Classification Project

(Assignment-5)

Student1: Sarthak Singhal(20171091)

Student2: Sajal Asati(20171183)

Data Cleaning

❑ We removed irrelevant/redundant columns and restricted our analysis based on only few columns.

❑ Columns such as Sl.No., Time, Location, Reference were removed as they don't have a major role in predicting the magnitude of an earthquake.

❑ Rows which didn't have any value in the magnitude field were also removed.

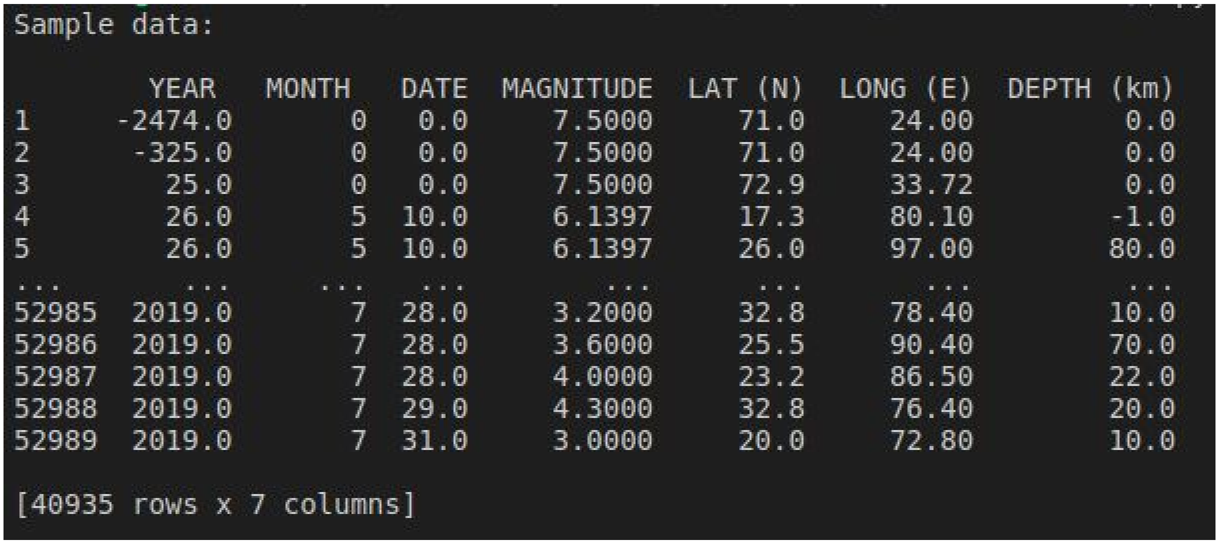
❑ Then we replaced empty values/strings suitably with a value decided according to that particular column.

❑ Fields such as 'LAT' and 'LONG' contained special characters which were also removed.

❑ Later all the values were converted to float so that they could be given to the decision tree.

❑ 80% of data is used for training the model and it is tested on the remaining 20% data.

❑ Sample data after cleaning looks like:



```
Sample data:

          YEAR  MONTH  DATE  MAGNITUDE  LAT (N)  LONG (E)  DEPTH (km)
1      -2474.0      0   0.0     7.5000     71.0     24.00         0.0
2       -325.0      0   0.0     7.5000     71.0     24.00         0.0
3         25.0      0   0.0     7.5000     72.9     33.72         0.0
4         26.0      5  10.0     6.1397     17.3     80.10        -1.0
5         26.0      5  10.0     6.1397     26.0     97.00        80.0
...        ...    ...   ...        ...      ...       ...         ...
52985   2019.0      7  28.0     3.2000     32.8     78.40        10.0
52986   2019.0      7  28.0     3.6000     25.5     90.40        70.0
52987   2019.0      7  28.0     4.0000     23.2     86.50        22.0
52988   2019.0      7  29.0     4.3000     32.8     76.40        20.0
52989   2019.0      7  31.0     3.0000     20.0     72.80        10.0

[40935 rows x 7 columns]
```

Classification Results

- For each dataset results are in the following order:
  - Classification based on magnitude >/< 3.
  - Classification based on magnitude >/< 4.
  - Classification based on magnitude >/< 5.

1. Combined Dataset

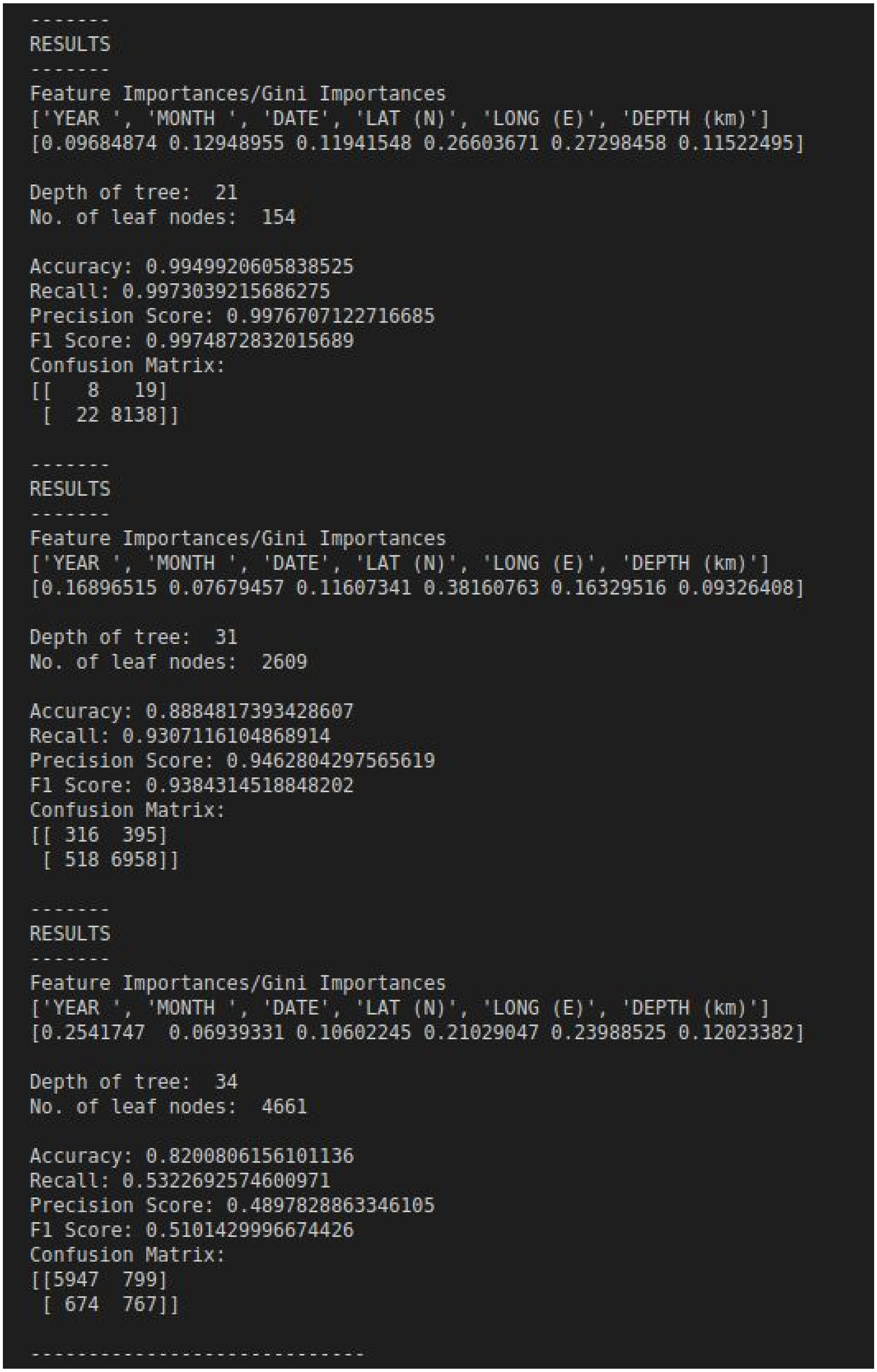
a. Classification based on magnitude >/< 3.
   i. Samples distribution : 27 vs 8160
   ii. Depth: 21
   iii. Leaf Nodes: 154
   iv. Important Attributes: Long(0.273), Lat(0.266)

b. Classification based on magnitude >/< 4.
   i. Samples distribution : 711 vs 7476
   ii. Depth: 31
   iii. Leaf Nodes: 2609
   iv. Important Attributes:Lat(0.3816), Long(0.1632)

c. Classification based on magnitude >/< 5.
   i. Samples distribution : 6746 vs 1441
   ii. Depth: 34
   iii. Leaf Nodes: 4661
   iv. Important Attributes:Long(0.2398), Lat(0.2102)

For this dataset:

- Accuracy decreases as we increase the magnitude since the complexity of decision tree increases as we increase the magnitude and the tree formed has more height and more leaf nodes.
- Attributes like latitude, longitude and depth play a major role in classification.
- When magnitude is 5 we get very low precision as there is a lot of complexity in the tree.



```
-------
RESULTS
-------
Feature Importances/Gini Importances
['YEAR ', 'MONTH ', 'DATE', 'LAT (N)', 'LONG (E)', 'DEPTH (km)']
[0.09684874 0.12948955 0.11941548 0.26603671 0.27298458 0.11522495]

Depth of tree:  21
No. of leaf nodes:  154

Accuracy: 0.9949920605838525
Recall: 0.9973039215686275
Precision Score: 0.9976707122716685
F1 Score: 0.9974872832015689
Confusion Matrix:
[[   8   19]
 [  22 8138]]

-------
RESULTS
-------
Feature Importances/Gini Importances
['YEAR ', 'MONTH ', 'DATE', 'LAT (N)', 'LONG (E)', 'DEPTH (km)']
[0.16896515 0.07679457 0.11607341 0.38160763 0.16329516 0.09326408]

Depth of tree:  31
No. of leaf nodes:  2609

Accuracy: 0.8884817393428607
Recall: 0.9307116104868914
Precision Score: 0.9462804297565619
F1 Score: 0.9384314518848202
Confusion Matrix:
[[ 316  395]
 [ 518 6958]]

-------
RESULTS
-------
Feature Importances/Gini Importances
['YEAR ', 'MONTH ', 'DATE', 'LAT (N)', 'LONG (E)', 'DEPTH (km)']
[0.2541747  0.06939331 0.10602245 0.21029047 0.23988525 0.12023382]

Depth of tree:  34
No. of leaf nodes:  4661

Accuracy: 0.8200806156101136
Recall: 0.5322692574600971
Precision Score: 0.4897828863346105
F1 Score: 0.5101429996674426
Confusion Matrix:
[[5947  799]
 [ 674  767]]

----------------------------
```

2. 1990<=['YEAR']<2000

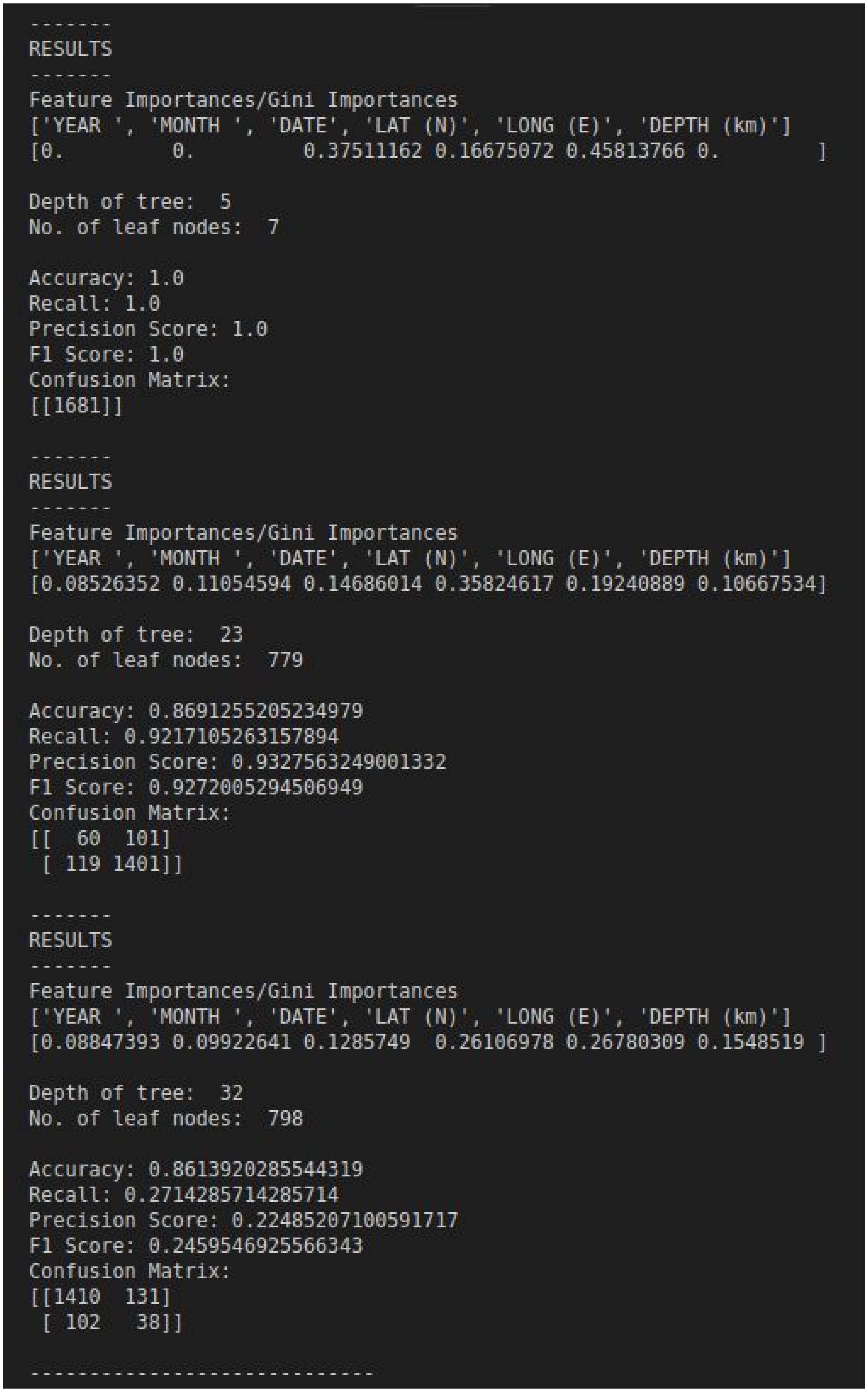
a. Classification based on magnitude >/< 3.
   i. Samples distribution : 1681
   ii. Depth: 5
   iii. Leaf Nodes: 7
   iv. Important Attributes:Long(0.45813), Lat(0.1667)

b. Classification based on magnitude >/< 4.
   i. Samples distribution : 161 vs 1520
   ii. Depth: 23
   iii. Leaf Nodes: 779
   iv. Important Attributes:Lat(0.3582), Long(0.1924)

c. Classification based on magnitude >/< 5.
   i. Samples distribution : 1541 vs 140
   ii. Depth: 32
   iii. Leaf Nodes: 798
   iv. Important Attributes:Long(0.2678), Lat(0.2611)

For this dataset:

- Accuracy decreases with increase in magnitude as the complexity of tree increases with increase in magnitude.
- There are no samples with magnitude less than 3.
- Attribute like latitude, longitude and depth were important in prediction except for the case when magnitude=3.
- When magnitude is 5 then we get very low precision as there is a lot of complexity in the tree.



```
-------
RESULTS
-------
Feature Importances/Gini Importances
['YEAR ', 'MONTH ', 'DATE', 'LAT (N)', 'LONG (E)', 'DEPTH (km)']
[0.         0.         0.37511162 0.16675072 0.45813766 0.        ]

Depth of tree:  5
No. of leaf nodes:  7

Accuracy: 1.0
Recall: 1.0
Precision Score: 1.0
F1 Score: 1.0
Confusion Matrix:
[[1681]]

-------
RESULTS
-------
Feature Importances/Gini Importances
['YEAR ', 'MONTH ', 'DATE', 'LAT (N)', 'LONG (E)', 'DEPTH (km)']
[0.08526352 0.11054594 0.14686014 0.35824617 0.19240889 0.10667534]

Depth of tree:  23
No. of leaf nodes:  779

Accuracy: 0.8691255205234979
Recall: 0.9217105263157894
Precision Score: 0.9327563249001332
F1 Score: 0.9272005294506949
Confusion Matrix:
[[  60  101]
 [ 119 1401]]

-------
RESULTS
-------
Feature Importances/Gini Importances
['YEAR ', 'MONTH ', 'DATE', 'LAT (N)', 'LONG (E)', 'DEPTH (km)']
[0.08847393 0.09922641 0.1285749  0.26106978 0.26780309 0.1548519 ]

Depth of tree:  32
No. of leaf nodes:  798

Accuracy: 0.8613920285544319
Recall: 0.2714285714285714
Precision Score: 0.22485207100591717
F1 Score: 0.2459546925566343
Confusion Matrix:
[[1410  131]
 [ 102   38]]

-----------------------------
```

3. 2000<=[‘YEAR’]<2010

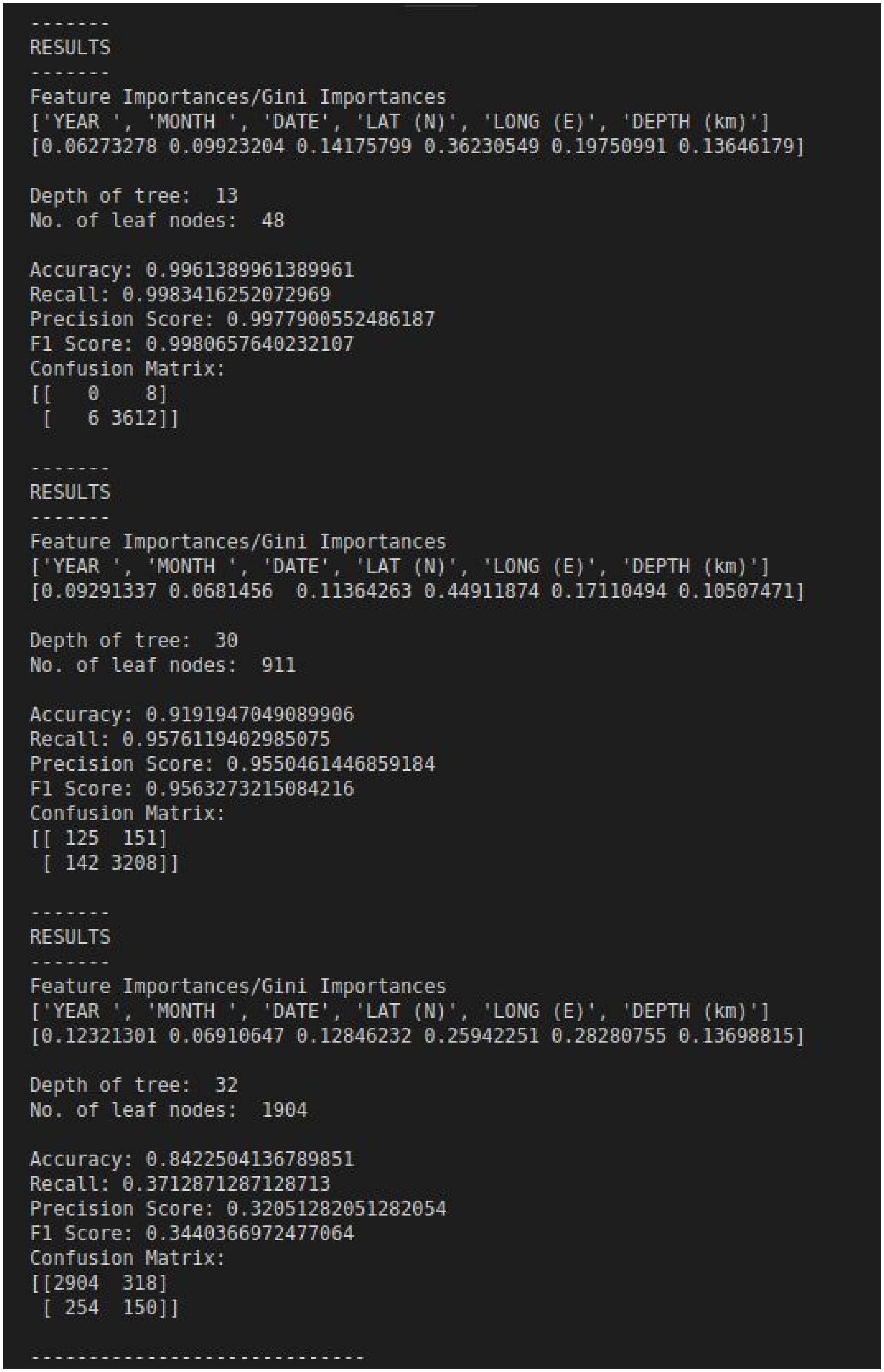
a. Classification based on magnitude >/< 3.
   i. Samples distribution : 8 vs 3618
   ii. Depth: 13
   iii. Leaf Nodes: 48
   iv. Important Attributes:Lat(0.4491), Long(0.1711)

b. Classification based on magnitude >/< 4.
   i. Samples distribution : 276 vs 3350
   ii. Depth: 30
   iii. Leaf Nodes: 911
   iv. Important Attributes:Long(0.2828), Lat(0.2594)

c. Classification based on magnitude >/< 5.
   i. Samples distribution : 3222 vs 404
   ii. Depth: 32
   iii. Leaf Nodes: 1904
   iv. Important Attributes:Long(0.2828), Lat(0.2594)

For this dataset:

- Accuracy decreases with increase in magnitude because of the same reasons as stated above. Here all the samples with magnitude less than 3, have been classified incorrectly by the classifier.
- Attribute like latitude, longitude and depth were important in prediction.
- When magnitude is 5 then we get very low precision as there is a lot of complexity in the tree.



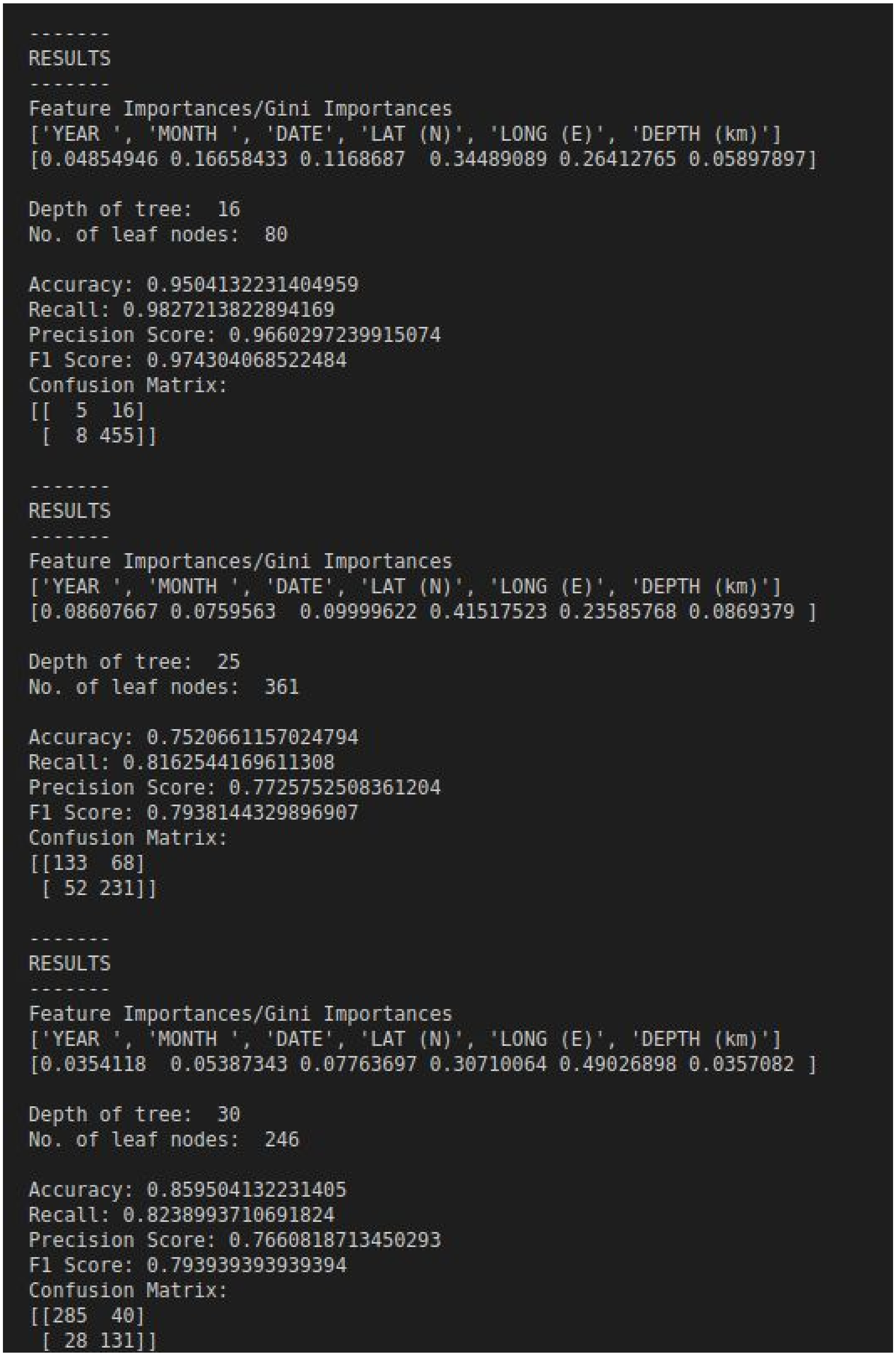
```
-------
RESULTS
-------
Feature Importances/Gini Importances
['YEAR ', 'MONTH ', 'DATE', 'LAT (N)', 'LONG (E)', 'DEPTH (km)']
[0.06273278 0.09923204 0.14175799 0.36230549 0.19750991 0.13646179]

Depth of tree:  13
No. of leaf nodes:  48

Accuracy: 0.9961389961389961
Recall: 0.9983416252072969
Precision Score: 0.9977900552486187
F1 Score: 0.9980657640232107
Confusion Matrix:
[[   0    8]
 [   6 3612]]

-------
RESULTS
-------
Feature Importances/Gini Importances
['YEAR ', 'MONTH ', 'DATE', 'LAT (N)', 'LONG (E)', 'DEPTH (km)']
[0.09291337 0.0681456  0.11364263 0.44911874 0.17110494 0.10507471]

Depth of tree:  30
No. of leaf nodes:  911

Accuracy: 0.9191947049089906
Recall: 0.9576119402985075
Precision Score: 0.9550461446859184
F1 Score: 0.9563273215084216
Confusion Matrix:
[[ 125  151]
 [ 142 3208]]

-------
RESULTS
-------
Feature Importances/Gini Importances
['YEAR ', 'MONTH ', 'DATE', 'LAT (N)', 'LONG (E)', 'DEPTH (km)']
[0.12321301 0.06910647 0.12846232 0.25942251 0.28280755 0.13698815]

Depth of tree:  32
No. of leaf nodes:  1904

Accuracy: 0.8422504136789851
Recall: 0.3712871287128713
Precision Score: 0.32051282051282054
F1 Score: 0.3440366972477064
Confusion Matrix:
[[2904  318]
 [ 254  150]]

-----------------------------
```

4. 2010<=[‘YEAR’]<2020

- a. Classification based on magnitude >/< 3.
    - i. Samples distribution : 21 vs 463
    - ii. Depth: 16
    - iii. Leaf Nodes: 80
    - iv. Important Attributes:Lat(0.34489), Long(0.26412)
- b. Classification based on magnitude >/< 4.
    - i. Samples distribution : 201 vs 283
    - ii. Depth: 25
    - iii. Leaf Nodes: 361
    - iv. Important Attributes:Long(0.49026), Lat(0.3071)
- c. Classification based on magnitude >/< 5.
    - i. Samples distribution : 325 vs 159
    - ii. Depth: 30
    - iii. Leaf Nodes: 246
    - iv. Important Attributes:Long(0.4902), Lat(0.3071)

For this dataset:

- We get very low accuracy when the magnitude is 4.
- Complexity of tree increases with increase in magnitude.
- Attribute like latitude, longitude and depth were important in prediction.



```
-------
RESULTS
-------
Feature Importances/Gini Importances
['YEAR ', 'MONTH ', 'DATE', 'LAT (N)', 'LONG (E)', 'DEPTH (km)']
[0.04854946 0.16658433 0.1168687  0.34489089 0.26412765 0.05897897]

Depth of tree:  16
No. of leaf nodes:  80

Accuracy: 0.9504132231404959
Recall: 0.9827213822894169
Precision Score: 0.9660297239915074
F1 Score: 0.974304068522484
Confusion Matrix:
[[  5  16]
 [  8 455]]

-------
RESULTS
-------
Feature Importances/Gini Importances
['YEAR ', 'MONTH ', 'DATE', 'LAT (N)', 'LONG (E)', 'DEPTH (km)']
[0.08607667 0.0759563  0.09999622 0.41517523 0.23585768 0.0869379 ]

Depth of tree:  25
No. of leaf nodes:  361

Accuracy: 0.7520661157024794
Recall: 0.8162544169611308
Precision Score: 0.7725752508361204
F1 Score: 0.7938144329896907
Confusion Matrix:
[[133  68]
 [ 52 231]]

-------
RESULTS
-------
Feature Importances/Gini Importances
['YEAR ', 'MONTH ', 'DATE', 'LAT (N)', 'LONG (E)', 'DEPTH (km)']
[0.0354118  0.05387343 0.07763697 0.30710064 0.49026898 0.0357082 ]

Depth of tree:  30
No. of leaf nodes:  246

Accuracy: 0.859504132231405
Recall: 0.8238993710691824
Precision Score: 0.7660818713450293
F1 Score: 0.793939393939394
Confusion Matrix:
[[285  40]
 [ 28 131]]
```

Analysis among different datasets

- Among all the classifiers, the accuracy decreases on increasing the magnitude upon increasing the magnitude because the complexity of the decision tree increases.
- This can also be verified by looking at the precision and recall values of the data points for magnitude=5 case.
- For the case when years lie in the range 2000-2010, the number of samples actually correctly predicted when magnitude=3 was zero, because there were just 8 samples in one class, while about 3618 samples in the other class i.e the class distribution was very much skewed.
- When we subsample the data we get pretty simple trees when magnitude is >/< 3 as number of samples are fairly less in this case.
- With increase in magnitude, number of misclassified points increases.