Decision Tree Results

Digit Recognition:

1) Accuracy :
   a) Training dataset : 100%
   b) Testing dataset without any pruning: 86%
   c) Training dataset with pruning (73% to 27% Split): 100%
   d) Validation dataset with pruning : 88.86%
   e) Testing dataset with pruning : 84.36%
2) References Used:
   a) Scikit-documentation
   b) Worked along with Nikhil Jonnalagadda (Student ID: 800975714), Chaitanya Sri Krishna (Student ID: 800960353), BhanuPraneeth Reddy (Student ID: 800953521) on implementation.

Data Exploration Results:

a. What is the number of attributes in each dataset?

There are 64 attributes where each attributerepresents each pixel in the given image which is represented as an 8x8 input matrix.

b. What is the number of observations?

3823 observations in training dataset, 1797 observations in testing dataset.

c. What is the mean and standard deviation of each attribute?

Attribute	Mean	Standard deviation
1	0	0
2	0.3013	0.8670
3	5.4818	4.6316
4	11.8059	4.2598
5	11.4515	4.5376
6	5.5054	5.6131
7	1.3874	3.3714
8	0.1423	1.0516
9	0.0021	0.0886
10	1.9605	3.0524
11	10.5773	5.4355
12	11.7154	4.0122
13	10.6249	4.7881
14	8.2956	5.9356
15	2.2001	4.0622

16	0.1520	0.9888
17	0.0050	0.1199
18	2.5959	3.4541
19	9.5807	5.8861
20	6.7350	5.9183
21	7.1865	6.1427
22	8.0484	6.2915
23	2.0460	3.5817
24	0.0492	0.4355
25	0.0010	0.0323
26	2.3356	3.0859
27	9.2391	6.1281
28	9.1337	5.9026
29	9.6733	6.2829
30	7.8676	6.0024
31	2.3403	3.6247
32	0.0031	0.0646
33	0.0013	0.0361
34	2.0429	3.2117
35	7.6594	6.2596
36	9.2380	6.1902
37	10.3476	5.9201
38	9.2001	5.8793
39	2.9126	3.4863
40	0	0
41	0.0275	0.3162
42	1.4057	2.9342
43	6.4567	6.5054
44	7.1873	6.4691
45	7.9215	6.3164
46	8.6749	5.8059
47	3.5103	4.3691
48	0.0199	0.2137
49	0.0178	0.2691
50	0.8200	2.0090
51	7.8690	5.6666
52	9.8857	5.1416
53	9.7648	5.3150
54	9.2833	5.9409
55	3.7439	4.9017
56	0.1483	0.7678
57	0.0003	0.0162
58	0.2830	0.9280
59	5.8559	4.9800
60	11.9430	4.3345
61	11.4612	4.9919
62	6.7005	5.7758
63	2.1057	4.0283

64 65	0.2022 4.4973	1.1507 2.8698

Amazon Reviews:

a. What is the number of attributes in each dataset?

The dataset contains three attributes- Product, Review and Rating.

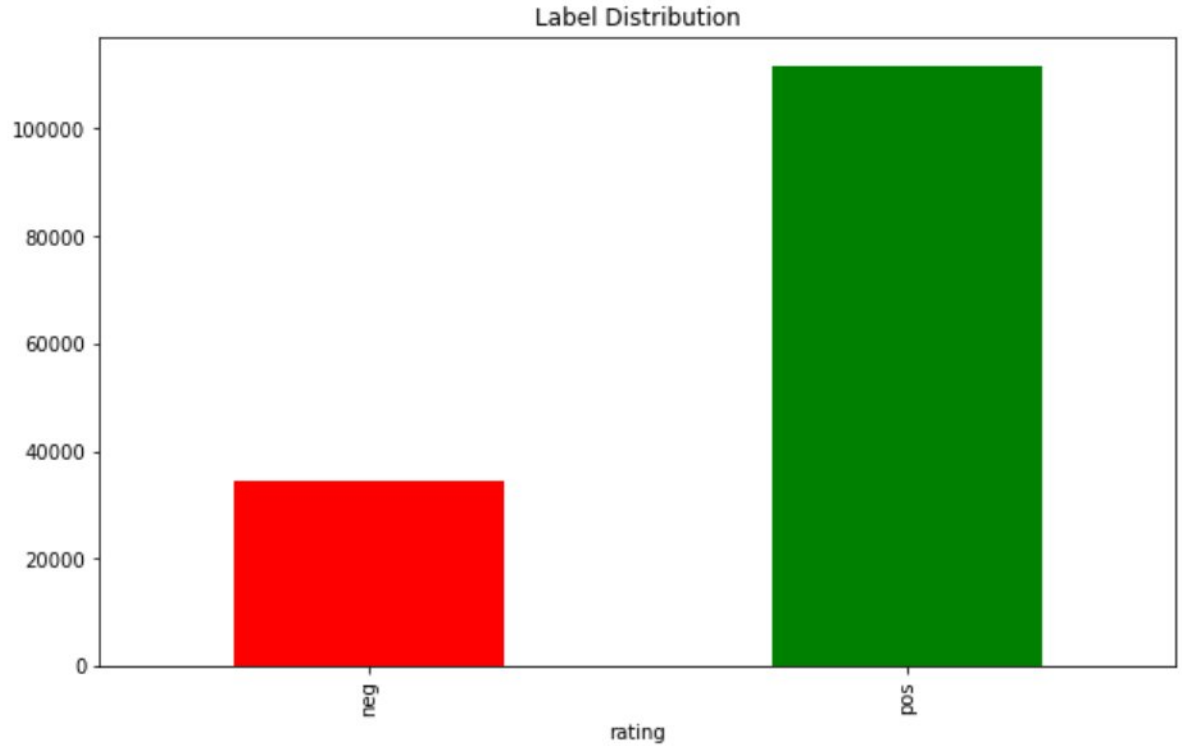
b. What is the number of observations?

146825 observations on Training dataset, 36708 on Testing Dataset.

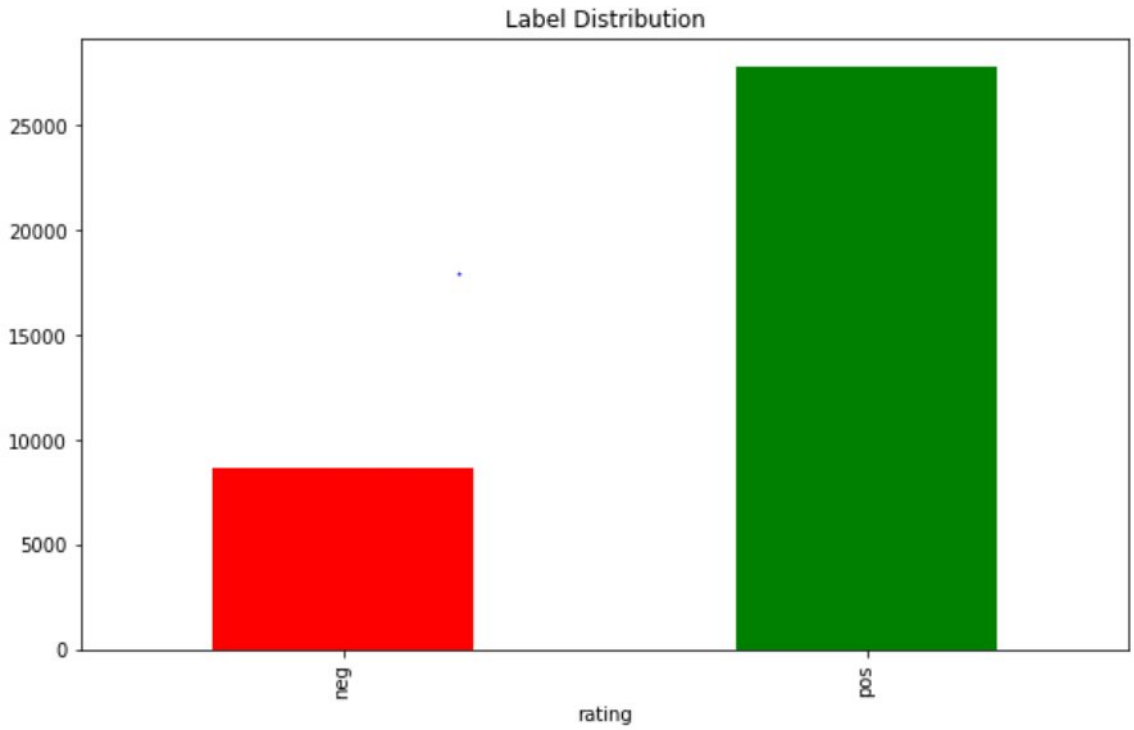
c. What is the mean and standard deviation of each attribute?

The mean and standard deviation for the ratings attribute is 4.120430078052725 and 1.2853703237434095 respectively.

d. What is the distribution of the different classes in each of the datasets?



Training Dataset



Test Dataset

Accuracy:

- Training dataset : 99%
- Testing dataset without any pruning : 63%