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.

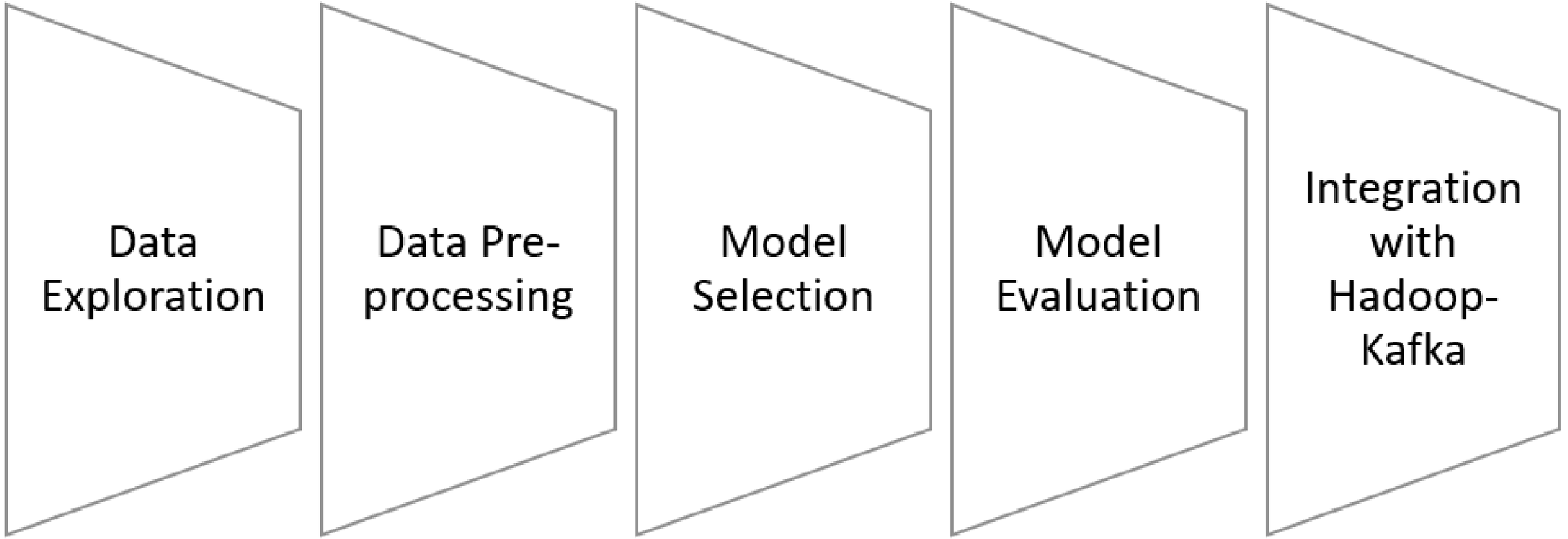
Goal # 2

Compare Machine Learning Models

Goal # 3

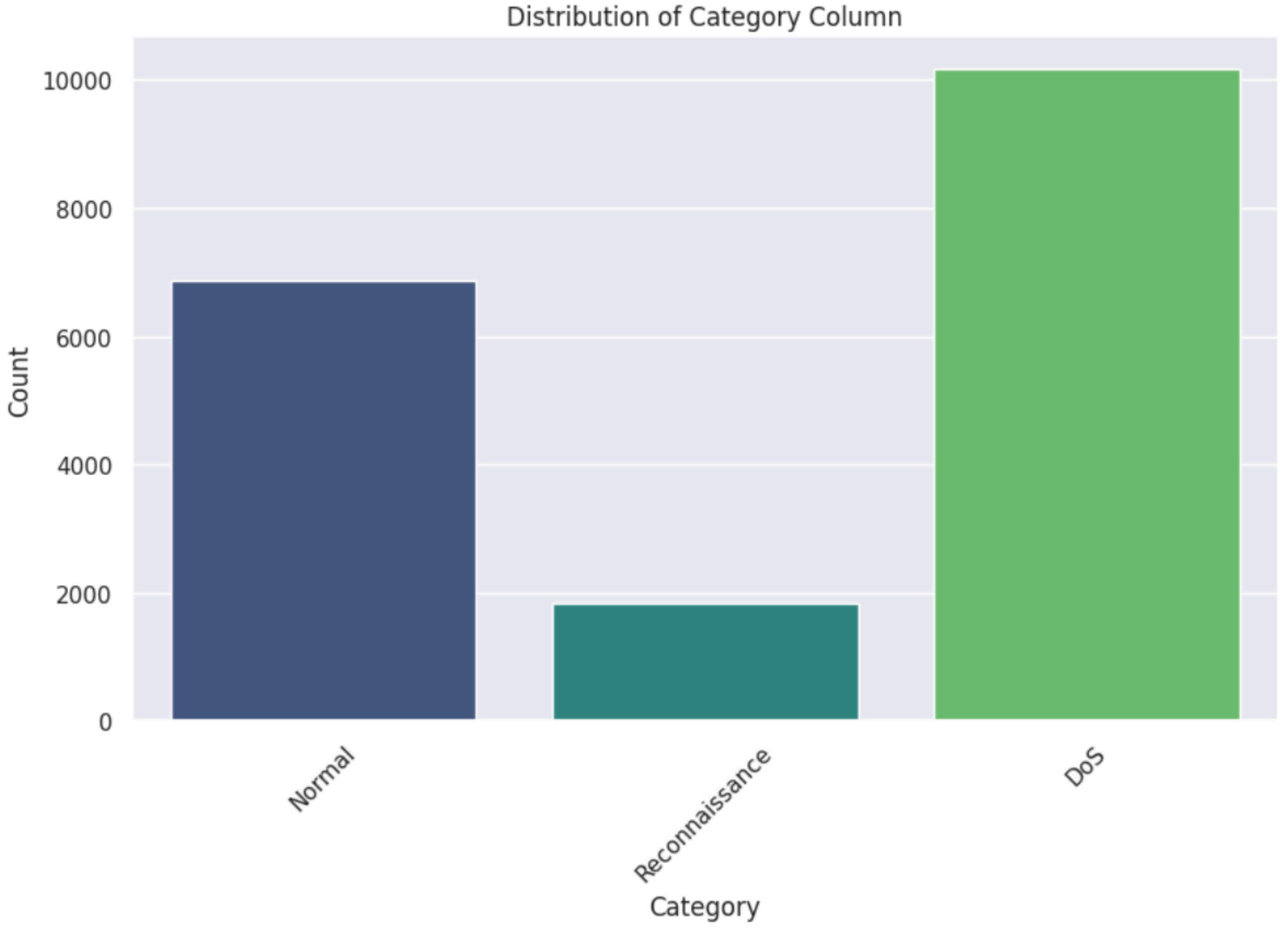
Monitor Real-time Detection

Methodolgy 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


Distribution of Category Column
Count
10000
8000
6000
4000
2000
0
Normal
Reconnaissance
DoS
Category

Data Exploration and Preprocessing

Description:

Dataset sourced from Kaggle.

Contains simulated network traffic for normal and malicious activities.

Key Features:

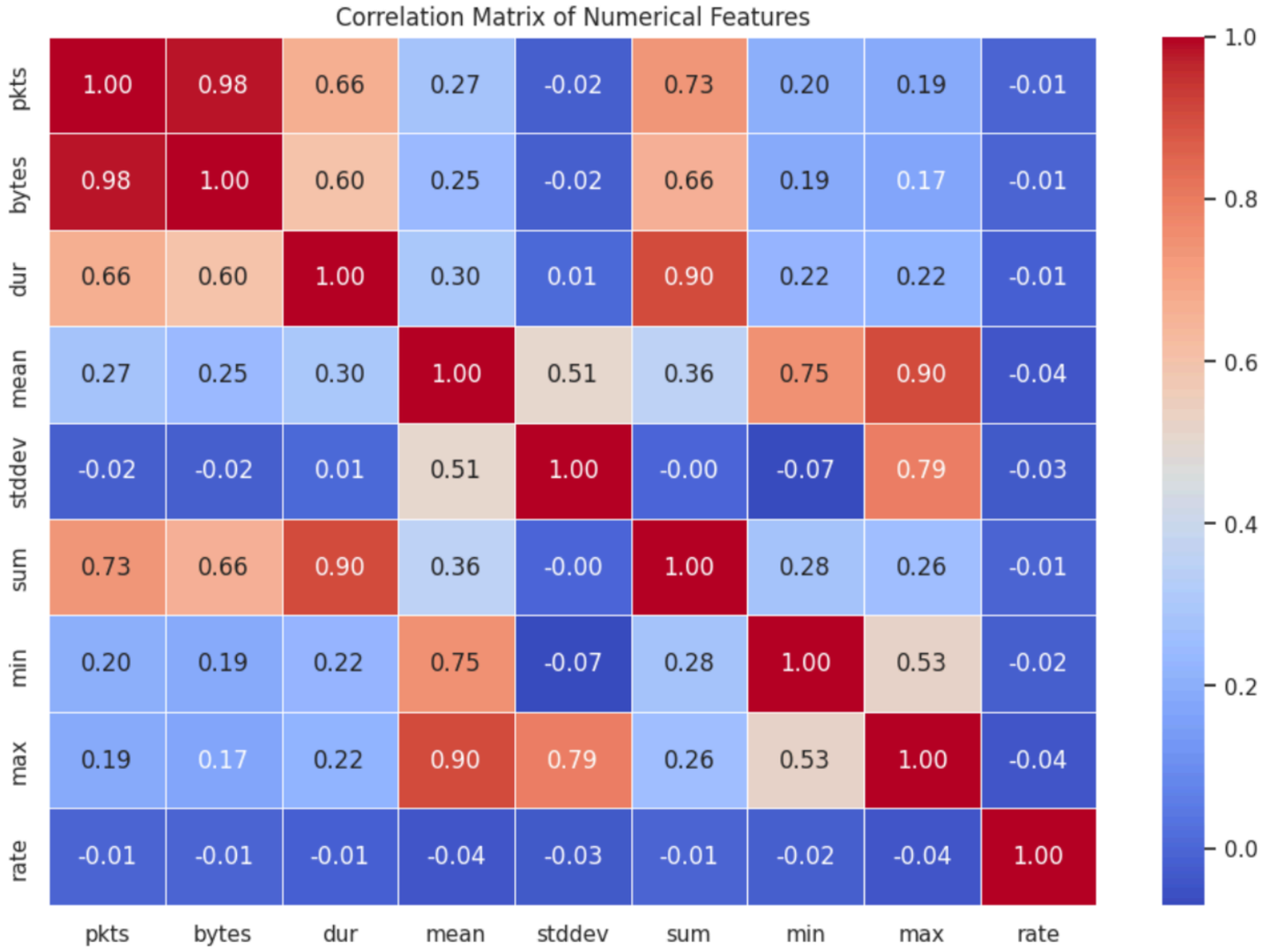
Covers multiple attack types like DoS and Reconnaissance

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

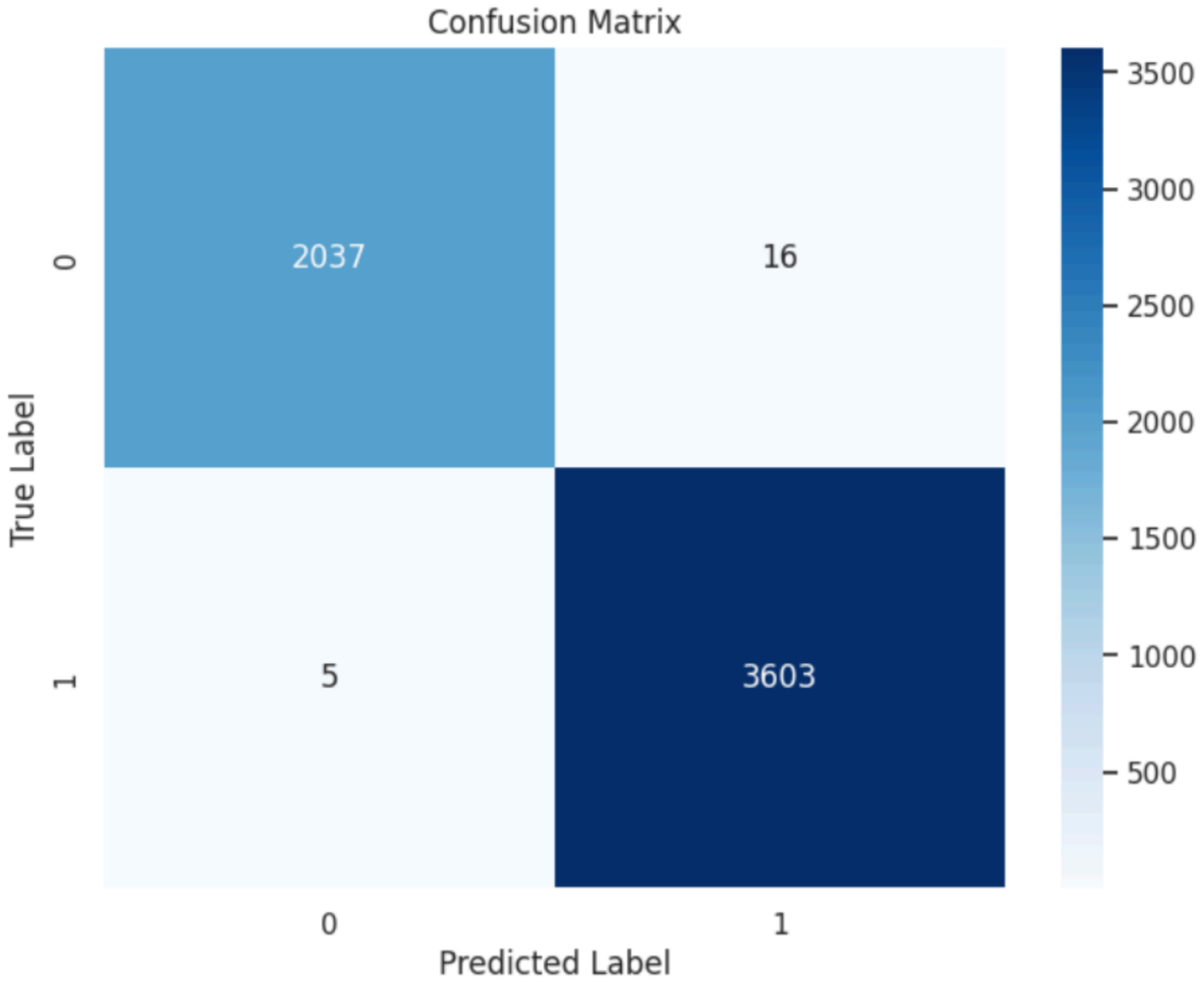
Training Setup:

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

Evaluation Metrics:

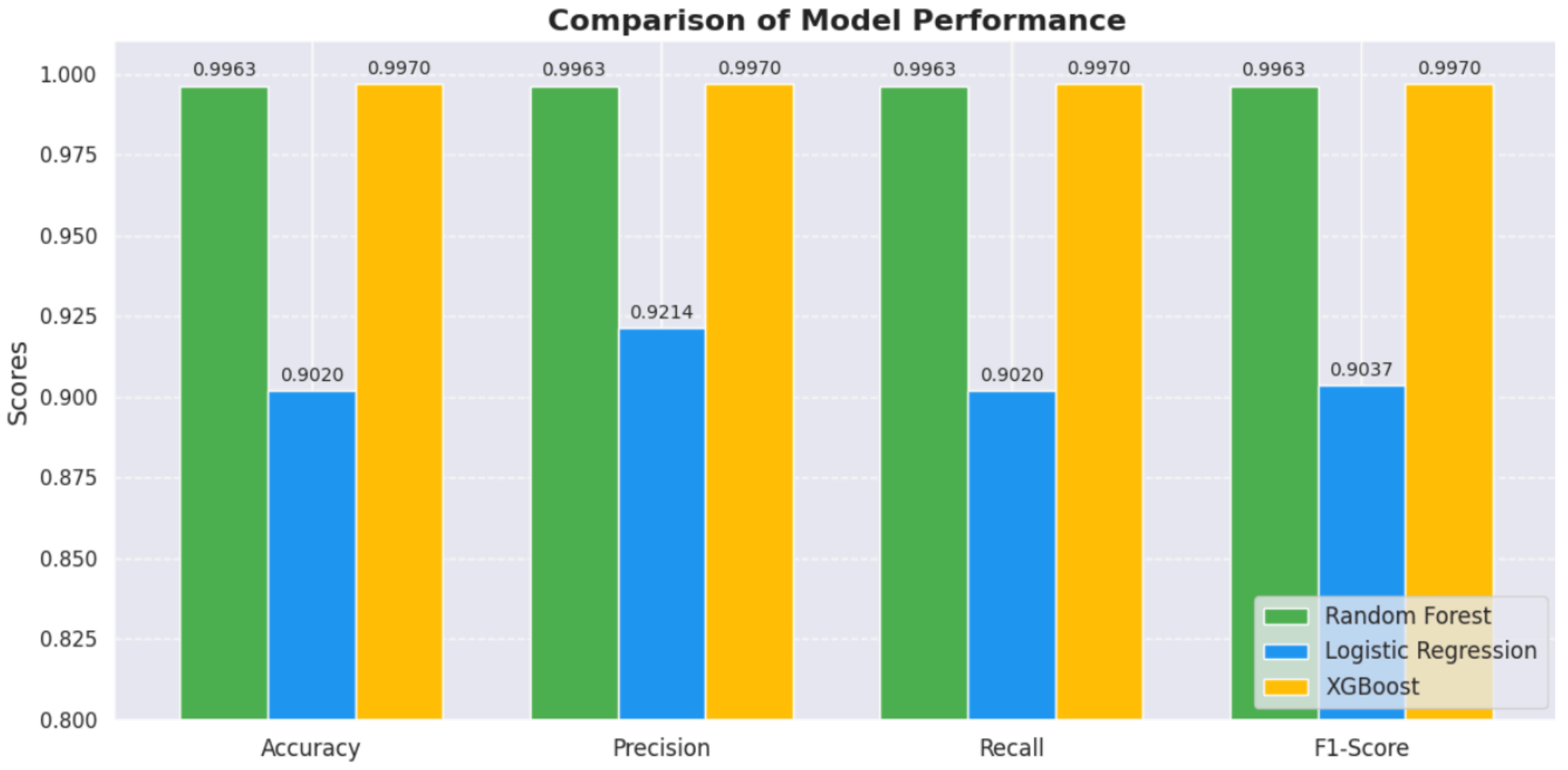
- Accuracy, Precision, Recall, F1-Score.
- Confusion Matrix to visualize classification results.

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
Scores
1.000
0.975
0.950
0.925
0.900
0.875
0.850
0.825
0.800
0.9963
0.9020
0.9970
0.9963
0.9214
0.9970
0.9963
0.9020
0.9970
0.9963
0.9037
0.9970
Accuracy
Precision
Recall
F1-Score
Random Forest
Logistic Regression
XGBoost

Integration with Hadoop-Kafka

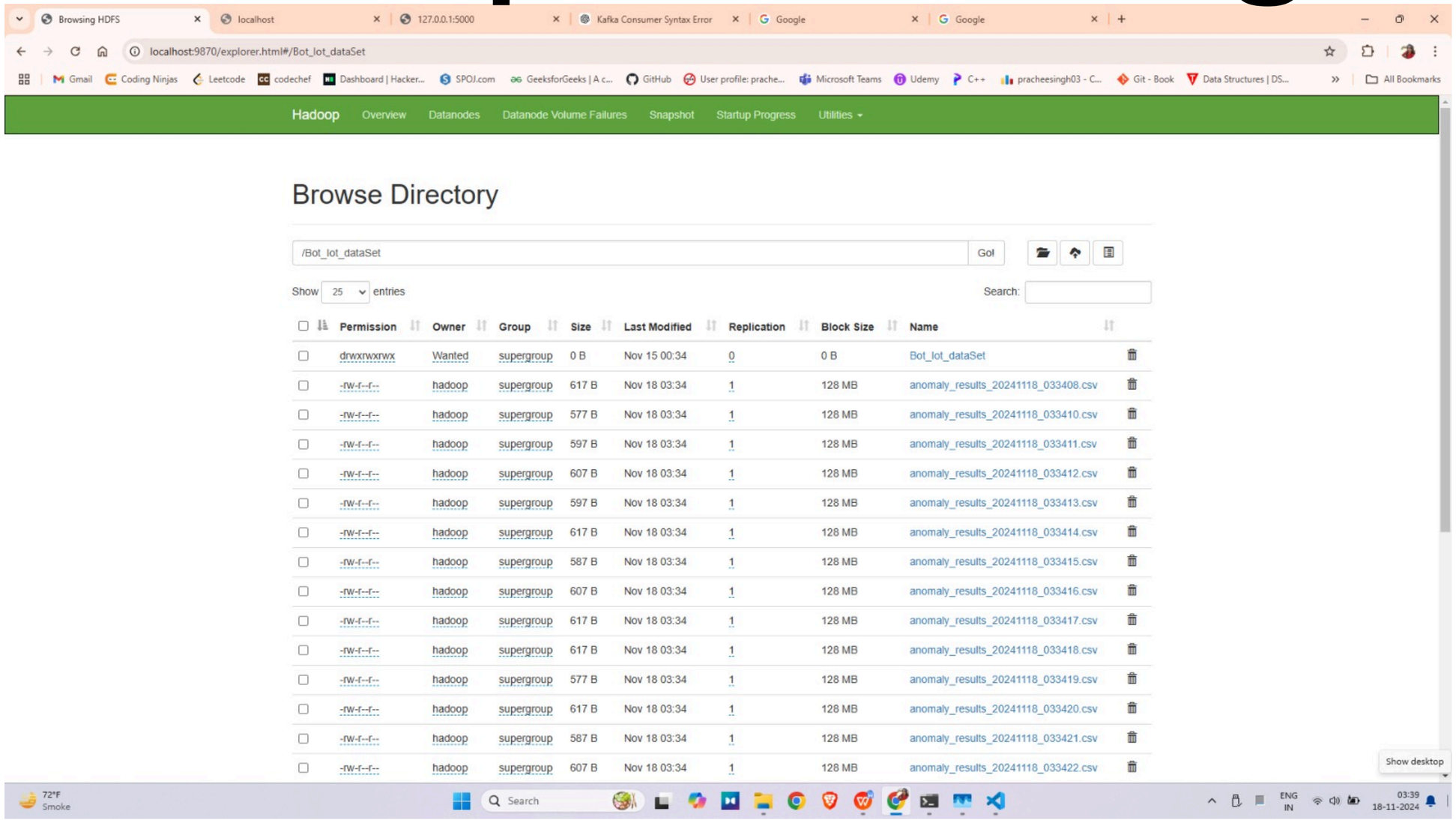
Architecture:

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

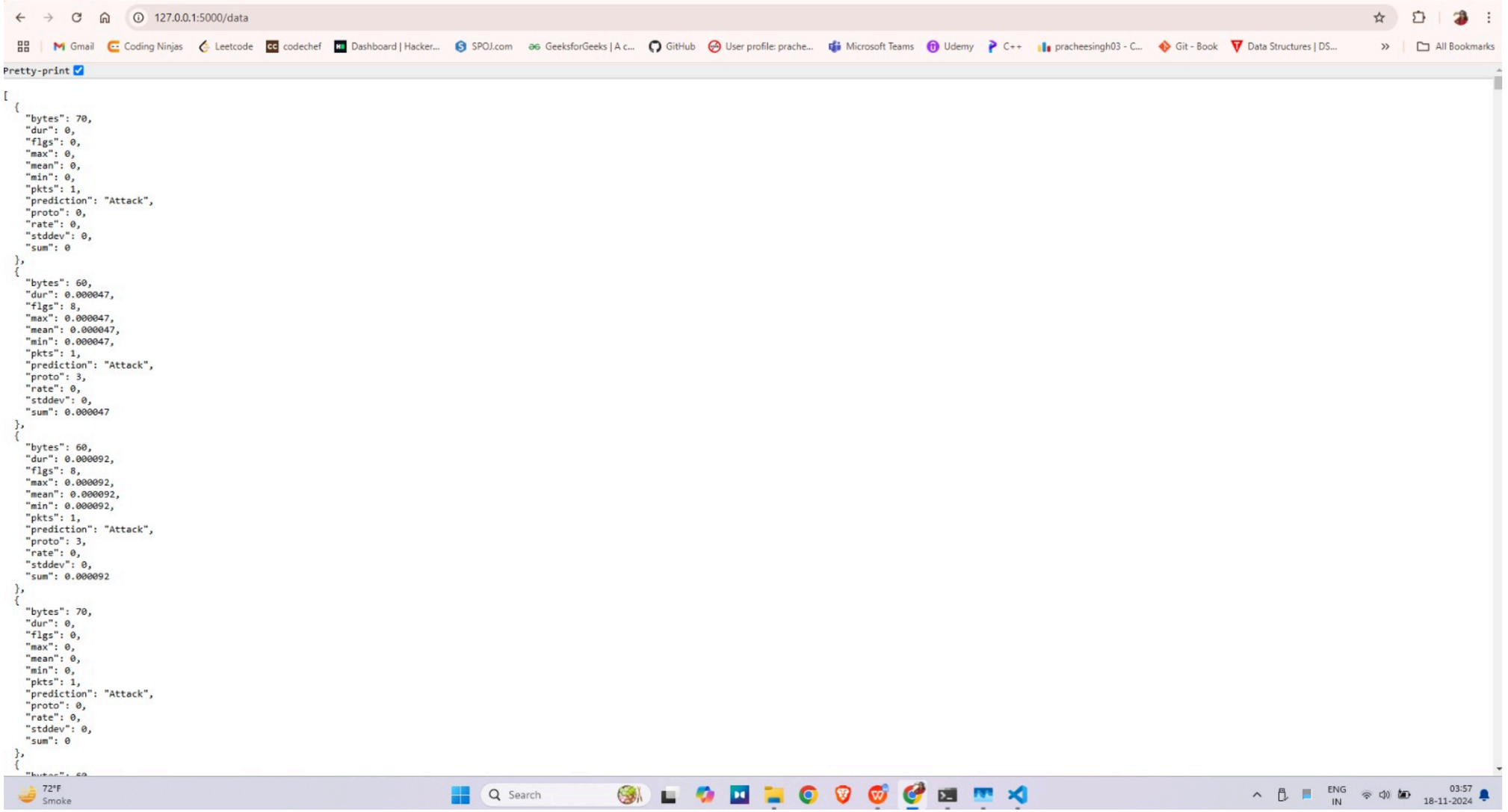
Workflow:

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

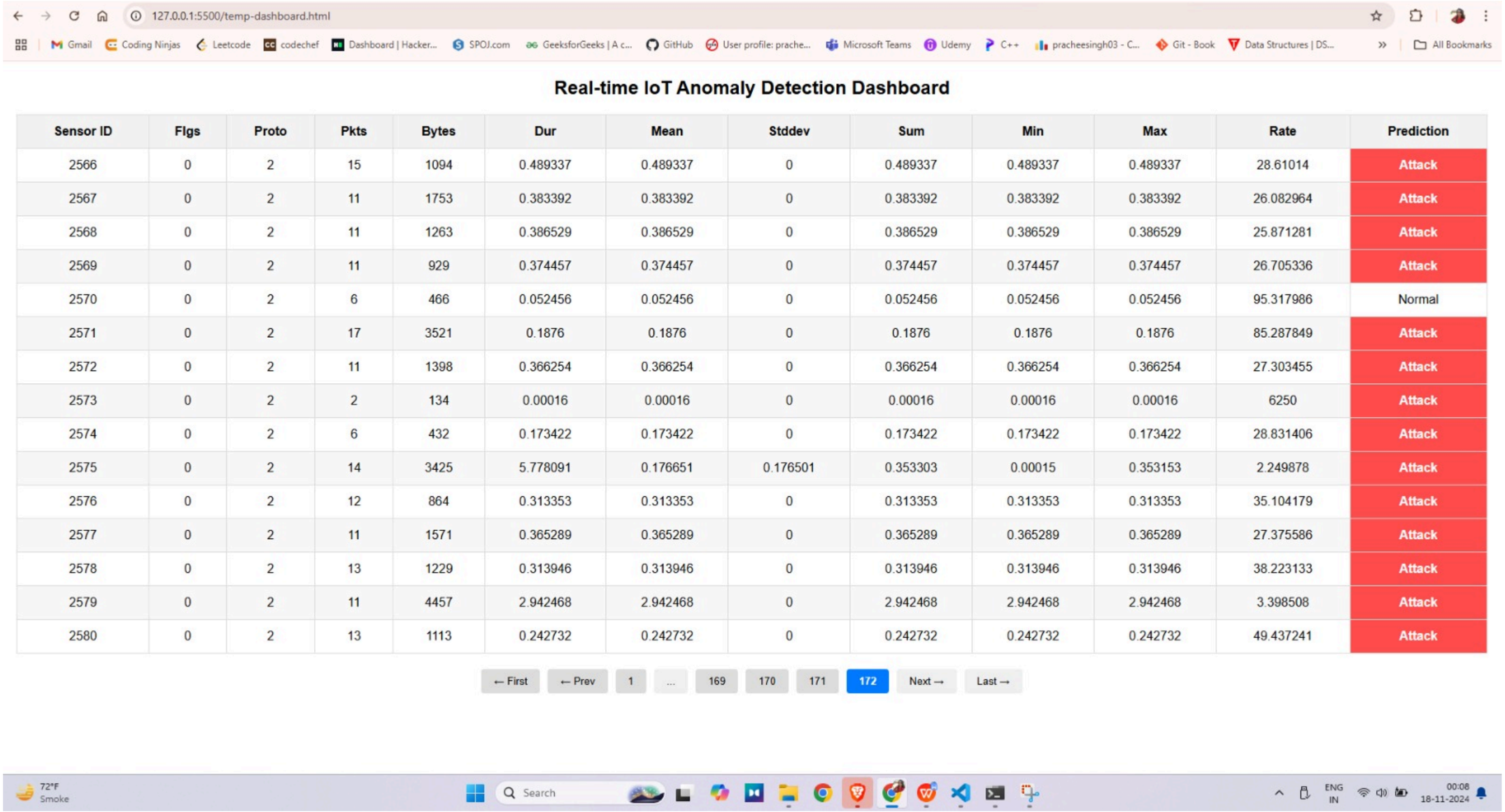
Hadoop Data Storage



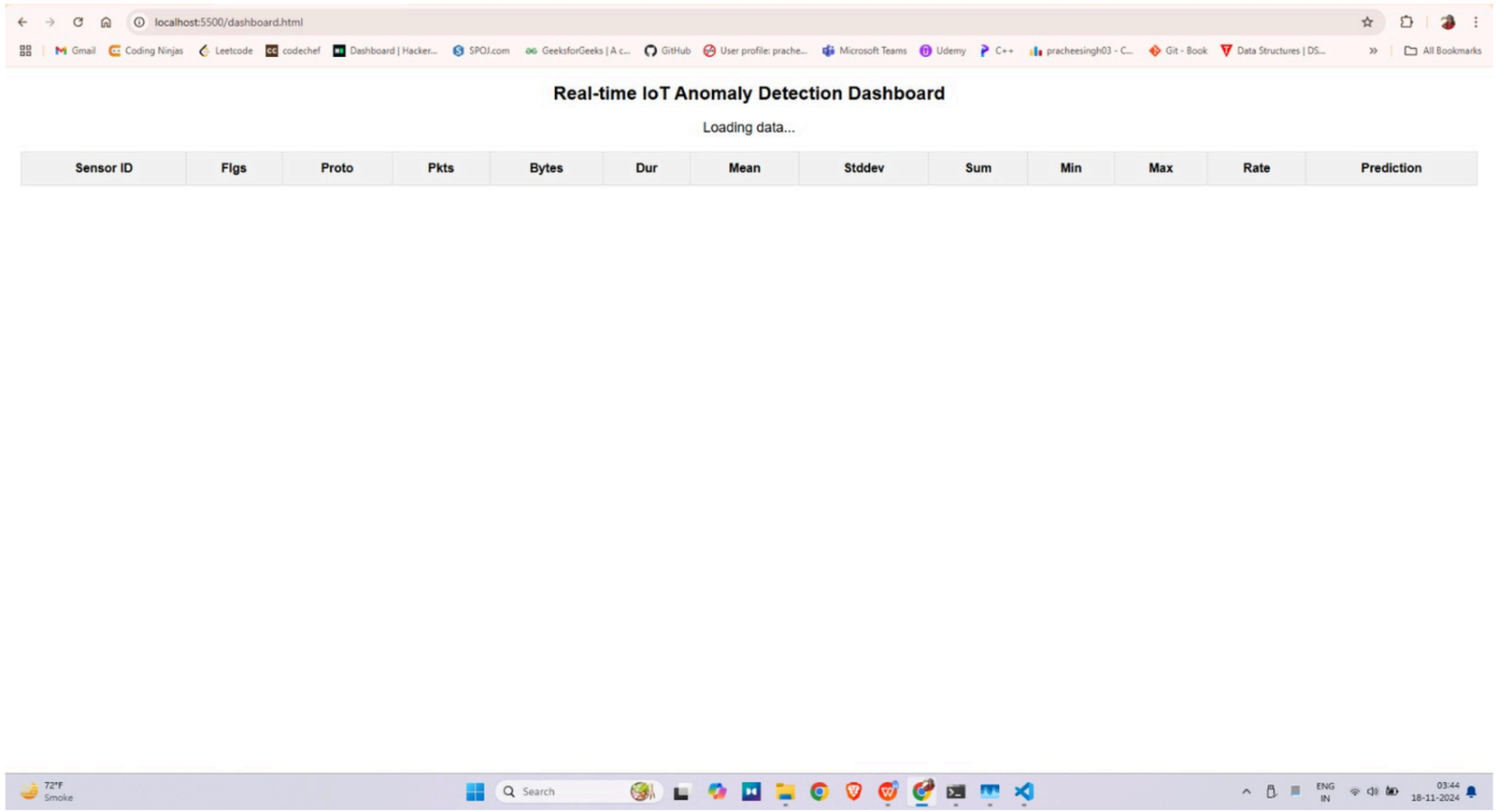
Flask app API



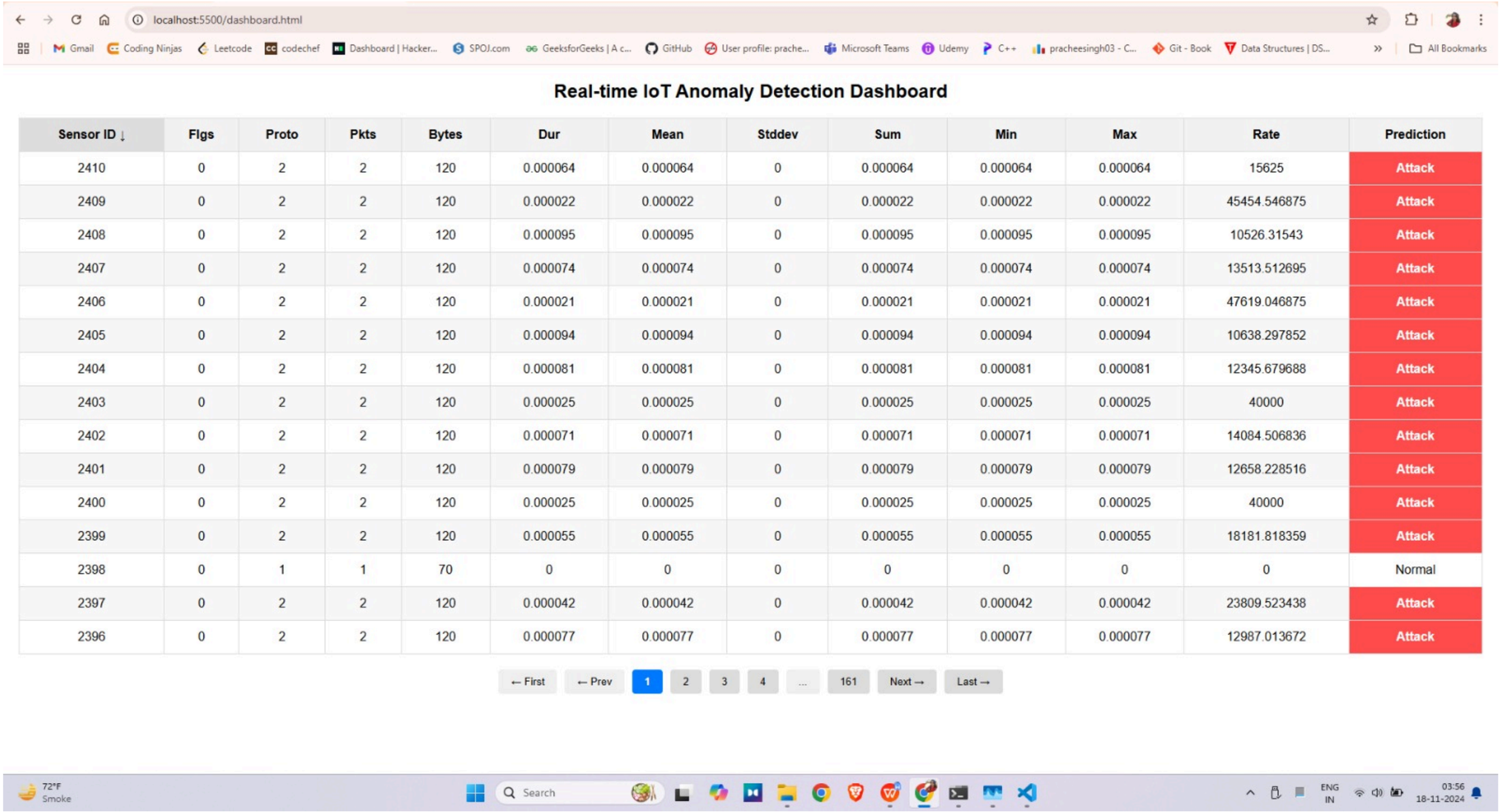
Real-time Dashboard



Dashboard..


localhost:5500/dashboard.html
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...
All Bookmarks
Real-time IoT Anomaly Detection Dashboard
Loading data...
Sensor ID
Flgs
Proto
Pkts
Bytes
Dur
Mean
Stddev
Sum
Min
Max
Rate
Prediction
72°F
Smoke
Search
ENG
IN
03:44
18-11-2024

Dashboard..


localhost:5500/dashboard.html
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...
All Bookmarks
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
← First
← Prev
1
2
3
4
...
161
Next →
Last →
72°F
Smoke
Search
ENG
IN
03:56
18-11-2024

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