Assignment 3 – Report

Predicting the Type of Activity (Label2)

Goal

To build a model that can predict the **type of network activity** (for example: Browsing, Chat, Audio-Streaming, etc.) based on all other features in the **Darknet.csv** file.

What was done

- **Data cleaning steps**:
    - Removed inf, NaN, and very large numbers.
    - Filled missing values with column medians.
    - Label-encoded all text columns.
- **Scaling** – Used StandardScaler so that all numeric features have similar range.
- **Model choice** – Used a **Random Forest Classifier** instead of a neural network. Random Forest works very well with mixed-type tabular data and needs less tuning.
- **Balanced training** – Set class_weight='balanced' so that smaller classes get equal importance.
- **Training and testing** – Split the data (80 % train / 20 % test) and trained the model.
- **Evaluation** – Measured overall accuracy and printed precision, recall, and F1 scores for every activity type.

Why these choices were made

- **Random Forest** is fast, strong, and interpretable; it can handle many numerical and categorical inputs easily.
- **Balancing** prevents large activity types (like Browsing or P2P) from overpowering smaller ones.
- **Scaling and cleaning** prevent errors and make the model learn more stable patterns.

Result

The model reached about **82% accuracy** on the test data. This means it correctly predicts most of the activity types. It also shows which features are most important for classification.

Simple summary

Cleaned the data, encoded text values, and trained a Random Forest model to recognize network activity types. This method is stable, handles mixed data well, and achieved around 82% accuracy. It can now be used to classify future network traffic into correct activity categories.

Code Location

https://github.com/sharaba22/cda01/tree/main/Assignment%203