Anomaly Detection in Network

Abstract: In our project, we build an intrusion detection model based on **CICIDS2017** dataset. Since the dataset is one of the largest datasets available online, to reduce the dimensionalities, we applied several feature selection process such as univariate feature selection, recursive features elimination. After getting most relevant features we applied random forest classification algorithm and we got **99**% accuracy.

1.Data:

The Dataset **CICIDS2017** used for this project was collected from kaggle[1]. It is one of the largest datasets with **225745** samples and **79** features. We split the dataset: 70% for training and 30% for testing.

1.1 Data Analysis :

a) ratio check of the Label`s column:

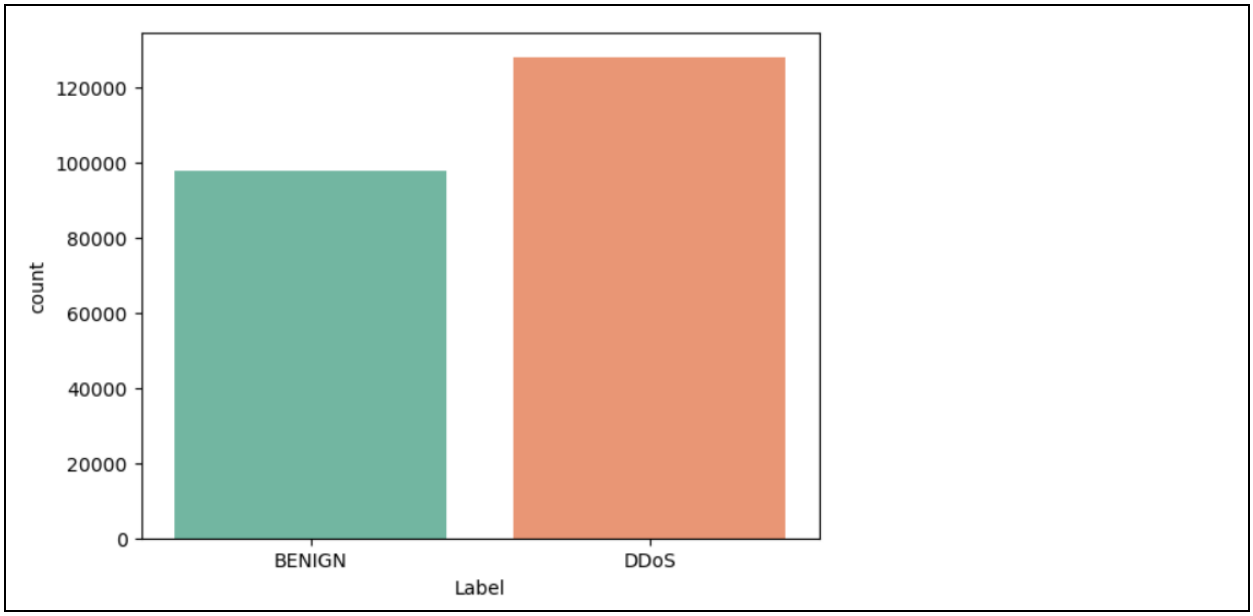


Fig4: ratio of the label

We can see that, it is a pretty much balanced dataset. So that, it is less possible that the predictions of the applied models will be biased.

b) correlation with each column: we find the correlations of each column with another and based on the correlation, we dropped some features which are highly correlated(0.9)

2. Models

2.1 Univariate features selection(SelectKBest)

Univariate feature selection methods works by selecting the best features based on univariate statistical tests like ANOVA. Statistical tests can be used to select those features that have the strongest relationship with the output variable.

Chi-Square Test for Feature Selection: A chi-square test is used in statistics to test the independence of two events. Given the data of two variables, we can get observed count O and expected count E. Chi-Square measures how expected count E and observed count O deviates each other.

The Formula for Chi Square Is

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where:

c = degrees of freedom

O = observed value(s)

E = expected value(s)

2.2 Recursive Feature Elimination

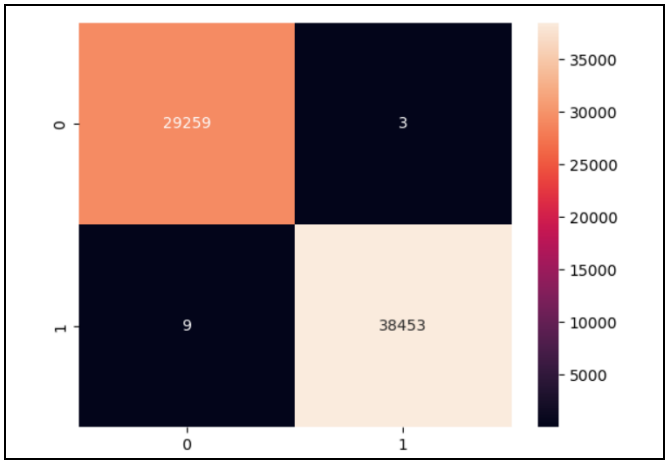
Recursive Feature Elimination, or RFE for short, is a feature selection algorithm.RFE is a wrapper-type feature selection algorithm. This means that a different machine learning algorithm is given and used in the core of the method, is wrapped by RFE, and used to help select features. This is in contrast to filter-based feature selections that score each feature and select those features with the largest (or smallest) score. We used random forest classification in the RFE.

3. Experiments and Results

3.1 Univariate Feature Selection: We selected most relevant 15 features and then apply random forest classification

Models	Features	Accuracy
SelectKBest	**15**	**99.98**

Confusion Matrix:



3.2 Recursive Feature Elimination(RFE): We selected most relevant 20 features and then apply random forest classification.

Models	Features	Accuracy
RFE	**20**	**99.98**

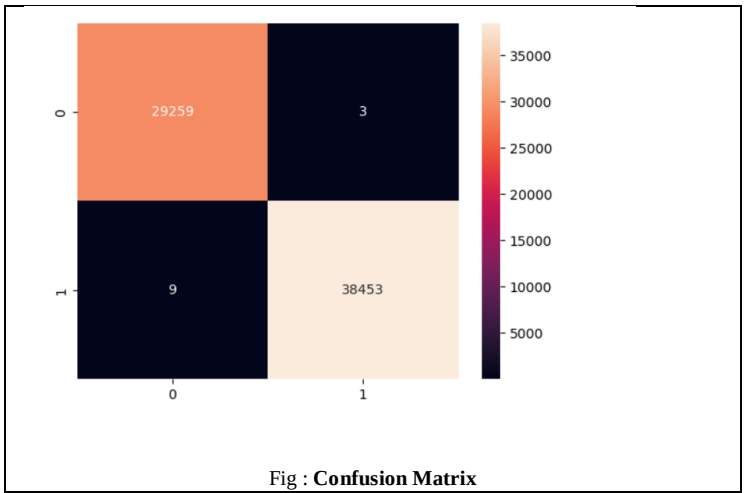


Fig : **Confusion Matrix**

3.3 Recursive Feature Elimination(RFE) with cross validation(CV): We got most relevant 30 features and then apply random forest classification.

Models	Features	Accuracy
RFE	**30**	**99.99**

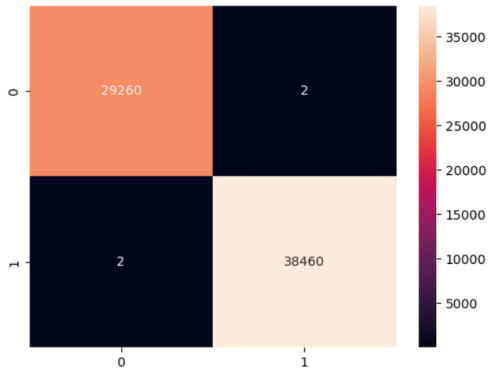


Fig: Confusion Matrix

References:

[1]. Dataset: CICIDS2017 (Link)

[2]. Serpil Ustebay, Zeynep Turgut, " Intrusion Detection System with Recursive Feature Elimination by using Random Forest and Deep Learning Classifier ", 2018.[Link]