**Network Intrusion Detection System**

**Provided by:**

Ziad Mahmoud abdeljawad 2205021

Abdallah Mohammed Hegazy 2205228

Ahmed Ibrahim Ghazal 2205004

**Introduction:**

We applied IDS using Decision tree model to detect the malicious traffic in the network by analyzing traffic’s packets and find the malicious ones.

**Dataset description:**

In that project, we used NSL\_KDD dataset. As that dataset is huge we work on subset for training and another for testing. The train subset contains of 125973 entities with 42 features and the test subset contains of 22544 entities with 41 features without the target (**normal / malicious**).Those subsets consist of mixed data types (**float & object**) data type .Those features divided into 4 categories:

1. Basic features: like protocol type (**TCP / UDP**), duration, service, …etc.
2. Time-based traffic features:used to capture those features which are mature over a 2 second temporal window
3. Content features: those features use domain kwonledge to access the payload of original TCP packet.
4. Host-based features: all attacks which span longer than 2 second intervals that have the same destination host as the current connection are access using these features.

For target feature, the attack types is categorized into 4 types:

1. DOS attack: all attacks that aim to block or make the system utilities down.
2. Prob attack: all attacks that aim to scan the system vulnerabilities or system open ports.
3. Remote to Local (R2L): the intruder remotely gains unauthorized access to a system device over the network by sending malicious data packets to it.
4. User to Root(U2R): the intruder gains a normal user access and try to access a user with root privilages.

**Data preprocessing:**

**Step 1: Data cleaning:**

In data cleaning step we remove the duplicate and messing features and add labels to the data in the dataset. After data cleaning we transform the categorical data to numerical. To apply that we use LabelEncoder() function that replace the object data type with numerical data type.

**Step 2: Features scaling:**

Feature scale is an important step that applied to avoid any bais in prediction step. After scaling all features will have zero average with a standard deviation of one.

**Step 3: Features selection:**

Feature selection is a technique used to eliminate the redundant and irrelevant data. That technique used to choose the most appropriate 13 features depending on its ANOVA F-test score.

**Model building:**

A decision tree model was built to partition the data using information gain until instances in each leaf node have uniform class labels. A Decision Tree is made up internal decision nodes and terminal leaves. A test function is implemented by each decision node with a discrete results labelling the branches. Providing an input, at every node, a test is constructed and based on the outcome, one of the branches will be considered. Here the learning algorithm starts at the root and until a leaf node is reached, the process will be done recursively at which moment the value represented in the leaf node is the output. Every leaf node possesses an outcomes label, which it is the class target in case of classification and numeric value for regression. A leaf node can describe a localized space or region where instances finding in this input space (region) possess the same labels for classification and similar numeric value for regression.

**Prediction before selection features :**

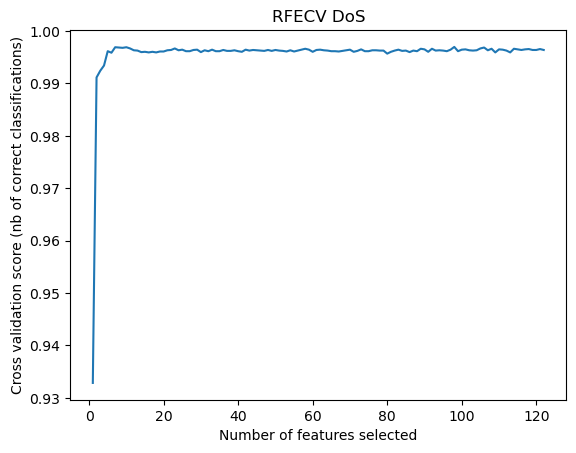
* **DOS :**

**Accuracy: 0.99639 (+/- 0.00341)**

**Precision: 0.99505 (+/- 0.00477)**

**Recall: 0.99665 (+/- 0.00483)**

**F-measure: 0.99585 (+/- 0.00392)**



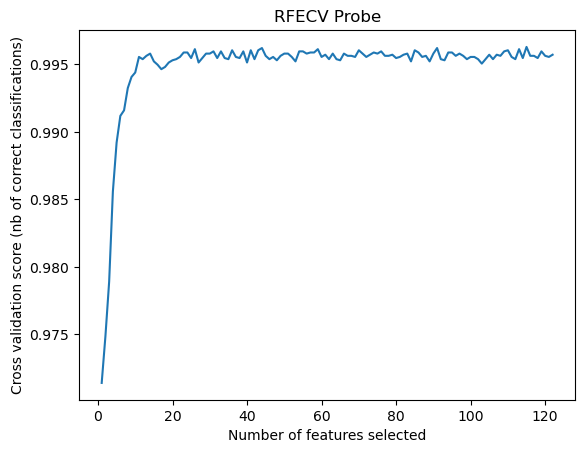
* **Probe :**

Accuracy: 0.99571 (+/- 0.00328)

Precision: 0.99392 (+/- 0.00684)

Recall: 0.99267 (+/- 0.00405)

F-measure: 0.99329 (+/- 0.00512)



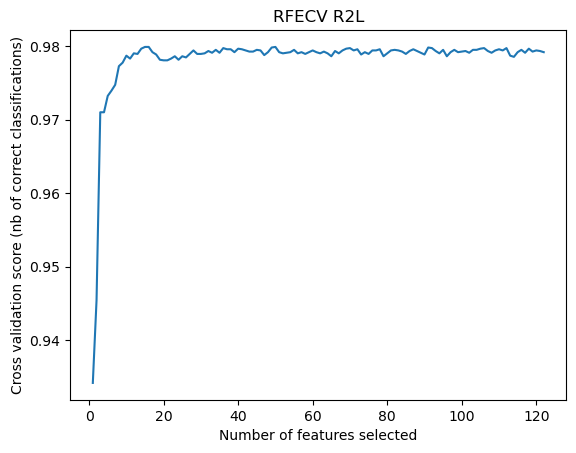
* **R2L :**

Accuracy: 0.97920 (+/- 0.01053)

Precision: 0.97151 (+/- 0.01736)

Recall: 0.96958 (+/- 0.01379)

F-measure: 0.97051 (+/- 0.01478)



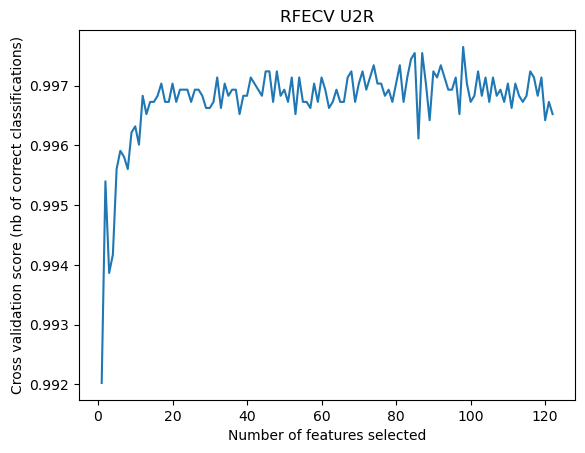
* **U2R :**

Accuracy: 0.99652 (+/- 0.00228)

Precision: 0.86295 (+/- 0.08961)

Recall: 0.90958 (+/- 0.09211)

F-measure: 0.88210 (+/- 0.06559)



**Prediction Using 13 Features for each category :**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Accuracy | Precision | Recall | F-measure | N of Features | Class |
| 99.90 | 99.69 | 99.79 | 99.74 | 12 | Dos |
| 99.80 | 99.37 | 99.37 | 99.37 | 15 | Probe |
| 99.88 | 97.40 | 97.41 | 97.40 | 13 | R2L |
| 99.95 | 99.70 | 99.69 | 99.70 | 11 | U2R |