UNIVERSITI TEKNOLOGI MARA

CSP760

PREDICTING VULNERABILITY SUSCEPTIBILITY IN MALAYSIAN BANK USING SUPERVISED MACHINE LEARNING

STUDENT NOR ADANI BINTI KAMAL MOHAMAD NASIR (2024782087)

SUPERVISOR DR SITI ARPAH BINTI AHMAD

Date: 26 October 2025

UNIVERSITI TEKNOLOGI MARA

TABLE OF CONTENT

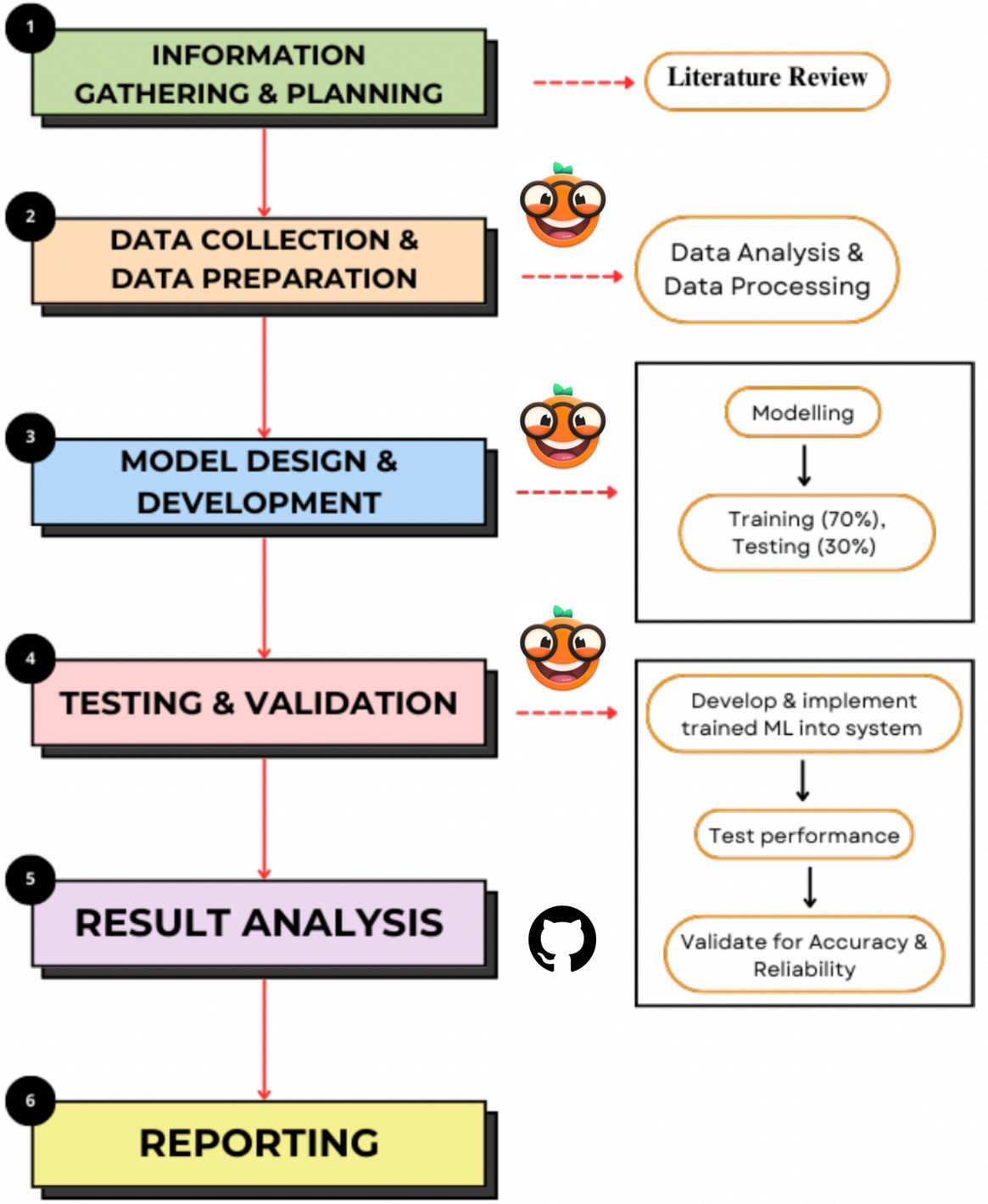
UPDATE WEEK 3

No	Item	Input	Process	Output / Expected Outcome	Status
1	**Dataset Integration**	Kaggle dataset + CVE → data (CSV)	- Merged datasets into single file (merged_vulnerability_data.csv). - Aligned key columns: **CVE, CVSS Score, Severity,** Plugin ID, Family, Label. - Ensured data consistency for ML use.	Unified, structured dataset ready for cleaning and feature engineering.	Preliminary Result
2	**Data Cleaning & Preprocessing**	Raw merged dataset	- Removed duplicates, because of merged dataset. - Handled missing “Severity” (Low–Critical). - Encoded categorical data for Orange compatibility.	Clean dataset (cleaned_vulnerability_data.csv) free of duplicates and missing values.	Preliminary Result
3	**Exploratory Data Analysis (EDA)**	Cleaned dataset	- Visualized CVSS distribution. - Plotted severity proportion and correlation heatmap. - Identified key predictive features.	Insights on severity trends and CVSS correlation. Initial visualization ready for presentation.	Preliminary Result
4	**Model Setup in Orange**	Orange Data Mining software	Imported dataset → Data → Select Columns → Random Forest & Neural Network Regression → Test & Score → Confusion Matrix.- Tested pipeline using sample data.	Preliminary Random Forest & NNR results available (accuracy, MAE, RMSE).	Preliminary Result
5	**Preliminary Result (Output)**	Clean dataset + Orange model	- Random Forest - Neural Network Regression - Observed “**Severity**” as most influential feature.	Initial performance metrics obtained — ready for further tuning.	Early Findings
6	**Report Update (Chapter 4)**	Chapter 4 (Implementation)	- Documented preliminary results.	Updated draft report with early findings and visuals.	In Progress

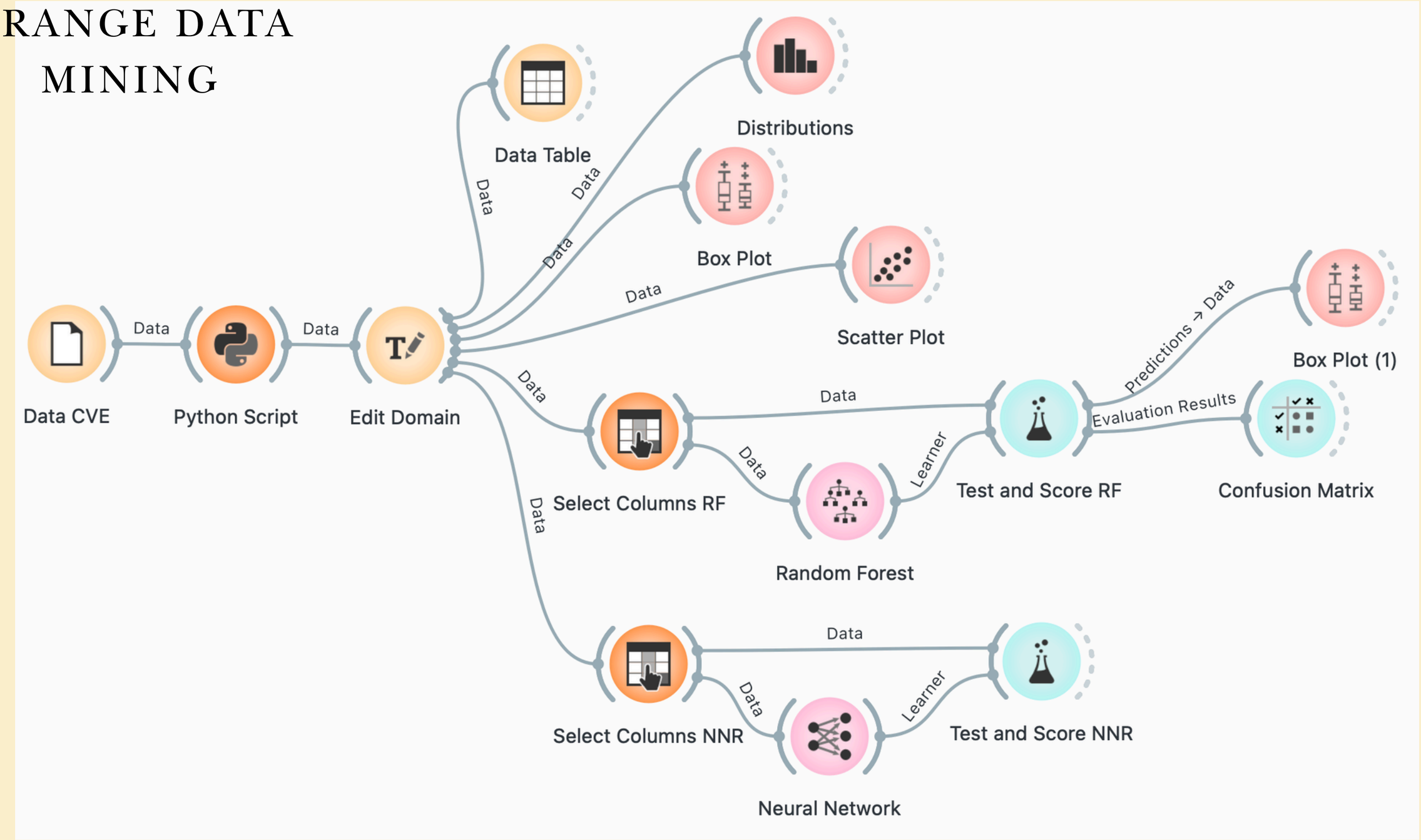
No	Title	Main Issue	Objective	Dataset	Algorithms	Solution
1	A cyber risk prediction model using common vulnerabilities and exposures (Negahdari Kia et al., 2023)	Predicting cyber risks using CVE data with supervised ML models	Eliminate expert bias and predict cyber risks through ML	CVE Database with topic mapping	Random Forest, Time Series Analysis	Generate a time-series risk prediction model, CyRiPred
2	A Hybrid Machine Learning System for Vulnerability Detection in Web Applications (Oliveira, 2023)	Hybrid ML approach for detecting vulnerabilities in web applications	Develop a hybrid ML model combining NLP and anomaly detection	Software Assurance Reference Database (SARD)	OCSVM, Random Forest, Logistic Regression	Propose a hybrid model integrating NLP and ML
3	A Vulnerability Analysis and Prediction Framework (Williams et al., 2020)	Predicting and analyzing vulnerability evolution over time	Develop a predictive framework for vulnerability trends	National Vulnerability Database (NVD)	Deep Neural Networks, Regression	Use topic modelling and storytelling techniques for vulnerability forecasting
4	Comprehensive Survey of different Machine Learning Algorithms used for Software Defect Prediction (K et al., 2022)	Addressing software defects using various ML techniques	Survey and analyze different ML algorithms for software defect prediction	PROMISE Repository, Software defect datasets	Random Forest, Naive Bayes, SVM, Decision Tree, ANN, K-Means Clustering	Comprehensive evaluation of supervised and unsupervised ML methods for defect prediction
5	Time series forecast modelling of vulnerabilities in the android operating system using ARIMA and deep learning methods (Gencer & Başçiftçi, 2021)	Forecasting future vulnerabilities in Android OS	Use time series and deep learning for vulnerability prediction	National Vulnerability Database (NVD) filtered for Android	ARIMA, LSTM, CNN	Apply deep learning models to predict Android vulnerabilities
6	Integrating Machine Learning for Sustaining Cybersecurity in Digital Banks (Asmar & Alia Tuqan, 2024)	Cybersecurity threats in digital banking and the need for ML-based solutions	Strengthen cybersecurity defenses using ML in digital banking	Literature review, cybersecurity threat reports	SVM, RNN, HMM, LOF	Develop an ML-driven cybersecurity framework for digital banks
7	Predicting Vulnerability Type in CVE Database with ML Classifiers (Yosifova et al., 2021)	Automating the classification of vulnerability types in CVE database	Enhance automated classification of CVE vulnerability types	CVE Database	Linear SVM, Naive Bayes, Random Forest	Train ML classifiers for improved CVE classification
8	Predicting Vulnerability Susceptibility in Malaysian Bank using Supervised Machine Learning	Current VA tools in Malaysian banks are reactive, lack predictive insights	Develop a machine learning model to predict cyberattack susceptibility & improve remediation efficiency.	Kaggle, Tenable, CVE	Random Forest, Neural Networks, Regression	Implement an AI-driven system to analyze VA data, forecast emerging threats, and provide real-time vulnerability insights for proactive mitigation.

PIPELINE

Phase	Type of Result
Phase 2 – Data Collection & Preparation	Screenshot of data table / cleaned CSV
Phase 2 – EDA (Exploratory Data Analysis)	Visual graphs that show patterns or imbalance
Phase 3 – Early Model Training	Show Accuracy / RMSE / Confusion Matrix. Table format of Orange evaluation results



ORANGE DATA MINING



PHASE 2

DATA COLLECTION & PREPARATION

Column can be easily **add** using
“Orange Data Mining” using python code

Data Table

	CVE ID	Description	Attack Vector	Affected OS	CVSS Score	Severity
1	CVE-2024-2...	FlyCms thro...	CVSS:3.1/AV:...	N/A	6.1	Medium
2	CVE-2023-5...	The affiliate-...	CVSS:3.1/AV:...	N/A	9.8	Critical
3	CVE-2023-6...	The Popup B...	CVSS:3.1/AV:...	N/A	6.1	Medium
4	CVE-2023-6...	The WP Trip...	CVSS:3.1/AV:...	N/A	4.8	Medium
5	CVE-2023-6...	The PayHere...	CVSS:3.1/AV:...	N/A	7.5	High
6	CVE-2023-6...	The WP STA...	CVSS:3.1/AV:...	N/A	7.5	High
7	CVE-2023-6...	The Backup ...	CVSS:3.1/AV:...	N/A	7.5	High
8	CVE-2023-6...	The Downloa...	CVSS:3.1/AV:...	N/A	7.5	High
9	CVE-2023-6...	The Html5 Vi...	CVSS:3.1/AV:...	N/A	5.4	Medium
10	CVE-2024-0...	A vulnerabilit...	CVSS:3.1/AV:...	N/A	2.4	Low
11	CVE-2023-5...	reNgine befo...	CVSS:3.1/AV:...	N/A	8.8	High
12	CVE-2023-5...	STMicroelec...	CVSS:3.1/AV:...	N/A	7.5	High

Python Script

Editor

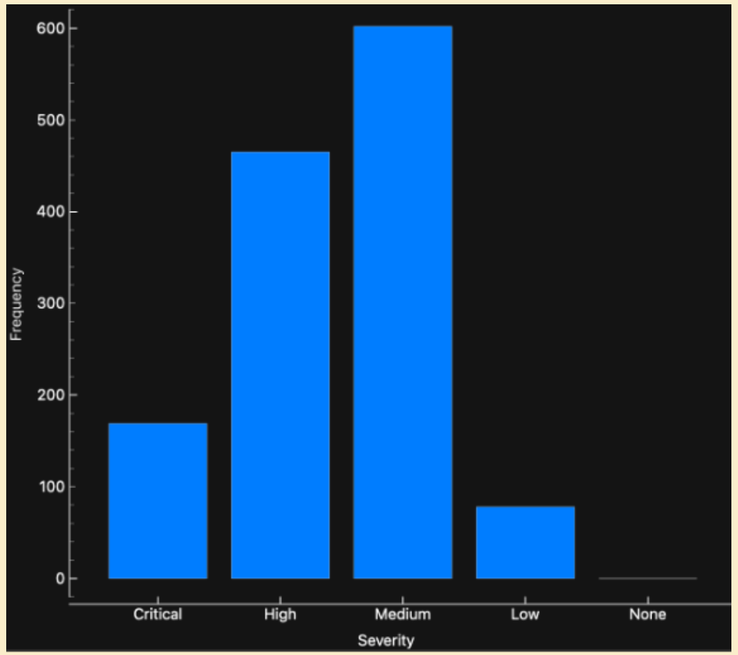
```
def python_script(in_data):
            return float(x)
        except Exception:
            return np.nan
    cvss_vals = np.array([safe_float(x) for x in cv], dtype=float)

# 3) Map CVSS -> Severity index (0..4)
sev_names = ["None", "Low", "Medium", "High", "Critical"]

def map_severity(s):
    if s != s or s is None:   # NaN check
        return 0  # "None"
    if s >= 9.0:  return 4   # "Critical"
    if s >= 7.0:  return 3   # "High"
    if s >= 4.0:  return 2   # "Medium"
    if s >  0.0:  return 1   # "Low"
    return 0

sev_idx = np.array([map_severity(float(x)) for x in cvss_vals],
dtype=float).reshape(-1, 1)

# 4) Build new domain with the new Discrete attribute (not class yet
    return out_data, out_learner, out_classifier, out_object
```

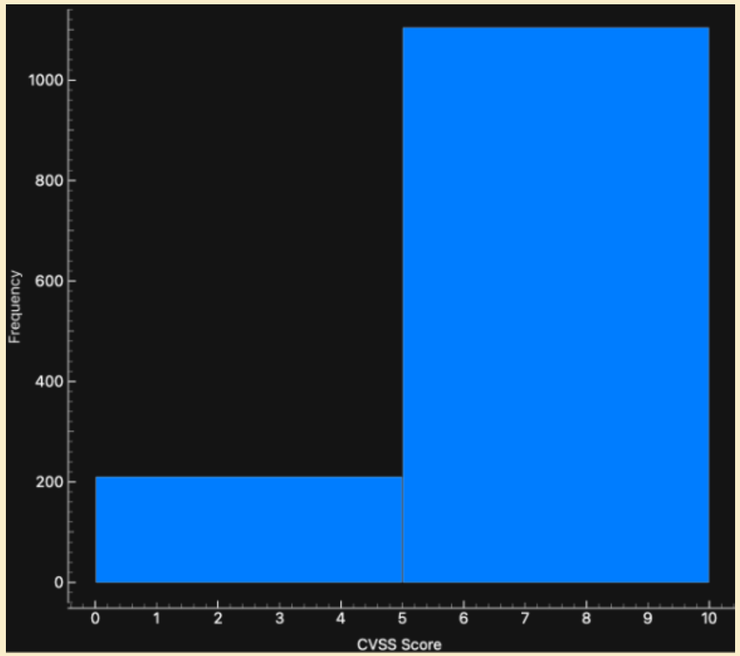
PHASE 2

EXPLORATORY DATA ANALYSIS

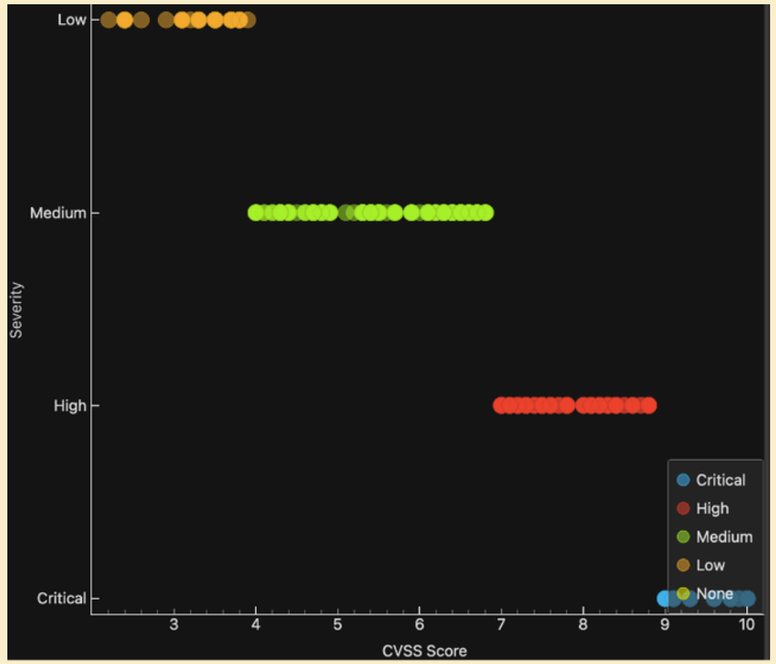
TO SHOW EARLY VISUAL RELATIONSHIPS AND DETECT OUTLIER

DISTRIBUTIONS (TO VISUALIZE DATA)

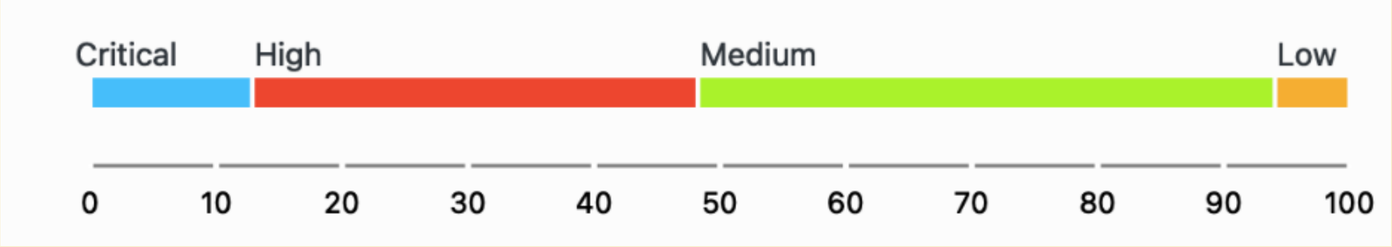




SCATTER PLOT



BOX PLOT (TO DETECT PATTERN OR OUTLIERS)



PHASE 3

EARLY MODEL TRAINING

RANDOM FOREST

Model	AUC	CA	F1	Prec	Recall	MCC
Random Forest	1.000	1.000	1.000	1.000	1.000	1.000

NEURAL NETWORK

Model	MSE	RMSE	MAE	MAPE	R2
Neural Network	0.386	0.621	0.476	8.229	0.878

CONFUSION MATRIX

Predicted

Actual	Critical	High	Medium	Low	None	Σ
Critical	494	0	0	0	0	494
High	0	1434	0	0	0	1434
Medium	0	0	1802	0	0	1802
Low	0	0	1	219	0	220
None	0	0	0	0	0	0
Σ	494	1434	1803	219	0	3950

CSP760

THANK YOU