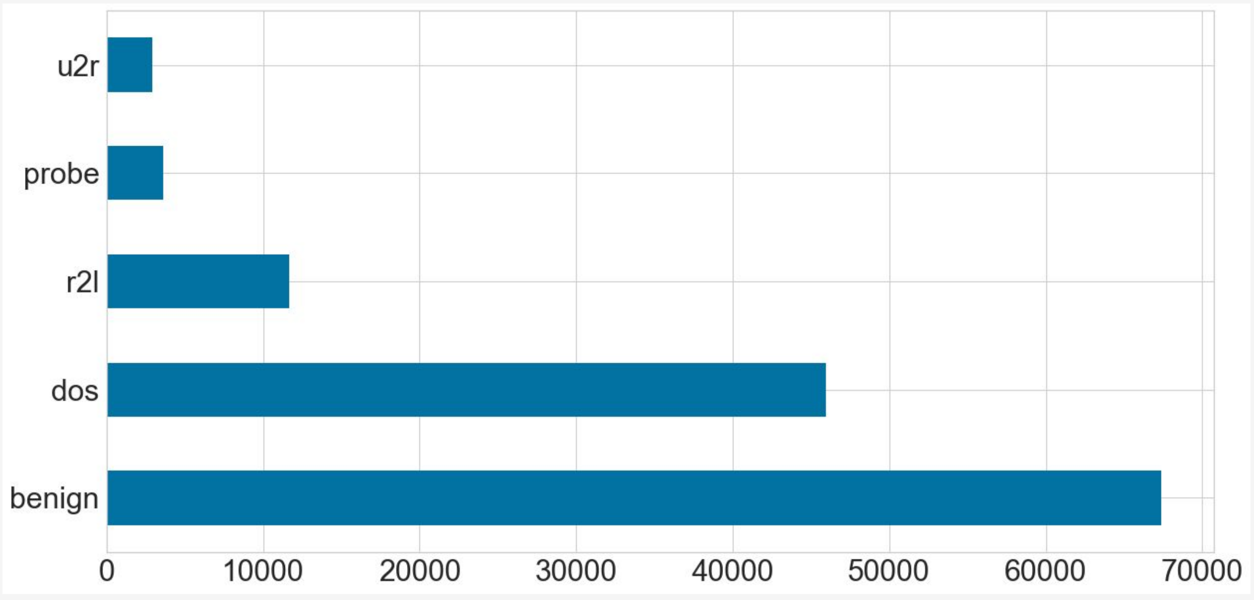
SIT719 Security and Privacy Issues in Analytics Distinction/Higher Distinction Task 5.1 End-to-end project delivery on cyber-security data analytics

Dataset used:

NSL-KDD is a data set suggested to solve some of the inherent problems of the KDD'99 data set. Although this new version of the KDD data set still suffers from some of the problems discussed by McHugh and may not be a perfect representative of existing real networks, because of the lack of public data sets for network-based IDSs, we believe it still can be applied as an effective benchmark data set to help researchers compare different intrusion detection methods.

Furthermore, the number of records in the NSL-KDD train and test sets are reasonable. This advantage makes it affordable to run the experiments on the complete set without the need to randomly select a small portion.



ML models used:

1. **Decision Trees (DTs)** are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.
2. A **Random Forest Classifier** is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over-fitting.
3. **Support Vector Classification**. The implementation is based on libsvm. The fit time scales at least quadratically with the number of samples and maybe impractical beyond tens of thousands of samples.
4. **GaussianNB** implements the Gaussian Naive Bayes algorithm for classification.
5. **Multi-layer Perceptron classifier**. This model optimizes the log-loss function using LBFGS or stochastic gradient descent.

ALgorithms	Precision	Recall	F1-score	Accuracy
.DecisionTree Classifier	0.82	0.78	0.75	77.8%
Random Forest Classifier	0.82	0.77	0.73	76.9%
Support Vector Classification	0.82	0.76	0.73	76.1%
Multi-layer Perceptron classifier	0.77	0.75	0.72	74.8%
GaussianNB	0.65	0.49	0.50	49.4%

Algorithms:

1.**DecisionTreeClassifier** [Best model (77.89)]:

DecisionTreeClassifier is a class capable of performing multi-class classification on a dataset.

As with other classifiers, **DecisionTreeClassifier** takes as input two arrays: an array X, sparse or dense, of size [n_samples, n_features] holding the training samples, and an array Y of integer values, size [n_samples], holding the class labels for the training samples:

Parameters:

criterion*{“gini”, “entropy”}, default=”gini”*

The function to measure the quality of a split. Supported criteria are “gini” for the Gini impurity and “entropy” for the information gain.

splitter*{“best”, “random”}, default=”best”*

The strategy used to choose the split at each node. Supported strategies are “best” to choose the best split and “random” to choose the best random split.

max_depth*int, default=None*

The maximum depth of the tree. If None, then nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples.

min_samples_split*int or float, default=2*

The minimum number of samples required to split an internal node:

- If int, then consider min_samples_split as the minimum number.
- If float, then min_samples_split is a fraction and ceil(min_samples_split * n_samples) are the minimum number of samples for each split.

Advantages of decision trees are:

- Simple to understand and to interpret. Trees can be visualised.
- Requires little data preparation. Other techniques often require data normalisation, dummy variables need to be created and blank values to be removed. Note however that this module does not support missing values.

- The cost of using the tree (i.e., predicting data) is logarithmic in the number of data points used to train the tree.
- Able to handle both numerical and categorical data. Other techniques are usually specialised in analysing datasets that have only one type of variable. See algorithms for more information.
- Able to handle multi-output problems.
- Uses a white-box model. If a given situation is observable in a model, the explanation for the condition is easily explained by boolean logic. By contrast, in a black-box model (e.g., in an artificial neural network), results may be more difficult to interpret.
- Possible to validate a model using statistical tests. That makes it possible to account for the reliability of the model.
- Performs well even if its assumptions are somewhat violated by the true model from which the data were generated.

Classifier with parameters updated:

```
classifier=DecisionTreeClassifier(random_state=50,max_depth=35,max_features=25,min_impurity_split=0,max_leaf_nodes=None)
```

Classification Report:

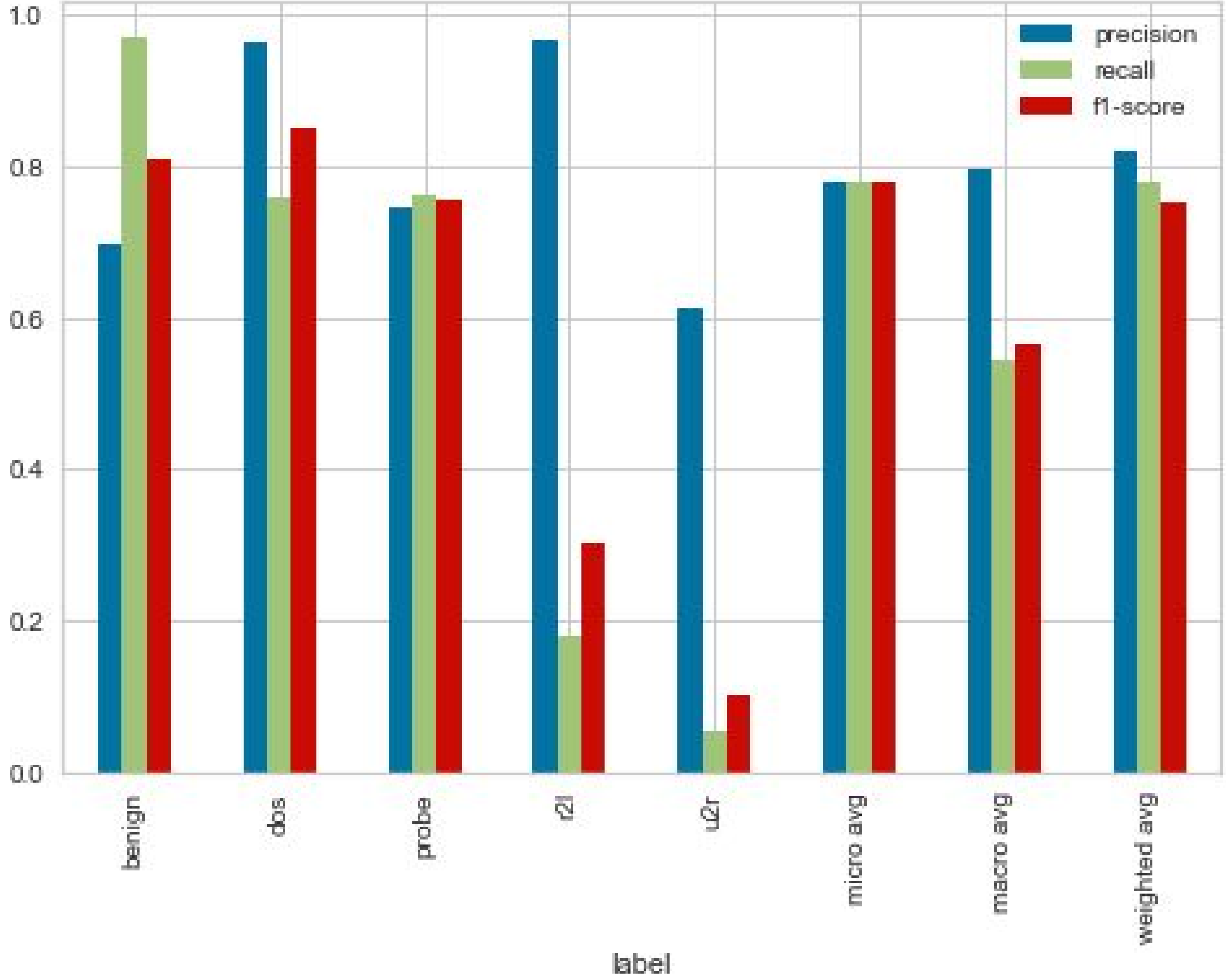
Classes	Precision	Recall	F1-score	Support
benign	0.70	0.97	0.81	9711
dos	0.96	0.76	0.85	7636
probe	0.74	0.76	0.75	2423
r2l	0.97	0.18	0.30	2574
u2r	0.61	0.06	0.10	200

weighted avg: 0.82 0.78 0.75 22544

```
Accuracy Score:0.7789212207239177

Confusion Matrix:

        [[9428   57  213   11    2]
        [1765 5806   65    0    0]
        [ 406  165 1852    0    0]
         [1755    0  351  463    5]
         [ 177    0    7      5      11]]
```



2.**Random Forest Classifier:**

Classification Report:

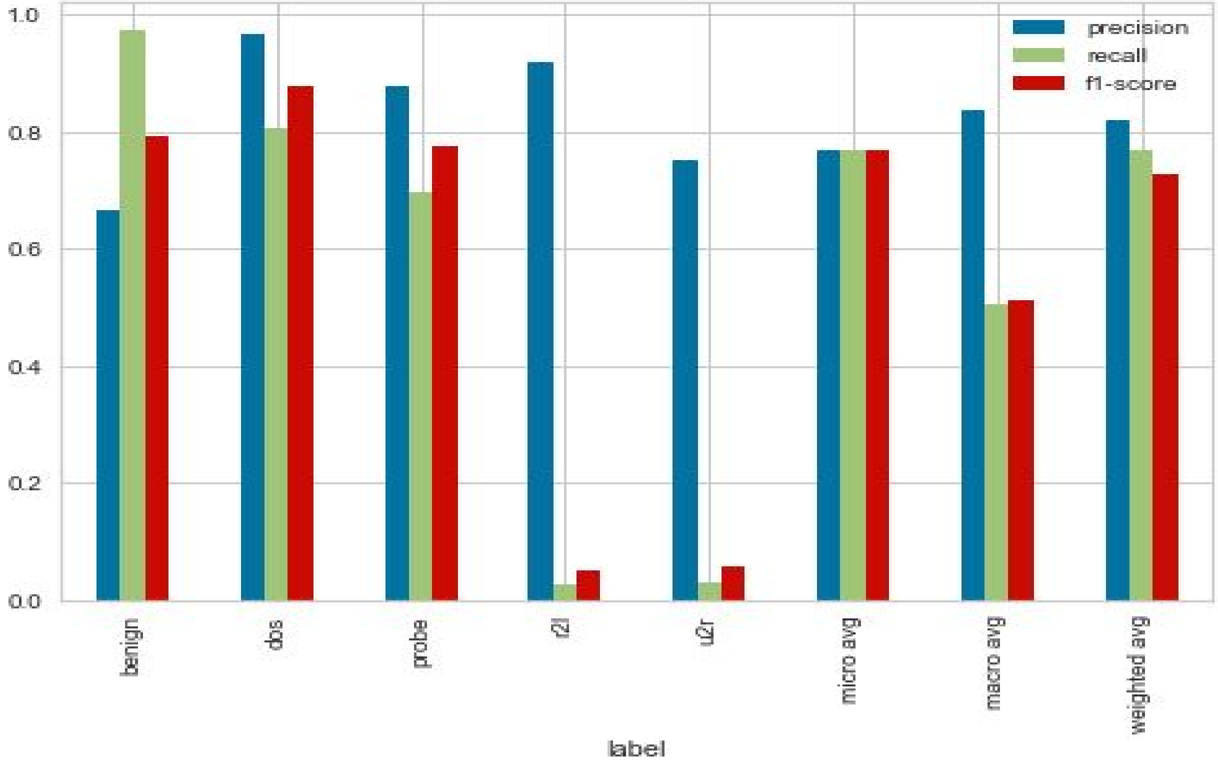
Classes	Precision	Recall	F1-score	Support
benign	0.67	0.97	0.79	9711
dos	0.96	0.80	0.88	7636
probe	0.88	0.69	0.77	2423
r2l	0.92	0.03	0.05	2574
u2r	0.75	0.03	0.06	200

weighted avg: 0.82 0.77 0.73 22544

Confusion Matrix:

```
[[9447   66  197    1    0]
 [1469 6136   31    0    0]
 [ 583  158 1682    0    0]
 [2503    2    1   66    2]
 [ 178    1   10    5    6]]
```

Accuracy Score:0.7690294535131299



3.**Support Vector Classification:**

Classification Report:

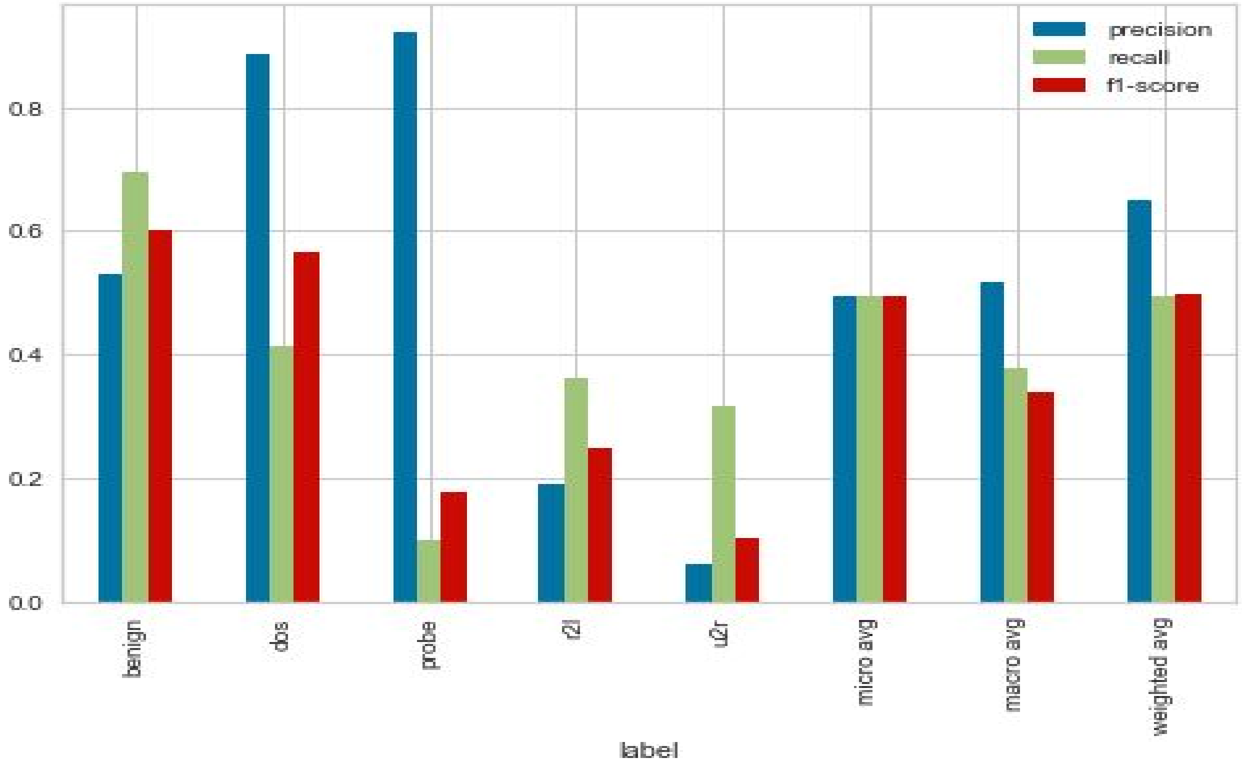
Classes	Precision	Recall	F1-score	Support
benign	0.66	0.97	0.79	9711
dos	0.96	0.76	0.85	7636
probe	0.86	0.71	0.78	2423
r2l	0.99	0.09	0.16	2574
u2r	0.0	0.0	0.0	200

weighted avg: 0.82 0.76 0.73 22544

Confusion Matrix:

```
[[9429   62  220    0    0]
 [1800 5789   47    0    0]
 [ 543  165 1715    0    0]
 [2337    0    6  231    0]
 [ 194    3    0    3    0]]
```

Accuracy Score:0.7613555713271823



4.**GaussianNB:**

Classification Report:

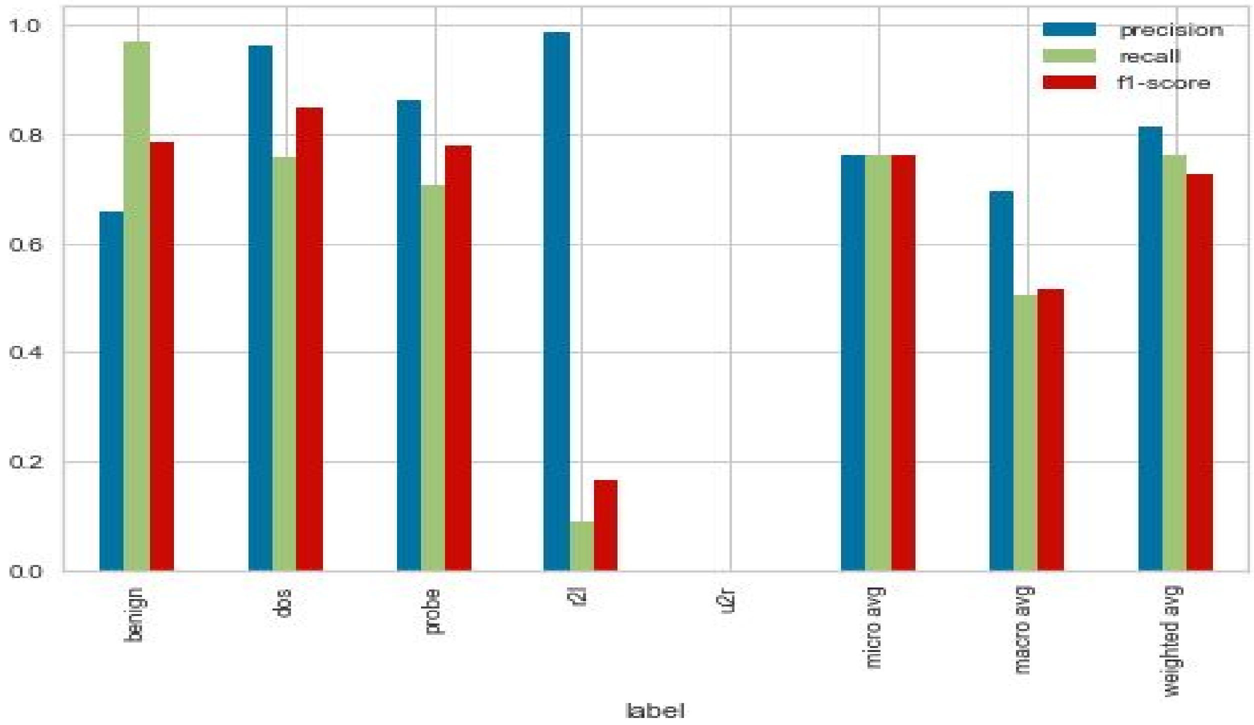
Classes	Precision	Recall	F1-score	Support
benign	0.53	0.70	0.60	9711
dos	0.89	0.41	0.56	7636
probe	0.92	0.10	0.18	2423
r2l	0.19	0.36	0.25	2574
u2r	0.06	0.32	0.10	200

weighted avg: 0.65 0.49 0.50 22544

Confusion Matrix:

```
[[6756   62   10 2599  284]
 [3400 3162    4 1041   29]
 [1382  345  237  323  136]
 [1113    3    6  933  519]
 [ 115    0    0   22   63]]
```

Accuracy Score:0.4946327182398864



5.Multi-layer Perceptron classifier:

Classification Report:

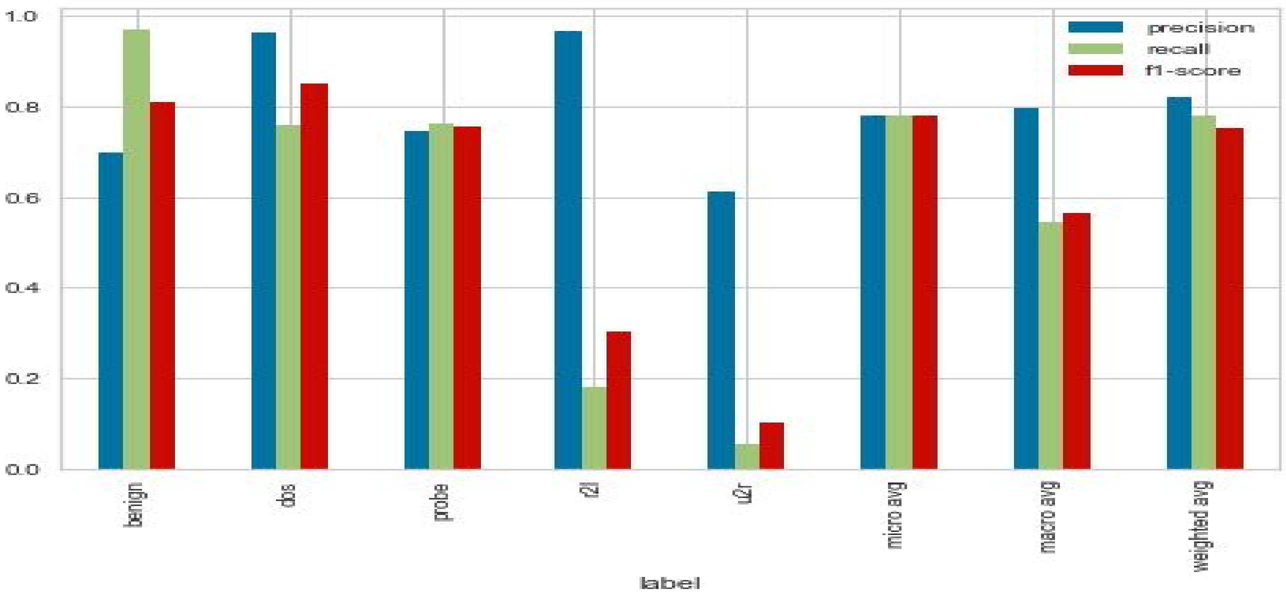
Classes	Precision	Recall	F1-score	Support
benign	0.68	0.93	0.78	9711
dos	0.89	0.75	0.81	7636
probe	0.75	0.76	0.75	2423
r2l	0.85	0.11	0.20	2574
u2r	0.0	0.0	0.0	200

weighted avg: 0.77 0.75 0.72 22544

Confusion Matrix:

```
[[9051  177  445   38    0]
 [1774 5697  164    1    0]
 [ 422  162 1839    0    0]
 [1925  360    3  286    0]
 [ 178   10    2   10    0]]
```

Accuracy Score:0.7484474804826118


precision
recall
f1-score
benign
dos
probe
r2l
u2r
micro avg
macro avg
weighted avg
label