Data Science Project Report

1. Background & Context

Malicious ips are IP addresses from which a malicious action is taken.
Watchers are software that reports malicious ips .
Autonomous Systems are entities that assign ips

2. Data :

Datasets :

- malicious_ips: Contains information about 23977 ip addresses
- watchers: contain information about 2121 watchers.
- autonomous_systems: contains information about 15161 AS.

3. Approach:

a. Model 1: Supervised Multiclass-Classification

Objective: Build an Intelligent model that can learn the relationship between the types of attacks and a set of information (features) about IP addresses.

Machine Learning problem :

- Build a model that can Classify the types of attacks based on features
- Target Variable: attack_type
- Features: 'id', 'malicious', 'AS_num'

Data Cleaning & Preprocessing :

- The main dataset is malicious IPS
- Using Malicious_IPs dataset :

Two data frames were created, one from ID columns as some missing features in the original data were inside the ID column.

The data frames were processed and concatenated to form the df_malicious_ips which is used for the model.

- **Features were Renamed to more significant names**
- **Missing Data :**

Data contains Missing values for column ‘is_validated’:’malicious’

7.56 % only of data exists

92.44 % of the data is NaN

Since ‘malicious’ is not the target variable: I choose to work only with labeled data because

Malicious is not the target variable -it’s a feature)

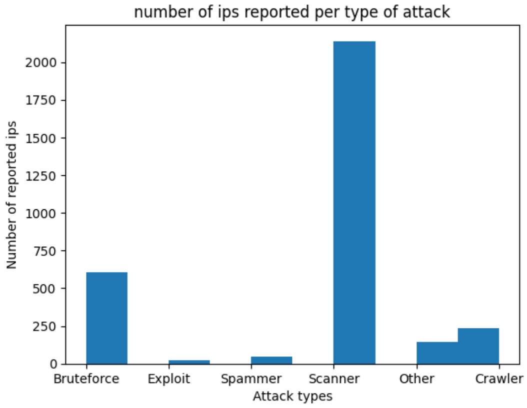
Imputation techniques like KNN imputation or even mode will bias the results since the number of missing values is huge.

The best way to do it is: Drop rows for which malicious is NaN.

Exploratory Data Analysis:

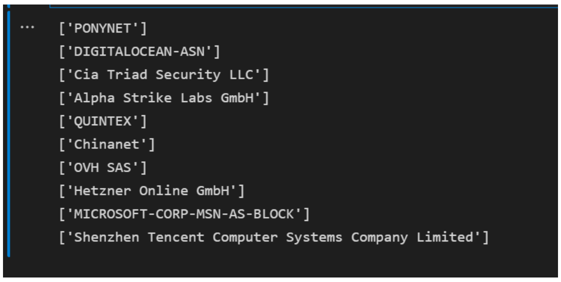
Questions asked:

Number of IPs reported by type of attack:



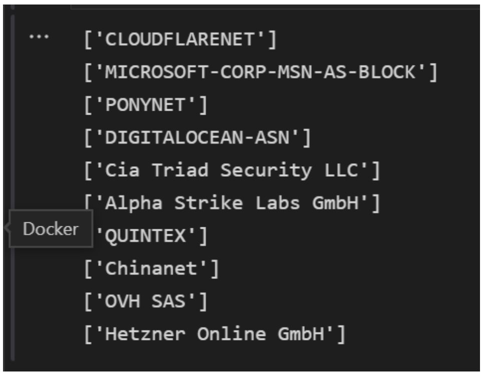
Interpretation: most attacks are Scanner attacks, this also shows that the Classes are imbalanced for types of attacks so we'll have more accuracy classifying these types of attacks at the end.

Names of top 10 AS hostings in malicious ips number



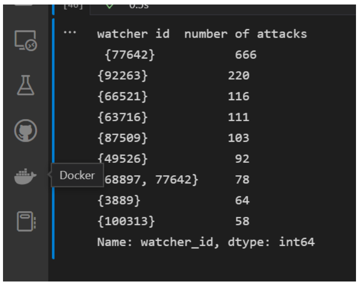
```
...  ['PONYNET']
     ['DIGITALOCEAN-ASN']
     ['Cia Triad Security LLC']
     ['Alpha Strike Labs GmbH']
     ['QUINTEX']
     ['Chinanet']
     ['OVH SAS']
     ['Hetzner Online GmbH']
     ['MICROSOFT-CORP-MSN-AS-BLOCK']
     ['Shenzhen Tencent Computer Systems Company Limited']
```

Top 10 AS for ips that perform the highest number of attacks



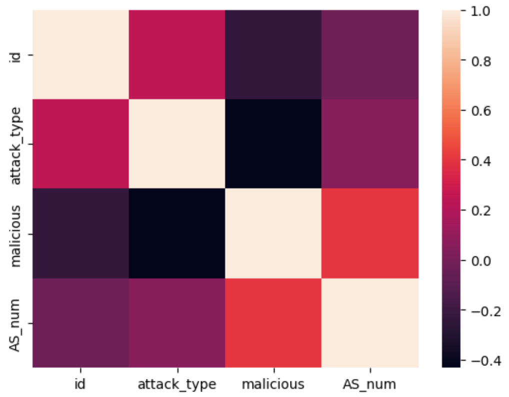
```
...  ['CLOUDFLARENET']
     ['MICROSOFT-CORP-MSN-AS-BLOCK']
     ['PONYNET']
     ['DIGITALOCEAN-ASN']
     ['Cia Triad Security LLC']
     ['Alpha Strike Labs GmbH']
      'QUINTEX']
     ['Chinanet']
     ['OVH SAS']
     ['Hetzner Online GmbH']
```

10 watcher IDs that suffered from the highest number of attacks



```
...  watcher id  number of attacks
      {77642}            666
     {92263}            220
     {66521}            116
     {63716}            111
     {87509}            103
     {49526}             92
      68897, 77642}      78
     {3889}              64
     {100313}            58
     Name: watcher_id, dtype: int64
```

Correlation Analysis :



Interpretation: Attack type is highly correlated with whether the attack is malicious and the autonomous system that generated it and its id.

Negatively correlated with the maliciousness and AS_num and positively correlated with id.

The correlation of id and AS_num might be due to the fact that ids are usually identifiers and unique to their sources, so they have representative information about their sources.

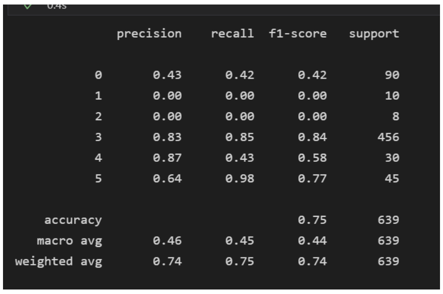
Data Preparation:

- train_test_split: 20 % of the data was test data
- Standardization: MinMAxScaler

Modeling & Evaluation:

Manual Model:

- Classifier: Support Vector Machine
- I chose to train an SVM with 'rbf' kernel first.
- The accuracy was 75 %
- The confusion matrix shows that Scanner attacks are the ones classified better which is logical since it's the most occurent in dataset .
- **Classification report :**



```
              precision    recall  f1-score   support

           0       0.43      0.42      0.42        90
           1       0.00      0.00      0.00        10
           2       0.00      0.00      0.00         8
           3       0.83      0.85      0.84       456
           4       0.87      0.43      0.58        30
           5       0.64      0.98      0.77        45

    accuracy                           0.75       639
   macro avg       0.46      0.45      0.44       639
weighted avg       0.74      0.75      0.74       639
```

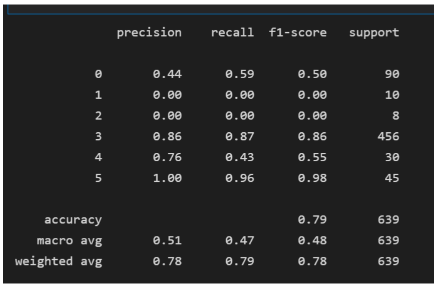
Grid Search Cross Validation Model :

- GridsearchCV :

  Grid Search cross-validation is a technique to select the best of the machine learning model, parameterized by a grid of hyperparameters.

  It combines both Gridsearch and Cross Validation

- The SVC was trained on several parameters , and output the best parameters as **{'C': 100, 'gamma': 1, 'kernel': 'rbf'}**
- Best Classifier is : SVC(C=100, gamma=1)
- Classification accuracy was relatively better :



```
              precision    recall  f1-score   support

           0       0.44      0.59      0.50        90
           1       0.00      0.00      0.00        10
           2       0.00      0.00      0.00         8
           3       0.86      0.87      0.86       456
           4       0.76      0.43      0.55        30
           5       1.00      0.96      0.98        45

    accuracy                           0.79       639
   macro avg       0.51      0.47      0.48       639
weighted avg       0.78      0.79      0.78       639
```

- Model is saved using joblib .

b. Model 2 : Semi-Supervised Clustering-Classification : Classification of attacks to malicious or not

Objective: Build an Intelligent model that can learn the relationship between the attack being malicious or not and a set of information (features) about IP addresses, generating AS systems and watchers activities

Machine Learning problem :

Build a model that can Classify the ips to malicious or not .

- Target Variable:Malicious
- Features:'AS_num','attack_type','number_of_watchers','n_ips_by_AS'

Data Cleaning and Preprocessing :

- Similar cleaning to first dataframe
- In this model , we will use both labled and unlabeled values for malicious
- the labeled values for classification and labeling and the second for clustering
- Here : Two dataframes were created
  - One for Labled Data
  - One for Lnlabeled data
- New features were created and used for training
  - 'number_of_watchers': Number of watchers for an IP address.
  - 'n_ips_by_AS' : number of ips generated by Autonomous Systems

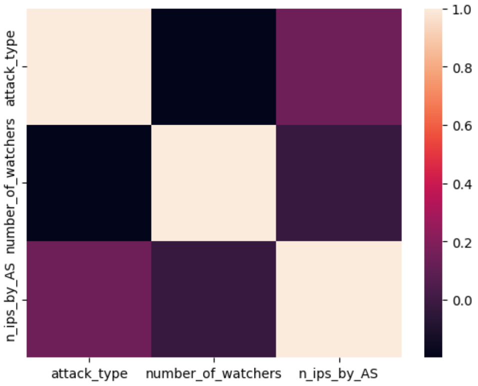
Data preparation :

- StandardScaler was used for scaling

Exploratory Data Aalysis :

Summary Statistics : Information about Data were given .

Correlation Analysis for features:



Interpretation : features are highly uncorrelated which makes them perfect for a linear model as logistic regression's success requires a high non correlation of features

Modeling and Evaluation :

Logistic Rergession :

- A logistic Regression model was trained on labled data only.
- Accuracy was : 0.91

Kmeans Clustering :

- Kmeans Clsuetring technique was applied for unlabeled data to cluster them .
- After Clustering :
  - Manual Labeling was performed on a few of the data
  - Label propagation was performed for entire dataset