**CSE523 Machine Learning**

**Prof. Mehul Raval**

**Anomaly detection in computer networks to identify unusual activity or potential security threats**

**Week 1 Report**

| **Name** | **Enrolment Number** |
| --- | --- |
| Vatsal Patel | AU2040043 |
| Shreya Karia | AU2040076 |
| Digant Patel | AU2040086 |
| Samarth Chauhan | AU2040097 |

The Canadian Institute for Cybersecurity has provided the following preprocessed dataset. We also found the research paper associated with the dataset.

**Dataset:** <https://www.kaggle.com/datasets/cicdataset/cicids2017>

**Related Work:**

* D. Xu, Y. Wang, Y. Meng, and Z. Zhang, "An Improved Data Anomaly Detection Method Based on Isolation Forest," *2017 10th International Symposium on Computational Intelligence and Design (ISCID)*, Hangzhou, China, 2017, pp. 287-291, doi: 10.1109/ISCID.2017.202.
* Sharafaldin, I., Habibi Lashkari, A., & Ghorbani, A.A. (2018). Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization. *International Conference on Information Systems Security and Privacy*.

**Discussion about the dataset and its research paper:**

The paper "Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization" explains the process of how the dataset was created by the authors. The dataset was curated over a week and consists of a combination of malicious and benign traffic. This dataset was created to provide a more comprehensive and diverse representation of intrusion traffic than the other existing datasets.

The analysis of the dataset was done by various methods such as packet-size distribution analysis, traffic type analysis, packet rate analysis, etc.

For the packet size distribution analysis, it was found that intrusion traffic tends to have larger packet sizes compared to benign traffic. This is consistent with the observation that intrusions often involve the transfer of large amounts of data, such as malware.

The packet rate analysis revealed that intrusion traffic generates a higher rate of packets per second than benign traffic. This can be seen as an indicator of the increased activity and malicious intent of intrusions.

The traffic type analysis showed that most intrusions in the dataset were caused by network scanning, followed by access attempts to vulnerable services and denial-of-service attacks. This information can be useful for developing intrusion detection systems tailored to specific types of intrusions.

They have also listed a set of features specific to each type of intrusion detection.

**Discussion about Algorithms:**

We will be using python’s *numpy*, *pandas,* and *scikit-learn* libraries for algorithm implementation on Google Collab. From work already done, we found that kNN, Random forest algorithm, QDA, etc., can be used to detect intrusion. Moreover, we can use k-fold cross-validation techniques as well.

Next week we plan to first implement the kNN algorithm for data classification. kNN has a good prediction rate and F-measure. It uses the proximity of data for classification. We will use normalized parameters to implement the kNN algorithm. It is not time efficient, so we plan to move to other more optimized approaches, like QDA, RF, etc., later.