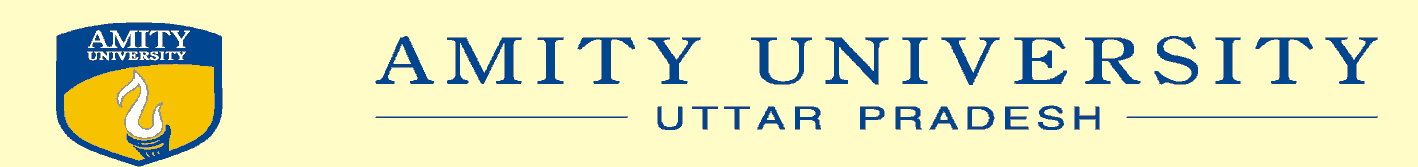
****

**AMITY SCHOOL OF ENGINEERING AND TECHNOLOGY**

**INTERNAL**

**NTCC**

**Weekly Draft**

**Topic:**

Introduction to the topics “Big Data Analytics, IOT, AI and Cloud Computing Technologies for Intelligent Network Performance and QoS”

**Problem Statement**:

Problem statement for this week is to

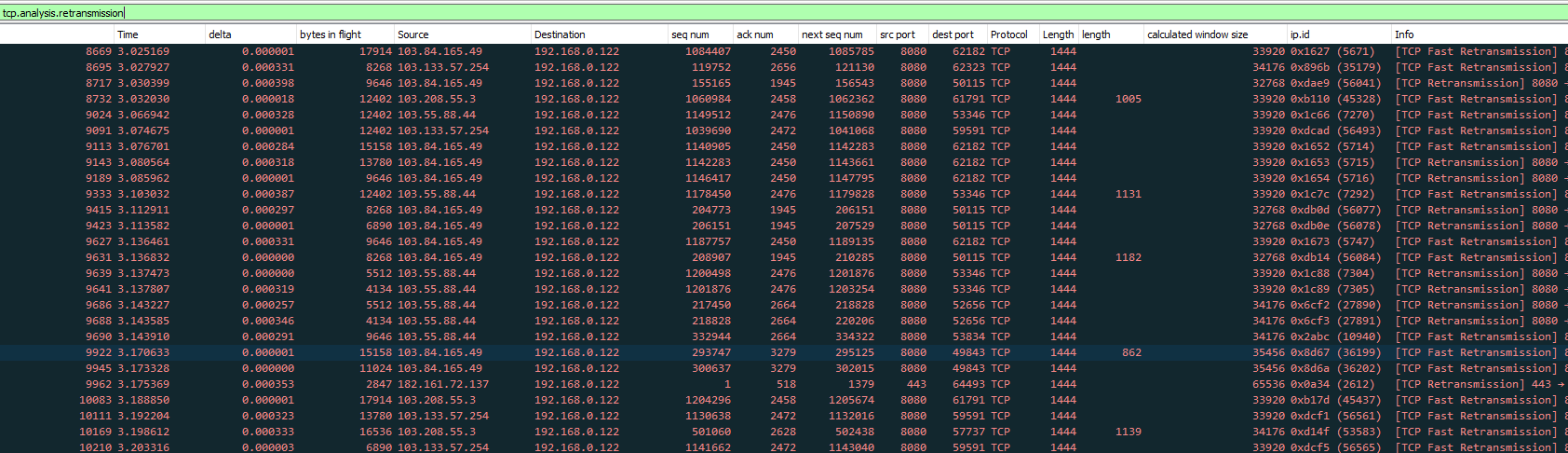
* Collect data-set regarding the various internet traffic components like jitter, throughput, source ip, destination ip, tcp traces etc. and use the learned AI classifiers
* Understanding prediction models and trying practical application on the said dataset.
* Try to predict packet loss

**My Work**:

Wireshark is a free open-source packet analyzer. It is used for troubleshooting network, analysis, and software and communication protocol development.

Wireshark lets the user put network interface controllers into promiscuous mode, so they can see all the traffic visible on that interface including unicast traffic not sent to that network interface controller's MAC address. However, when capturing with a packet analyzer in promiscuous mode on a port on a network switch, not all traffic through the switch is necessarily sent to the port where the capture is done, so capturing in promiscuous mode is not necessarily sufficient to see all network traffic. Port mirroring or various network taps extend capture to any point on the network. Simple passive taps are extremely resistant to tampering

For this week I have tried to get hands over some already existing network traffic datasets which was not an easy task as getting the required dependent and independent fields is a tough task. I have used data set from ‘Tstat’ networking tool and also dataset from wireshark which is already filtered to give out only lost or unreached packets of data. This was achieved by using the following display filter known as ‘tcp.analysis.retransmission’ this filter gives out the tcp packets that were retransmitted because of packet loss. The data set is given in the github repository [https://github.com/Rc17git/NTCC]



TYPES OF CLASSIFIERS

There are different types of classifiers, a classifier is an algorithm that maps the input data to a specific category.

Classification is the process of predicting the class of given data points. Classes are sometimes called as targets/ labels or categories. Classification predictive modeling is the task of approximating a mapping function (f) from input variables (X) to discrete output variables (y).

1. Perceptron
2. Naive Bayes
3. Decision Tree
4. Logistic Regression
5. K-Nearest Neighbor
6. Artificial Neural Networks/Deep Learning
7. Support Vector Machine

TYPES OF DATA ANALYTICS MODELS

**Linear Regression**

Linear regression is one of the most basic algorithms of advanced analytics. This also makes it one of the most widely used. People can easily visualize how it is working and how the input data is related to the output data.

Linear regression uses the relationship between two sets of continuous quantitative measures. The first set is called the predictor or independent variable. The other is the response or dependent variable. The goal of linear regression is to identify the relationship in the form of a formula that describes the dependent variable in terms of the independent variable. Once this relationship is quantified, the dependent variable can be predicted for any instance of an independent variable.

**Logistic Regression**

Logistic regression sounds similar to linear regression but is actually focused on problems involving categorization instead of quantitative forecasting. Here the output variable values are discrete and finite rather than continuous and with infinite values as with linear regression.

The goal of logistic regression is to categorize whether an instance of an input variable either fits within a category or not. The output of logistic regression is a value between 0 and 1. Results closer to 1 indicate that the input variable more clearly fits within the category. Results closer to 0 indicate that the input variable likely does not fit within the category.

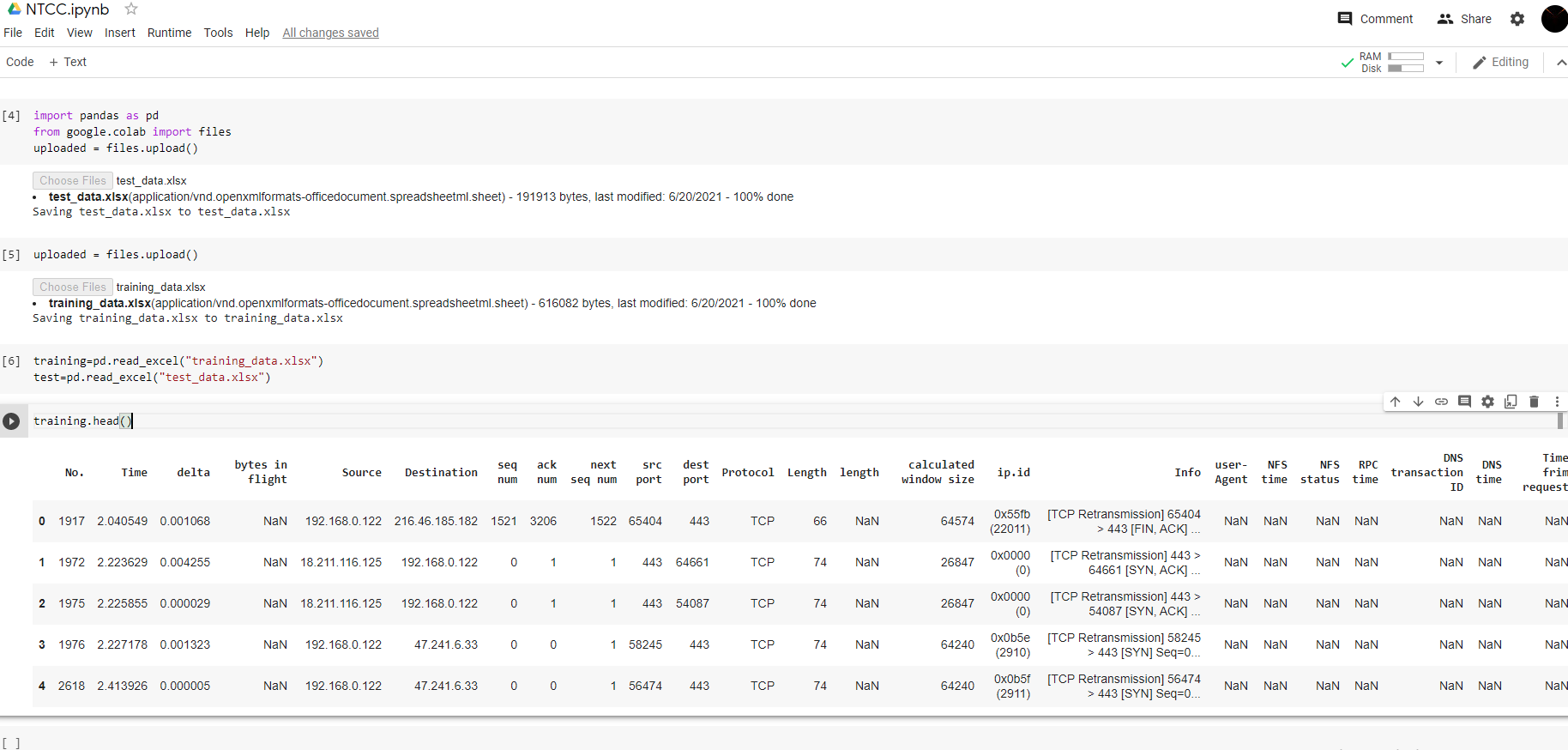
**Classification and Regression Trees**

Classification and regression trees use a decision to categorize data. Each decision is based on a question related to one of the input variables. With each question and corresponding response, the instance of data gets moved closer to being categorized in a specific way. This set of questions and responses and subsequent divisions of data create a tree-like structure. At the end of each line of questions is a category. This is called the leaf node of the classification tree.

These classification trees can become quite large and complex. One method of controlling the complexity is through pruning the tree or intentionally removing levels of questioning to balance between exact fit and abstraction. A model that works well with all instances of input values, both those that are known in training and those that are not, is paramount. Preventing overfitting of this model requires a delicate balance between exact fit and abstraction.

A variant of classification and regression trees is called random forests. Instead of constructing a single tree with many branches of logic, a random forest is a culmination of many small and simple trees that each evaluate the instances of data and determine a categorization. Once all of these simple trees complete their data evaluation, the process merges the individual results to create a final prediction of the category based on the composite of the smaller categorizations. This is commonly referred to as an ensemble method. These random forests often do well at balancing exact fit and abstraction and have been implemented successfully in many business cases.

WORK ON GOOGLE COLLAB



Trying to apply a classifier or a predictive model over this data set is quiet a difficult task and hence has not been achieved yet

Prediction of packet loss with the present data set is not achievable because of lack of information about the “packet loss dependent variables”. I am unable to identify those variables, the ones I can are actually hard to record

Throughput is one of the variable on which packet loss is dependent. Throughput is the amount of data or packets that are received successfully. The following throughput graph tells about the lost packets.

