**Data Preprocessing of KDDCup99 dataset for Network Intrusion Detection**

***Libraries Used:***

***Numpy:*  Numpy**is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays. It is the fundamental package for scientific computing with Python.

***Pandas:*** Pandas is the most popular python library that is used for data analysis. It provides highly optimized performance with back-end source code is purely written in C or Python.

***Matplotlib:*** Matplotlib is an amazing visualization library in Python for making plots of arrays. Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack. It was introduced by John Hunter in the year 2002.

***Scikit-Learn:*** Scikit-learn is a machine learning library written in python. Most of the learning algorithm implement in scikit-learn required data to be stored in a two-dimensional array or matrix.Sckit-learn provides a range of supervised and unsupervised learning algorithms via consistent interface in Python.

***Methodology:***

***Step 1: Data Cleaning and Pre-processing:***

Basically, in this step the dataset has to go through a cleaning process to remove duplicate records. In order to decrease the number of columns we have to find the correlated data and delete them.

Next a Pre-processing operation has to be taken in place because the dataset contains numerical and non-numerical instances. Label Encoder has to be used to encode the categorical data. The technique will transform each categorical feature with different numbers.

First open the csv file with pandas.read\_csv function and assign it to a variable named ***data***.

Now I need to make the correlation matrix and delete the correlated features, but for making a correlation matrix all the data should be in a numeric format. For this I created a duplicate variable of ***data*** named ***data1.*** In ***data1*** I dropped all the categorical features in the ***data.*** To find the correlation matrix there is a function given in pandas named *corr().*

**Pearson’s Correlation :**

The Pearson’s correlation coefficient is calculated as the covariance of the two variables divided by the product of the standard deviation of each data sample.

Pearson's correlation coefficient = covariance(*x, y*) / (stdv(*x*) \* stdv(*y*))

The result of the calculation, the correlation coefficient can be interpreted to

***About corr() function :***

Dataframe.corr() is used to find the pairwise correlation of all columns in the dataframe.

**Parameters:**

* **method:**

*pearson :* standard correlation coefficient

*kendall :* Kendall Tau correlation coefficient

*spearman:* Spearman rank correlation

* **min\_period :** Minimum number of observations required per pair of columns to have a valid result. Currently only available for pearson and spearman correlation.

Pearson’s Method is the default method of corr() function.

Let’s see how Pearson’s Correlation Method works.

It makes the correlation matrix of features. All the values in the matrix will be between -1 and 1. Values between -0.5 and 0.5 are less correlated and values greater than 0.5 or below -0.5 indicates a notable correlation. I have selected the columns to drop having correlation value greater than or equal to 0.7 and less than or equal to -0.7.

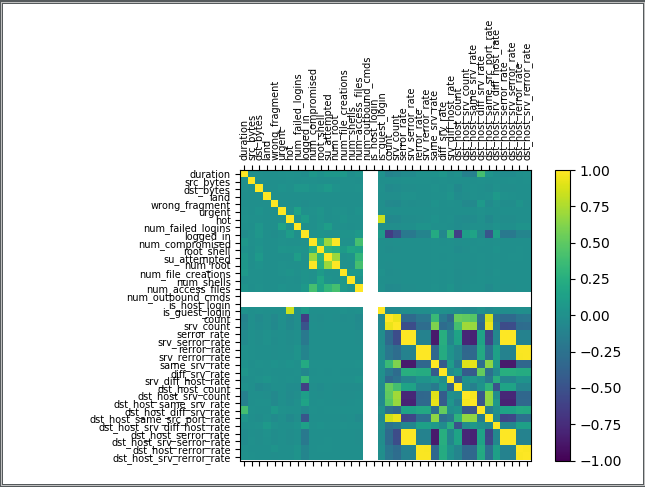
understand the relationship.

The coefficient returns a value between -1 and 1 that represents the limits of correlation from a full negative correlation to a full positive correlation. A value of zero means no correlation. The values below -0.5 or above 0.5 indicates a notable correlation, and values below those values suggests a less notable correlation.

This can be done by calculating a matrix of the relationships between each pair of variables in the dataset. The result is a symmetric matrix called a correlation matrix with a value of 1.0 along the diagonal as each column always perfectly correlates with itself.

I am giving the plot of correlation matrix below:

Fig1. Plot of Correlation Matrix



Below is the set of 14 column names selected to drop from the original dataset.

{'dst\_host\_srv\_rerror\_rate', 'num\_root', 'same\_srv\_rate', 'dst\_host\_srv\_serror\_rate', 'serror\_rate', 'srv\_rerror\_rate', 'dst\_host\_same\_srv\_rate', 'is\_guest\_login', 'dst\_host\_same\_src\_port\_rate', 'dst\_host\_srv\_count', 'dst\_host\_serror\_rate', 'num\_compromised', 'count', 'dst\_host\_rerror\_rate'}

Now we must drop these columns from original ***data*** to remove the collinearity between the features. Number of columns before dropping the correlated features was 42 and after deleting, we are left with 28 features.

For training the model, dataset must not have categorical variables. The columns {‘protocol\_type’, ‘service’, ‘flag’, ‘label’} in our dataset are having categorical variables so we

must convert these variables to numeric. A class named *LabelEncoder* is very useful for this purpose. It encodes the variables from categorical to numeric.

***LabelEncoder() :***

In machine learning, we usually deal with datasets which contains multiple labels in one or more than one column. These labels can be in the form of words or numbers. To make the data understandable or in human readable form, the training data is often labeled in words*.*

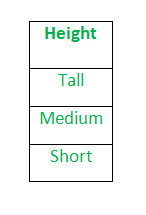
Label Encoding refers to converting the labels into numeric form to convert it into the machine-readable form. Machine learning algorithms can then decide in a better way on how those labels must be operated. It is an important pre-processing step for the structured dataset in supervised learning.

Label encoding convert the data in machine readable form, it assigns a unique number (starting from zero) to each class of data.

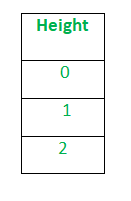
**For Example:**

Suppose we have a column Height in some dataset.

Now, convert the dataframe in the into a numpy matrix of rows and columns. Separate the input features and the label features from the matrix. I assigned ***dataX*** to the input features and ***dataY*** to the label feature.



After applying label encoding, the Height column is converted into:



In input features ***dataX*** column indexes 1, 2, 3 are categorical so I encoded them in numerical from using label encoder and the label ( ***dataY*** ) also encoded to numerical form.

***Step 2: Feature Selection***

Feature Selection is one of the core concepts in machine learning which hugely impacts the performance of your model. The data features that you use to train your machine learning models have a huge influence on the performance you can achieve. Irrelevant or partially relevant features can negatively impact model performance. Feature selection and Data cleaning should be the first and most important step of your model designing.

Feature Selection is the process where you automatically or manually select those features which contribute most to your prediction variable or output in which you are interested in. Having irrelevant features in your data can decrease the accuracy of the models and make your model learn based on irrelevant features.

***What are Benefits of performing feature selection before modeling your data?***

**· Reduces Overfitting:** Less redundant data means less opportunity to make decisions based on noise.

**· Improves Accuracy:** Less misleading data means modeling accuracy improves.

**· Reduces Training Time:** fewer data points reduce algorithm complexity and algorithms train faster.

Correlations Matrix is also a feature selection technique, but I have done this previously.

I have used 4 feature selection techniques:

* Correlation Matrix
* Univariate Feature Selection:
* Recursive Feature Elimination (RFE)
* Principle Component Analysis (PCA)
* Linear Discriminant Analysis (LDA)

**Univariate Feature Selection:**

Statistical tests can be used to select those features that have the strongest relationship with the output variable.The scikit-learn library provides the SelectKBest class that can be used with a suite of different statistical tests to select a specific number of features. I have used chi-squared(chi2) statistical test for non-negative features to select the best features from the dataset.

**Recursive Feature Elimination (RFE):**

Recursive feature elimination (RFE) is a feature selection method that fits a model and removes the weakest feature (or features) until the specified number of features is reached. Features are ranked by the model’s or feature\_importances\_ attributes, and by recursively eliminating a small number of features per loop, RFE attempts to eliminate dependencies and collinearity that may exist in the model. RFE requires a specified number of features to keep, however it is often not known in advance how many features are valid. RFE builds a model using the remaining attributes and calculates model accuracy. RFE is able to work out the combination of attributes that contribute to the prediction on the target variable (or class). Scikit Learn does most of the heavy lifting just import RFE from sklearn.feature\_selection and pass any classifier model to the RFE() method with the number of features to select. Using familiar Scikit Learn syntax, the .fit() method must then be called. I have used LogisticRegression classifier for RFE.

**Principle Component Analysis (PCA):**

Principal Component Analysis (PCA) is an unsupervised linear transformation technique that is widely used across different fields, most prominently for feature extraction and dimensionality reduction. Other popular applications of PCA include exploratory data analyses and de-noising of signals in stock market trading, and the analysis of genome data and gene expression levels in the field of bioinformatics.

PCA finds a new set of dimensions such that all the dimensions are orthogonal (and hence linearly independent) and ranked according to the variance of data along them. It means more important principle axis occurs first. (more important = more variance/more spread out data).

**How does PCA work -**

1. Calculate the covariance matrix X of data points.
2. Calculate eigen vectors and corresponding eigen values.
3. Sort the eigen vectors according to their eigen values in decreasing order.
4. Choose first k eigen vectors and that will be the new k dimensions.
5. Transform the original n dimensional data points into k dimensions.

[k is the number of features that are given to select.]

**Linear Discriminant Analysis (LDA):**

Linear Discriminant Analysis is a dimensionality reduction technique used as a preprocessing step in Machine Learning and pattern classification applications. The main goal of dimensionality reduction techniques is to reduce the dimensions by removing the redundant and dependent features by transforming the features from higher dimensional space to a space with lower dimensions. Linear Discriminant Analysis is a supervised classification technique which takes labels into consideration. This category of dimensionality reduction is used in biometrics, bioinformatics and chemistry.

The goal of Linear Discriminant Analysis is to project the features in higher dimension space onto a lower dimensional space.

This can be achieved in three steps:

The first step is to calculate the separability between different classes (i.e. the distance between the mean of different classes) also called as between-class variance.

Second Step is to calculate the distance between the mean and sample of each class, which is called the within class variance.

The third step is to construct the lower dimensional space which maximizes the between class variance and minimizes the within class variance.

***Step 3: Model***

Here, a function *acuu()* was made to find the accuracies by implementing the different algorithms on the dataset. This function takes two arguments *x, y.* ***x*** is input data and ***y*** is the label. Inside the function data was divided into training and testing set using train\_test\_split. 33% of data was separated as testing data from the original data given as arguments to the function.

Next step inside the function was feature scaling. Features scaling is a common requirement of machine learning methods, to avoid that features with large values may weight too much on the final results. For feature scaling Normalization was used. Normalization rescales the value into a range of [0,1] but the pattern between the values doesn’t change.

***Normalization:***

Normalization is a technique often applied as part of data preparation for machine learning. The goal of normalization is to change the values of numeric columns in the dataset to a common scale, without distorting differences in the ranges of values. For machine learning, every dataset does not require normalization. It is required only when features have different ranges.

Inside this function 5 different classifiers were used for experiment. Below is the list of all the algorithms that were used in the function:

1. Support Vector Machine
2. Perceptron
3. K-Nearest Neighbors
4. Stochastic Gradient Descent
5. XGBoost

Then, for model evaluation 2 methods were used-

1. Accuracy Score
2. K-fold Cross Validation

With each classifier *accuracy\_score()* function was used for finding the accuracy of that classifier and then *cross\_val\_score()* was also used for model evaluation with 10 folds.

***K-Fold Cross Validation:***

Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample.

The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross-validation. When a specific value for k is chosen, it may be used in place of k in the reference to the model, such as k=10 becoming 10-fold cross-validation.

Cross-validation is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data. That is, to use a limited sample in order to estimate how the model is expected to perform in general when used to make predictions on data not used during the training of the model.

It is a popular method because it is simple to understand and because it generally results in a less biased or less optimistic estimate of the model skill than other methods, such as a simple train/test split.

The general procedure is as follows:

1. Shuffle the dataset randomly.
2. Split the dataset into k groups
3. For each unique group:
4. Take the group as a hold out or test data set
5. Take the remaining groups as a training data set

1. Fit a model on the training set and evaluate it on the test set
2. Retain the evaluation score and discard the model.

4. Summarize the skill of the model using the sample of model evaluation scores

Importantly, each observation in the data sample is assigned to an individual group and stays in that group for the duration of the procedure. This means that each sample is given the opportunity to be used in the hold out set 1 time and used to train the model k-1 times.

***Experiments and Results:***

In this project 5 feature selection algorithms were used and the function *acuu()* was called 11 times.

First time, the dataset after deleting the correlation was given (which was having 28 features) to the function *acuu().* Function *acuu()* performed all the algorithms on the dataset and gave the accuracies and the graphs of accuracies given by the functions accuracy\_score() and cross\_val\_score() .

Then, on a random basis 17 features were selected from the dataset (which was having 28 features) by each of the four feature scaling algorithms and passed into the function one by one.

After that, 11 features were selected from the dataset (which was having 28 features) by each of the four feature scaling algorithms and passed into the function one by one.

Then, all the four feature scaling algorithms were used as a stack i.e., first 23 features were selected from the dataset (which was having 28 features) by Univariate Feature Selection and then from those 23 features, 20 best features were selected by Recursive Feature Elimination and then from those 20 features, 16 features were selected by Principal Component Analysis and then from those 16 features, 11 features were selected by Linear Discriminant Analysis and then those 11 features were passed as argument to the function *acuu().*

At last, 2 algorithms Recursive Feature Elimination and Linear Discriminant Analysis where chosen to perform feature scaling as a stack. First 17 columns from the dataset (which was having 28 features) were selected by Recursive Feature Elimination and then from those 17 features, 11 best features were selected by Linear Discriminant Analysis and those 11 features were passed as an argument to the function *acuu().*

Below, all the results from the function *acuu()* are given:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ***Classifier Used* Line arrow Straight**  ***Method Used*Line arrow Straight** | | **SVM** | **Perceptron** | **KNN** | **SGD** | **XGBoost** |
| After Dropping Correlations | ***No. of Columns*** | 78.53% | 78.28% | 99.88% | 92.31% | 10.23% |
| 28 |
| By Using Univariate Feature Selection | 17 | 78.85% | 80.65% | 99.86% | 92.21% | 22.29% |
| By Using RFE | 17 | 58.32% | 88.8% | 99.9% | 70.16% | 20.16% |
| By Using PCA | 17 | 22.6% | 22.17% | 99.87% | 24.97% | 8.2% |
| By Using LDA | 17 | 98.08% | 79.54% | 99.87% | 99.02% | 77.79% |
| By Using Univariate Feature Selection | 11 | 79.35% | 79.19% | 99.84% | 88.46% | 18.99% |
| By Using RFE | 11 | 27.78% | 27.79% | 99.28% | 19.74% | 76.86% |
| By Using PCA | 11 | 22.16% | 22.17% | 99.87% | 24.98% | 11.02% |
| By Using LDA | 11 | 97.73% | 79.08% | 99.87% | 98.98% | 74.8% |

**Accuracies by using mehod *sklearn.metrics.*accuracy\_score():**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| By using all the selection algorithms as a stack | 11 | 80.26% | 74.01% | 99.91% | 98.5% | 59.44% |
| By using RFE and LDA as a stack | 11 | 92.43% | 79.26% | 99.88% | 98.95% | 75.49% |

References:

<https://medium.com/@urvashilluniya/why-data-normalization-is-necessary-for-machine-learning-models-681b65a05029>

<https://machinelearningmastery.com/k-fold-cross-validation/>