Phase-2 Submission Template – Data Analytics

Student Name: Aarthi.R
Register Number: 212923205001
Institution: St.Joseph college of engineering
Department: Information Technology
Date of Submission: 08.05.25
GitHub Repository Link: "https://github.com/aarthi2008/NM.git"

---

1. Problem Statement

This project focuses on identifying potential cyber threats by analyzing network traffic for **anomalous behavior**. In modern networks, a large volume of data flows continuously between devices, systems, and users. Most of this traffic follows normal patterns—but malicious actors (like hackers or malware) may cause **unusual or unexpected patterns** in the traffic.

This project aims to build a system that can **automatically detect those unusual patterns (anomalies)**, which might indicate activities such as:

- Unauthorized access
- Malware communication
- Data leaks
- Network scans
- DDoS attacks

Instead of relying solely on known attack signatures (as traditional intrusion detection systems do), this system will use **anomaly detection** techniques to discover new, unknown, or evolving threats.

2. Project Objectives

The goal of this project is to **develop an intelligent system that detects potential cyber threats by identifying anomalies in network traffic data using machine learning and statistical techniques**. The system will analyze network behavior, recognize patterns that deviate from the norm, and flag suspicious activities that may indicate intrusions, malware, or other cyber-attacks—enabling early threat detection and enhanced network security.

3. Flowchart of the Project Workflow

Network traffic data collection→ Data preprocessing→ Anomaly detection model→ Anomaly classifications → Threat Aleart → Reporting]

4. Data Description

[Provide an overview of the dataset(s) used.
Include:

- CICIDS 2017/18, UNSW-NB15 (Kaggle)
- Data type: Both structured and unstructured
- Number of rows and columns 2000 rows
- Dynamic dataset
- Src IP, Dst IP

5. Data Preprocessing

- Collection of data
- Data cleaning process and remove duplicates and outliers

- Feature extraction
- Nomalization and scaling
- Data labelling, splitting, handling
- Final data format

6. Exploratory Data Analysis (EDA)

[Detail the exploration performed to understand the data.
Include:

- **Univariate Analysis**: Distribution of single variables using plots
- **Bivariate/Multivariate Analysis**: Heatmaps, pairplots, grouped bars, etc.
- Analysis of key metrics or KPIs

7. Tools and Technologies Used

[Mention all tools used during the analysis.

- **Programming Language:** Python
- **Notebook/IDE:** Google Colab, Jupyter Notebook
- **Libraries:** pandas, numpy, matplotlib, seaborn, plotly,tcdump

- **Optional Automation Tools:** pandas-profiling]

8. Team Members and Contributions

Name	Contribution
[S.keshavarthini]	Data cleaning
[M.Nandhini]	Data collection, Insights
[R. Aarthi]	Flowchart, documentation