big data analytics, the Internet of Things (IoT), Artificial Intelligence (AI), and large-scale cloud computing, the reliance on the Internet for day-to-day operations has become ubiquitous. This dependence, however, escalates data protection challenges, given the pervasive nature of cyber threats. A mid-year threat landscape report for 2023 by CyberSecurity Malaysia, based on data collected over six months, underscores the severity of this issue, documenting a total of 3989 incidents primarily dominated by malware and phishing attacks (CyberSecurity Malaysia, 2023).

Cyber-attacks can significantly harm objectives, operations, reputations, national cyber assets, and individuals, manifesting through unauthorised access, destruction, or alteration of information. Although primarily financially driven, these attacks occasionally bear military or political motives, highlighting the importance of robust cybersecurity measures (Li, Y., & Liu, Q., 2021).

Against this backdrop, the practice of good cybersecurity hygiene becomes paramount in mitigating risks and safeguarding personal and sensitive data. A pivotal tool in this preventive strategy is the cyber risk scoring model, which quantifies the effectiveness of cybersecurity measures, thereby enhancing the cyber risk posture of individuals and organisations. Notably, the Common Vulnerability Scoring System (CVSS), first introduced between 2003 and 2004 by the National Infrastructure Advisory Council and later managed by the Forum of Incident Response and Security Teams (FIRST) since April 2005, remains a critical component in assessing and prioritising computer security issues. This model offers a free, open industry standard that generates a numerical score to reflect the severity of vulnerabilities, aiding in the efficient allocation of remedial resources (Mell, P., Scarfone, K., & Romanosky, S., 2006; Jonathan Risto, 2023).

Despite these advancements, traditional risk-scoring models often fail to accommodate real-world attack vectors and the dynamic nature of cyber threats. They frequently overlook critical temporal aspects of vulnerabilities, such as the time-to-exploit or patch availability, which are vital for timely threat assessment (Bozorgi et al., 2010; Farris et al., 2018). In light of an escalating threat landscape and the frequent evolution of cyberattack methodologies, this research seeks to develop, evaluate, and

optimise vulnerability risk-scoring models to more effectively address these challenges, particularly in the Malaysian context. The urgency of this effort is underscored by recent high-profile data leaks.

This paper aims to introduce a simplified yet versatile risk score feature derived from existing datasets, encompassing network traffic characteristics linked to specific cyber threats. This approach promises broad utility, allowing various stakeholders to generate tailored risk scores for their datasets and develop customised risk-scoring models, thereby fostering a more resilient cybersecurity environment.

2. Materials and Methods

2.1 Data Collection

The data collected is from Kaggle, related to the Network Traffic Analysis Dataset for Cybersecurity, which provides a comprehensive collection of network traffic data designed to simulate diverse communication scenarios among network entities, with a specific focus on varying protocols and potential security threats (Zunxhi Samniea, 2023). The dataset contains 750 rows and 11 columns. Table 1 shows the data description. The dataset can be accessed through the reference provided (Zunxhi Samniea, 2023).

Table 1. Data Description of the dataset

Column/Variable	Description	Data Type
Protocol	Communication protocol used for the packet (e.g., TCP or UDP).	Categorical
Flag	Flag associated with the packet (e.g., SYN, ACK, RST, FIN).	Categorical
Packet	Type of packet exchanged (e.g., HTTP, DNS, SSH, FTP, NTP).	Categorical
Sender ID	Unique identifier for the sender entity.	String
Receiver ID	Unique identifier for the receiver entity.	String
Source IP Address	IP address of the source entity.	String
Destination IP Address	IP address of the destination entity.	String
Source Port	Port number on the source entity.	String
Destination Port	Port number on the destination entity.	String
Packet Size	Size of the packet in bytes.	Numeric
Target Variable	Potential security threat associated with the packet (e.g., Phishing, DoS, Man-in-the-Middle).	Categorical

By examining and studying the existing vulnerability risk scoring models, specifically CVSS version 2 and version 3, the essential characteristics of the independent variables that should be included in a dataset are identified to ensure a comprehensive and accurate risk assessment. These models highlight the need for integrating a variety of data types to effectively evaluate and mitigate security threats. Consequently, this research integrates variables that align well with the CVSS framework. The Kaggle dataset was selected for this study due to its open-source nature, which is particularly beneficial

given that security-related data is often not readily available through other open-source channels. The availability of such a dataset on Kaggle facilitates access to relevant information that supports the research objectives and enhances the vulnerability risk scoring model.

Table 2. Mapping of CVSS Model Components to Variables and Their Explanations

CVSS Model Component	Variable	Explanation
Attack Vector (AV)	Protocol	The protocol used can affect the attack vector, indicating how vulnerability can be exploited (e.g., network-based or local-based attacks).
Impact Metrics (IM)	Flag	Flags in network traffic can indicate the state of a connection, which may be relevant for determining the impact on confidentiality, integrity, and availability.
Exploitability Metrics (EM)	Packet	The number of packets or the nature of the packets can influence the exploitability of the vulnerability.
Scope (S)	Sender ID	Identifies the origin of the attack, helping to determine whether the scope of the impact is limited or widespread.
Scope (S)	Receiver ID	Identifies the target of the attack, helping to assess the potential impact on systems within the scope of the vulnerability.
Attack Vector (AV)	Source IP Address	The source IP address is part of determining the attack vector, as it indicates where the attack originated from (e.g., internal or external to the network).
Attack Vector (AV)	Destination IP Address	The destination IP address helps in understanding the target of the attack, relevant to assessing how the vulnerability can be exploited.
Attack Complexity (AC)	Source Port	The source port can indicate the complexity of the attack, particularly if the attack requires specific or unusual network conditions.
Attack Complexity (AC)	Destination Port	The destination port is relevant to understanding the network layer's exposure, which influences the attack complexity.
Exploitability Metrics (EM)	Packet Size	Packet size can affect how easily an exploit can be carried out, particularly in buffer overflow attacks or similar vulnerabilities.

No additional sampling techniques were necessary given the dataset size of 750 rows. The dataset size was deemed appropriate for the scope of this study due to several reasons. First, the dataset contains a sufficient number of rows to represent a diverse range of network traffic scenarios, which is essential for the study. Besides, the dataset includes various categories of potential security threats, including Phishing, DoS, and Man-in-the-Middle, with normal traffic. Ensuring that these classes are well-represented is crucial for developing accurate models. The existing size of the dataset is adequate to capture the distribution of these classes without the need for oversampling or under-sampling techniques. Last but not least, the scope of this study is focused on understanding and analysing network

traffic to identify security threats. The dataset size aligns well with this objective, allowing for a comprehensive examination of patterns and anomalies in network traffic. For exploratory and proof-of-concept studies, a dataset of this size is typically sufficient to derive meaningful insights.

2.2 Data Preparation

Data preparation is a foundational step in ensuring the integrity and effectiveness of the vulnerability risk scoring model. It involves several critical procedures that transform the raw data into a format suitable for analysis and modelling. Initially, the data undergoes a thorough cleaning and preprocessing phase. This involves addressing missing values and inconsistencies within the dataset. Missing data are handled either through imputation techniques, which estimate and fill in the gaps based on existing data, or by removing incomplete entries if they are too numerous or if imputation is not feasible. Ensuring consistency in variable formats and values is also crucial during this phase, as discrepancies can lead to inaccuracies in subsequent analyses.

Following data cleaning, categorical variables in the dataset are encoded to facilitate their use in statistical models. Categorical variables are converted into numerical representations using methods such as label encoding and one-hot encoding. Label encoding assigns a unique integer to each category within a variable, simplifying its integration into models that require numerical input. One-hot encoding, on the other hand, creates binary columns for each category, allowing the model to interpret the presence or absence of each category more explicitly. This transformation is essential for models that do not inherently handle categorical data.

Feature transformation is another important step in data preparation. This process adjusts the scale and distribution of features to ensure they contribute equally to the model. Normalisation is used to scale features to a range between 0 and 1, which can improve the performance of certain models by ensuring that no feature dominates due to its scale. Alternatively, standardisation transforms features to have a mean of zero and a standard deviation of one, which helps in normalising the impact of each feature on the model.

Although the dataset expands to 56 variables due to one-hot encoding, various methods can be used to assess multicollinearity. According to James et al. (2013), a simple way to detect collinearity is by examining the correlation matrix of the predictors. High absolute values in the matrix indicate pairs of strongly correlated variables, signalling potential multicollinearity issues (James et al., 2013). Figure 2 presents the correlation matrix of the dataset, calculated using Spearman's rank correlation, as defined in Eq. (1).

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2-1)} \quad (1)$$

where:
r_s = Spearman's Rank Correlation Coefficient
d_i = The difference between the ranks of two variables for the *i*-th observation
n = The number of observations

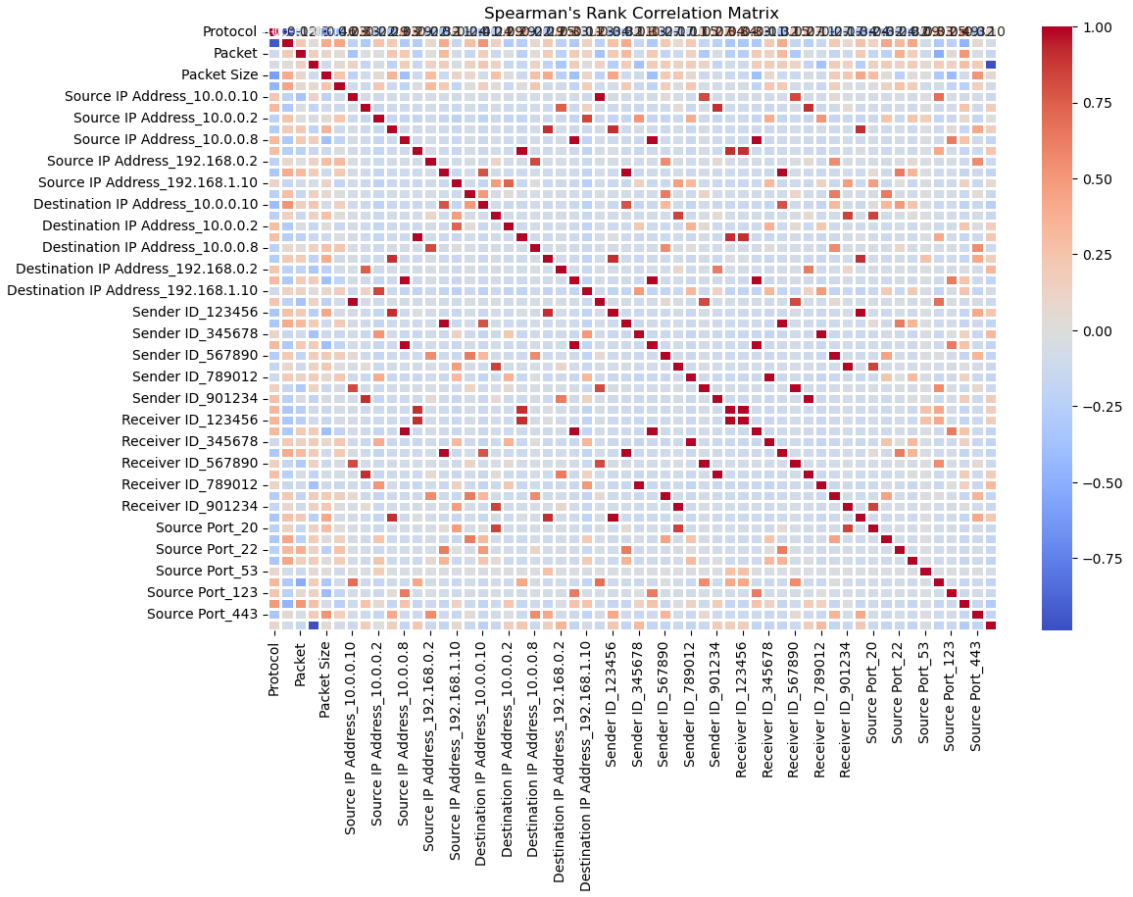

Figure 2. Correlation Matrix

Spearman's correlation is particularly suited for ordinal or non-parametric data, as it evaluates monotonic relationships. Based on the matrix, most features exhibit low correlations with one another, as indicated by the predominance of light-coloured squares. This implies that multicollinearity is not a significant concern for the majority of features in this dataset. However, there are a few notable correlations, such as between specific Source IP Addresses and Destination IP Addresses or between Sender IDs and Receiver IDs, which exhibit stronger positive correlations (red/orange squares). It is important to acknowledge that collinearity issues cannot always be fully identified using a correlation matrix, as multicollinearity may exist among three or more variables, even if no pair of variables exhibits a strong correlation. Additionally, the weak correlations observed for most variables suggest they may contribute independently to the analysis, reducing redundancy in the dataset.

To more thoroughly address potential multicollinearity, we use the Variance Inflation Factor (VIF), which is calculated using the formula in Eq. (2). The VIF measures the ratio of the variance of a regression coefficient (β_j) in the full multiple regression model to the variance of β_j when the predictor is fitted independently. While the Spearman correlation matrix provides valuable insights into pairwise monotonic relationships, it does not fully capture multicollinearity that may exist among three or more variables. Thus, the VIF analysis becomes essential to detect and quantify such hidden multicollinearity in the dataset.

The response variable is the one being predicted, while the predictor variables are the independent variables used in the regression model. A VIF value of 1 indicates no multicollinearity, while higher values suggest increasing levels of multicollinearity. By identifying variables with high VIF values, we can reduce redundancy, improve model interpretability, and ensure stable coefficient estimates. The results of the VIF calculation are shown in Figure 3.

$$VIF\ (X_i)\ =\ \frac{1}{1-R_i^2} \tag{2}$$

where:
X_i = The predictor variable
R_i^2 = The coefficient of determination

	feature	VIF
0	Protocol	inf
1	Flag	inf
2	Packet	inf
3	Destination Port	inf
4	Packet Size	118.4736
5	Source IP Address_10.0.0.10	inf
6	Source IP Address_10.0.0.15	inf
7	Source IP Address_10.0.0.2	inf
8	Source IP Address_10.0.0.5	inf
9	Source IP Address_10.0.0.8	inf
10	Source IP Address_192.168.0.1	inf
11	Source IP Address_192.168.0.2	inf
12	Source IP Address_192.168.0.5	inf
13	Source IP Address_192.168.1.10	inf
14	Source IP Address_192.168.1.5	inf
15	Destination IP Address_10.0.0.10	inf
16	Destination IP Address_10.0.0.15	inf
17	Destination IP Address_10.0.0.2	inf
18	Destination IP Address_10.0.0.5	inf
19	Destination IP Address_10.0.0.8	inf
20	Destination IP Address_192.168.0.1	inf
21	Destination IP Address_192.168.0.2	inf
22	Destination IP Address_192.168.0.5	inf
23	Destination IP Address_192.168.1.10	inf
24	Destination IP Address_192.168.1.5	inf
25	Sender ID_123456	inf
26	Sender ID_234567	inf
27	Sender ID_345678	inf
28	Sender ID_456789	inf
29	Sender ID_567890	inf
30	Sender ID_678901	inf
31	Sender ID_789012	inf
32	Sender ID_890123	inf
33	Sender ID_901234	inf
34	Sender ID_987654	inf
35	Receiver ID_123456	inf
36	Receiver ID_234567	inf
37	Receiver ID_345678	inf
38	Receiver ID_456789	inf
39	Receiver ID_567890	inf
40	Receiver ID_678901	inf
41	Receiver ID_789012	inf
42	Receiver ID_890123	inf
43	Receiver ID_901234	inf
44	Receiver ID_987654	inf
45	Source Port_20	inf
46	Source Port_21	inf
47	Source Port_22	inf
48	Source Port_25	inf
49	Source Port_53	inf
50	Source Port_67	inf
51	Source Port_123	inf
52	Source Port_161	inf
53	Source Port_443	inf
54	Source Port_12345	inf

Figure 3. VIF analysis

When VIF shows `INF` (infinity), it indicates severe multicollinearity in the dataset. One common reason for this is the dummy variable trap resulting from one-hot encoding categorical variables. When one hot encodes a categorical variable and includes all the resulting binary columns in the model, the sum of these columns can perfectly predict the original categorical variable. This perfect correlation leads to multicollinearity, which results in infinite VIF for these columns. Another reason for infinite VIF is the presence of redundant features. If multiple columns in a dataset are perfectly correlated or identical, this redundancy causes a severe multicollinearity problem. In such cases, the variance of the estimated regression coefficient becomes infinite, resulting in an infinite VIF.

To address these issues, one of the dummy variables is excluded to avoid the dummy variable trap and remove redundant features to prevent perfect correlation, thus mitigating multicollinearity and ensuring finite VIF values. However, even after excluding one of the dummy variables and removing redundant features, infinite VIF values persist, as shown in Figure 4.

	feature	VIF
0	Protocol	inf
1	Flag	inf
2	Packet	inf
3	Destination Port	inf
4	Packet Size	118.4736
5	Source IP Address_10.0.0.15	inf
6	Source IP Address_10.0.0.2	inf
7	Source IP Address_10.0.0.5	inf
8	Source IP Address_10.0.0.8	inf
9	Source IP Address_192.168.0.1	inf
10	Source IP Address_192.168.0.2	inf
11	Source IP Address_192.168.0.5	inf
12	Source IP Address_192.168.1.10	inf
13	Source IP Address_192.168.1.5	inf
14	Destination IP Address_10.0.0.15	inf
15	Destination IP Address_10.0.0.2	inf
16	Destination IP Address_10.0.0.5	inf
17	Destination IP Address_10.0.0.8	inf
18	Destination IP Address_192.168.0.1	inf
19	Destination IP Address_192.168.0.2	inf
20	Destination IP Address_192.168.0.5	inf
21	Destination IP Address_192.168.1.10	inf
22	Destination IP Address_192.168.1.5	inf
23	Sender ID_234567	inf
24	Sender ID_345678	inf
25	Sender ID_456789	inf
26	Sender ID_567890	inf
27	Sender ID_678901	inf
28	Sender ID_789012	inf
29	Sender ID_890123	inf
30	Sender ID_901234	inf
31	Sender ID_987654	inf
32	Receiver ID_234567	inf
33	Receiver ID_345678	inf
34	Receiver ID_456789	inf
35	Receiver ID_567890	inf
36	Receiver ID_678901	inf
37	Receiver ID_789012	inf
38	Receiver ID_890123	inf
39	Receiver ID_901234	inf
40	Receiver ID_987654	inf
41	Source Port_20	inf
42	Source Port_22	inf
43	Source Port_25	inf
44	Source Port_53	inf
45	Source Port_67	inf
46	Source Port_123	inf
47	Source Port_161	inf
48	Source Port_443	inf
49	Source Port_12345	inf

Figure 4. VIF analysis after excluding one of the dummy variables

After conducting both Spearman's Correlation and VIF analyses, it became evident that certain features exhibit high or perfect multicollinearity, as indicated by strong correlations and infinite VIF values. Although this typically suggests redundancy among variables, these features are retained because they align directly with the core objectives of this analysis. Each feature represents a distinct and critical aspect of the underlying data, essential for meeting the specific requirements of the study. Therefore, the dataset remains valuable when appropriate modelling techniques and validation strategies are applied. Secondly, the observed correlations between certain variables reflect the inherent relationships within the data. These relationships are integral to the predictive power of the model.

Although multicollinearity may affect the stability of coefficient estimates, the primary goal of this model is prediction rather than the interpretation of individual coefficients. The predictive performance of the model is of utmost importance, and initial tests indicate that the inclusion of these features enhances the model's ability to generalise. It is important to recognise that the challenges of multicollinearity do not necessarily invalidate the dataset for this research. While multicollinearity can complicate model interpretation, it does not render the dataset unusable. As discussed in a study by James and his team (2013) named "An Introduction to Statistical Learning", multicollinearity can be managed through regularisation techniques such as Ridge regression, which effectively handles multicollinear features by adding a penalty to the regression coefficients (James et. al., 2013).

In this study, although multicollinearity is present in the dataset, the choice of predictive models has been intentionally made to address this issue. Unlike linear regression models, which are highly sensitive to multicollinearity and may produce unstable coefficient estimates, the selected models do not rely on the assumption of linearity and are inherently more robust to multicollinear features. For example, decision trees and ensemble methods like AdaBoost focus on splitting the data based on feature importance rather than estimating coefficients, thereby minimising the impact of multicollinear predictors. Similarly, KNN operates by measuring distances between data points in feature space, which is not directly affected by multicollinearity. Therefore, instead of applying regularisation techniques like Ridge regression, which are designed to mitigate multicollinearity in linear models, this study leverages the strengths of non-linear models that naturally manage multicollinearity, ensuring that predictive performance remains the primary focus.

Hence, the decision to proceed with the current feature set, despite the presence of multicollinearity, is a calculated one. It is driven by the need to fully capture the complexity of the dataset, which is crucial for the success of this analysis.

Finally, the dataset is split into training and testing subsets in a 70:30 ratio to facilitate model training and performance evaluation. This partitioning enables the training of models on one portion of the data while assessing their performance on an unseen subset. This approach helps validate the model's effectiveness and prevent overfitting, ensuring that the model performs well not only on the training data but also on new, unseen data.

2.3 Creating New Feature

The development of a risk score is a systematic process designed to quantify and evaluate potential risks based on various contributing factors. The flowchart in Figure 5 outlines the key steps involved in creating a new feature named "Risk Score" that can be integrated into predictive models. This structured approach ensures that the risk score is both data-driven and reflective of underlying patterns within the dataset.

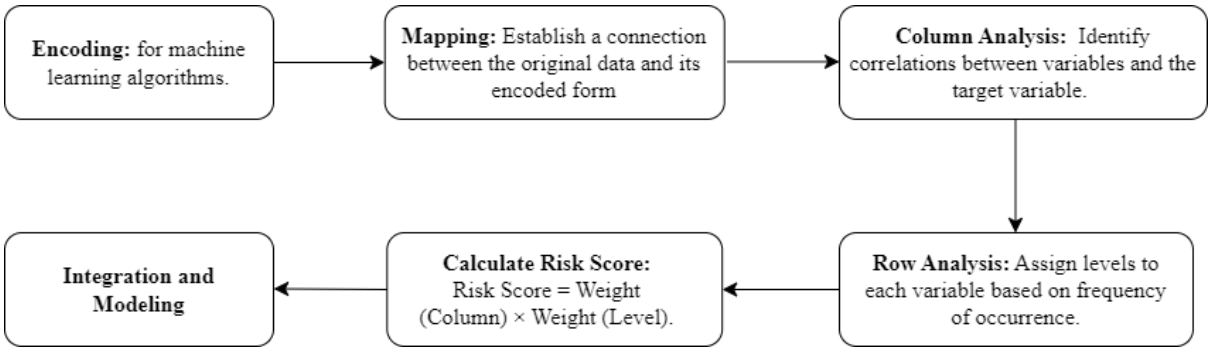

Figure 5. Steps on developing the new feature and risk score

To construct a precise and logical feature, all columns are considered as input variables, with 'Risk Score' designated as the target variable. The initial step involves assigning weights to each input variable based on their correlation with the original target variable, 'Target Variable' which represents the type of cyber threats in the dataset. These weights are normalised to ensure they collectively sum up to 1.

Subsequently, a detailed examination of each variable is conducted. The assignment of weights at this level is determined by the frequency of occurrence in the dataset. This approach assumes that more frequently appearing categories are less distinctive or more common, warranting a lower weightage, as they may have more available solutions to address the associated risk. This paper stems

from the viewpoint that, due to the uncertainty and potential risks associated with infrequently occurring unknown vulnerabilities, they should be prioritised for immediate attention. The normalisation of these weights guarantees that they collectively sum to 1.

Next, the risk score is calculated using the weights. To begin, a new column named 'Risk Score' is created and initialised to zero for all records. The subsequent loop iterates through each categorical column in the dataset. For one-hot encoded columns, the code multiplies the weights by the correlation weight for each unique value. These products are then added to the 'Risk Score' column, effectively incorporating the impact of each value on the risk assessment. In the case of non-one-hot encoded columns, the code multiplies the normalised weightage and the corresponding correlation weight. The cumulative result for each column contributes to the final 'Risk Score.' Finally, the 'Risk Score' values are normalised to a range from 0 to 10 by scaling based on the minimum and maximum scores in the column. This normalisation ensures that the 'Risk Score' aligns with a standardised scale, making it more interpretable and comparable.

The calculated risk score is reintegrated into the dataset, serving as a critical variable for subsequent analysis. This integration allows for the seamless transition into regression modelling, where the risk score can be utilised as a predictor of outcomes or as a key component in evaluating the overall risk landscape. By incorporating the risk score back into the dataset, the model is enriched with a quantifiable measure of risk that has been systematically derived from the underlying data. This step ensures that the developed model not only considers the raw input data but also leverages the nuanced insights captured by the risk score, thereby enhancing the model's predictive accuracy and its ability to inform risk management decisions.

2.4 Regression Modelling

In this paper, a range of machine learning models are trained on the training dataset to construct a risk-scoring system. The employed models encompass Decision Tree, XGBoost, Random Forest, k-Nearest Neighbors Regression (KNN), Support Vector Regression (SVR), AdaBoost, and Artificial Neural Network (ANN), with the formula of these models are stated from Eq. (3) to Eq. (9) respectively.

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{3}$$

where:
y_i = The target values
N = The number of samples at the leaf node

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} \hat{y}_t \tag{4}$$

where:
$\hat{y}_t$ = The prediction from the *t*-th tree
T = The number of trees

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta f_t(x_i) \tag{5}$$

where:
$\hat{y}_i^{(t)}$= The prediction at iteration *t*
$\hat{y}_i^{(t-1)}$ = The prediction at the previous iteration
$\eta =$ The learning rate
$f_t(x_i)$ = The *t*-th tree's prediction

$$\hat{y} = \frac{1}{k}\sum_{i=1}^{k} y_i \quad (6)$$

where:
y_i = The target values of the *k*-nearest neighbours

$$min\frac{1}{2}||w||^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*) \quad (7)$$

where:
ξ_i, ξ_i^* = The slack variables
C= The regularization parameter
w = weight vector

$$\hat{y} = \sum_{i=1}^{T} \alpha_t f_t(x) \quad (8)$$

where:
α_t = The weight for the *t*-th weak learner
$f_t(x)$ = The *t*-th weak learner's prediction

$$z^{(l)} = W^{(l)}a^{(t-l)} + b^{(l)} \quad (9)$$
$$a^{(l)} = \sigma(z^{(l)})$$

where:
$z^{(l)}$ = The weighted input to layer *l*
$W^{(l)}$ = The weighted matrix to layer *l*
$b^{(l)}$ = The bias vector to layer *l*
$a^{(t-l)}$ = Activation from the previous layer
$\sigma =$ Activation function

The selected machine learning models are well-suited to handle multicollinearity in various ways. Decision Trees and Random Forests are resilient due to their hierarchical structure and ensemble approach, which reduce the impact of correlated features by focusing on the most informative features and averaging predictions from multiple trees. XGBoost enhances this by applying regularisation to penalise complex models, which helps in managing multicollinearity effectively. KNN is less affected by multicollinearity because it relies on distance metrics for predictions, though it requires careful feature scaling. SVR addresses multicollinearity by using kernel functions to transform data into higher dimensions, aiding in the separation of collinear features. AdaBoost iteratively corrects errors, thereby mitigating the influence of multicollinear features. ANN, with its flexible architecture, can adapt to multicollinearity but needs regularisation to prevent overfitting. This combination of models ensures robust handling of multicollinearity through various complementary techniques.

According to Gupta, B. (2023), optimising a machine learning model's hyperparameters is a far better strategy to improve the model. Hyperparameters are the different setups and choices made throughout a machine learning model's training phase. They significantly impact a machine learning model's accuracy. Therefore, the error on the testing dataset can be further decreased through hyperparameter adjustment. In this case, GridSearchCV is utilised (Gupta, B., 2023).

2.5 Model Evaluation

Subsequently, an extensive evaluation will be conducted on the testing data to compare the effectiveness of each model. A comprehensive analysis will consider various metrics, such as Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2), to identify the model that exhibits the highest performance. The aim is to identify the model that excels in accuracy and precision, providing the best predictive capabilities for the risk-scoring system. This detailed evaluation ensures the selection of a reliable and robust model tailored to

the specific nuances of the dataset, ensuring accurate and effective risk assessments.

3. Results and Discussion

3.1 Development of New Feature

To develop a new feature, Risk Score as the target variable, correlations between the variables and the original target variable in the dataset are considered and contribute to a numeric value for the weights. The final output is shown in Eq. (10).

$$\begin{aligned} Risk\ Score = {} & 0.053910x_1 + 0.049013x_2 + 0.004481x_3 + 0.001561x_4 + 0.028179x_5 \\ & + 0.115255x_6 + 0.011482x_7 + 0.011674x_8 + 0.023882x_9 \\ & + 0.011603x_{10} + 0.024146x_{11} + 0.000770x_{12} + 0.034926x_{13} \\ & + 0.023248x_{14} + 0.020795x_{15} + 0.001662x_{16} + 0.018795x_{17} \\ & + 0.003260x_{18} + 0.018825x_{19} + 0.000770x_{20} + 0.026332x_{21} \\ & + 0.010399x_{22} + 0.009524x_{23} + 0.024146x_{24} + 0.018583x_{25} \\ & + 0.011482x_{26} + 0.006788x_{27} + 0.023248x_{28} + 0.018749x_{29} \\ & + 0.024146x_{30} + 0.019507x_{31} + 0.001534x_{32} + 0.025257x_{33} \\ & + 0.008044x_{34} + 0.007932x_{35} + 0.004381x_{36} + 0.004381x_{37} \\ & + 0.024146x_{38} + 0.025257x_{39} + 0.023248x_{40} + 0.008044x_{41} \\ & + 0.007932x_{42} + 0.018749x_{43} + 0.019507x_{44} + 0.001534x_{45} \\ & + 0.006788x_{46} + 0.001662x_{47} + 0.015765x_{48} + 0.028828x_{49} \\ & + 0.020754x_{50} + 0.009404x_{51} + 0.009180x_{52} + 0.022308x_{53} \\ & + 0.027724x_{54} + 0.018259x_{55} + 0.008236x_{56} \end{aligned} \quad (10)$$

where:

x_1 = Protocol
x_2 = Flag
x_3 = Packet
x_4 = Destination Port
x_5 = Packet Size
x_6 =Target Variable
x_7 = Source IP Address_10.0.0.10
x_8 = Source IP Address_10.0.0.15
x_9 = Source IP Address_10.0.0.2
x_{10} = Source IP Address_10.0.0.5
x_{11} = Source IP Address_10.0.0.8
x_{12} = Source IP Address_192.168.0.1
x_{13} = Source IP Address_192.168.0.2
x_{14} = Source IP Address_192.168.0.5
x_{15} = Source IP Address_192.168.1.10
x_{16} = Source IP Address_192.168.1.5
x_{17} = Destination IP Address_10.0.0.10
x_{18} = Destination IP Address_10.0.0.15
x_{19} = Destination IP Address_10.0.0.2
x_{20} = Destination IP Address_10.0.0.5
x_{21} = Destination IP Address_10.0.0.8
x_{22} = Destination IP Address_192.168.0.1

x_{23} = Destination IP Address_192.168.0.2
x_{24} = Destination IP Address_192.168.0.5
x_{25} = Destination IP Address_192.168.1.10
x_{26} = Destination IP Address_192.168.1.5
x_{27} = Sender ID_123456
x_{28} = Sender ID_234567
x_{29} = Sender ID_345678
x_{30} = Sender ID_456789
x_{31} = Sender ID_567890
x_{32} = Sender ID_678901
x_{33} = Sender ID_789012
x_{34} = Sender ID_890123
x_{35} = Sender ID_901234
x_{36} = Sender ID_987654
x_{37} = Receiver ID_123456
x_{38} = Receiver ID_234567
x_{39} = Receiver ID_345678
x_{40} = Receiver ID_456789
x_{41} = Receiver ID_567890
x_{42} = Receiver ID_678901
x_{43} = Receiver ID_789012
x_{44} = Receiver ID_890123
x_{45} = Receiver ID_901234
x_{46} = Receiver ID_987654
x_{47} = Source Port_20
x_{48} = Source Port_21
x_{49} = Source Port_22
x_{50} = Source Port_25
x_{51} = Source Port_53
x_{52} = Source Port_67
x_{53} = Source Port_123
x_{54} = Source Port_161
x_{55} = Source Port_443
x_{56} = Source Port_12345

Next, after assigning the weightage for each variable, a weightage for each variable level is also assigned. However, instead of correlation, it is assigned based on the frequency of occurrence, as mentioned in Section 2.3. The weights are shown in Table 3. Note that for variables encoded using one hot encoding, the weights are assigned to 0.000000 for 0 and 1.000000 for 1.

Table 3. Weightage for each level of each variable

Variable	Level (Encoded)	Weightage
Protocol	0	0.496000
Protocol	1	0.504000
Flag	0	0.079570
Flag	1	01411114
Flag	2	0.379490

Flag	3	0.129826
Packet	0	0.053336
Packet	1	0.049233
Packet	2	0.025601
Packet	3	0.640026
Packet	4	0.053336
Packet	5	0.049233
Packet	6	0.053336
Packet	7	0.026668
Packet	8	0.049233
Packet Size	0	0.244276
Packet Size	1	0.039766
Packet Size	2	0.055159
Packet Size	3	0.142494
Packet Size	4	0.244276
Packet Size	5	0.131533
Packet Size	6	0.142494
Destination Port	0	0.104561
Destination Port	1	0.104561
Destination Port	2	0.052280
Destination Port	3	0.627364
Destination Port	4	0.104561
Destination Port	5	0.006674
Target Variable	0	0.077670
Target Variable	1	0.388350
Target Variable	2	0.072816
Target Variable	3	0.077670
Target Variable	4	0.077670
Target Variable	5	0.072816
Target Variable	6	0.077670
Target Variable	7	0.077670
Target Variable	8	0.077670

Therefore, once the weights are determined, the risk scores are calculated by inserting the value for each level into Eq. (10).

3.2 Model Development

The new dataset mentioned in Figure 6 is used to develop the regression modelling and Figure 7 shows the evaluation of the models.

Source Port_20	Source Port_21	Source Port_22	Source Port_25	Source Port_53	Source Port_67	Source Port_123	Source Port_161	Source Port_443	Source Port_12345	Risk Score
0	0	0	0	0	0	0	0	0	1	2.591990
0	0	0	0	1	0	0	0	0	0	0.120429
0	0	1	0	0	0	0	0	0	0	0.606199
0	0	0	0	0	0	1	0	0	0	9.508131
1	0	0	0	0	0	0	0	0	0	3.919613
...	...	...	...	...	...	...	...	...	...	...
0	1	0	0	0	0	0	0	0	0	0.680016
0	0	0	0	0	0	0	0	0	1	0.164765
0	0	1	0	0	0	0	0	0	0	0.606199
0	0	0	0	0	0	1	0	0	0	1.320266
1	0	0	0	0	0	0	0	0	0	3.793646

Figure 6. Dataset with new feature

Model	MSE	MAE	RMSE	R2
K-Nearest Neighbors Regression (KNN)	2.136498e-33	1.430954e-17	4.622227e-17	1.000000
Decision Tree	3.601896e-31	2.862525e-16	6.001579e-16	1.000000
XGBoost	7.363440e-07	4.319542e-04	8.581049e-04	1.000000
Random Forest	5.205950e-04	5.205950e-04	2.281655e-02	0.999759
AdaBoost	1.009633e-01	2.544546e-01	3.177471e-01	0.953166
Support Vector Regression (SVR)	1.506785e-01	3.440583e-01	3.881733e-01	0.930105
Artificial Neural Network (ANN)	1.527066e-01	1.268014e-01	3.907769e-01	0.929164

Figure 7. Model Evaluation

The evaluation of various regression models applied to this dataset is depicted in Figure 7, showcasing performance metrics including MSE, MAE, RMSE and R^2. The analysis highlights that KNN, Decision Tree, XGBoost, and Random Forest achieved exceptional accuracy, characterised by minimal error across all metrics. KNN, in particular, demonstrates a notably low MSE and MAE, reflecting its high precision. The Decision Tree and XGBoost models exhibit extremely low RMSE, indicating their accuracy in predicting values in the original unit of the target variable. Furthermore, all models, including AdaBoost, SVR, and ANN, show high R^2 values, suggesting that a substantial proportion of variance in the target variable is explained by the models.

The choice of models reflects their ability to handle the dataset's characteristics effectively. KNN, Decision Tree, XGBoost, and Random Forest have emerged as the most robust performers. Their selection is based on their ability to minimise errors and provide reliable risk scores, which can be critical for practical applications. For instance, the precision of KNN makes it suitable for scenarios where detailed, localised predictions are necessary. Decision Trees and XGBoost, with their accuracy in predicting values and handling complex relationships, are advantageous for dynamic environments

where model adaptability is essential. Random Forest's ensemble approach helps in reducing variance, making it robust in varied conditions.

In practical terms, organisations can utilise these models to develop customised risk-scoring systems tailored to their specific needs. For example, companies can apply the most accurate models to assess their cybersecurity risks more effectively, identify high-risk areas, and allocate resources more efficiently. The flexibility of the proposed model means it can be adapted to various contexts, such as different industries or organisational environments, enhancing its relevance and utility.

Moreover, integrating these models into existing cybersecurity frameworks can significantly improve risk management strategies. By providing more accurate and relevant risk assessments, organisations can better prioritise their security measures and address vulnerabilities proactively. This approach ensures that the risk-scoring system is not only precise but also practical, offering actionable insights that align with the unique requirements of different organisational contexts.

The comprehensive performance evaluation confirms that the proposed model, supported by these robust regression techniques, provides a valuable tool for developing effective risk-scoring systems. This approach enhances the accuracy of risk assessments, contributes to more proactive and effective cybersecurity strategies, and demonstrates the model's adaptability to various scenarios, ultimately advancing the overall effectiveness of cybersecurity measures.

4. Conclusion

In conclusion, this research has utilised a well-defined framework and dataset to develop a robust risk-scoring model, highlighting the importance of a thorough data science process. This model, crafted through a comprehensive approach from data preparation to model evaluation, provides a foundational tool for addressing cybersecurity challenges. A key contribution of this research is the demonstration of how such a risk-scoring model can be adapted to various contexts, including different industries, companies, or specific situations. This adaptability underscores the model's versatility, enabling it to be customised to reflect the unique characteristics and threats of different organisational environments.

While CVSS scores are accurate and standardised, they often require detailed data that may not always be available due to constraints such as time or resource limitations. In these cases, a customised risk-scoring model can offer a practical solution by focusing on the data that is accessible and relevant to the specific context. This approach ensures more accurate and actionable risk assessments, improving vulnerability identification and prioritisation even in scenarios where comprehensive CVSS data is lacking. By integrating local data characteristics and contextual factors, this model enhances the effectiveness of cybersecurity measures and provides a valuable tool for tailored risk management.

This approach not only optimises resource allocation by focusing on high-risk areas but also aligns with broader objectives such as fostering innovation and enhancing cybersecurity resilience. By integrating principles from Sustainable Development Goal (SDG) 9, which emphasises Industry, Innovation, and Infrastructure, the model promotes inclusive and innovative cybersecurity solutions. It underscores a commitment to user-centric design and cost considerations, making advanced cybersecurity measures accessible and scalable.

In addition to its practical applications, the research significantly contributes to the academic field by deepening the understanding of how cyber threats and vulnerabilities impact various systems. By evaluating the severity of these threats, researchers can acquire critical insights into their evolving dynamics. Integrating the findings from this model into cybersecurity education offers a valuable opportunity for educators to present students with up-to-date knowledge and real-world examples. This approach bridges the gap between theoretical concepts and practical application, allowing students to engage with the model directly, analyse relevant data, and develop strategies for addressing similar

threats. Consequently, this hands-on experience reinforces the applicability of theoretical concepts and enhances students' preparedness for real-world cybersecurity challenges.

5. Acknowledgements

The authors thank the Universiti Malaysia Pahang Al-Sultan Abdullah (UMPSA) for providing financial support under the Social Grant UMP RDU223419 and the UMPSA facilities. The authors also would like to thank the reviewers for the valuable comments for improvements to this paper.

6. References

Ahmad, T., Yunos, Z., & Ariffin, K. A. Z. (2021). Cybersecurity challenges in Malaysia: A strategic outlook. Malaysian Journal of Computer Science, 34(1), 1-18.

Allodi, L., & Massacci, F. (2014). Security events and vulnerability data for cybersecurity risk estimation. Risk Analysis.

Bozorgi, M., Saul, L., Savage, S., & Voelker, G. M. (2010). Beyond heuristics: Learning to classify vulnerabilities and predict exploits. In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD).

Chatzipoulidis, A. (2015). Enterprise management and software risk prediction based on security metrics. ResearchGate. https://doi.org/10.13140/RG.2.2.35608.85768

Farris, K. A., Shah, A., Cybenko, G., Ganesan, R., & Jajodia, S. (2018). An empirical analysis of vulnerability lifecycle and patch timeliness. Journal of Cybersecurity, 4(1), tyx011.

Gupta, B. (2023). Grid search in machine learning. Scaler Topics. https://www.scaler.com/topics/machine-learning/grid-search-in-machine-learning/

James et. al. (2013). An introduction to statistical learning. https://www.researchgate.net/publication/278405332_An_Introduction_to_Statistical_Learning

Kim, J., Malhotra, A., Desmet, L., & Rieck, K. (2020). Using machine learning to predict real-world exploits. ACM Transactions on Information and System Security (TISSEC), 23(2), Article 13.

Kemp, S. (2023). Digital 2023: Malaysia — DataReportal – Global Digital Insights. DataReportal – Global Digital Insights. https://datareportal.com/reports/digital-2023-malaysia#:~:text=There%20were%2033.03%20million%20internet%20users%20in%20Malaysia%20in%20January,percent%20between%202022%20and%202023

Li, Y., & Liu, Q. (2021). A comprehensive review study of cyber-attacks and cyber security; Emerging trends and recent developments. Energy Reports, 7, 8176–8186. https://www.sciencedirect.com/science/article/pii/S2352484721007289

Malaysia. Ministry of Communications and Multimedia. (2020). Malaysia Cyber Security Strategy 2020-2024. https://asset.mkn.gov.my/wpcontent/uploads/2020/10/MalaysiaCyberSecurityStrategy2020-2024.pdf

Mell, P., Scarfone, K., & Romanosky, S. (2007). A complete guide to the Common Vulnerability Scoring System version 2.0.

Samniea, Z. (2023). Cyber threat data for new malware attacks. Kaggle. https://www.kaggle.com/datasets/zunxhisamniea/cyber-threat-data-for-new-malware-attacks

Smith, R., Smith, T., & Fischbach, J. (2022). Behavioral analytics for enhancing risk scoring models. IEEE Security & Privacy, 20(3), 42-51.

Statista Research Department. (2023). Number of internet users in Malaysia 2013-2028. https://www.statista.com/statistics/553752/number-of-internet-users-in-malaysia/

Zunxhi Samniea. (2023). *Cyber threat data for new malware attacks* [Data set]. Kaggle. https://www.kaggle.com/datasets/zunxhisamniea/cyber-threat-data-for-new-malware-attacks