**Week 1: Understanding the problem statement, data analysis, preprocessing and understanding data / features from the domain perspective**

The problem statement is on binary classification. The dataset used here was of NSL-KDD. It contained the following files:

1. **KDDTrain+.ARFF**: The full NSL-KDD train set with binary labels in ARFF format
2. **KDDTrain+.TXT**: The full NSL-KDD train set including attack-type labels and difficulty level in CSV format
3. **KDDTrain+\_20Percent.ARFF**: A 20% subset of the KDDTrain+.arff file
4. **KDDTrain+\_20Percent.TXT**: A 20% subset of the KDDTrain+.txt file
5. **KDDTest+.ARFF**: The full NSL-KDD test set with binary labels in ARFF format
6. **KDDTest+.TXT**: The full NSL-KDD test set including attack-type labels and difficulty level in CSV format
7. **KDDTest-21.ARFF**: A subset of the KDDTest+.arff file which does not include records with difficulty level of 21 out of 21
8. **KDDTest-21.TXT**: A subset of the KDDTest+.txt file which does not include records with difficulty level of 21 out of 21

I have used KDDTrain+.ARFF and KDDTest+.ARFF data files as part of understanding the data and performing the Exploratory Data Analysis (EDA). KDDTest-21.ARF is a part of KDDTest+.ARFF data.

**Data**: The dataset consists of 42 features out of which 41 features are input traffic and the last attribute is target variable. i.e. class (labeled as normal and anomaly). There are 3 features of categorical type. i.e. protocol\_type, service, flag. All the features are as follows:

1. Duration: The length of time of the connection.
2. Protocol type: The type of protocol used in the connection (e.g., TCP, UDP, ICMP).
3. Service: The network service on the destination machine (e.g., http, ftp, telnet).
4. Flag: The status of the connection (e.g., normal, error, intrusion).
5. Source bytes: The number of data bytes sent by the source host.
6. Destination bytes: The number of data bytes sent by the destination host.
7. Land: Indicates if the connection is from/to the same host/port.
8. Wrong fragment: Indicates if there is a wrong fragment in the packet.
9. Urgent: Indicates if the connection has urgent data.
10. Hot: Indicates if the connection is flagged as "hot."
11. Num failed logins: The number of failed login attempts.
12. Logged in: Indicates if a user is successfully logged in.
13. Num compromised: The number of compromised conditions.
14. Root shell: Indicates if a root shell is obtained.
15. Su attempted: Indicates if the "su" command is attempted.
16. Num root: The number of root accesses.
17. Num file creations: The number of file creation operations.
18. Num shells: The number of shell prompts.
19. Num access files: The number of access file operations.
20. Num outbound cmds: The number of outbound commands in an ftp session.
21. Is host login: Indicates if the login belongs to the "host" category.
22. Is guest login: Indicates if the login belongs to the "guest" category.
23. Count: The number of connections to the same host as the current connection.
24. Srv count: The number of connections to the same service as the current connection.
25. Serror rate: The percentage of connections that have "SYN" errors.
26. Srv serror rate: The percentage of connections to the same service that have "SYN" errors.
27. Rerror rate: The percentage of connections that have "REJ" errors.
28. Srv rerror rate: The percentage of connections to the same service that have "REJ" errors.
29. Same srv rate: The percentage of connections to the same service.
30. Diff srv rate: The percentage of connections to different services.
31. Srv diff host rate: The percentage of connections to different hosts.
32. Dst host count: The number of connections to the same destination host.
33. Dst host srv count: The number of connections to the same service on the same destination host.
34. Dst host same srv rate: The percentage of connections to the same service on the same destination host.
35. Dst host diff srv rate: The percentage of connections to different services on the same destination host.
36. Dst host same src port rate: The percentage of connections from the same source port to the same destination host.
37. Dst host srv diff host rate: The percentage of connections to the same service but different hosts on the same destination host.
38. Dst host serror rate: The percentage of connections to the same destination host that have "SYN" errors.
39. Dst host srv serror rate: The percentage of connections to the same service on the same destination host that have "SYN" errors.
40. Dst host rerror rate: The percentage of connections to the same destination host that have "REJ" errors.
41. Dst host srv rerror rate: The percentage of connections to the same service on the same destination host that have "REJ" errors.
42. Class: Type of attack: normal or anomaly

Steps followed to perform EDA:

1. Tools used for development: Anaconda prompt, Jupyter notebook.
2. Importing the necessary packages/python libraries and loading the test and train datasets into the Jupyter notebook.
3. Generate dataframe for each of these datasets and check for the information, data types, columns, shape, size, describe the dataset and check for count of object occurring in a list.
4. Verify the KDDTrain .arff file and KDDTrain+\_20Percent.arff file.
5. Verify the KDDTest .arff file and KDDTest-21.arff file.
6. Check if any null values are present in both the datasets. There are no null values present in both the datasets.
7. Find the number of unique values in each column on the dataset. dst\_bytes: Destination bytes have a greater number of unique values followed by src\_bytes: source bytes
8. Find the correlation between different features.
9. The variables like is\_guest\_login and hot , srv\_serror\_rate and serror\_rate, dst\_host\_error\_rate and serror\_rate, dst\_host\_rerror\_rate and serror\_rate, dst\_host\_error\_rate and srv\_serror\_rate, dst\_host\_rerror\_rate and srv\_serror\_rate, dst\_host\_rerror\_rate and rerror\_rate, dst\_host\_srv\_rerror\_rate and rerror\_rate, etc. have a highly positive correlation. dst\_host\_serror\_rate and same\_srv\_rate, dst\_host\_srv\_serror\_rate and same\_srv\_rate, serror\_rate and same\_srv\_rate, srv\_serror\_rate and same\_srv\_rate, etc. have a highly negative correlation.
10. Sort the correlated features inside train and test dataset.
11. Removing the highly positive correlations. Removing features with high positive correlation is necessary to improve the efficiency and accuracy of the prediction model. It can help improve the performance of your machine-learning models.
12. Classify the categorical features present in train and test dataset.
13. Performing the visualization for different insights from the train and test dataset.

Steps to get insights from Data exploration:

1. Collecting the count of different categorical variables like service, protocol\_type, flag, class for train and test dataset.  Here the values of TCP protocol, http network service and SF flag are maximum. Also normal class (good connections) is having more percentage compared to attacks/intrusions (bad connections).
2. Visualization for these categorical variables was carried out using bar plots, line plots present in matplotlib package.
3. Finding the count of the number of connections to the same host as the current connection for different services (count). Here http service is having maximum value for the number of connections to the same host as the current connection.
4. Finding the count of the percentage of connections that have "SYN" errors for different services (serror\_rate). Here http service is having most percentage of connections with SYN errors
5. Finding the count for the status of connection with respect to the percentage of connections to the same service (same\_srv\_rate). Here flag SF is having more percentage of connections to the same service.
6. Finding the count of the number of connections to the same service on the same destination host with respect to the type of protocol used in the connection, the status of connection and the network service on the destination machine (dst\_host\_srv\_count). Here TCP, flag SF, http service is having maximum value for the number of connections to the same service on the same destination host.
7. Finding the count of the percentage of connections to the same destination host that have "REJ" errors with respect to the type of protocol used in the connection, the status of connection and the network service on the destination machine (dst\_host\_rerror\_rate). Here TCP protocol, http service, SF flag are having maximum entries for dst\_host\_rerror\_rate.
8. Finding the count of the percentage of connections to the same service on the same destination host with respect to the type of protocol used in the connection, the status of connection and the network service on the destination machine (dst\_host\_same\_srv\_rate). Here TCP protocol, http service, SF flag are having maximum value for the percentage of connections to the same service on the same destination host.
9. Finding the count of the percentage of connections to different services on the same destination host with respect to the type of protocol used in the connection, the status of connection and the network service on the destination machine (dst\_host\_diff\_srv\_rate). Here TCP protocol, http service, SF flag are having maximum value for the percentage of connections to the same service on the same destination host.
10. Finding the count of the percentage of connections from the same source port to the same destination host with respect to the type of protocol used in the connection, the status of connection and the network service on the destination machine (dst\_host\_same\_src\_port\_rate).
11. Checking for the count if user logged in successfully with respect to the type of protocol used in the connection, the status of connection and the network service on the destination machine (logged\_in). Here the number of successful login attempts are less.
12. Checking for the count serror\_rate (percentage of connections that have "SYN" errors) to same\_srv\_rate (percentage of connections to the same service).
13. Checking for the count rerror\_rate (The percentage of connections that have "REJ" errors) to dst\_host\_rerror\_rate (The percentage of connections to the same destination host that have "REJ" errors)
14. Checking for the count dst\_host\_same\_srv\_rate (The percentage of connections to the same service on the same destination host) to dst\_host\_srv\_count (The number of connections to the same service on the same destination host)
15. Checking for the Count dst\_host\_diff\_srv\_rate (The percentage of connections to different services on the same destination host) to dst\_host\_same\_src\_port\_rate (The percentage of connections from the same source port to the same destination host)

After performing data exploratory steps and data visualization, we select important attributes for feature selection. The following features are selected:

1. service,
2. flag,
3. protocol\_type,
4. count,
5. logged\_in,
6. serror\_rate,
7. rerror\_rate,
8. same\_srv\_rate,
9. dst\_host\_srv\_count,
10. dst\_host\_rerror\_rate,
11. dst\_host\_same\_srv\_rate,
12. dst\_host\_diff\_srv\_rate,
13. dst\_host\_same\_src\_port\_rate