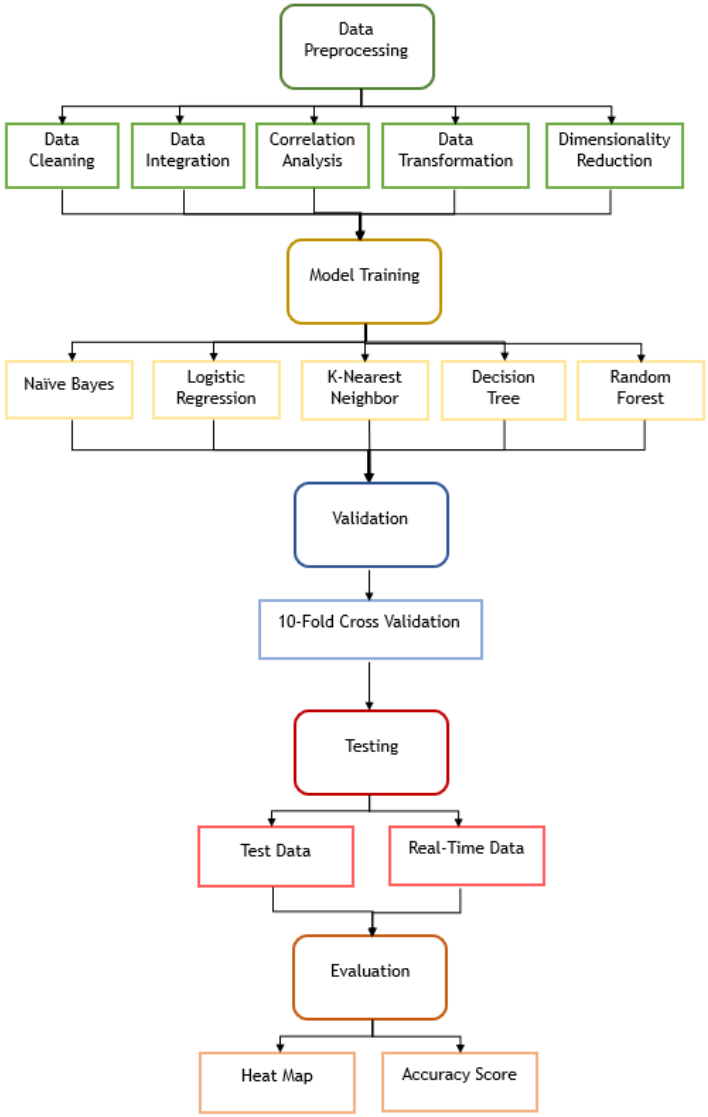
Smart SIEM – Machine Learning Based Detection

Contents

- Pipeline


Data Preprocessing
Data Cleaning
Data Integration
Correlation Analysis
Data Transformation
Dimensionality Reduction
Model Training
Naïve Bayes
Logistic Regression
K-Nearest Neighbor
Decision Tree
Random Forest
Validation
10-Fold Cross Validation
Testing
Test Data
Real-Time Data
Evaluation
Heat Map
Accuracy Score

Dataset Comparison

Dataset	NSL_KDD CIC Dataset	Intrusion Detection Evaluation Dataset (CIC-IDS2017) - Network Packet-Based Detection	Kaggle Dataset	CTU-13 Dataset - Network Flow-Based Detection	Iot-23 Dataset Small
Attacks Detected	• DoS • Probe • R2L • U2R	• DDoS • Probing • Brute Force Attacks • Web Attacks • Botnet Attack	DNS Tunnelling	• DDoS Zombie Attack • Spam Emails Attack • Password Attack • Distributed Brute Force Attacks • Data Exfiltration • Remote Attacks • Distributed Malware Propagation • Distributed Scanning • Click Fraud • Distributed Cryptocurrency Mining • Distributed Keylogging • DNS amplification attacks • Distributed web-based attacks	• Horizontal Port Scan • Okiru • DDoS • C&C • File Download

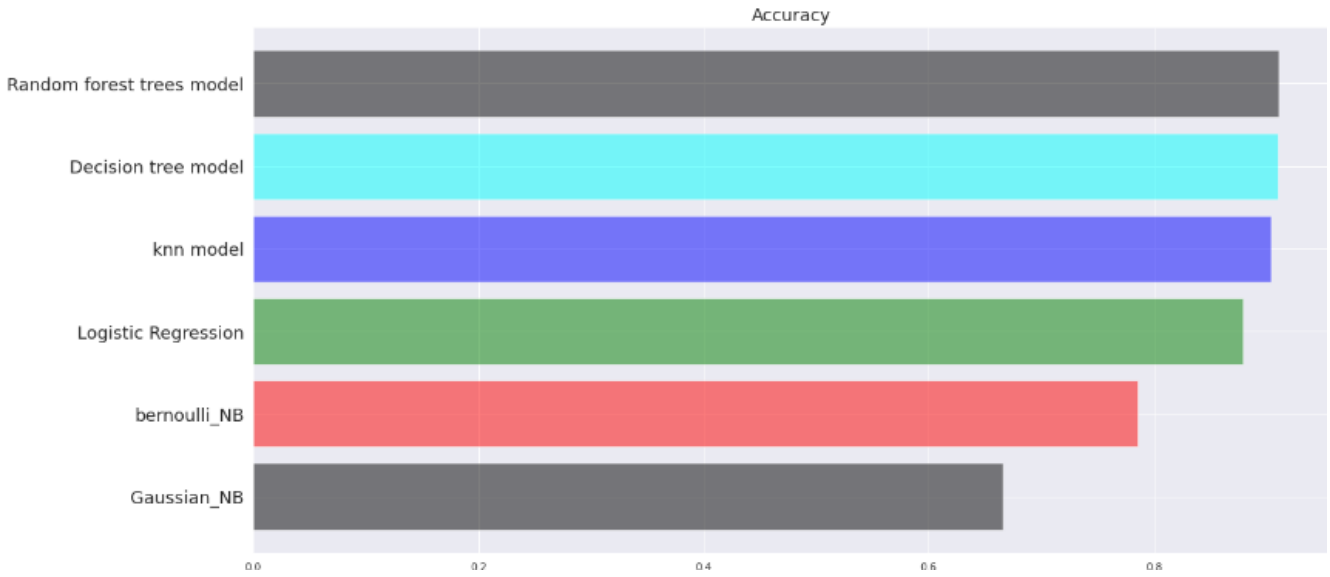
Dataset	NSL_KDD CIC Dataset	Intrusion Detection Evaluation Dataset (CIC-IDS 2017) - Network Packet-Based Detection	Kaggle Dataset	CTU-13 Dataset - Network Flow-Based Detection	Iot-23 Dataset Small
Total Entries	125,973	2,830,743	20,000	101,691	1,444,674
Number of Features (Except Label)	41	78	1	13	20
Number of Important Features	10	15	-	12	9
Decision Tree Criteria	Gini	Entropy	Entropy	Gini	Gini
Testing Accuracy	91%	99%	99.82%	99.9%	95.39%

Dataset selection

- Network Packets Based Dataset: **CIC-IDS2017** and **Kaggle DNS Tunnelling Dataset**
- Network Flows Based Dataset: **CTU-13 Dataset**

Model Selection - Virtual Environment

Model	Accuracies
Random Forest Trees	**91.09%**
Decision Trees (Entropy)	**91.01%**
K-Nearest Neighbours	90.37%
Logistic Regression	87.89%
Bernoulli Naïve Bayes	78.61%
Gaussian Naïve Bayes	66.56%



Model Selection - Real-Time Environment

Multilayer Perceptron (Neural Network)

```
from sklearn.neural_network import MLPClassifier
Neural_Net_model = MLPClassifier(hidden_layer_sizes=(10,), activation='relu',
                                 solver='adam', alpha=0.01, batch_size=1000,
                                 learning_rate='adaptive', learning_rate_init=0.01,
                                 max_iter=2)
```

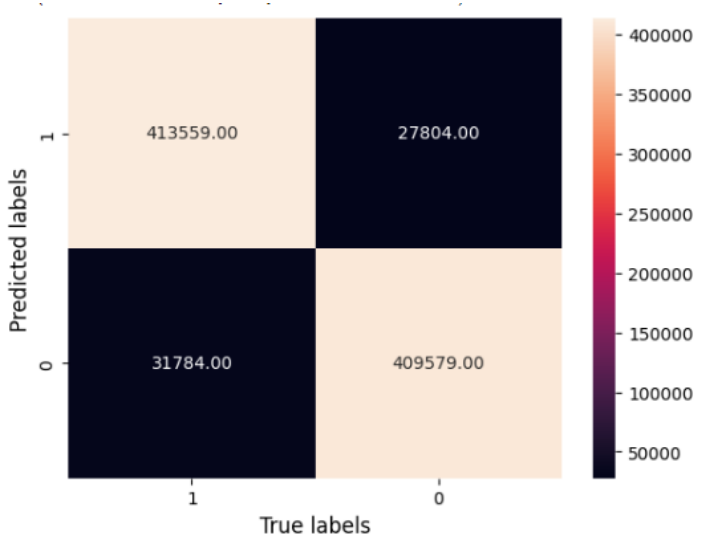
Hidden Layers = 10

Activation = ReLU

Batch Size = 1000

Learning Rate = 0.01

Accuracy: 93.25%

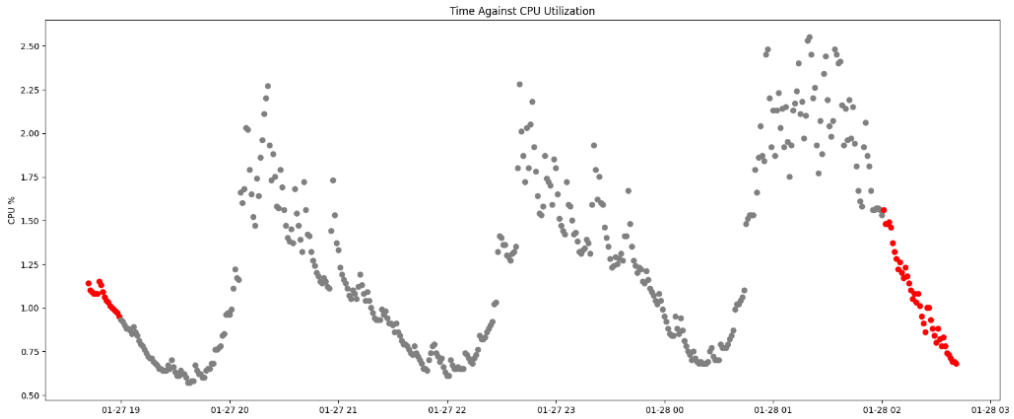


SPAM EMAILS FLOW

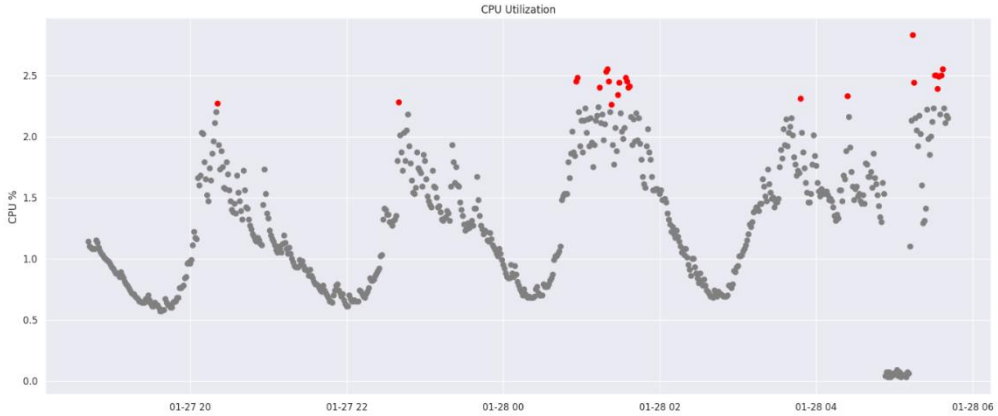
Organizational Routine Intelligence

Kaggle CPU Utilization Dataset

Heuristic Time Monitoring



Machine Learning Decision Tree Based Over CPU Utilization Detection, Accuracy: 95%

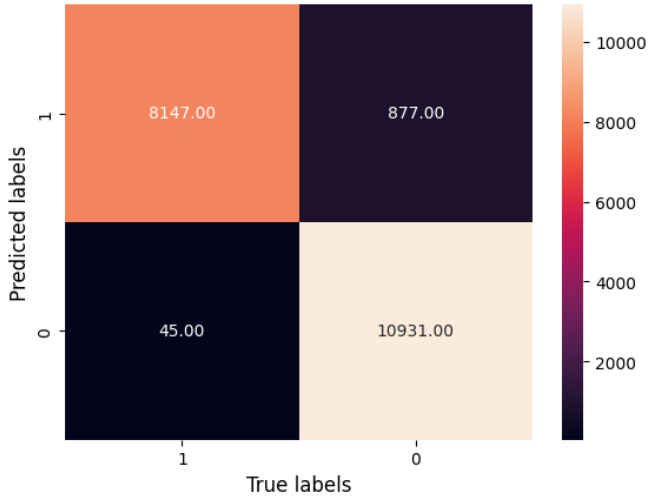


IoT-Based Smart Environment

Aposemat IoT-23 Dataset Small

	ts	uid	id.orig_h	id.orig_p	id.resp_h	id.resp_p	proto	service	duration	orig_bytes	...	conn_state	local_orig
0	1525879832.01624	CDe43c1PtgynajGl6	192.168.100.103	60905.0	131.174.215.147	23.0	tcp	-	2.998796	0	...	S0	-
1	1525879832.024985	CJaDcG3MZzvf1YVYl4	192.168.100.103	44301.0	91.42.47.63	23.0	tcp	-	-	-	...	S0	-
2	1525879832.044975	CMBrup3BLXivSp4Avc	192.168.100.103	50244.0	120.210.108.200	23.0	tcp	-	-	-	...	S0	-
3	1525879833.016171	CfHl9r3XMYtDQRrHnh	192.168.100.103	34243.0	147.7.65.203	49560.0	tcp	-	2.998804	0	...	S0	-
4	1525879833.044906	C7USrA15nFVkniMqC5	192.168.100.103	34840.0	145.164.35.6	21288.0	tcp	-	-	-	...	S0	-
...	...	...	...	...	...	...	...	...	...	...	...	...	...
99994	1532526102.004508	CMeH6R2aua5c5Dd65a	192.168.100.111	41762.0	221.182.209.127	23.0	tcp	-	-	-	...	S0	-
99995	1532526102.00451	CvqGx33hsXDpDVXa1i	192.168.100.111	58758.0	208.50.139.48	23.0	tcp	-	-	-	...	S0	-
99996	1532526102.004511	CC83RoUd9RLFuTL81	192.168.100.111	40400.0	40.95.136.51	23.0	tcp	-	-	-	...	S0	-
99997	1532526102.004752	C4ISld2cuSukEEuQtk	192.168.100.111	27117.0	122.37.183.236	23.0	tcp	-	-	-	...	S0	-
99998	1532526102.004756	C4U1azYmDx32faVY7	192.168.100.111	23227.0	189.62.234.179	23.0	tcp	-	-	-	...	S0	-

1444674 rows × 21 columns



- 23 Data Frames Concatenation
- Converting Categorical Features into Numerical Features
- Scaling Numerical Features
- Removing Missing Values
- Labelling Resulting Label as Benign and the Attack itself
- Finding 9 Important Features out of a Total of 20 Features using Random Forest Classifier
- Training Decision Tree Classifier with “Gini” criteria
- Predicting on Test Data and Calculating Cross-Validation Score

Accuracy: 95.39%