Internship Report

On

NETWORK ANALYSIS USING MACHINE LEARNING

Submitted in partial fulfilment of the requirements for the award of the degree

of

Bachelor of Technology

in

Computer Science & Engineering

by

Anil Kumar	**(Roll No 180101005)**
Kunal	**(Roll No 180101027)**
Ravi Rajesh Keer	**(Roll No 180101037)**
Sandeep Kumar	**(Roll No 180101040)**

Under the esteemed Supervision

of

Dr. Thejaswini M

Department of Computer Science & Engineering

Indian Institute of Information Technology Bhagalpur **October, 2021**

M. Thejaswini
14/11/21

Abstract

In this project, we were asked to experiment with a real world dataset, and to explore how machine learning algorithms can be used to find different network protocol based in data. We were expected to gain experience using a common data-mining and machine learning library and were expected to submit a report about the dataset and the algorithms used. After performing the required tasks on a dataset of my choice, herein lies my final report.

Keywords: Machine Learning, Network protocol, Classification, Supervised learning (SL), Random Forest Model

Motivation

The main aim of the project is to build a machine learning model which can predict types of protocol with the help of given input data set that contains all most 84 different features like packet length, flow duration, segment size and many more. We have done this project to gain some real life experience on machine learning and to get practical knowledge on various aspect of machine learning models like SVM kernel, Decision Tree and Random Forest. With such a wide area covered in the project, it is going to help us later in many industrial project. Thus , this project will provide flexibility and experience in development of future machine learning models.

Index

Learning Objectives/Internship Objectives

- Internships are generally thought of to be reserved for college students looking to gain experience in a particular field. However, a wide array of people can benefit from Training Internships in order to receive real world experience and develop their skills.
- An objective for this position should emphasize the skills you already possess in the area and your interest in learning more.
- Internships are utilized in a number of different career fields, including architecture, engineering, healthcare, economics, advertising and many more.
- Some internship is used to allow individuals to perform scientific research while others are specifically designed to allow people to gain first-hand experience working.
- Utilizing internships is a great way to build your resume and develop skills that can be emphasized in your resume for future jobs. When you are applying for a Training Internship, make sure to highlight any special skills or talents that can make you stand apart from the rest of the applicants so that you have an improved chance of landing the position.

Chapter 1: Introduction

1.1 Problem Definition

This project is all about predicting the types of a new protocol with the help of given input data set that contains all most 84 different features like packet length, flow duration, segment size, header length and many more.

1.2 Task

We were assigned with the task to come up with an efficient machine learning model which will result in maximum accuracy on a given data set.

1.3 Action

To do this work in an organized way, we enlisted the entire relevant classification machine learning algorithm on a paper which fits the problem like SVM kernal, Decision Tree and Random forest. After that we started to analyze the performance of all those models meticulously

1.4 Result

Eventually we got maximum accuracy of 98% on both training set and test set with help of random forest.

Chapter 2: Requirements

2.1 Technology Used

Python

Python is an interpreted high-level general- purpose programming language. Its design philosophy emphasizes code readability with its use of significant indentation. Its language constructs as well as its object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects.

Machine Learning

As our project title is network analysis using machine learning we used random forest machine learning algorithm to build this model**.**

Data Science

Data science is an **interdisciplinary field** that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from noisy, structured and unstructured data, and apply knowledge and actionable insights from data across a broad range of application domains.

2.2 Libraries

Numpy

Numpy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.

Pandas

A panda is a software library written for the Python programming language for data manipulation and analysis.

Sklearn

Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python.

Matplotlib

Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy.

Seaborn

Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

2.3 Software used:

- Notepad++
- Microsoft Office Word
- Google Colab
- Window
- Ubuntu

Chapter 3: Graph Exploratory Analysis

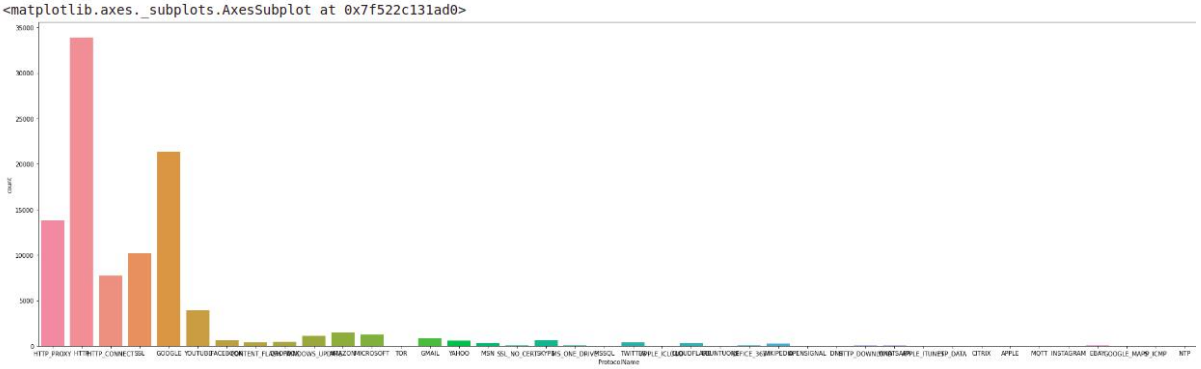
- It's also known as Exploratory Data analysis (EDA).
- It refers to the critical process of performing Initial investigations on data.
- It's used to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

3.1 Data Visualization

- Count plot of protocol name or protocol labels.

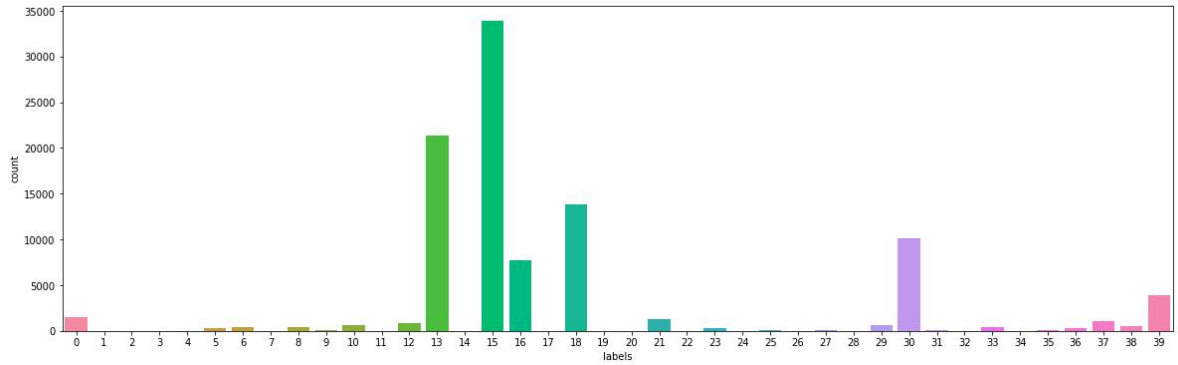
```
plt.figure(figsize=(35, 10))
sns.countplot(x="ProtocolName", data=df1)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f522c131ad0>
```



```
plt.figure(figsize=(20, 6))
sns.countplot(x='labels', data=df1)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f522b2a1cd0>
```

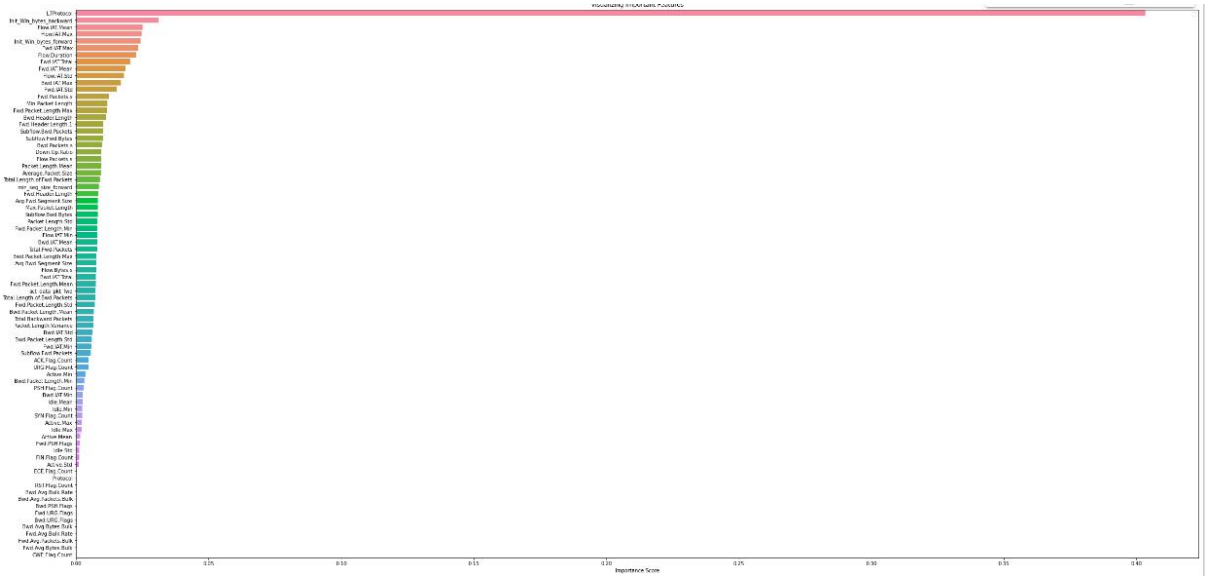


- Visualizing important feature (importance score vs attribute names)

```
feature_imp = pd.Series(clf.feature_importances_,index=df2.columns[:-1]).sort_values(ascending=False)

print(feature_imp)
print()

# Creating a bar
plt.figure(figsize=(40, 20))
sns.barplot(x=feature_imp, y=feature_imp.index)
# Add labels to your graph

plt.xlabel('Importance Score')
plt.ylabel('Attribute Names')

plt.title("Visualizing Important Features")
plt.legend()

plt.show()
```

```
L7Protocol                0.403257
Init_Win_bytes_backward   0.031210
Flow.IAT.Mean             0.025007
Flow.IAT.Max              0.024691
Init_Win_bytes_forward    0.024266
                            ...
Bwd.Avg.Bytes.Bulk        0.000000
Fwd.Avg.Bulk.Rate         0.000000
Fwd.Avg.Packets.Bulk      0.000000
Fwd.Avg.Bytes.Bulk        0.000000
```

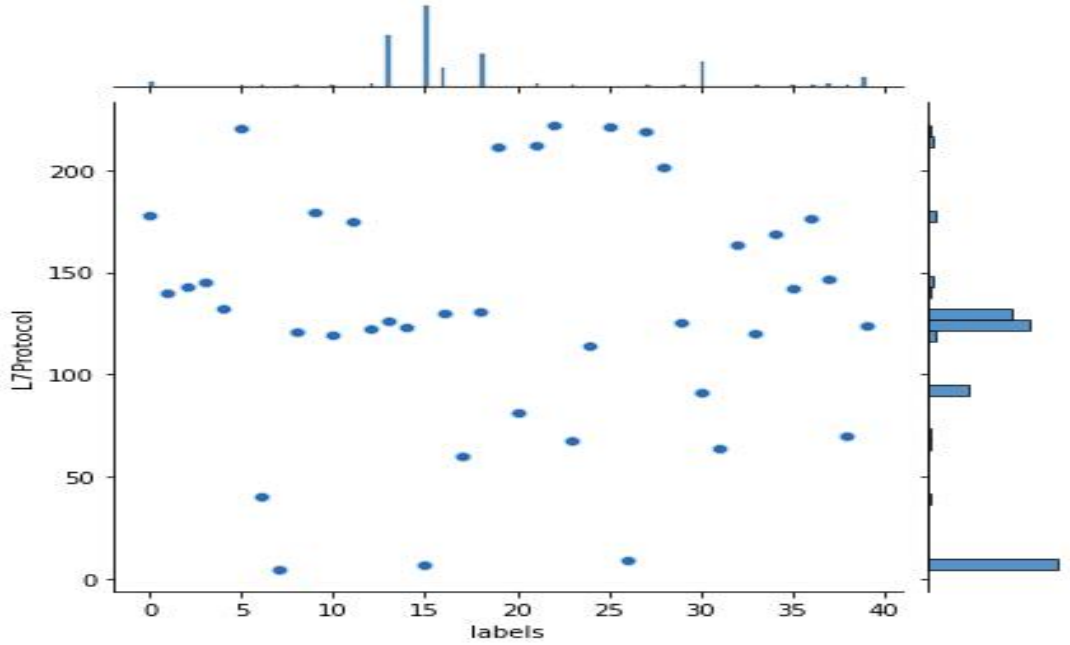
2s completed at 22:36



- Joint plot between labels and L7 protocol

```
sns.jointplot(x='labels', y="L7Protocol", data=df2)
```

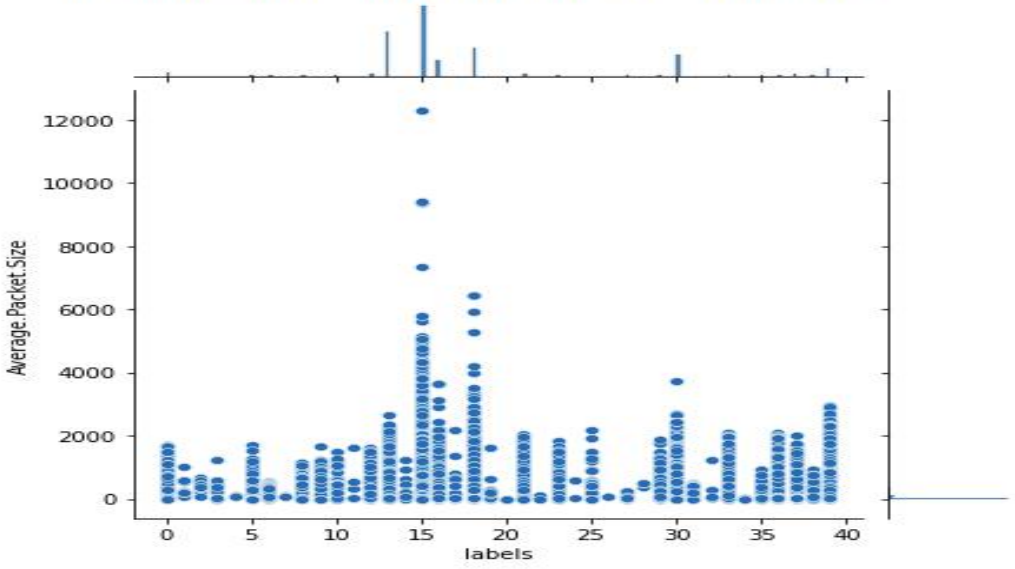
```
<seaborn.axisgrid.JointGrid at 0x7f522aa6f310>
<Figure size 720x432 with 0 Axes>
```



- Joint plot between labels and avg.packet.size

```
sns.jointplot(x="labels", y="Average.Packet.Size", data=df2)
```

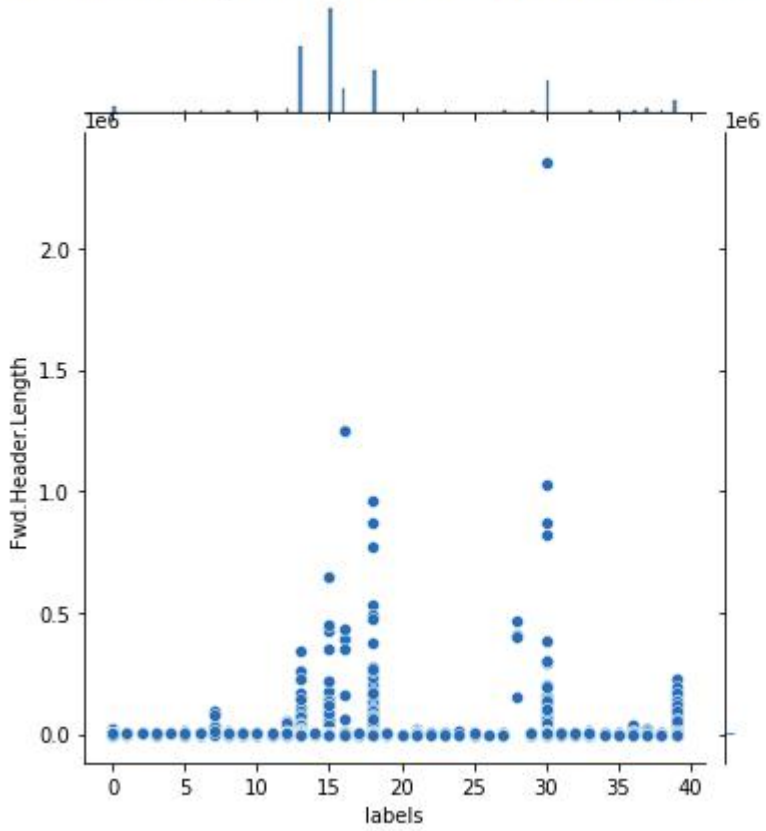
```
<seaborn.axisgrid.JointGrid at 0x7f522addf250>
```



- Joint plot between labels and forward.header.length

```
sns.jointplot(x='labels', y='Fwd.Header.Length', data=df2)
```

```
<seaborn.axisgrid.JointGrid at 0x7fc5c2e258d0>
```



Chapter 4: Model description and Intuition

4.1 Model description (Random forest):

- What is random forest?

Random forest is a supervised learning algorithm. The "forest" it builds, is an ensemble of decision trees, usually trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result.

- How Random Forest Works?

Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction

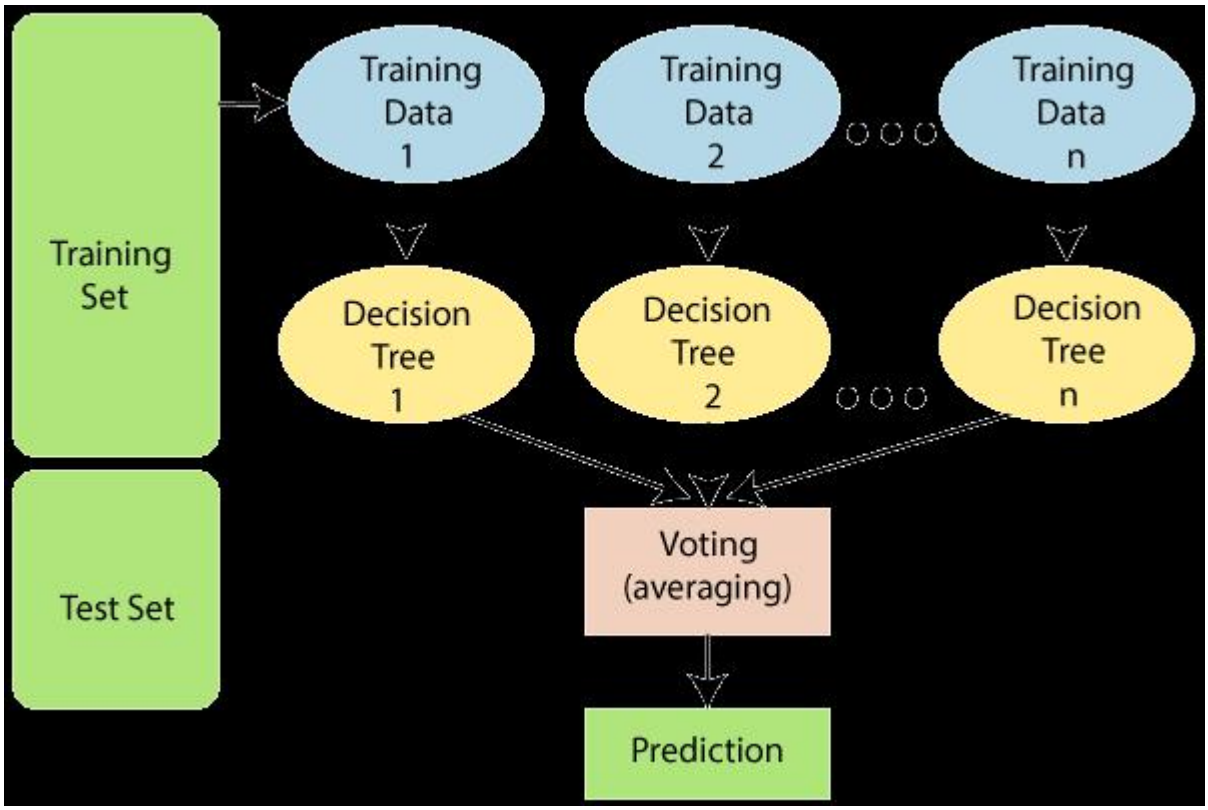
4.2 Model Intiution (Random Forest)

STEP 1: Pick at random K data points from the Training set.

STEP 2: Build the Decision Tree associated to these K data points.

STEP 3: Choose the number Ntree of trees you want to build and repeat STEPS 1 & 2

STEP 4: For a new data point, make each one of your Ntree trees predict the category to which the data points belongs, and assign the new data point to the category that wins the majority vote.



Chapter 5: Implementations

5.1 Importing Libraries and Data set

We imported a few libraries of python which are necessary to train that model like numpy, panda, matplotlib, sklearn etc.

After importing the libraries, we needed a dataset to train our model. So, we took the dataset from the open source website kaggle and then we imported the dataset into a data frame using the read_csv function of panda library. That dataset is a really large dataset. In this dataset there are 87 features like source ip, source port, destination ip, destination port, protocol, flow duration and many more

5.2 Converting the Categorical Data into Numeric Representation

After importing the dataset, we are required to convert the categorical i.e. text features to its numeric representation. Because machine learning algorithms perform better in terms of accuracy and other performance metrics when the data is represented as a number instead of categorical to a model for training and testing. So, we did label encoding of the feature protocol name. After this we removed some text features like source ip, destination ip, timestamp which are not required or you can say they can give a bad impact on the accuracy of the model.

5.3 Splitting the dataset

We needed to split our data set into a training set and test set. Generally we separate our data set into a training set and test set, most of the data is used for training, and a smaller portion of the data set is used for testing. So, we have used 75% of data for training purposes and 25% data for testing purposes for better accuracy.

5.4 Feature Scaling

After splitting the data set, we did feature scaling. In this large data set there are high differences b/w two values for some features and this can give a bad impact on accuracy. So, we needed to minimize that difference. To do the same we did feature scaling using standard deviation algorithms to get better accuracy.

5.5 Training Model

After feature scaling, we were all set to train our model using a training set using a random classifier algorithm. We had a large dataset so we needed to increase the no. of decision trees in our random forest to overcome the problem of over fitting. And we also prevented the under fitting problem by not building too many decision trees. We built a forest of 100 decision trees instead of 10 or 1000. So that it will not fall into the category of over fitting problem as well as under fitting problem and will give better prediction.

5.6 Predicting Result and Finding Accuracy

And after completion of training of that model, first we predicted the protocol for a particular data using that trained model and then we predicted the protocol of all the data of the test set and stored the predicted results. And we also stored the actual results.

And with the help of these predicted results and actual results, we built the confusion matrix. After building the confusion matrix we checked the accuracy of that model with the help of the confusion matrix and we got 98% accuracy.

5.7 Visualization

After this, we did data visualization. In data visualization we were visualizing which feature is contributing maximum to predict the result for a particular data. And after that we are sorting the all features in the order that which feature is contributing maximum in more no. of data. So we got that the feature L7 protocol is contributing maximum in majority of the data and we also got the order and plotted it.

5.8 Conclusion

We have successfully implemented the random forest machine learning model on a given network data set, which predict the types of protocol with an accuracy of 98 percentages.

References

1. Kaggle

https://towardsdatascience.com/random-forests-algorithm-explained-with-a-real-life-example-and-some-python-code-affbfa5a942c

2. Medium

https://www.kaggle.com/ar9avgupta/little-bit-cleaning-and-fitting-to-random-forest

3. Built In

https://builtin.com/data-science/random-forest-algorithm