FlexiSAF Final Project - Integrated Machine Learning Approaches

Description:

This project demonstrates the implementation of both supervised and unsupervised machine learning techniques to solve practical problems using publicly available datasets.

SUPERVISED LEARNING

Dataset:

Microsoft Malware Classification dataset (Kaggle)

Objective:

Predict the malware class based on extracted features.

Methodology:

1. Loaded and pre-processed the dataset.

2. Split data into training and test sets (stratified).

3. Standardized features using StandardScaler.

4. Trained a baseline Random Forest Classifier.

5. Tuned hyperparameters using GridSearchCV.

6. Evaluated performance using classification report, confusion matrix, and log loss.

UNSUPERVISED LEARNING

Dataset: Online Retail II dataset (Kaggle)

Objective:

Segment customers into groups based on purchasing behaviour.

Methodology:

1. Removed cancelled transactions and missing customer IDs.

2. Created RFM (Recency, Frequency, Monetary) features.

3. Standardized features.

4. Applied K-Means clustering.

5. Evaluated clusters using Silhouette Score and Davies–Bouldin Index.

6. Visualized clusters using PCA and interactive Plotly charts.

CONCLUSION

This project shows how both supervised and unsupervised learning can be applied to different domains:

- Classification for cybersecurity threat detection.
- Clustering for customer segmentation and marketing insights.