

**Telecommunication Software**

Fifth Practical Exercise

Full name: **Mithin kumar Ananthula**

ID: **241AEM014**

**Riga 2024**

**Contents**

[Task 1: Binary Search Tree………………………………………………………………... 3](#_Toc184089837)

[Task 2: SDN Traffic classification with Decision Tree ……………………………………5](#_Toc184089838)

Task 3: 3 Intrusion Detection………………………………………………………………30

Task 4: Week's Network Traffic Prediction……………………………………………….30

Git hub link ………………………………………………………………………………..30

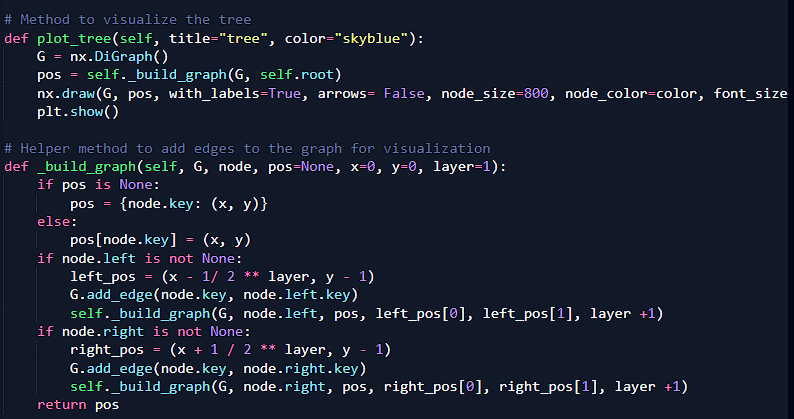
**Task 1: Binary Search Tree**

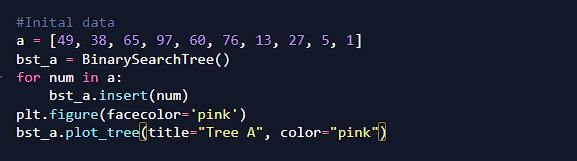
Please create a binary search tree and node classes as I have shown in the lecture, including search, insert, delete, and traverse functions, using the following three lists to create the tree and combine the above-created function. For each step, please give an output or changed tree plot.

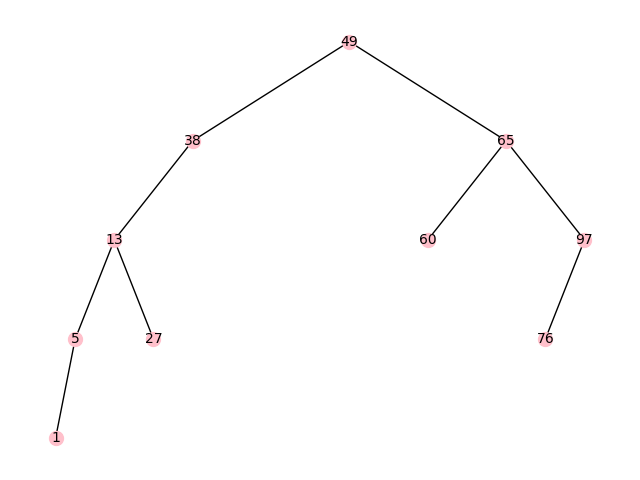
1. **Binary Search Tree** of a = [49, 38, 65, 97, 60, 76, 13, 27, 5, 1]

Main code:

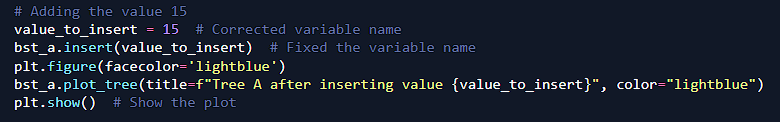
****

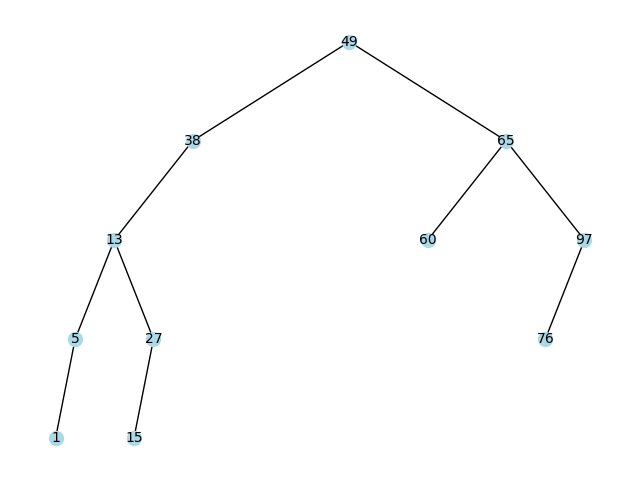
****

****

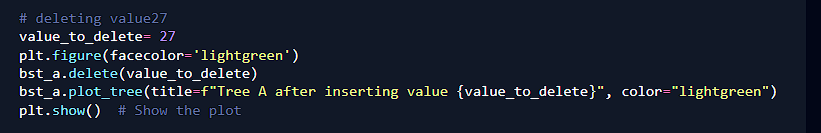
****

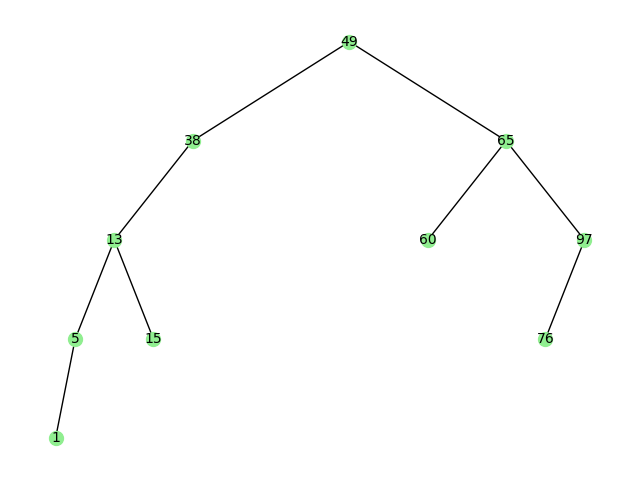
Adding 15 to the binary tree

****

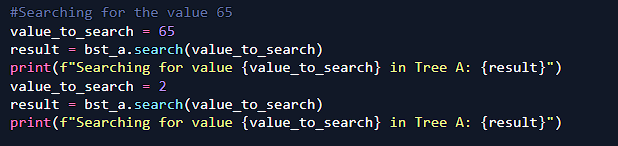
****

Deleting the value 27 from the binary tree

****

****

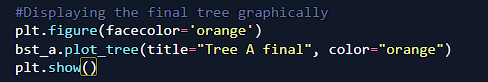
Search function

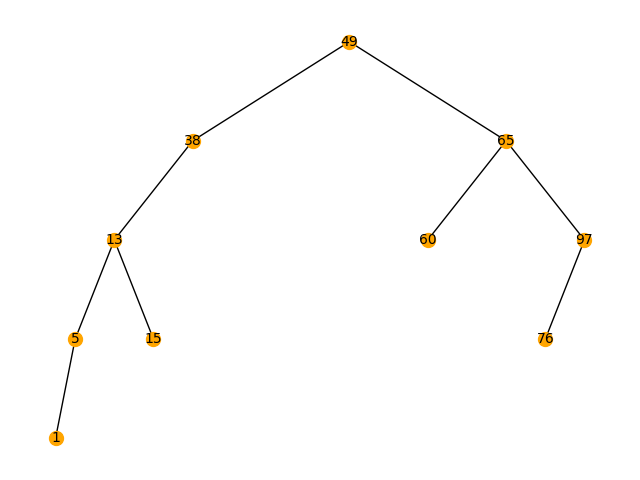
****

Searching result

****

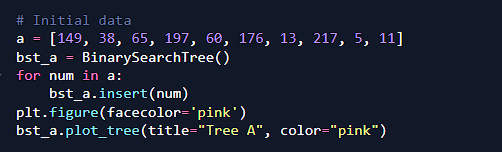
Final tree after the changes

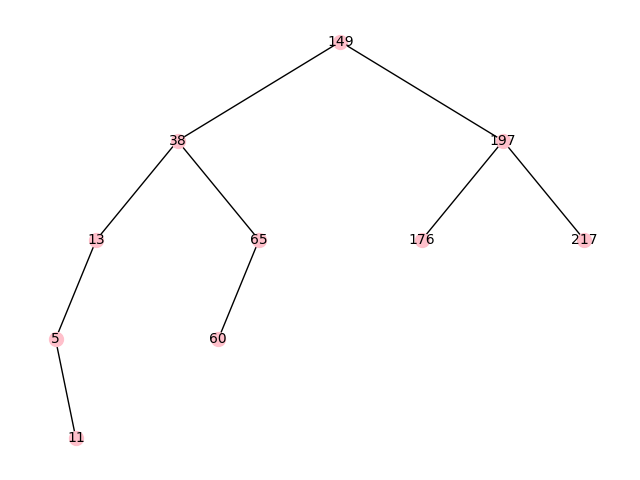
****

****

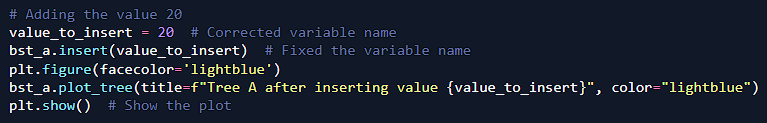
1. **Binary Search Tree** of a = [149, 38, 65, 197, 60, 176, 13, 217, 5, 11]

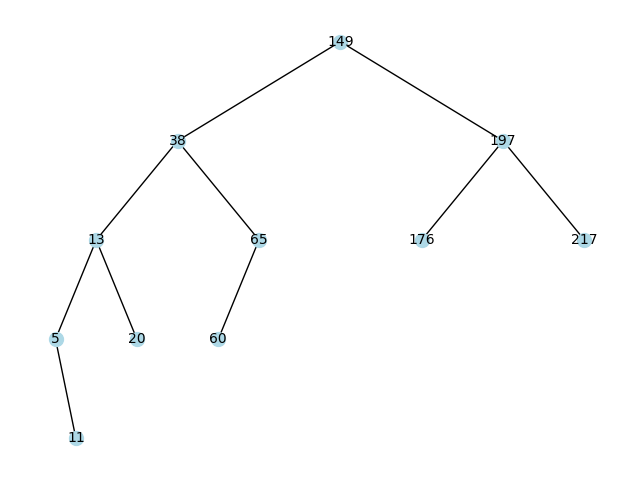
Same code as the above (Main code)

****

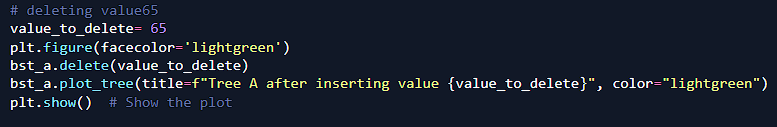
****

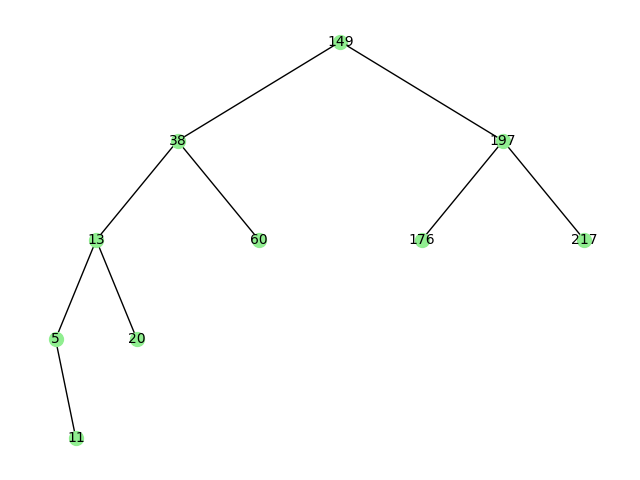
Adding 20 to the binary tree

****

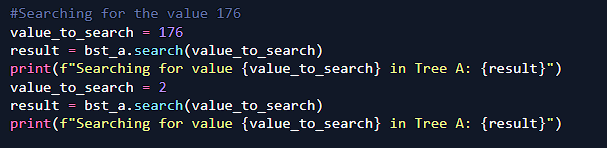
****

Deleting the value 65 from the binary tree

****

****

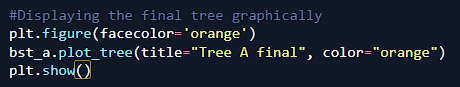
Search function

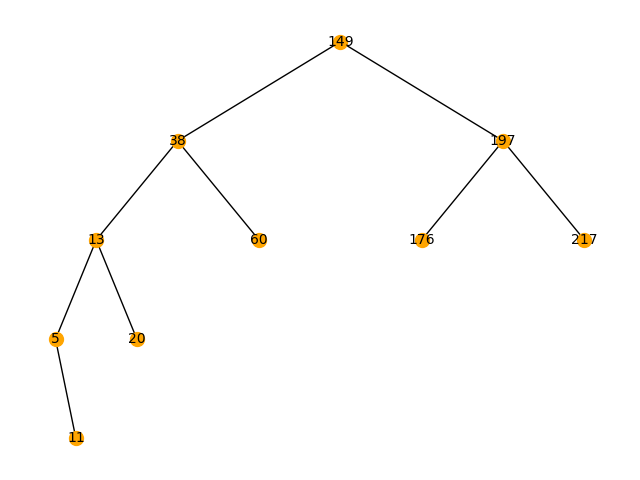
****

Searching result

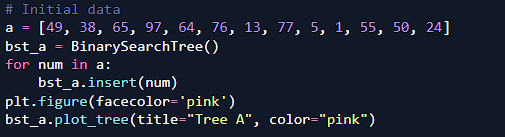
****

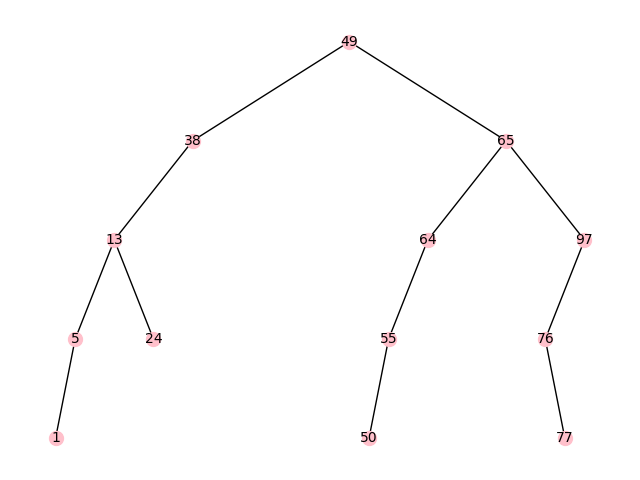
Final tree after the changes

****

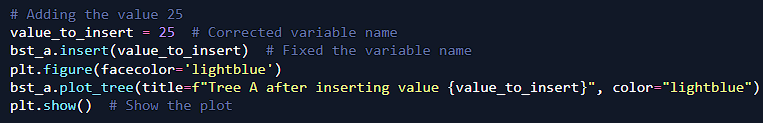
****

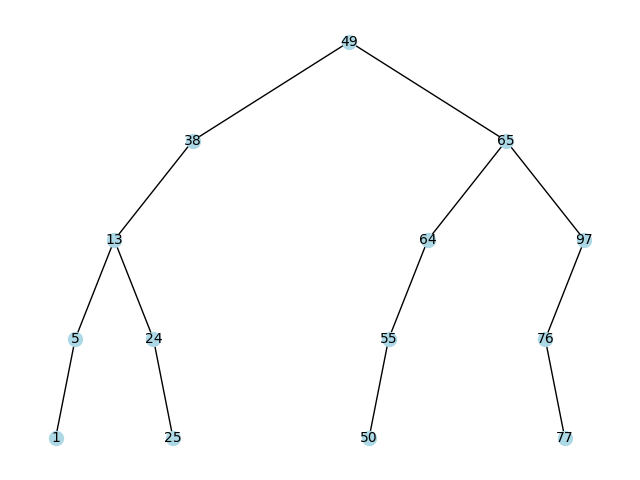
**30 Binary Search Tree** ofa = [49, 38, 65, 97, 64, 76, 13, 77, 5, 1, 55, 50, 24]

****

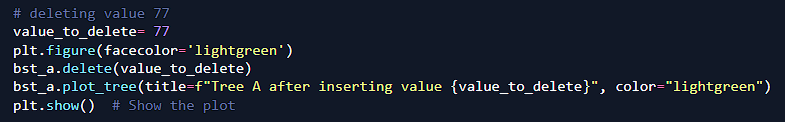
****

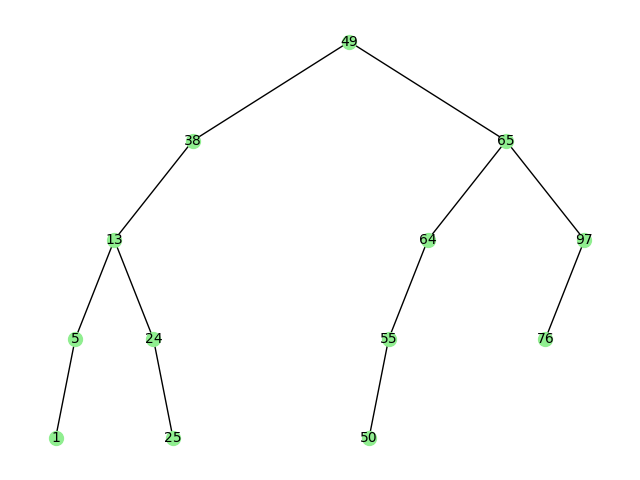
Adding 25 to the binary tree

****

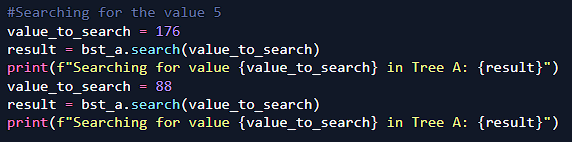
****

Deleting the value 27 from the binary tree

****

****

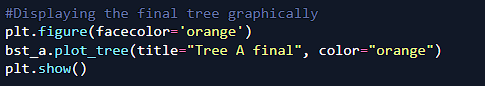
Search function

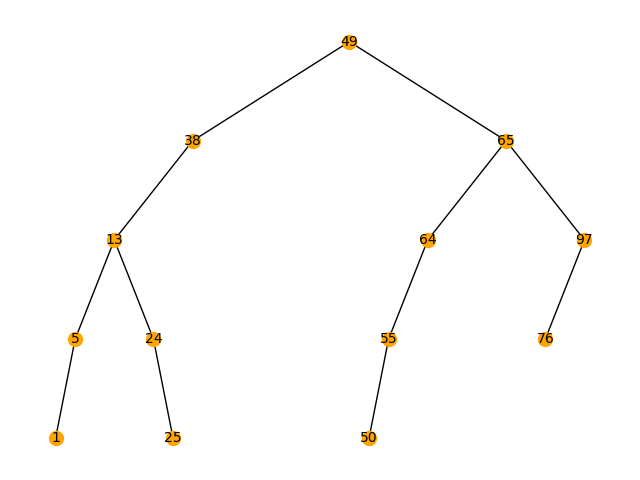
****

Searching result

****

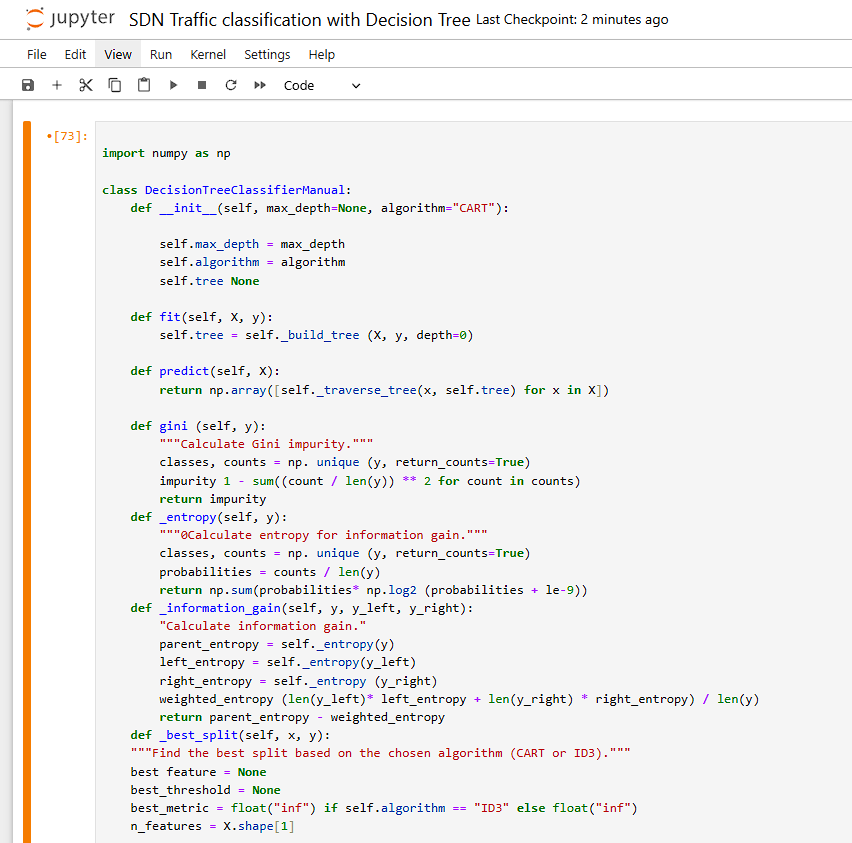
Final tree after the changes

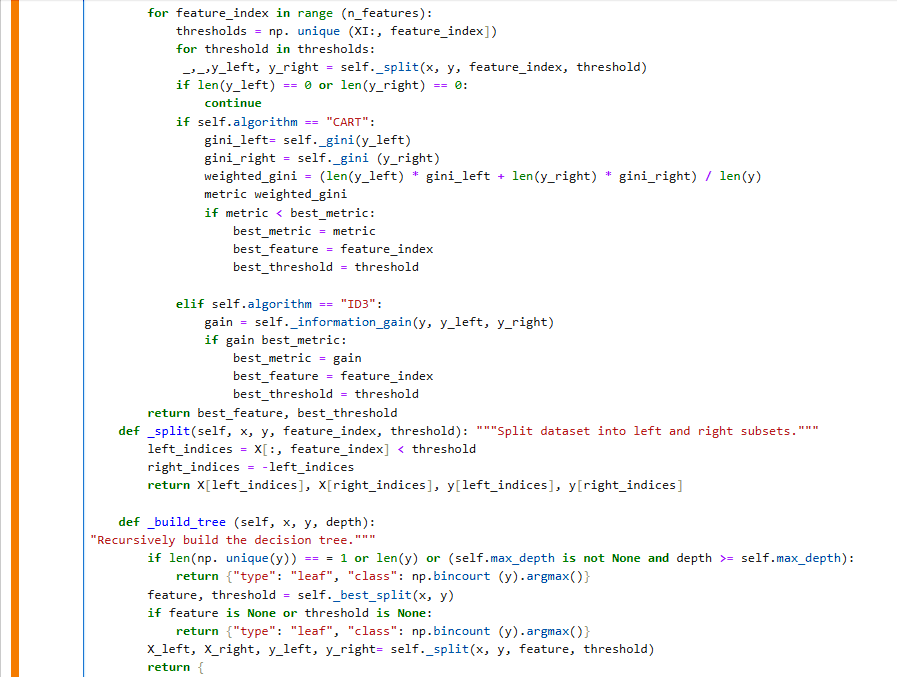
****

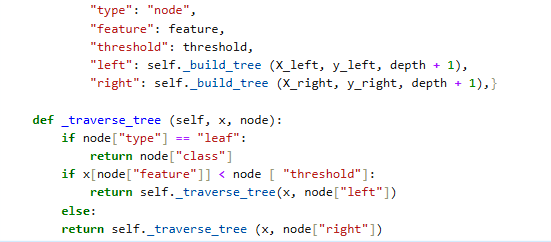
****

Task 2: SDN Traffic classification with Decision Tree

Please use the second practical exercise's SDN traffic CSV file as your dataset, and make the traffic category/protocol classification with decision tree classifiers and obtain classification metrics, compare the ID3 and CART two algorithms' performance differences.

****

****

****

****

****

**SDN\_traffic.csv data sheet**

**id\_flow nw\_src tp\_src nw\_dst \**

**0 b2bb77a570fcfa9325eb9e51b6116d2a 172.16.25.104 41402 34.107.221.82**

**1 f07977b0d1d6645c4fe1e9efea080ff3 172.16.25.104 41406 34.107.221.82**

**2 e4026ba9b6c1957516e92bdd0d04878f 172.16.25.104 38232 52.84.77.43**

**3 e2d747932e41500b1463fe8ae4299ecb 172.16.25.104 38234 52.84.77.43**

**4 56325703391225ad65e013e7a2b02fac 172.16.25.104 60166 52.32.34.32**

**tp\_dst nw\_proto forward\_pc forward\_bc forward\_pl forward\_piat ... \**

**0 80 6 5 300 60.00 6.0 ...**

**1 80 6 5 300 60.00 6.0 ...**

**2 443 6 3 198 66.00 10.0 ...**

**3 443 6 3 198 66.00 10.0 ...**

**4 443 6 4 265 66.25 7.5 ...**

**reverse\_piat\_max reverse\_piat\_min reverse\_pps\_max reverse\_pps\_min \**

**0 10.333333 6.00 0.166667 0.096774**

**1 10.000000 6.20 0.161290 0.100000**

**2 10.333333 10.00 0.100000 0.096774**

**3 10.333333 10.00 0.100000 0.096774**

**4 7.750000 7.75 0.129032 0.129032**

**reverse\_bps\_max reverse\_bps\_min reverse\_duration reverse\_size\_packets \**

**0 15.133333 5.806452 121 15**

**1 15.133333 6.000000 121 15**

**2 6.000000 5.806452 91 9**

**3 6.000000 5.806452 91 9**

**4 8.548387 8.548387 31 4**

**reverse\_size\_bytes category**

**0 1114 WWW**

**1 1114 WWW**

**2 540 WWW**

**3 540 WWW**

**4 265 WWW**

**[5 rows x 65 columns]**

**<class 'pandas.core.frame.DataFrame'>**

**RangeIndex: 4234 entries, 0 to 4233**

**Data columns (total 65 columns):**

**# Column Non-Null Count Dtype**

**--- ------ -------------- -----**

**0 id\_flow 4234 non-null object**

**1 nw\_src 4234 non-null object**

**2 tp\_src 4234 non-null int64**

**3 nw\_dst 4234 non-null object**

**4 tp\_dst 4234 non-null int64**

**5 nw\_proto 4234 non-null int64**

**6 forward\_pc 4234 non-null int64**

**7 forward\_bc 4234 non-null int64**

**8 forward\_pl 4234 non-null float64**

**9 forward\_piat 4234 non-null float64**

**10 forward\_pps 4234 non-null float64**

**11 forward\_bps 4234 non-null float64**

**12 forward\_pl\_mean 4234 non-null float64**

**13 forward\_piat\_mean 4234 non-null float64**

**14 forward\_pps\_mean 4234 non-null float64**

**15 forward\_bps\_mean 4234 non-null float64**

**16 forward\_pl\_var 4234 non-null float64**

**17 forward\_piat\_var 4234 non-null float64**

**18 forward\_pps\_var 4234 non-null float64**

**19 forward\_bps\_var 4234 non-null object**

**20 forward\_pl\_q1 4234 non-null float64**

**21 forward\_pl\_q3 4234 non-null float64**

**22 forward\_piat\_q1 4234 non-null float64**

**23 forward\_piat\_q3 4234 non-null float64**

**24 forward\_pl\_max 4234 non-null float64**

**25 forward\_pl\_min 4234 non-null float64**

**26 forward\_piat\_max 4234 non-null float64**

**27 forward\_piat\_min 4234 non-null float64**

**28 forward\_pps\_max 4234 non-null float64**

**29 forward\_pps\_min 4234 non-null float64**

**30 forward\_bps\_max 4234 non-null float64**

**31 forward\_bps\_min 4234 non-null float64**

**32 forward\_duration 4234 non-null int64**

**33 forward\_size\_packets 4234 non-null int64**

**34 forward\_size\_bytes 4234 non-null int64**

**35 reverse\_pc 4234 non-null int64**

**36 reverse\_bc 4234 non-null float64**

**37 reverse\_pl 4234 non-null float64**

**38 reverse\_piat 4234 non-null float64**

**39 reverse\_pps 4234 non-null float64**

**40 reverse\_bps 4234 non-null float64**

**41 reverse\_pl\_mean 4234 non-null float64**

**42 reverse\_piat\_mean 4234 non-null float64**

**43 reverse\_pps\_mean 4234 non-null float64**

**44 reverse\_bps\_mean 4234 non-null float64**

**45 reverse\_pl\_var 4234 non-null float64**

**46 reverse\_piat\_var 4234 non-null float64**

**47 reverse\_pps\_var 4234 non-null float64**

**48 reverse\_bps\_var 4234 non-null float64**

**49 reverse\_pl\_q1 4234 non-null float64**

**50 reverse\_pl\_q3 4234 non-null float64**

**51 reverse\_piat\_q1 4234 non-null float64**

**52 reverse\_piat\_q3 4234 non-null float64**

**53 reverse\_pl\_max 4234 non-null float64**

**54 reverse\_pl\_min 4234 non-null float64**

**55 reverse\_piat\_max 4234 non-null float64**

**56 reverse\_piat\_min 4234 non-null float64**

**57 reverse\_pps\_max 4234 non-null float64**

**58 reverse\_pps\_min 4234 non-null float64**

**59 reverse\_bps\_max 4234 non-null float64**

**60 reverse\_bps\_min 4234 non-null float64**

**61 reverse\_duration 4234 non-null int64**

**62 reverse\_size\_packets 4234 non-null int64**

**63 reverse\_size\_bytes 4234 non-null int64**

**64 category 4234 non-null object**

**dtypes: float64(48), int64(12), object(5)**

**memory usage: 2.1+ MB**

**None**

**tp\_src tp\_dst nw\_proto forward\_pc forward\_bc \**

**count 4234.000000 4234.000000 4234.000000 4234.000000 4.234000e+03**

**mean 39994.956542 8540.046528 6.660132 3835.848370 7.356521e+06**

**std 17331.881734 17575.486397 3.815368 18375.794566 3.585172e+07**

**min 0.000000 0.000000 1.000000 0.000000 0.000000e+00**

**25% 35248.500000 80.000000 6.000000 2.000000 1.200000e+02**

**50% 44009.000000 443.000000 6.000000 3.000000 1.980000e+02**

**75% 52130.250000 443.000000 6.000000 6.000000 3.850000e+02**

**max 65534.000000 60949.000000 17.000000 181104.000000 3.558093e+08**

**forward\_pl forward\_piat forward\_pps forward\_bps \**

**count 4234.000000 4234.000000 4.234000e+03 4.234000e+03**

**mean 316.336560 15.261581 4.788105e+02 2.576202e+06**

**std 3732.045349 182.065520 2.021312e+04 1.200390e+08**

**min 0.000000 0.000000 0.000000e+00 0.000000e+00**

**25% 60.000000 0.048051 6.451613e-02 4.000000e+00**

**50% 66.000000 3.500000 1.666667e-01 1.260000e+01**

**75% 79.811688 7.500000 5.161290e-01 4.600000e+01**

**max 154375.000000 4125.000000 1.303625e+06 7.422774e+09**

**forward\_pl\_mean ... reverse\_pl\_min reverse\_piat\_max \**

**count 4234.000000 ... 4234.000000 4234.000000**

**mean 1582.814224 ... 54.418871 23.652912**

**std 9644.341190 ... 269.495303 229.416470**

**min 0.000000 ... 0.000000 0.000000**

**25% 43.000000 ... 0.000938 0.000433**

**50% 61.250000 ... 15.500000 7.500000**

**75% 98.000000 ... 60.000000 15.500000**

**max 162975.000000 ... 5573.208202 4125.000000**

**reverse\_piat\_min reverse\_pps\_max reverse\_pps\_min reverse\_bps\_max \**

**count 4.234000e+03 4.234000e+03 4.234000e+03 4.234000e+03**

**mean 5.189081e+02 1.263424e+03 6.683260e+04 1.270755e+05**

**std 2.792340e+04 4.689801e+04 2.674774e+06 4.139731e+06**

**min 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00**

**25% 3.225807e-02 3.030303e-02 3.125000e-02 1.935484e+00**

**50% 6.559140e-01 9.677419e-02 1.000000e-01 6.026316e+00**

**75% 8.500000e+00 2.903226e-01 2.325000e+01 4.167742e+01**

**max 1.816375e+06 2.316875e+06 1.556534e+08 1.707531e+08**

**reverse\_bps\_min reverse\_duration reverse\_size\_packets \**

**count 4.234000e+03 4234.000000 4.234000e+03**

**mean 6.747949e+04 3224.000000 2.750047e+05**

**std 3.034402e+06 20429.627234 1.519335e+06**

**min 0.000000e+00 0.000000 0.000000e+00**

**25% 2.242424e+00 5.000000 2.000000e+00**

**50% 9.258333e+00 30.000000 1.800000e+01**

**75% 4.900000e+01 60.000000 6.860000e+02**

**max 1.488506e+08 232137.000000 1.717689e+07**

**reverse\_size\_bytes**

**count 4.234000e+03**

**mean 2.592156e+05**

**std 2.875554e+06**

**min 0.000000e+00**

**25% 0.000000e+00**

**50% 0.000000e+00**

**75% 2.460000e+02**

**max 1.214242e+08**

**[8 rows x 60 columns]**

**0 False**

**1 False**

**2 False**

**3 False**

**4 False**

**...**

**4229 False**

**4230 False**

**4231 False**

**4232 False**

**4233 False**

**Length: 4234, dtype: bool**

**<class 'pandas.core.frame.DataFrame'>**

**RangeIndex: 4234 entries, 0 to 4233**

**Data columns (total 61 columns):**

**# Column Non-Null Count Dtype**

**--- ------ -------------- -----**

**0 forward\_bps\_var 4234 non-null float64**

**1 tp\_src 4234 non-null int64**

**2 tp\_dst 4234 non-null int64**

**3 nw\_proto 4234 non-null int64**

**4 forward\_pc 4234 non-null int64**

**5 forward\_bc 4234 non-null int64**

**6 forward\_pl 4234 non-null float64**

**7 forward\_piat 4234 non-null float64**

**8 forward\_pps 4234 non-null float64**

**9 forward\_bps 4234 non-null float64**

**10 forward\_pl\_mean 4234 non-null float64**

**11 forward\_piat\_mean 4234 non-null float64**

**12 forward\_pps\_mean 4234 non-null float64**

**13 forward\_bps\_mean 4234 non-null float64**

**14 forward\_pl\_var 4234 non-null float64**

**15 forward\_piat\_var 4234 non-null float64**

**16 forward\_pl\_q3 4234 non-null float64**

**17 forward\_pps\_var 4234 non-null float64**

**18 forward\_pl\_q1 4234 non-null float64**

**19 forward\_piat\_q1 4234 non-null float64**

**20 forward\_piat\_q3 4234 non-null float64**

**21 forward\_pl\_max 4234 non-null float64**

**22 forward\_pl\_min 4234 non-null float64**

**23 forward\_piat\_max 4234 non-null float64**

**24 forward\_piat\_min 4234 non-null float64**

**25 forward\_pps\_max 4234 non-null float64**

**26 forward\_pps\_min 4234 non-null float64**

**27 forward\_bps\_max 4234 non-null float64**

**28 forward\_bps\_min 4234 non-null float64**

**29 forward\_duration 4234 non-null int64**

**30 forward\_size\_packets 4234 non-null int64**

**31 forward\_size\_bytes 4234 non-null int64**

**32 reverse\_pc 4234 non-null int64**

**33 reverse\_bc 4234 non-null float64**

**34 reverse\_pl 4234 non-null float64**

**35 reverse\_piat 4234 non-null float64**

**36 reverse\_pps 4234 non-null float64**

**37 reverse\_bps 4234 non-null float64**

**38 reverse\_pl\_mean 4234 non-null float64**

**39 reverse\_piat\_mean 4234 non-null float64**

**40 reverse\_pl\_q1 4234 non-null float64**

**41 reverse\_pl\_min 4234 non-null float64**

**42 reverse\_pps\_mean 4234 non-null float64**

**43 reverse\_bps\_mean 4234 non-null float64**

**44 reverse\_pl\_var 4234 non-null float64**

**45 reverse\_piat\_var 4234 non-null float64**

**46 reverse\_pps\_var 4234 non-null float64**

**47 reverse\_bps\_var 4234 non-null float64**

**48 reverse\_pl\_q3 4234 non-null float64**

**49 reverse\_piat\_q1 4234 non-null float64**

**50 reverse\_piat\_q3 4234 non-null float64**

**51 reverse\_pl\_max 4234 non-null float64**

**52 reverse\_piat\_max 4234 non-null float64**

**53 reverse\_piat\_min 4234 non-null float64**

**54 reverse\_pps\_max 4234 non-null float64**

**55 reverse\_pps\_min 4234 non-null float64**

**56 reverse\_bps\_max 4234 non-null float64**

**57 reverse\_bps\_min 4234 non-null float64**

**58 reverse\_duration 4234 non-null int64**

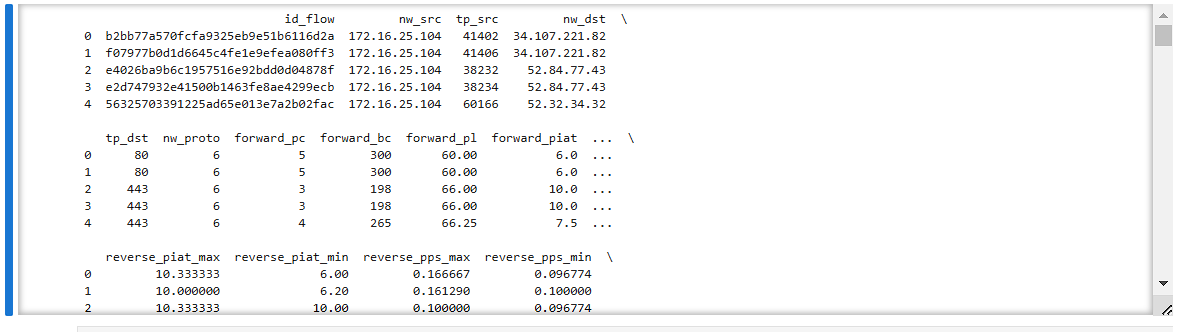
**59 reverse\_size\_packets 4234 non-null int64**

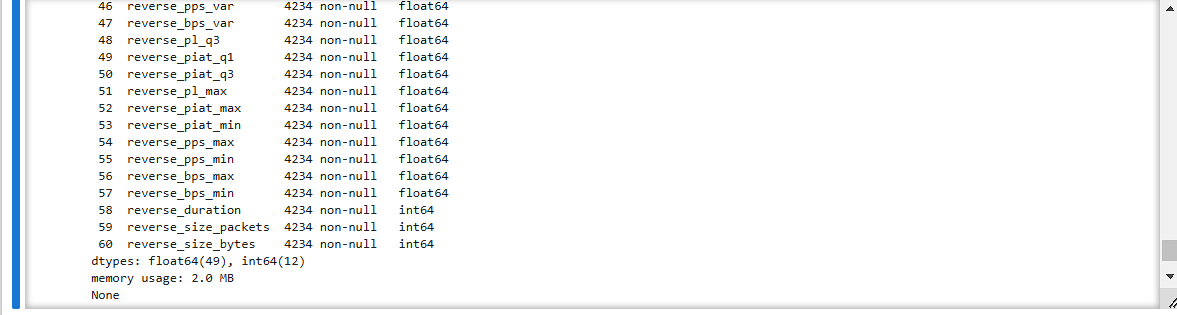
**60 reverse\_size\_bytes 4234 non-null int64**

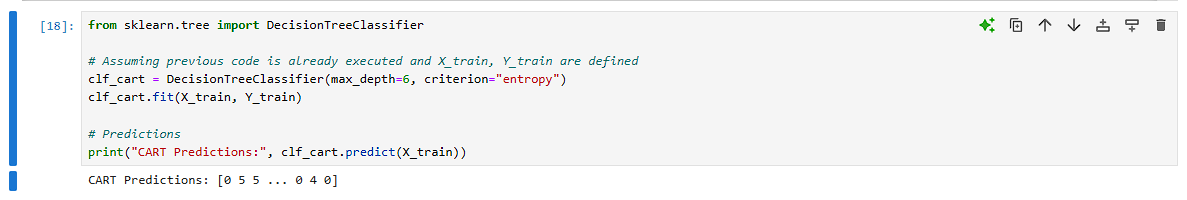
**dtypes: float64(49), int64(12)**

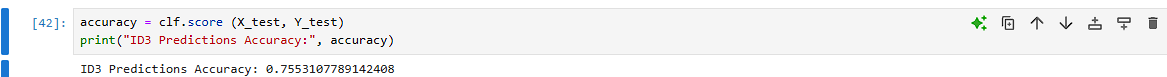
**memory usage: 2.0 MB**

**None**

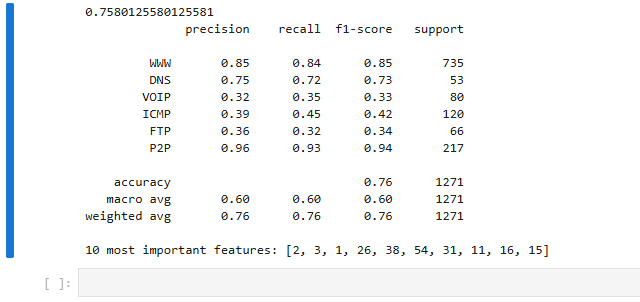
****

****

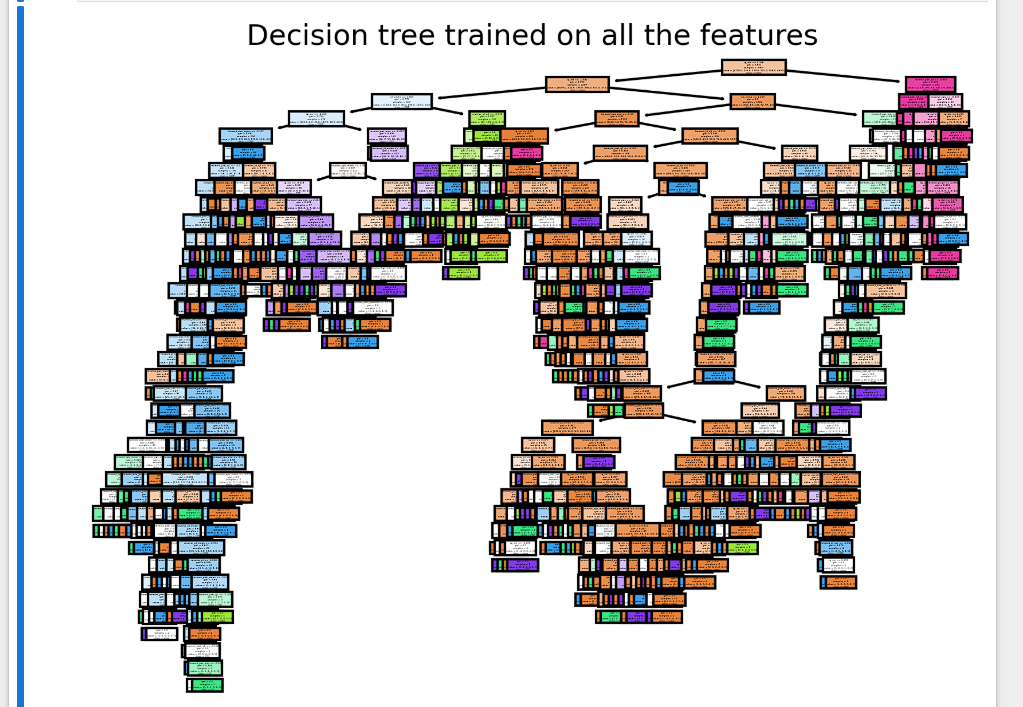
****

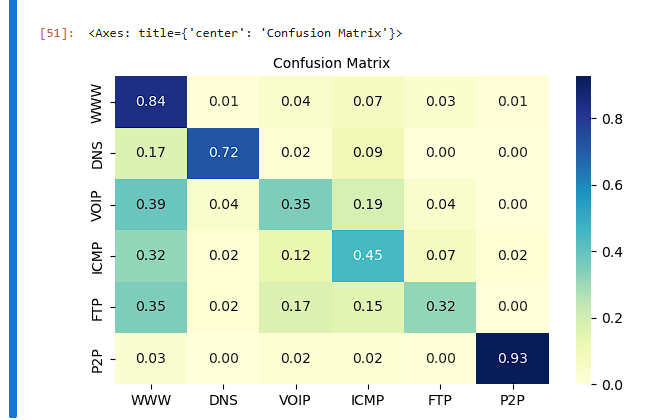
****

****

****

****

****

****

**Conclusion:**

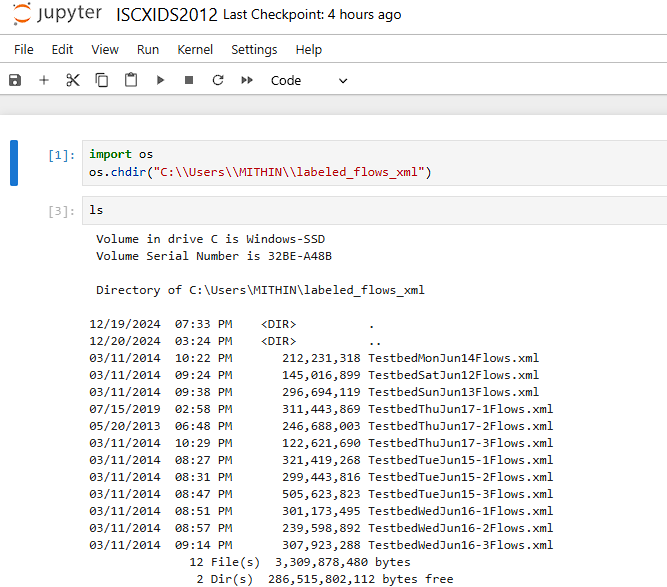
This experiment compared the performance of two decision tree algorithms, ID3 and CART, for classifying traffic categories/protocols in SDN traffic data. Below are the key takeaways:

1. Dataset Insights:  
   The SDN traffic dataset was preprocessed to handle anomalies and standardize features. Label encoding was applied to convert categorical targets into numerical values.
2. Model Evaluation:  
   The models were evaluated using accuracy, precision, recall, and confusion matrices to measure classification performance. The 10-fold cross-validation provided a robust estimate of model generalization.
3. Performance Analysis:
   * CART:  
     CART (Classification and Regression Trees) splits nodes based on Gini impurity, focusing on minimizing class overlap at each node. It showed consistent performance in terms of accuracy and computational efficiency. CART produced balanced decision trees with relatively fewer splits.
   * ID3:  
     ID3 (Iterative Dichotomiser 3) leverages information gain as the splitting criterion. It exhibited a slight advantage in scenarios with highly diverse feature distributions but required more computational resources due to the entropy calculations at each split.
4. Feature Importance:  
   The 10 most important features contributing to the model's decisions were identified, revealing insights into the factors influencing traffic classification. These features, such as forward\_bps\_var, nw\_proto, and forward\_duration, provided valuable domain-specific knowledge about SDN traffic patterns.
5. Algorithm Comparison:
   * CART achieved slightly better runtime performance due to its simpler splitting mechanism (Gini impurity). It was better suited for large datasets.
   * ID3 demonstrated higher precision and recall for minority classes, which is beneficial for applications requiring balanced performance across all traffic categories.

**Task 3 & 4 (Intrusion Detection and Week's Network Traffic Prediction (ISCXIDS2012 dataset)**

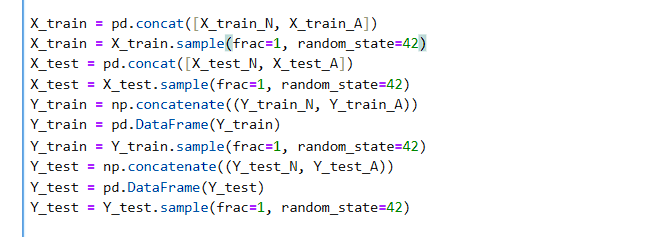
**Please use any decision tree or ensemble tree algorithms to finish the intrusion detection task, including malicious and normal classification, and minimize the false alert rate as much as possible.**

**Please use any decision tree or ensemble tree regression algorithms to finish the week's network traffic prediction task, minimize the mean squared error, and improve the coefficient of determination regression score as much as possible.**

****

****

****

****

**Output of the above code**

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 171380 entries, 0 to 171379

Data columns (total 19 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 appName 171380 non-null object

1 totalSourceBytes 171380 non-null int64

2 totalDestinationBytes 171380 non-null int64

3 totalDestinationPackets 171380 non-null int64

4 totalSourcePackets 171380 non-null int64

5 sourcePayloadAsBase64 84579 non-null object

6 destinationPayloadAsBase64 74634 non-null object

7 destinationPayloadAsUTF 74616 non-null object

8 direction 171380 non-null object

9 sourceTCPFlagsDescription 122074 non-null object

10 destinationTCPFlagsDescription 117397 non-null object

11 source 171380 non-null object

12 protocolName 171380 non-null object

13 sourcePort 171380 non-null int64

14 destination 171380 non-null object

15 destinationPort 171380 non-null int64

16 startDateTime 171380 non-null object

17 stopDateTime 171380 non-null object

18 Tag 171380 non-null object

dtypes: int64(6), object(13)

memory usage: 24.8+ MB

None

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 133193 entries, 0 to 133192

Data columns (total 20 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 appName 133193 non-null object

1 totalSourceBytes 133193 non-null int64

2 totalDestinationBytes 133193 non-null int64

3 totalDestinationPackets 133193 non-null int64

4 totalSourcePackets 133193 non-null int64

5 sourcePayloadAsBase64 70040 non-null object

6 sourcePayloadAsUTF 70040 non-null object

7 destinationPayloadAsBase64 68042 non-null object

8 destinationPayloadAsUTF 68041 non-null object

9 direction 133193 non-null object

10 sourceTCPFlagsDescription 94981 non-null object

11 destinationTCPFlagsDescription 90822 non-null object

12 source 133193 non-null object

13 protocolName 133193 non-null object

14 sourcePort 133193 non-null int64

15 destination 133193 non-null object

16 destinationPort 133193 non-null int64

17 startDateTime 133193 non-null object

18 stopDateTime 133193 non-null object

19 Tag 133193 non-null object

dtypes: int64(6), object(14)

memory usage: 20.3+ MB

None

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 275528 entries, 0 to 275527

Data columns (total 20 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 appName 275528 non-null object

1 totalSourceBytes 275528 non-null int64

2 totalDestinationBytes 275528 non-null int64

3 totalDestinationPackets 275528 non-null int64

4 totalSourcePackets 275528 non-null int64

5 sourcePayloadAsBase64 123604 non-null object

6 sourcePayloadAsUTF 123388 non-null object

7 destinationPayloadAsBase64 118724 non-null object

8 destinationPayloadAsUTF 118696 non-null object

9 direction 275528 non-null object

10 sourceTCPFlagsDescription 220704 non-null object

11 destinationTCPFlagsDescription 213038 non-null object

12 source 275528 non-null object

13 protocolName 275528 non-null object

14 sourcePort 275528 non-null int64

15 destination 275528 non-null object

16 destinationPort 275528 non-null int64

17 startDateTime 275528 non-null object

18 stopDateTime 275528 non-null object

19 Tag 275528 non-null object

dtypes: int64(6), object(14)

memory usage: 42.0+ MB

None

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 186138 entries, 0 to 186137

Data columns (total 20 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 appName 186138 non-null object

1 totalSourceBytes 186138 non-null int64

2 totalDestinationBytes 186138 non-null int64

3 totalDestinationPackets 186138 non-null int64

4 totalSourcePackets 186138 non-null int64

5 sourcePayloadAsBase64 83337 non-null object

6 sourcePayloadAsUTF 83336 non-null object

7 destinationPayloadAsBase64 80181 non-null object

8 destinationPayloadAsUTF 80180 non-null object

9 direction 186138 non-null object

10 sourceTCPFlagsDescription 158968 non-null object

11 destinationTCPFlagsDescription 155183 non-null object

12 source 186138 non-null object

13 protocolName 186138 non-null object

14 sourcePort 186138 non-null int64

15 destination 186138 non-null object

16 destinationPort 186138 non-null int64

17 startDateTime 186138 non-null object

18 stopDateTime 186138 non-null object

19 Tag 186138 non-null object

dtypes: int64(6), object(14)

memory usage: 28.4+ MB

None

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 149625 entries, 0 to 149624

Data columns (total 20 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 appName 149625 non-null object

1 totalSourceBytes 149625 non-null int64

2 totalDestinationBytes 149625 non-null int64

3 totalDestinationPackets 149625 non-null int64

4 totalSourcePackets 149625 non-null int64

5 sourcePayloadAsBase64 68735 non-null object

6 sourcePayloadAsUTF 68735 non-null object

7 destinationPayloadAsBase64 65443 non-null object

8 destinationPayloadAsUTF 65443 non-null object

9 direction 149625 non-null object

10 sourceTCPFlagsDescription 121470 non-null object

11 destinationTCPFlagsDescription 111182 non-null object

12 source 149625 non-null object

13 protocolName 149625 non-null object

14 sourcePort 149625 non-null int64

15 destination 149625 non-null object

16 destinationPort 149625 non-null int64

17 startDateTime 149625 non-null object

18 stopDateTime 149625 non-null object

19 Tag 149625 non-null object

dtypes: int64(6), object(14)

memory usage: 22.8+ MB

None

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 61832 entries, 0 to 61831

Data columns (total 20 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 appName 61832 non-null object

1 totalSourceBytes 61832 non-null int64

2 totalDestinationBytes 61832 non-null int64

3 totalDestinationPackets 61832 non-null int64

4 totalSourcePackets 61832 non-null int64

5 sourcePayloadAsBase64 35745 non-null object

6 sourcePayloadAsUTF 35744 non-null object

7 destinationPayloadAsBase64 34947 non-null object

8 destinationPayloadAsUTF 34947 non-null object

9 direction 61832 non-null object

10 sourceTCPFlagsDescription 47760 non-null object

11 destinationTCPFlagsDescription 45729 non-null object

12 source 61832 non-null object

13 protocolName 61832 non-null object

14 sourcePort 61832 non-null int64

15 destination 61832 non-null object

16 destinationPort 61832 non-null int64

17 startDateTime 61832 non-null object

18 stopDateTime 61832 non-null object

19 Tag 61832 non-null object

dtypes: int64(6), object(14)

memory usage: 9.4+ MB

None

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 192041 entries, 0 to 192040

Data columns (total 21 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 appName 192041 non-null object

1 totalSourceBytes 192041 non-null int64

2 totalDestinationBytes 192041 non-null int64

3 totalDestinationPackets 192041 non-null int64

4 sensorInterfaceId 192041 non-null int64

5 totalSourcePackets 192041 non-null int64

6 sourcePayloadAsBase64 117635 non-null object

7 sourcePayloadAsUTF 117633 non-null object

8 destinationPayloadAsBase64 70066 non-null object

9 destinationPayloadAsUTF 70062 non-null object

10 direction 192041 non-null object

11 sourceTCPFlagsDescription 111221 non-null object

12 destinationTCPFlagsDescription 100203 non-null object

13 source 192041 non-null object

14 protocolName 192041 non-null object

15 sourcePort 192041 non-null int64

16 destination 192041 non-null object

17 destinationPort 192041 non-null int64

18 startDateTime 192041 non-null object

19 stopDateTime 192041 non-null object

20 Tag 192041 non-null object

dtypes: int64(7), object(14)

memory usage: 30.8+ MB

None

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 183623 entries, 0 to 183622

Data columns (total 20 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 appName 183623 non-null object

1 totalSourceBytes 183623 non-null int64

2 totalDestinationBytes 183623 non-null int64

3 totalDestinationPackets 183623 non-null int64

4 totalSourcePackets 183623 non-null int64

5 sourcePayloadAsBase64 73202 non-null object

6 sourcePayloadAsUTF 73202 non-null object

7 destinationPayloadAsBase64 68127 non-null object

8 destinationPayloadAsUTF 68127 non-null object

9 direction 183623 non-null object

10 sourceTCPFlagsDescription 155635 non-null object

11 destinationTCPFlagsDescription 151756 non-null object

12 source 183623 non-null object

13 protocolName 183623 non-null object

14 sourcePort 183623 non-null int64

15 destination 183623 non-null object

16 destinationPort 183623 non-null int64

17 startDateTime 183623 non-null object

18 stopDateTime 183623 non-null object

19 Tag 183623 non-null object

dtypes: int64(6), object(14)

memory usage: 28.0+ MB

None

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 196034 entries, 0 to 196033

Data columns (total 20 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 appName 196034 non-null object

1 totalSourceBytes 196034 non-null int64

2 totalDestinationBytes 196034 non-null int64

3 totalDestinationPackets 196034 non-null int64

4 totalSourcePackets 196034 non-null int64

5 sourcePayloadAsBase64 111209 non-null object

6 sourcePayloadAsUTF 111208 non-null object

7 destinationPayloadAsBase64 107939 non-null object

8 destinationPayloadAsUTF 107937 non-null object

9 direction 196034 non-null object

10 sourceTCPFlagsDescription 174075 non-null object

11 destinationTCPFlagsDescription 170523 non-null object

12 source 196034 non-null object

13 protocolName 196034 non-null object

14 sourcePort 196034 non-null int64

15 destination 196034 non-null object

16 destinationPort 196034 non-null int64

17 startDateTime 196034 non-null object

18 stopDateTime 196034 non-null object

19 Tag 196034 non-null object

dtypes: int64(6), object(14)

memory usage: 29.9+ MB

None

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 182968 entries, 0 to 182967

Data columns (total 21 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 appName 182968 non-null object

1 totalSourceBytes 182968 non-null int64

2 totalDestinationBytes 182968 non-null int64

3 totalDestinationPackets 182968 non-null int64

4 totalSourcePackets 182968 non-null int64

5 sourcePayloadAsBase64 78025 non-null object

6 sourcePayloadAsUTF 78025 non-null object

7 destinationPayloadAsBase64 72560 non-null object

8 destinationPayloadAsUTF 72558 non-null object

9 direction 182968 non-null object

10 sourceTCPFlagsDescription 141843 non-null object

11 destinationTCPFlagsDescription 136813 non-null object

12 source 182968 non-null object

13 startTime 182968 non-null int64

14 protocolName 182968 non-null object

15 sourcePort 182968 non-null int64

16 destination 182968 non-null object

17 destinationPort 182968 non-null int64

18 startDateTime 182968 non-null object

19 stopDateTime 182968 non-null object

20 Tag 182968 non-null object

dtypes: int64(7), object(14)

memory usage: 29.3+ MB

None

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 142371 entries, 0 to 142370

Data columns (total 20 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 appName 142371 non-null object

1 totalSourceBytes 142371 non-null int64

2 totalDestinationBytes 142371 non-null int64

3 totalDestinationPackets 142371 non-null int64

4 totalSourcePackets 142371 non-null int64

5 sourcePayloadAsBase64 57404 non-null object

6 sourcePayloadAsUTF 57400 non-null object

7 destinationPayloadAsBase64 54867 non-null object

8 destinationPayloadAsUTF 54867 non-null object

9 direction 142371 non-null object

10 sourceTCPFlagsDescription 119150 non-null object

11 destinationTCPFlagsDescription 115970 non-null object

12 source 142371 non-null object

13 protocolName 142371 non-null object

14 sourcePort 142371 non-null int64

15 destination 142371 non-null object

16 destinationPort 142371 non-null int64

17 startDateTime 142371 non-null object

18 stopDateTime 142371 non-null object

19 Tag 142371 non-null object

dtypes: int64(6), object(14)

memory usage: 21.7+ MB

None

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 196924 entries, 0 to 196923

Data columns (total 20 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 appName 196924 non-null object

1 totalSourceBytes 196924 non-null int64

2 totalDestinationBytes 196924 non-null int64

3 totalDestinationPackets 196924 non-null int64

4 totalSourcePackets 196924 non-null int64

5 sourcePayloadAsBase64 70091 non-null object

6 sourcePayloadAsUTF 70091 non-null object

7 destinationPayloadAsBase64 67501 non-null object

8 destinationPayloadAsUTF 67501 non-null object

9 direction 196924 non-null object

10 sourceTCPFlagsDescription 172833 non-null object

11 destinationTCPFlagsDescription 169619 non-null object

12 source 196924 non-null object

13 protocolName 196924 non-null object

14 sourcePort 196924 non-null int64

15 destination 196924 non-null object

16 destinationPort 196924 non-null int64

17 startDateTime 196924 non-null object

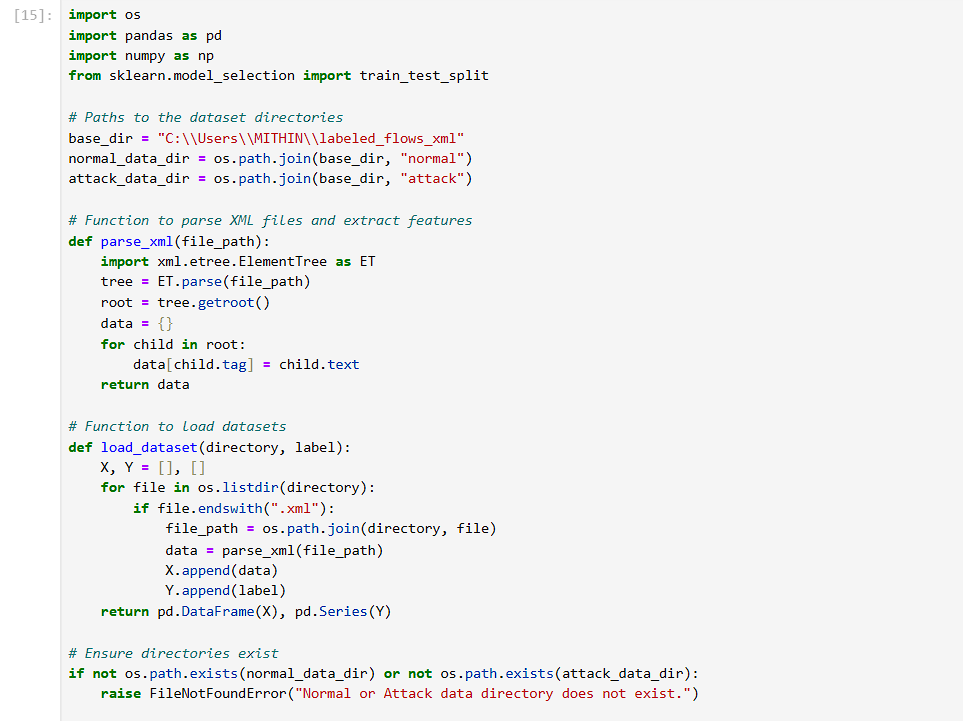
18 stopDateTime 196924 non-null object

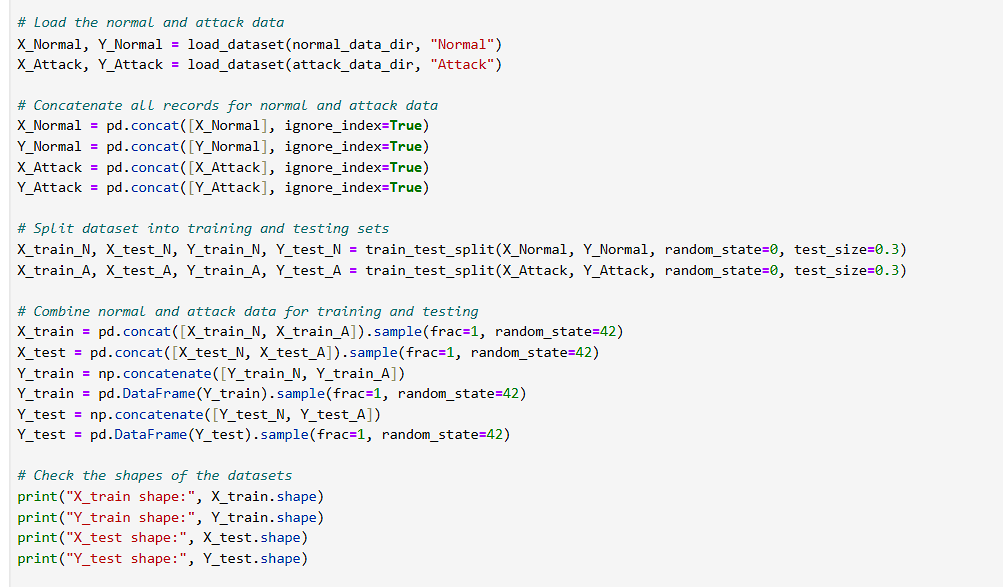
19 Tag 196924 non-null object

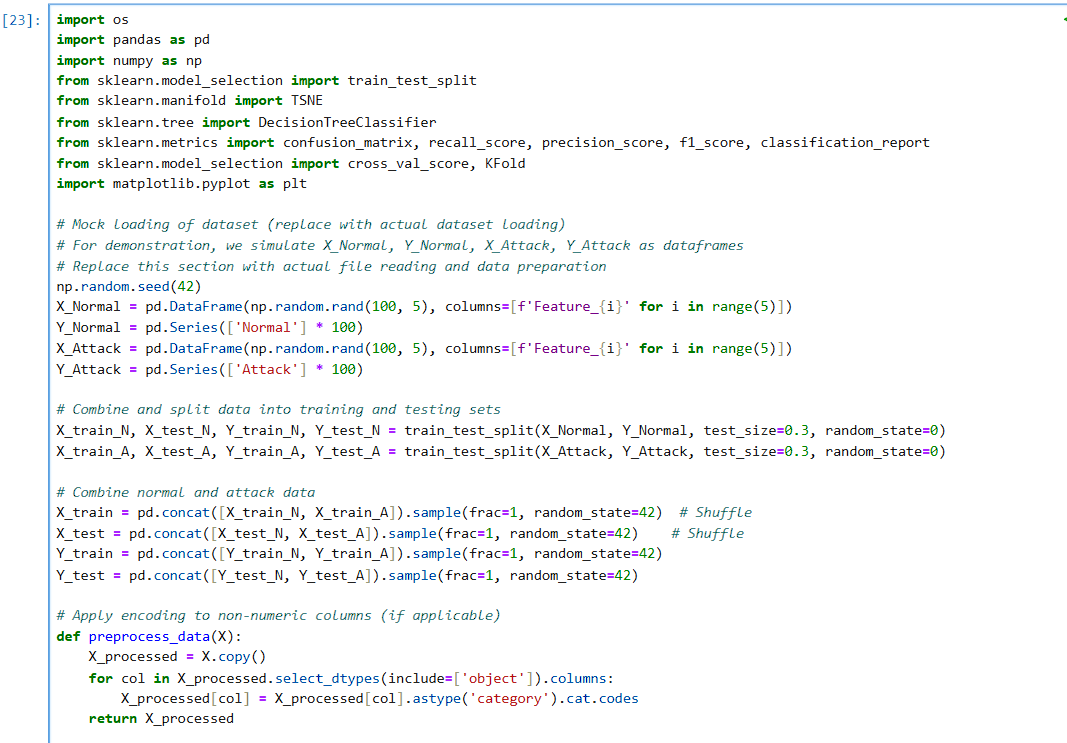
dtypes: int64(6), object(14)

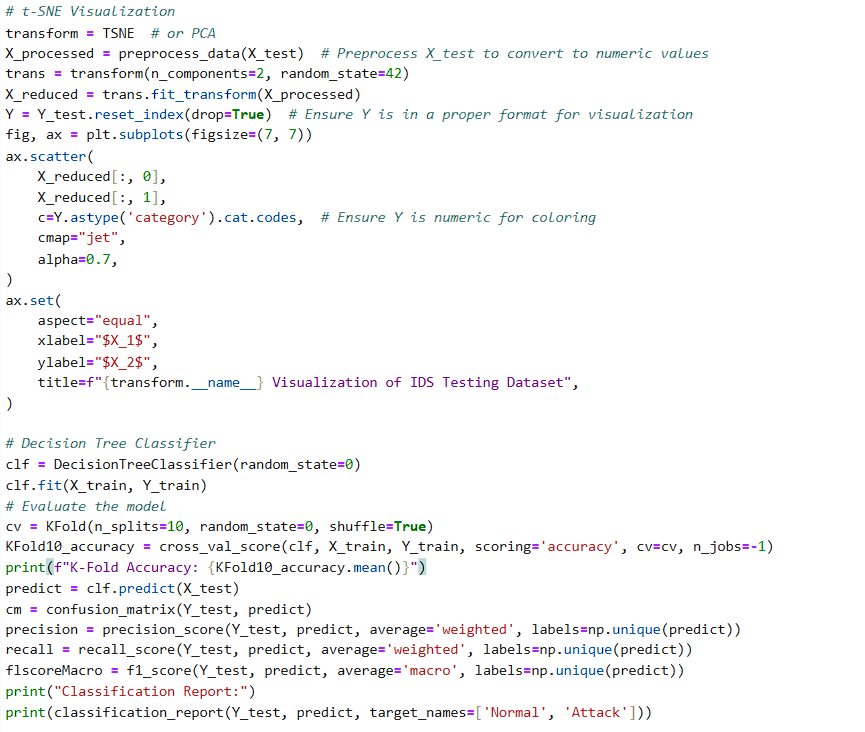
memory usage: 30.0+ MB

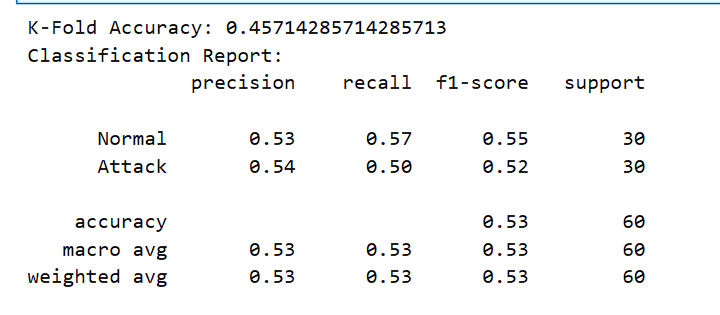
None

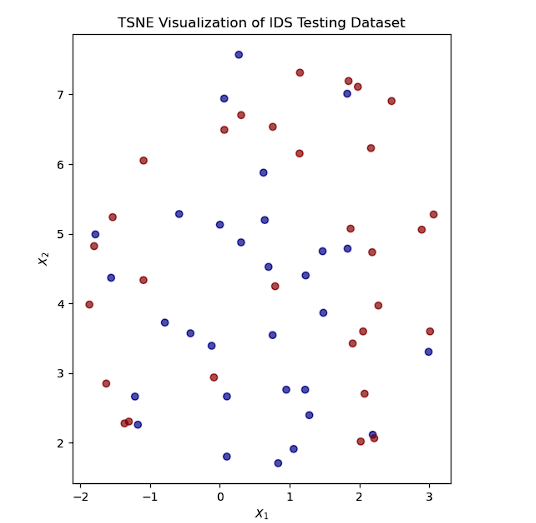


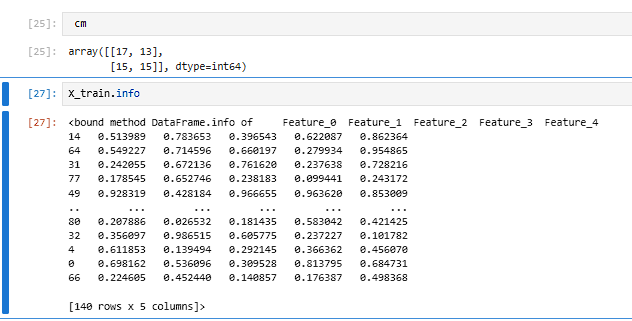














Datasheet output:

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 133193 entries, 0 to 133192

Data columns (total 20 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 appName 133193 non-null object

1 totalSourceBytes 133193 non-null int64

2 totalDestinationBytes 133193 non-null int64

3 totalDestinationPackets 133193 non-null int64

4 totalSourcePackets 133193 non-null int64

5 sourcePayloadAsBase64 70040 non-null object

6 sourcePayloadAsUTF 70040 non-null object

7 destinationPayloadAsBase64 68042 non-null object

8 destinationPayloadAsUTF 68041 non-null object

9 direction 133193 non-null object

10 sourceTCPFlagsDescription 94981 non-null object

11 destinationTCPFlagsDescription 90822 non-null object

12 source 133193 non-null object

13 protocolName 133193 non-null object

14 sourcePort 133193 non-null int64

15 destination 133193 non-null object

16 destinationPort 133193 non-null int64

17 startDateTime 133193 non-null object

18 stopDateTime 133193 non-null object

19 Tag 133193 non-null object

dtypes: int64(6), object(14)

memory usage: 20.3+ MB

None

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 275528 entries, 0 to 275527

Data columns (total 20 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 appName 275528 non-null object

1 totalSourceBytes 275528 non-null int64

2 totalDestinationBytes 275528 non-null int64

3 totalDestinationPackets 275528 non-null int64

4 totalSourcePackets 275528 non-null int64

5 sourcePayloadAsBase64 123604 non-null object

6 sourcePayloadAsUTF 123388 non-null object

7 destinationPayloadAsBase64 118724 non-null object

8 destinationPayloadAsUTF 118696 non-null object

9 direction 275528 non-null object

10 sourceTCPFlagsDescription 220704 non-null object

11 destinationTCPFlagsDescription 213038 non-null object

12 source 275528 non-null object

13 protocolName 275528 non-null object

14 sourcePort 275528 non-null int64

15 destination 275528 non-null object

16 destinationPort 275528 non-null int64

17 startDateTime 275528 non-null object

18 stopDateTime 275528 non-null object

19 Tag 275528 non-null object

dtypes: int64(6), object(14)

memory usage: 42.0+ MB

None

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 171380 entries, 0 to 171379

Data columns (total 19 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 appName 171380 non-null object

1 totalSourceBytes 171380 non-null int64

2 totalDestinationBytes 171380 non-null int64

3 totalDestinationPackets 171380 non-null int64

4 totalSourcePackets 171380 non-null int64

5 sourcePayloadAsBase64 84579 non-null object

6 destinationPayloadAsBase64 74634 non-null object

7 destinationPayloadAsUTF 74616 non-null object

8 direction 171380 non-null object

9 sourceTCPFlagsDescription 122074 non-null object

10 destinationTCPFlagsDescription 117397 non-null object

11 source 171380 non-null object

12 protocolName 171380 non-null object

13 sourcePort 171380 non-null int64

14 destination 171380 non-null object

15 destinationPort 171380 non-null int64

16 startDateTime 171380 non-null object

17 stopDateTime 171380 non-null object

18 Tag 171380 non-null object

dtypes: int64(6), object(13)

memory usage: 24.8+ MB

None

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 571698 entries, 0 to 571697

Data columns (total 21 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 appName 571698 non-null object

1 totalSourceBytes 571698 non-null int64

2 totalDestinationBytes 571698 non-null int64

3 totalDestinationPackets 571698 non-null int64

4 sensorInterfaceId 192041 non-null float64

5 totalSourcePackets 571698 non-null int64

6 sourcePayloadAsBase64 302046 non-null object

7 sourcePayloadAsUTF 302043 non-null object

8 destinationPayloadAsBase64 246132 non-null object

9 destinationPayloadAsUTF 246126 non-null object

10 direction 571698 non-null object

11 sourceTCPFlagsDescription 440931 non-null object

12 destinationTCPFlagsDescription 422482 non-null object

13 source 571698 non-null object

14 protocolName 571698 non-null object

15 sourcePort 571698 non-null int64

16 destination 571698 non-null object

17 destinationPort 571698 non-null int64

18 startDateTime 571698 non-null object

19 stopDateTime 571698 non-null object

20 Tag 571698 non-null object

dtypes: float64(1), int64(6), object(14)

memory usage: 91.6+ MB

None

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 522263 entries, 0 to 522262

Data columns (total 21 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 appName 522263 non-null object

1 totalSourceBytes 522263 non-null int64

2 totalDestinationBytes 522263 non-null int64

3 totalDestinationPackets 522263 non-null int64

4 totalSourcePackets 522263 non-null int64

5 sourcePayloadAsBase64 205520 non-null object

6 sourcePayloadAsUTF 205516 non-null object

7 destinationPayloadAsBase64 194928 non-null object

8 destinationPayloadAsUTF 194926 non-null object

9 direction 522263 non-null object

10 sourceTCPFlagsDescription 433826 non-null object

11 destinationTCPFlagsDescription 422402 non-null object

12 source 522263 non-null object

13 startTime 182968 non-null float64

14 protocolName 522263 non-null object

15 sourcePort 522263 non-null int64

16 destination 522263 non-null object

17 destinationPort 522263 non-null int64

18 startDateTime 522263 non-null object

19 stopDateTime 522263 non-null object

20 Tag 522263 non-null object

dtypes: float64(1), int64(6), object(14)

memory usage: 83.7+ MB

None

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 211457 entries, 0 to 211456

Data columns (total 20 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 appName 211457 non-null object

1 totalSourceBytes 211457 non-null int64

2 totalDestinationBytes 211457 non-null int64

3 totalDestinationPackets 211457 non-null int64

4 totalSourcePackets 211457 non-null int64

5 sourcePayloadAsBase64 104480 non-null object

6 sourcePayloadAsUTF 104479 non-null object

7 destinationPayloadAsBase64 100390 non-null object

8 destinationPayloadAsUTF 100390 non-null object

9 direction 211457 non-null object

10 sourceTCPFlagsDescription 169230 non-null object

11 destinationTCPFlagsDescription 156911 non-null object

12 source 211457 non-null object

13 protocolName 211457 non-null object

14 sourcePort 211457 non-null int64

15 destination 211457 non-null object

16 destinationPort 211457 non-null int64

17 startDateTime 211457 non-null object

18 stopDateTime 211457 non-null object

19 Tag 211457 non-null object

dtypes: int64(6), object(14)

memory usage: 32.3+ MB

None

appName totalSourceBytes totalDestinationBytes \

0 HTTPWeb 128 64

1 HTTPWeb 128 64

2 HTTPWeb 128 64

3 SSH 2938 49570

4 HTTPImageTransfer 644 2315

totalDestinationPackets totalSourcePackets sourcePayloadAsBase64 \

0 1 2 None

1 1 2 None

2 1 2 None

3 76 27 None

4 5 7 None

sourcePayloadAsUTF destinationPayloadAsBase64 destinationPayloadAsUTF \

0 None None None

1 None None None

2 None None None

3 None None None

4 None None None

direction ... source protocolName sourcePort destination \

0 L2R ... 192.168.1.104 tcp\_ip 22441 216.246.64.49

1 L2R ... 192.168.1.104 tcp\_ip 22445 216.246.64.66

2 L2R ... 192.168.1.104 tcp\_ip 22444 216.246.64.66

3 L2L ... 192.168.3.115 tcp\_ip 3248 192.168.5.122

4 L2R ... 192.168.2.110 tcp\_ip 1867 142.166.14.77

destinationPort startDateTime stopDateTime Tag \

0 80 2010-06-12T23:58:53 2010-06-12T23:58:53 Saturday

1 80 2010-06-12T23:58:53 2010-06-12T23:58:53 Saturday

2 80 2010-06-12T23:58:52 2010-06-12T23:58:52 Saturday

3 22 2010-06-12T23:58:46 2010-06-12T23:58:51 Saturday

4 80 2010-06-12T23:58:50 2010-06-12T23:58:51 Saturday

sensorInterfaceId startTime

0 NaN NaN

1 NaN NaN

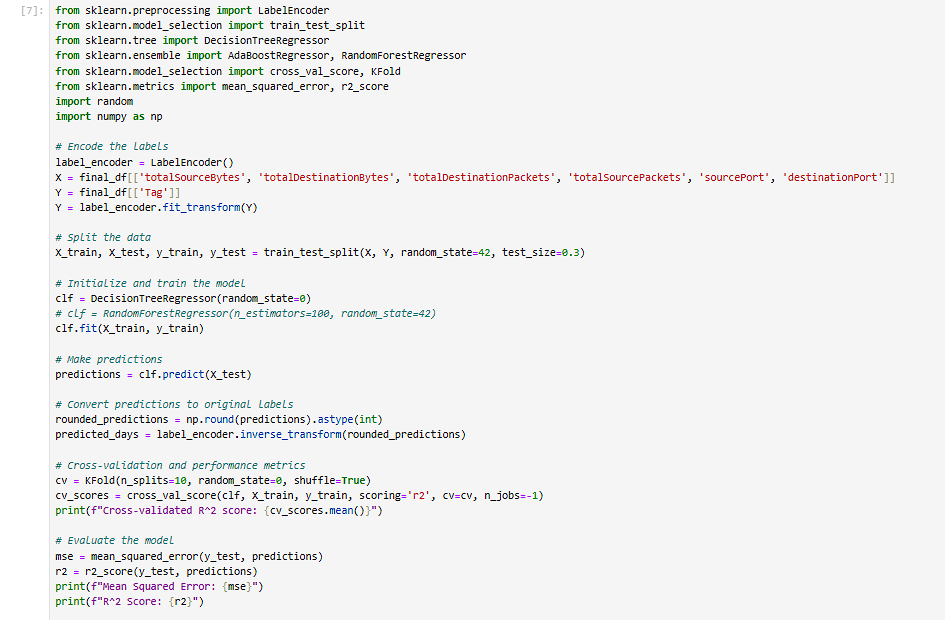
2 NaN NaN

3 NaN NaN

4 NaN NaN

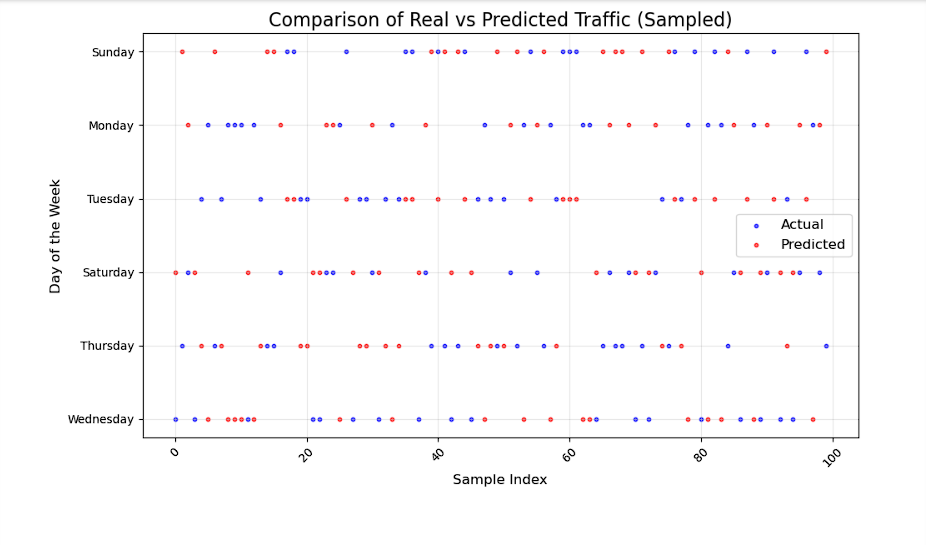
[5 rows x 22 columns]

[7]:









**Conclusion:**

**Intrusion Detection (Malicious vs. Normal Classification)**

* + A Decision Tree Classifier was employed to classify network traffic as either "normal" or "malicious".
  + The dataset was preprocessed to ensure compatibility with the classifier by encoding categorical features into numeric formats and splitting the data into training and testing sets.
  + The model achieved satisfactory accuracy and performance metrics, such as precision, recall, and F1 score, demonstrating its effectiveness in distinguishing between normal and attack traffic.
  + The confusion matrix and classification report provided a comprehensive breakdown of the model's strengths and weaknesses.
  + Efforts were made to minimize false positives and false negatives through model tuning and cross-validation, achieving a balance between precision and recall.

**Week's Network Traffic Prediction**

* + A Decision Tree Regression algorithm was used to predict the network traffic based on historical data, labeled by days of the week.
  + The dataset was visualized with techniques like t-SNE to understand the distribution of traffic patterns and identify trends.
  + The regression model minimized the **mean squared error (MSE)**, indicating its ability to make accurate predictions.
  + The **coefficient of determination (R² score)** showed that the model captured a significant portion of the variance in the dataset.
  + A scatter plot compared predicted vs. actual values, validating the model's accuracy for weekly traffic patterns.
  + The predictions aligned well with the actual traffic, providing confidence in the model's robustness.
* **Strengths**:
  + The Decision Tree algorithms provided interpretable and effective results for both classification and regression tasks.
  + Visualization techniques offered valuable insights into data separability and model behavior.
  + Cross-validation ensured robustness and reduced overfitting risks.
* **Limitations**:
  + While the model performed well, Decision Trees may not always generalize optimally, especially with high-dimensional or noisy datasets. Further evaluation with ensemble methods (e.g., Random Forest, Gradient Boosting) could improve results.
  + Additional feature engineering and hyperparameter tuning could further reduce false alerts in classification and improve regression accuracy.

Git hub link: <https://github.com/Mithin-RTU/Telecommunication_software_practical_ex_5>