data collection:

The KDD Cup 1999 dataset, utilized for benchmarking intrusion

detection issues, is used in our experiments.it is collected from uci machine learning repository

KDD training dataset consists of approximately 4,900,000 single connection

vectors each of which contains 41 features and is labeled as either normal

or an attack, with exactly one specific attack type.

preprocessing:

Data preprocessing is a data mining technique which is used to transform the raw data

in a useful and efficient format.

Steps Involved in Data Preprocessing:

1. ... To handle this part, data cleaning is done.

It involves handling of missing data, noisy data ,null values & outliers etc

feature selection:

it enables the machine learning algorithm to train faster.

It reduces the complexity of a model and makes it easier to interpret.

It improves the accuracy of a model if the right subset is chosen.

It reduces overfitting.

Feature selection techniques that are easy to use and also gives good results.

Univariate Selection

using variance method

Feature Importance

Correlation Matrix with Heatmap

feature extraction:

Feature extraction is a process that identifies important features or attributes of the data.

If the data has a lot of dimensions then feature extraction can be used to produce a more concise description of the data.

Feature extraction can be useful to reduce the number of attributes that describe the data.

This can speed up the ml algorithm fastly.in features extraction we use standard scalar,normalization& pca used for dimensionality reduction

supervised learning algorithms

A supervised learning algorithm learns from labeled training data,

helps you to predict outcomes for unforeseen data.

In Supervised learning, you train the machine using data which is well "labeled."

It means some data is already tagged with the correct answer.

classfication:

Classification means to group the output inside a class. If the algorithm tries to label input into two distinct classes, it is called binary classification. Selecting between more than two classes is referred to as multiclass classification.

Example: Determining whether the network is normal or attack

Challenges in Supervised machine learning

Here, are challenges faced in supervised machine learning:

Irrelevant input feature present training data could give inaccurate results

Data preparation and pre-processing is always a challenge.

Accuracy suffers when impossible, unlikely, and incomplete values have been inputted as training data

If the concerned expert is not available, then the other approach is "brute-force." It means you need to think that the right features (input variables) to train the machine on. It could be inaccurate.

in classification supervised algorithm we use nb ,svm,sgd,decission tree,random forest,gbt,@lr to classify and get better output .

hyperparameter tuning

wo simple strategies to optimize/tune the hyperparameters:

Models can have many hyperparameters and finding the best combination of parameters can be treated as a search problem.

Although there are many hyperparameter optimization/tuning algorithms now, this post discusses two simple strategies: 1. grid search and 2. Random Search.

Grid search is arguably the most basic hyperparameter tuning method. With this technique, we simply build a model for each possible combination of all of the hyperparameter values provided, evaluating each model, and selecting the architecture which produces the best results.

For example, we would define a list of values to try for both n\_estimators and max\_depth and a grid search would build a model for each possible combination.

ml pipe line

For building any machine learning model, it is important to have a sufficient amount of data to train the model. The data is often collected from various resources and might be available in different formats. Due to this reason, data cleaning and

preprocessing become a crucial step in the machine learning project.

Whenever new data points are added to the existing data, we need to perform the same preprocessing steps again before we can use the machine learning model to make predictions. This becomes a tedious and time-consuming process!

An alternate to this is creating a machine learning pipeline that remembers the complete set of preprocessing steps in the exact same order. So that whenever any new data point is introduced, the machine learning pipeline performs the steps as defined and uses the machine learning model to predict the target variable.