

# BDA

## Real-time anomaly detection in IoT networks

Leveraging Hadoop-Spark for Scalable Detection

Sakshi Agrawal MSE2023004

Prachee Singh MSE2023006

# PROBLEM STATEMENT

The exponential growth of IoT devices necessitates efficient and scalable real-time anomaly detection systems to ensure network security. This project aims to develop such a system using the Hadoop-Kafka framework, capable of flagging anomalous and non-anomalous events in real-time.

# OBJECTIVES

## **Goal # 1**

Develop a Scalable Anomaly Detection Model.

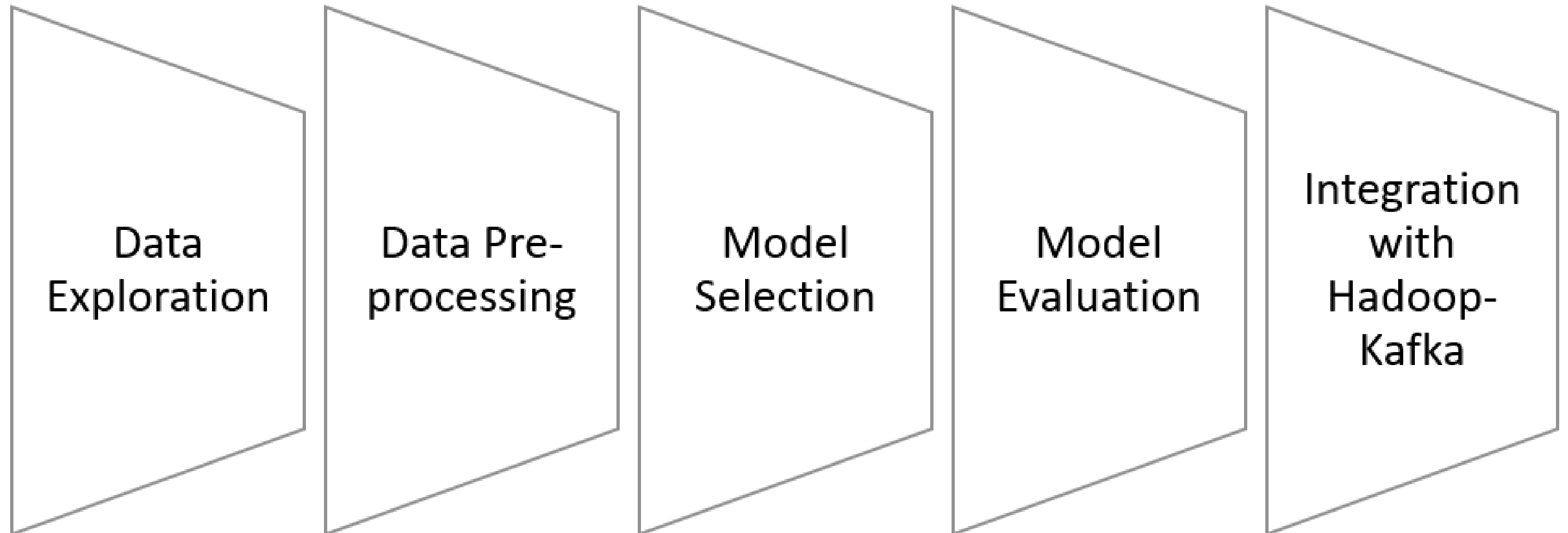
## **Goal # 2**

Compare Machine Learning Models

## **Goal # 3**

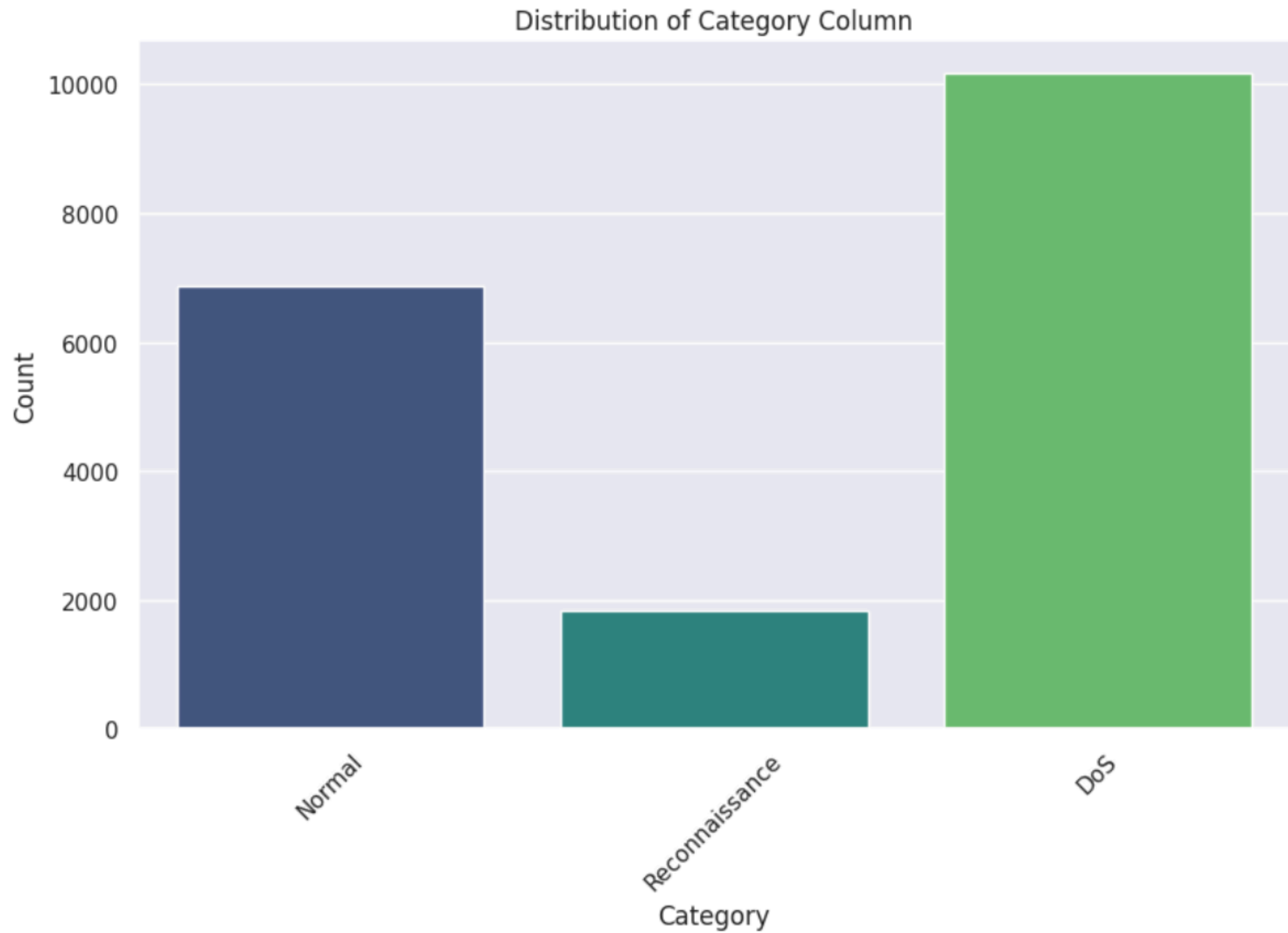
Monitor Real-time Detection

# Methodology and Workflow



# DATASET

	pkSeqID	stime	flgs	proto	saddr	sport	daddr	dport	pkts	bytes	...	spkts	dpkts	sbytes	dbytes	rate	srate	drate	attack	category
0	1	1.526344e+09	e	arp	192.168.100.1	NaN	192.168.100.3	NaN	4	240	...	2	2	120	120	0.002508	0.000836	0.000836	0	Normal
1	2	1.526344e+09	e	tcp	192.168.100.7	139	192.168.100.4	36390	10	680	...	5	5	350	330	0.006190	0.002751	0.002751	0	Normal
2	3	1.526344e+09	e	udp	192.168.100.149	51838	27.124.125.250	123	2	180	...	1	1	90	90	20.590960	0.000000	0.000000	0	Normal
3	4	1.526344e+09	e	arp	192.168.100.4	NaN	192.168.100.7	NaN	10	510	...	5	5	210	300	0.006189	0.002751	0.002751	0	Normal
4	5	1.526344e+09	e	udp	192.168.100.27	58999	192.168.100.1	53	4	630	...	2	2	174	456	0.005264	0.001755	0.001755	0	Normal



# Data Exploration and Preprocessing

## **Description:**

Dataset sourced from Kaggle.

Contains simulated network traffic for normal and malicious activities.

## **Key Features:**

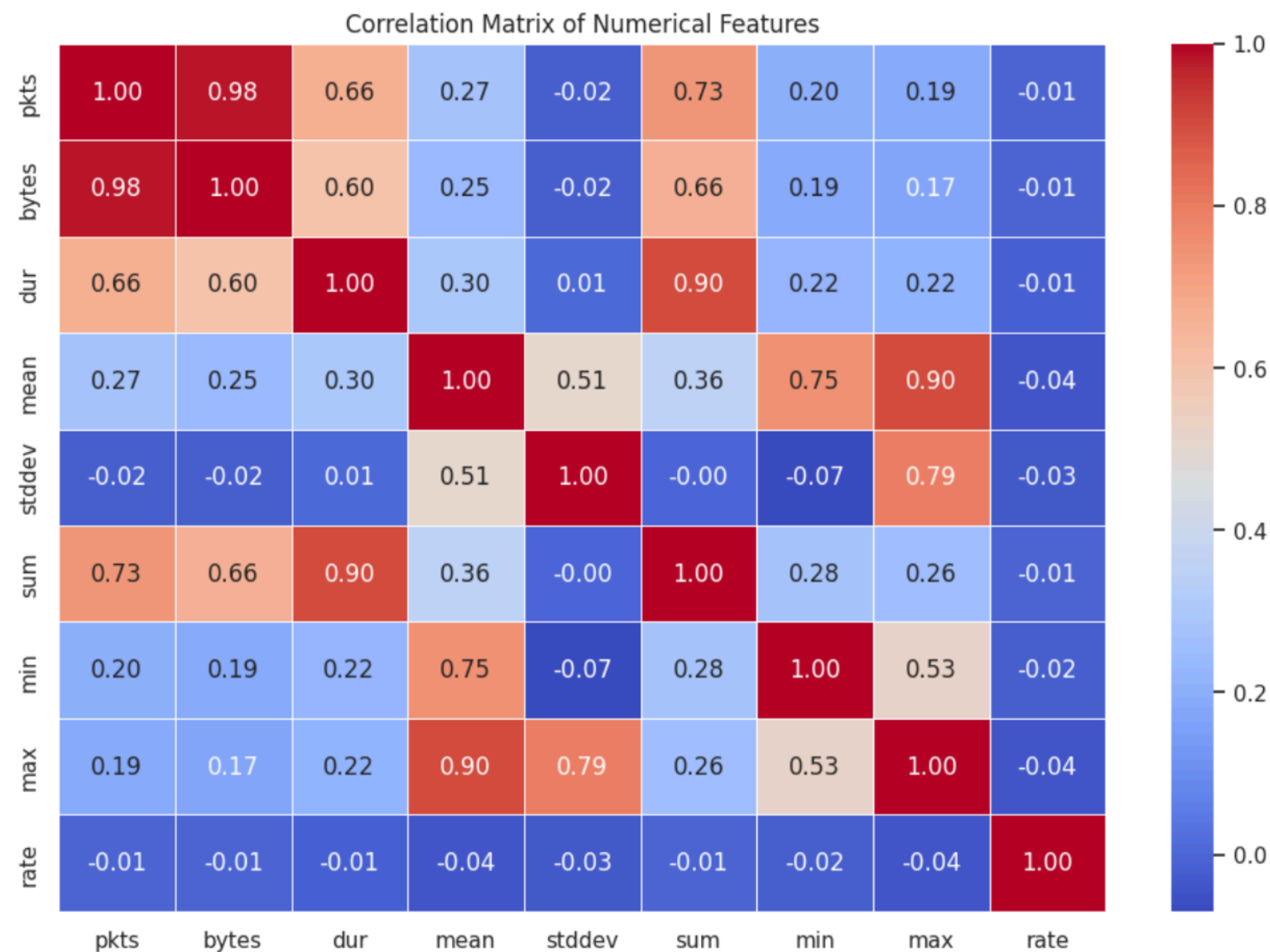
Covers multiple attack types like DoS and Reconnaissance

Over 70 million records in CSV format.

## **Challenges:**

Large dataset size, requiring efficient handling.

# Correlation Matrix of the features





# Algorithm Selection

## **Logistic Regression:**

- A simple and effective binary classifier.

## **Random Forest:**

- Ensemble learning method using decision trees.
- Good for handling large datasets.

## **XGBoost:**

- Gradient boosting algorithm.
- Optimized for performance and accuracy, especially on imbalanced data.

# Model Training and Evaluation

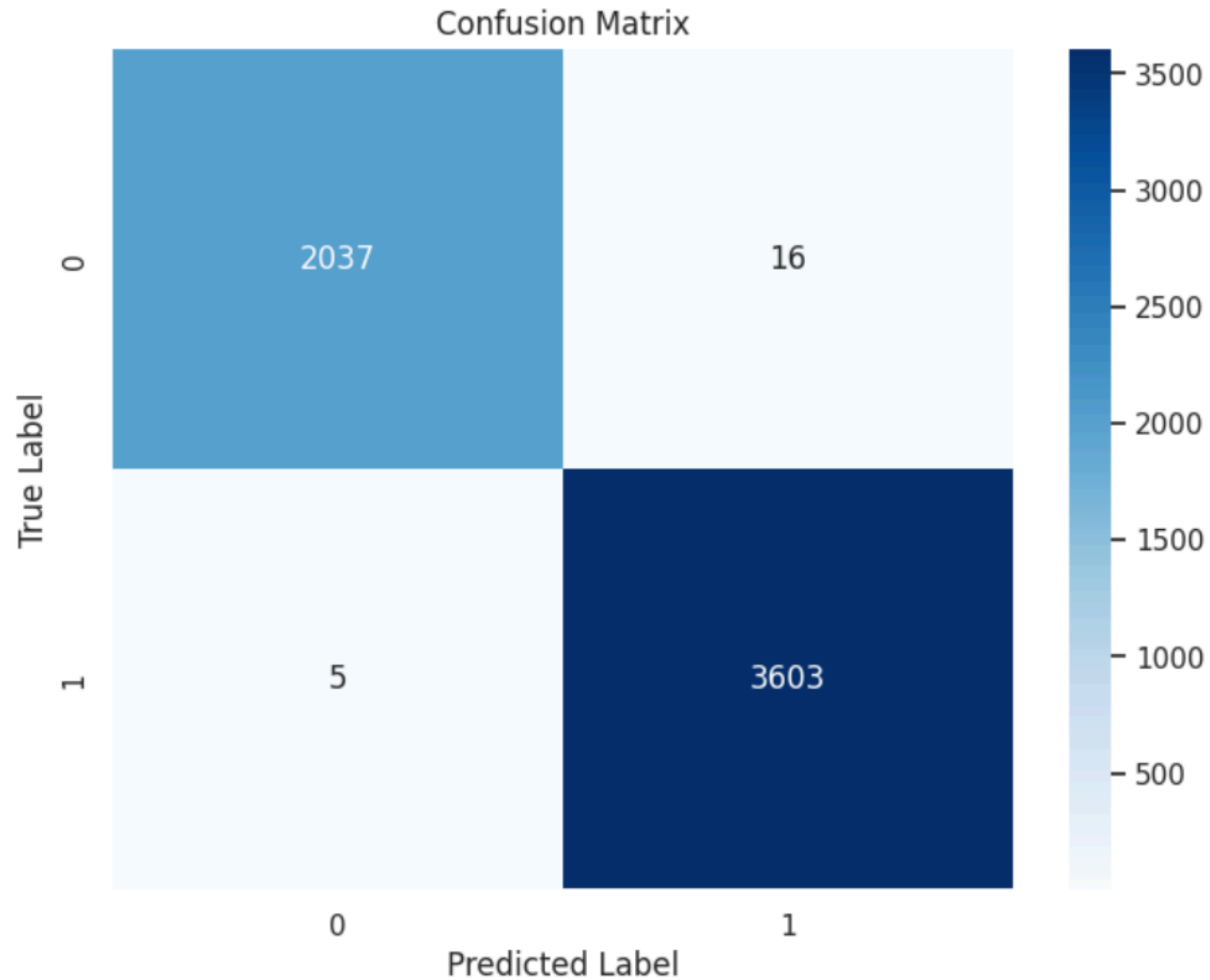
## Training Setup:

- Split dataset into training and testing sets.
- Cross-validation for model tuning.

## Evaluation Metrics:

- Accuracy, Precision, Recall, F1-Score.
- Confusion Matrix to visualize classification results.

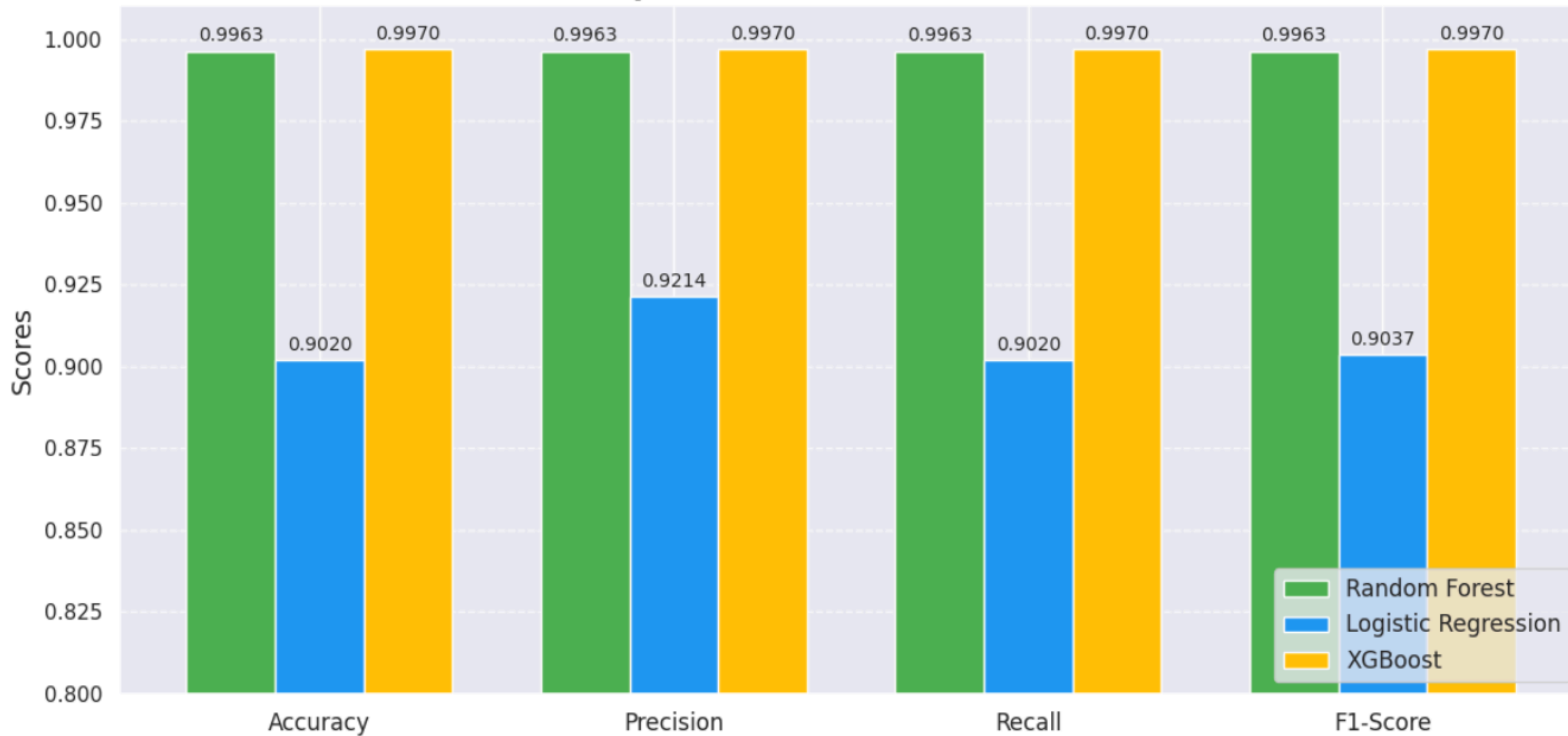
# Random Forest Confusion Matrix



# Models Comparison

Metric	Random Forest	Logistic Regression	XGBoost
Accuracy	0.9963	0.9020	0.9970
Precision	0.9963	0.9214	0.9970
Recall	0.9963	0.9020	0.9970
F1-Score	0.9963	0.9037	0.9970

**Comparison of Model Performance**



# Integration with Hadoop-Kafka

## Architecture:

- Kafka for data pipelining and batch processing.
- HDFS for data storage.

## Workflow:

- Ingest data using Kafka Producer.
- Process it using kafka Consumer and apply the trained Random Forest model.
- Real-time anomaly detection and alert generation.

# Hadoop Data Storage

The screenshot displays the Hadoop Distributed File System (HDFS) Explorer interface in a web browser. The browser's address bar shows the URL `localhost:9870/explorer.html#/Bot_lot_dataSet`. The interface features a green navigation bar at the top with the following menu items: Hadoop, Overview, Datanodes, Datanode Volume Failures, Snapshot, Startup Progress, and Utilities. Below the navigation bar, the main heading is "Browse Directory". A search bar at the top of the content area contains the path `/Bot_lot_dataSet` and a "Go!" button. To the right of the search bar are icons for file operations. Below the search bar, a dropdown menu indicates "Show 25 entries". A search input field is also present. The main content area displays a table of files and directories. The table has the following columns: Permission, Owner, Group, Size, Last Modified, Replication, Block Size, and Name. The first row shows a directory named `Bot_lot_dataSet` with permissions `drwxrwxrwx`, owner `Wanted`, group `supergroup`, size `0 B`, last modified `Nov 15 00:34`, replication `0`, and block size `0 B`. The subsequent rows list 15 CSV files, each with permissions `-rw-r--r--`, owner `hadoop`, group `supergroup`, size `617 B` or `577 B`, last modified `Nov 18 03:34`, replication `1`, and block size `128 MB`. The files are named `anomaly_results_20241118_033408.csv` through `anomaly_results_20241118_033422.csv`. Each row includes a checkbox on the left and a trash icon on the right. At the bottom right of the interface, there is a "Show desktop" button. The Windows taskbar at the bottom of the screen shows the system clock as 03:39 on 18-11-2024, along with various system icons and application shortcuts.

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxrwxrwx	Wanted	supergroup	0 B	Nov 15 00:34	0	0 B	Bot_lot_dataSet
-rw-r--r--	hadoop	supergroup	617 B	Nov 18 03:34	1	128 MB	anomaly_results_20241118_033408.csv
-rw-r--r--	hadoop	supergroup	577 B	Nov 18 03:34	1	128 MB	anomaly_results_20241118_033410.csv
-rw-r--r--	hadoop	supergroup	597 B	Nov 18 03:34	1	128 MB	anomaly_results_20241118_033411.csv
-rw-r--r--	hadoop	supergroup	607 B	Nov 18 03:34	1	128 MB	anomaly_results_20241118_033412.csv
-rw-r--r--	hadoop	supergroup	597 B	Nov 18 03:34	1	128 MB	anomaly_results_20241118_033413.csv
-rw-r--r--	hadoop	supergroup	617 B	Nov 18 03:34	1	128 MB	anomaly_results_20241118_033414.csv
-rw-r--r--	hadoop	supergroup	587 B	Nov 18 03:34	1	128 MB	anomaly_results_20241118_033415.csv
-rw-r--r--	hadoop	supergroup	607 B	Nov 18 03:34	1	128 MB	anomaly_results_20241118_033416.csv
-rw-r--r--	hadoop	supergroup	617 B	Nov 18 03:34	1	128 MB	anomaly_results_20241118_033417.csv
-rw-r--r--	hadoop	supergroup	617 B	Nov 18 03:34	1	128 MB	anomaly_results_20241118_033418.csv
-rw-r--r--	hadoop	supergroup	577 B	Nov 18 03:34	1	128 MB	anomaly_results_20241118_033419.csv
-rw-r--r--	hadoop	supergroup	617 B	Nov 18 03:34	1	128 MB	anomaly_results_20241118_033420.csv
-rw-r--r--	hadoop	supergroup	587 B	Nov 18 03:34	1	128 MB	anomaly_results_20241118_033421.csv
-rw-r--r--	hadoop	supergroup	607 B	Nov 18 03:34	1	128 MB	anomaly_results_20241118_033422.csv

# Flask app API

A screenshot of a web browser window displaying a JSON array of network traffic data. The browser's address bar shows the URL '127.0.0.1:5000/data'. The JSON data is as follows:

```
[
  {
    "bytes": 70,
    "dur": 0,
    "flgs": 0,
    "max": 0,
    "mean": 0,
    "min": 0,
    "pkts": 1,
    "prediction": "Attack",
    "proto": 0,
    "rate": 0,
    "stddev": 0,
    "sum": 0
  },
  {
    "bytes": 60,
    "dur": 0.000047,
    "flgs": 8,
    "max": 0.000047,
    "mean": 0.000047,
    "min": 0.000047,
    "pkts": 1,
    "prediction": "Attack",
    "proto": 3,
    "rate": 0,
    "stddev": 0,
    "sum": 0.000047
  },
  {
    "bytes": 60,
    "dur": 0.000092,
    "flgs": 8,
    "max": 0.000092,
    "mean": 0.000092,
    "min": 0.000092,
    "pkts": 1,
    "prediction": "Attack",
    "proto": 3,
    "rate": 0,
    "stddev": 0,
    "sum": 0.000092
  },
  {
    "bytes": 70,
    "dur": 0,
    "flgs": 0,
    "max": 0,
    "mean": 0,
    "min": 0,
    "pkts": 1,
    "prediction": "Attack",
    "proto": 0,
    "rate": 0,
    "stddev": 0,
    "sum": 0
  },
  {
    "bytes": 60,
    "dur": 0.000047,
    "flgs": 8,
    "max": 0.000047,
    "mean": 0.000047,
    "min": 0.000047,
    "pkts": 1,
    "prediction": "Attack",
    "proto": 3,
    "rate": 0,
    "stddev": 0,
    "sum": 0.000047
  }
]
```

The browser's taskbar at the bottom shows various application icons and the system clock indicating 03:57 on 18-11-2024.



# Real-time Dashboard



Real-time IoT Anomaly Detection Dashboard

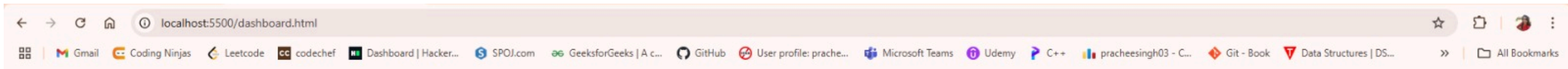
Sensor ID	Flgs	Proto	Pkts	Bytes	Dur	Mean	Stddev	Sum	Min	Max	Rate	Prediction
2566	0	2	15	1094	0.489337	0.489337	0	0.489337	0.489337	0.489337	28.61014	Attack
2567	0	2	11	1753	0.383392	0.383392	0	0.383392	0.383392	0.383392	26.082964	Attack
2568	0	2	11	1263	0.386529	0.386529	0	0.386529	0.386529	0.386529	25.871281	Attack
2569	0	2	11	929	0.374457	0.374457	0	0.374457	0.374457	0.374457	26.705336	Attack
2570	0	2	6	466	0.052456	0.052456	0	0.052456	0.052456	0.052456	95.317986	Normal
2571	0	2	17	3521	0.1876	0.1876	0	0.1876	0.1876	0.1876	85.287849	Attack
2572	0	2	11	1398	0.366254	0.366254	0	0.366254	0.366254	0.366254	27.303455	Attack
2573	0	2	2	134	0.00016	0.00016	0	0.00016	0.00016	0.00016	6250	Attack
2574	0	2	6	432	0.173422	0.173422	0	0.173422	0.173422	0.173422	28.831406	Attack
2575	0	2	14	3425	5.778091	0.176651	0.176501	0.353303	0.00015	0.353153	2.249878	Attack
2576	0	2	12	864	0.313353	0.313353	0	0.313353	0.313353	0.313353	35.104179	Attack
2577	0	2	11	1571	0.365289	0.365289	0	0.365289	0.365289	0.365289	27.375586	Attack
2578	0	2	13	1229	0.313946	0.313946	0	0.313946	0.313946	0.313946	38.223133	Attack
2579	0	2	11	4457	2.942468	2.942468	0	2.942468	2.942468	2.942468	3.398508	Attack
2580	0	2	13	1113	0.242732	0.242732	0	0.242732	0.242732	0.242732	49.437241	Attack

# Dashboard..

The image is a screenshot of a web browser window displaying a dashboard. The browser's address bar at the top shows the URL 'localhost:5500/dashboard.html'. Below the address bar, there is a row of bookmarks including 'Gmail', 'Coding Ninjas', 'Leetcode', 'codechef', 'Dashboard | Hacker...', 'SPOJ.com', 'GeeksforGeeks | A c...', 'GitHub', 'User profile: prache...', 'Microsoft Teams', 'Udemy', 'C++', 'pracheesingh03 - C...', 'Git - Book', 'Data Structures | DS...', and 'All Bookmarks'. The main content area of the browser has a title 'Real-time IoT Anomaly Detection Dashboard' in bold black text. Below the title, the text 'Loading data...' is displayed. Underneath this, there is a table with 12 columns. The column headers are: 'Sensor ID', 'Figs', 'Proto', 'Pkts', 'Bytes', 'Dur', 'Mean', 'Stddev', 'Sum', 'Min', 'Max', 'Rate', and 'Prediction'. The table body is currently empty. At the bottom of the image, the Windows taskbar is visible, showing the system tray with a weather widget (72°F, Smoke), a search bar, and various application icons. The system clock in the bottom right corner shows the date '18-11-2024' and the time '03:44'.



# Dashboard..



Real-time IoT Anomaly Detection Dashboard

Sensor ID ↓	Flgs	Proto	Pkts	Bytes	Dur	Mean	Stddev	Sum	Min	Max	Rate	Prediction
2410	0	2	2	120	0.000064	0.000064	0	0.000064	0.000064	0.000064	15625	Attack
2409	0	2	2	120	0.000022	0.000022	0	0.000022	0.000022	0.000022	45454.546875	Attack
2408	0	2	2	120	0.000095	0.000095	0	0.000095	0.000095	0.000095	10526.31543	Attack
2407	0	2	2	120	0.000074	0.000074	0	0.000074	0.000074	0.000074	13513.512695	Attack
2406	0	2	2	120	0.000021	0.000021	0	0.000021	0.000021	0.000021	47619.046875	Attack
2405	0	2	2	120	0.000094	0.000094	0	0.000094	0.000094	0.000094	10638.297852	Attack
2404	0	2	2	120	0.000081	0.000081	0	0.000081	0.000081	0.000081	12345.679688	Attack
2403	0	2	2	120	0.000025	0.000025	0	0.000025	0.000025	0.000025	40000	Attack
2402	0	2	2	120	0.000071	0.000071	0	0.000071	0.000071	0.000071	14084.506836	Attack
2401	0	2	2	120	0.000079	0.000079	0	0.000079	0.000079	0.000079	12658.228516	Attack
2400	0	2	2	120	0.000025	0.000025	0	0.000025	0.000025	0.000025	40000	Attack
2399	0	2	2	120	0.000055	0.000055	0	0.000055	0.000055	0.000055	18181.818359	Attack
2398	0	1	1	70	0	0	0	0	0	0	0	Normal
2397	0	2	2	120	0.000042	0.000042	0	0.000042	0.000042	0.000042	23809.523438	Attack
2396	0	2	2	120	0.000077	0.000077	0	0.000077	0.000077	0.000077	12987.013672	Attack

# Future Scope

Optimize the model for other IoT datasets.

Implement deeper anomaly detection using neural networks.

Deploy the system on cloud or edge environments for scalability.

**THANK YOU**