Microsoft : Classifying Cybersecurity Incidents with Machine Learning

**1.Data Exploration and Understanding:**

Start by loading the train.csv dataset and perform an initial inspection to understand the structure of the data, including the number of features, types of variables (categorical, numerical), and the distribution of the target variable (TP, BP, FP).

Since it has more than 90 lakh data, check for missing values .If it’s more than 10,000 null cells in a column ,we could drop it since we don’t know how handle and all the columns are new to us.

Check for class imbalance since its classification class imbalance affects the model.

**2.Create a virtual environment:**

Since the data is of 2GB+ using a virtual environment holds good to process,insert into model,etc..

**3.Model build :**

Since the data is so huge just use sampling of each column and use simple classification models.

**4.Model evaluation:**

Check for confusion matrix and classification report and get the best model and have it for test.csv.