Ian Zimmer

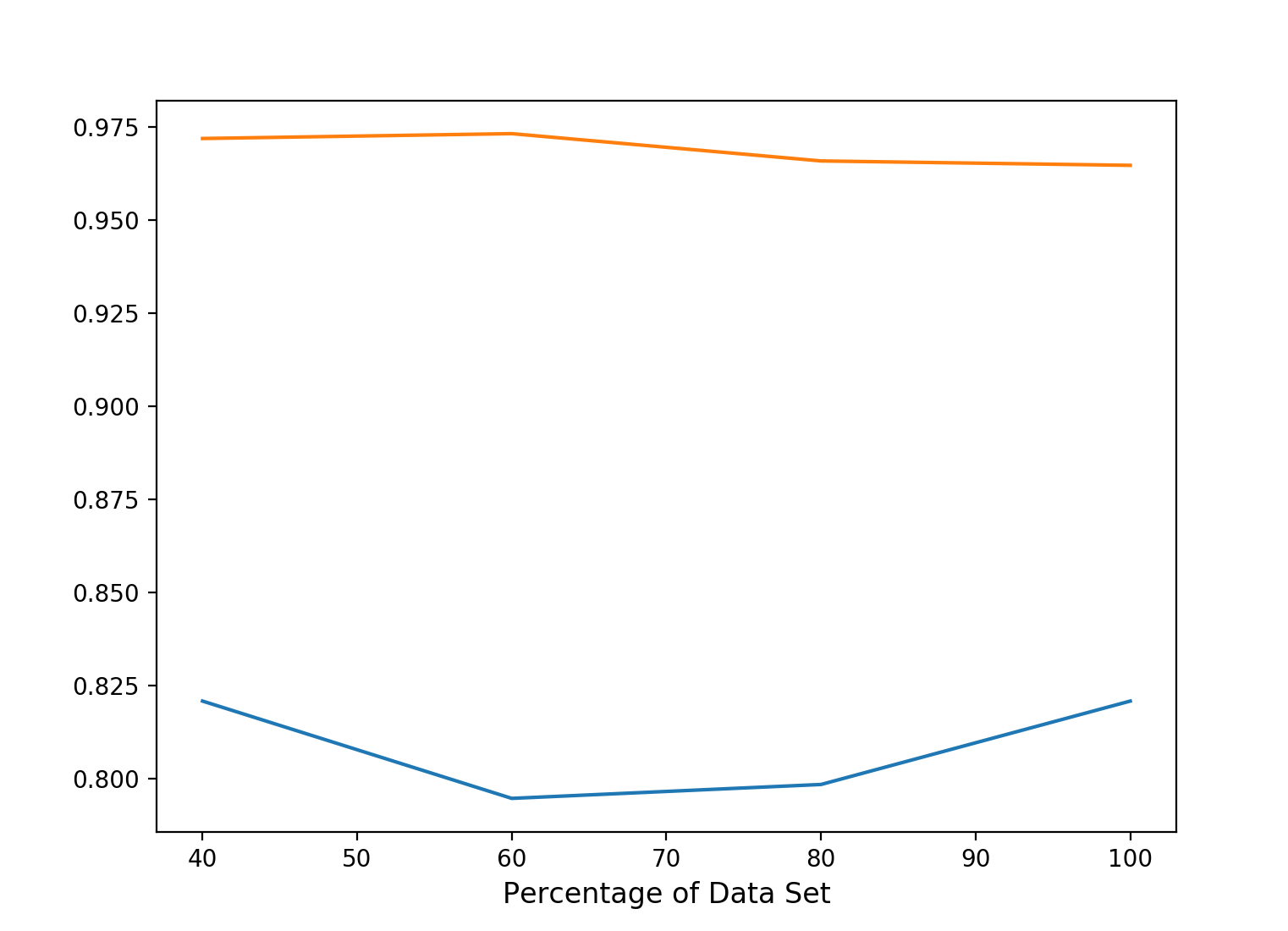
0028719394

Late Day Used: 1

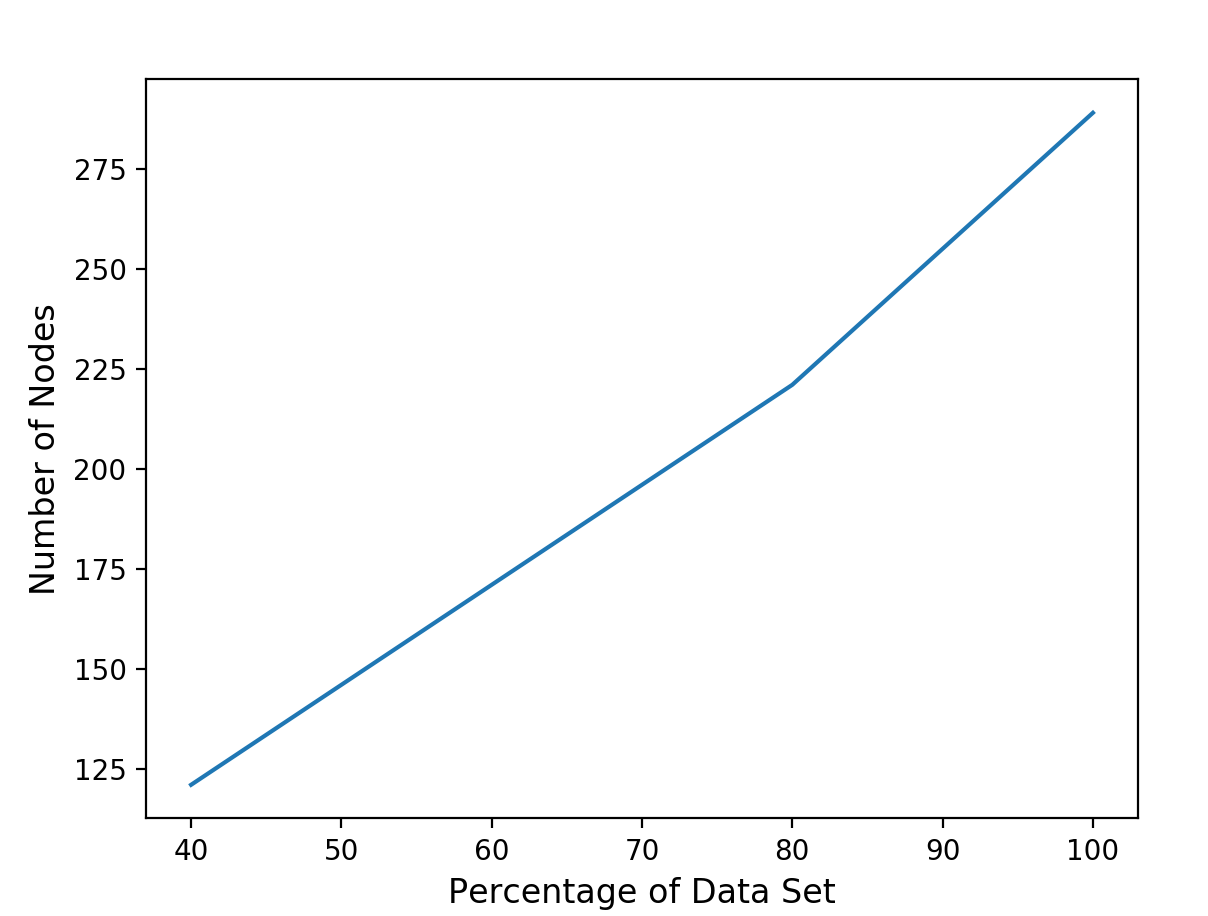
I collaborated with Oscar Dillman and Ishan Kaul. I affirm that I wrote the solutions in my own words and that I understand the solutions I am submitting.

1. Vanilla

|  |  |  |  |
| --- | --- | --- | --- |
| Training Set Percentage | Training Set Accuracy | Test Set Accuracy | Number of Nodes |
| 40 | 97.2 | 82.1 | 121 |
| 60 | 97.3 | 79.5 | 171 |
| 80 | 96.6 | 79.9 | 221 |
| 100 | 96.5 | 82.1 | 289 |

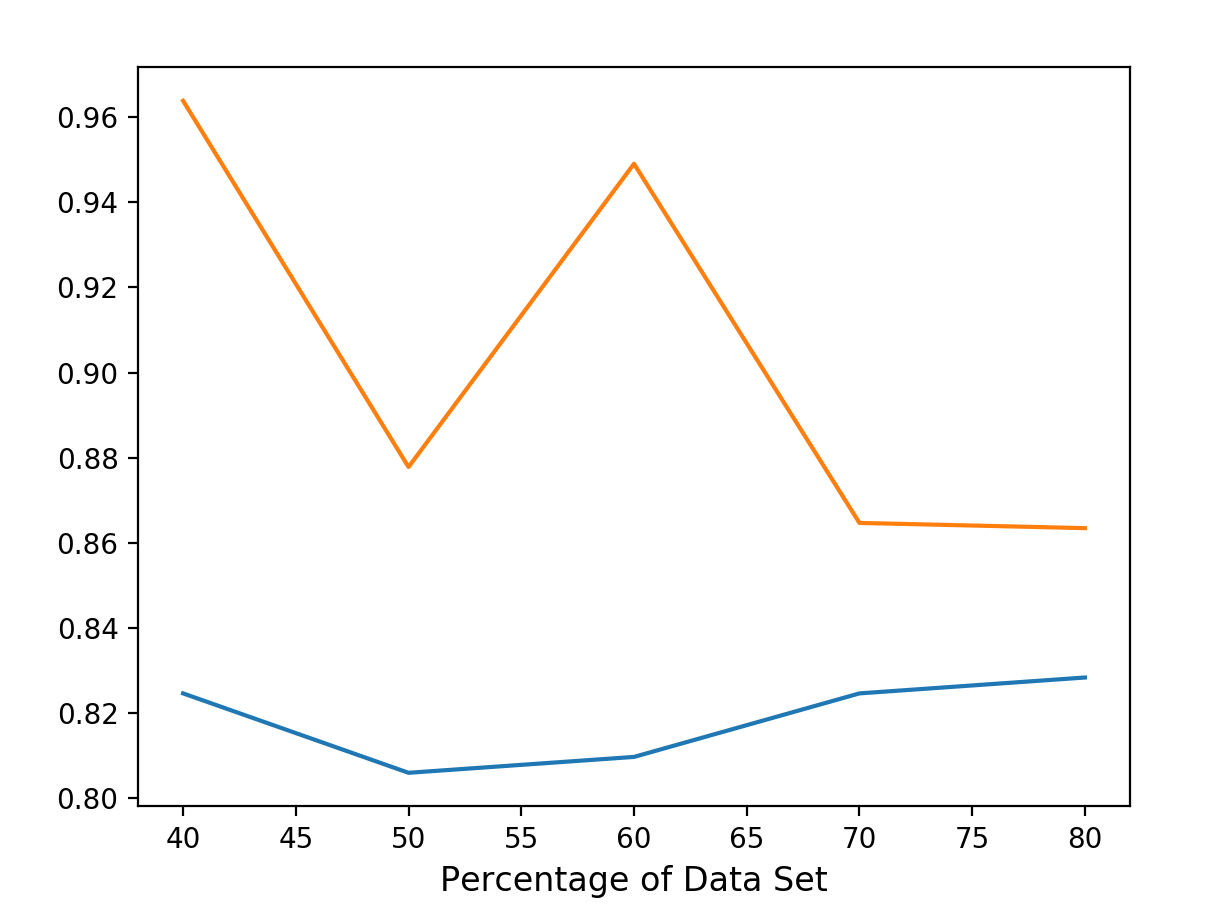


Training Accuracy is the orange line and the Test Accuracy is the blue line

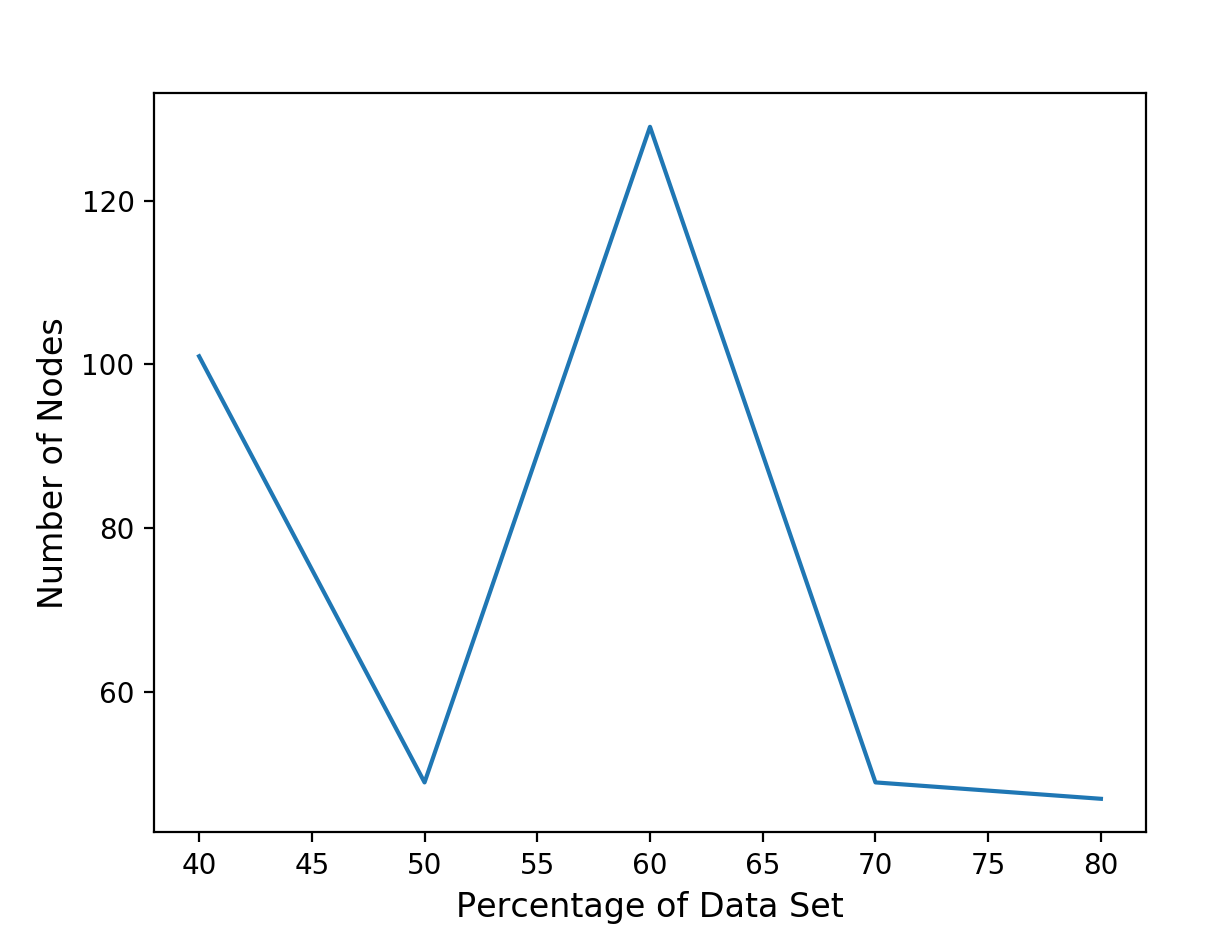


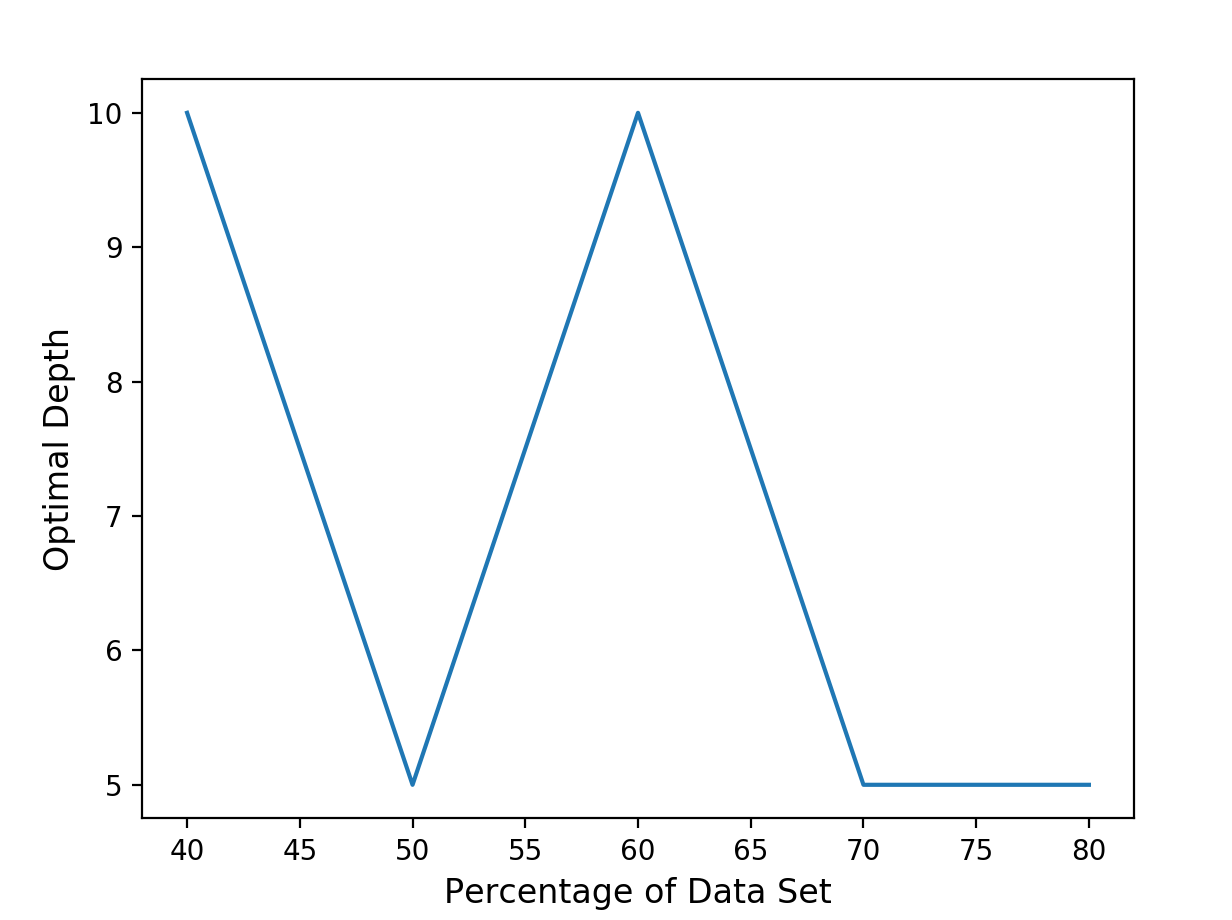
2. Max Depth

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Training Set Percentage | Best Depth Value | Training Set Accuracy | Test Set Accuracy | Number of Nodes |
| 40 | 10 | 96.4 | 82.5 | 101 |
| 50 | 5 | 87.8 | 81.0 | 49 |
| 60 | 10 | 94.9 | 80.6 | 129 |
| 70 | 5 | 86.5 | 82.5 | 49 |
| 80 | 5 | 86.3 | 82.8 | 47 |



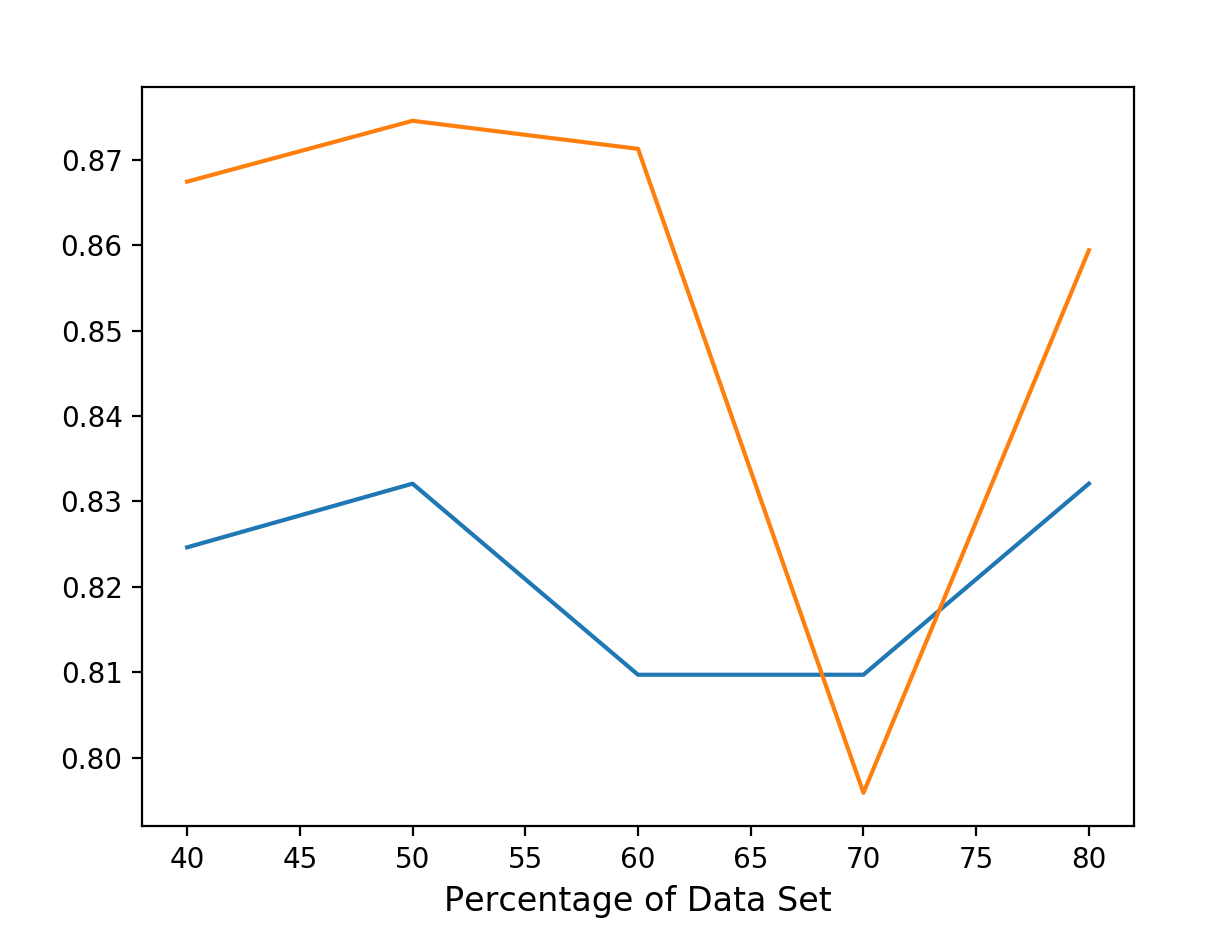
Training Accuracy is the orange line and the Test Accuracy is the blue line



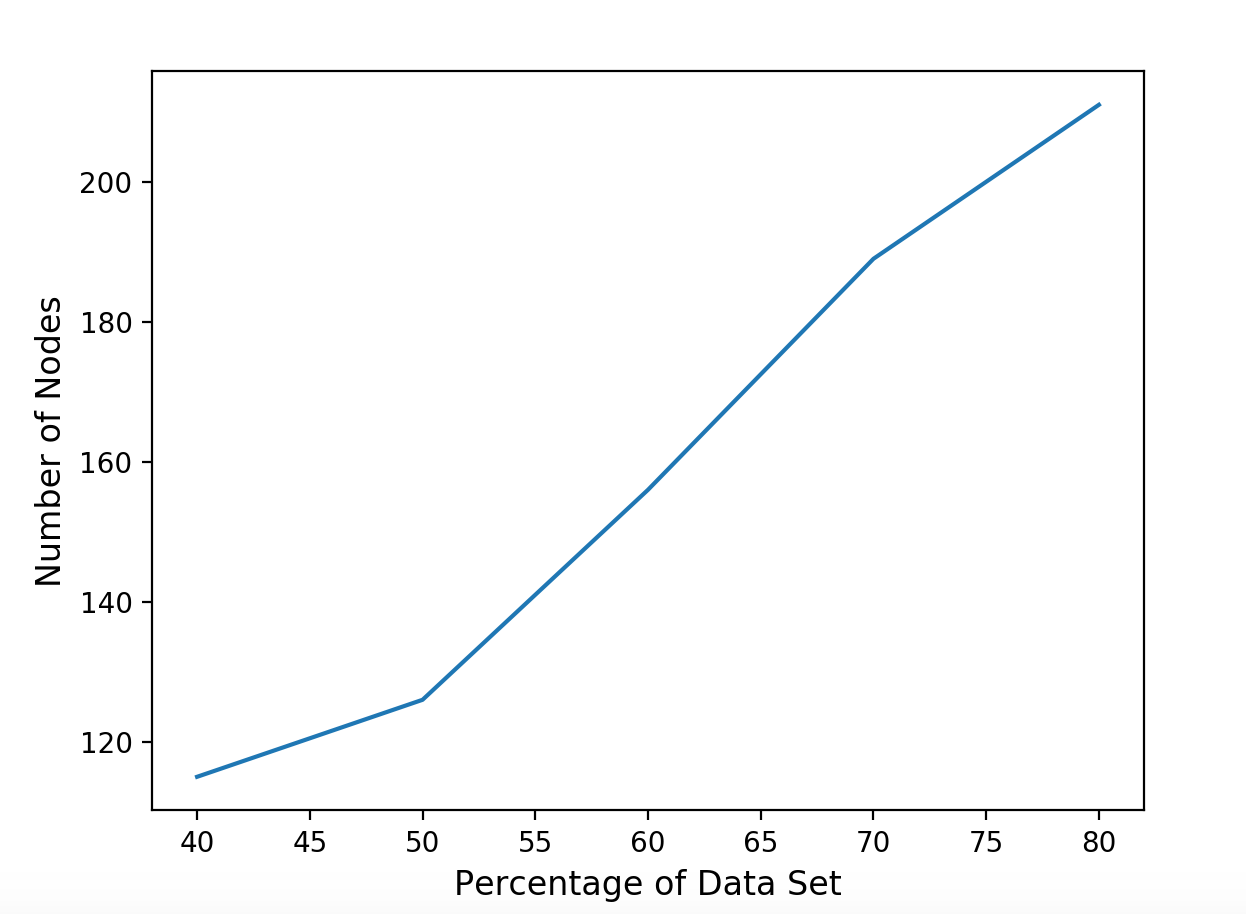


3.Prune

|  |  |  |  |
| --- | --- | --- | --- |
| Training Set Percentage | Training Set Accuracy | Test Set Accuracy | Number of nodes |
| 40 | 86.8 | 82.5 | 115 |
| 50 | 87.5 | 83.2 | 126 |
| 60 | 87.1 | 81.0 | 156 |
| 70 | 79.6 | 81.0 | 189 |
| 80 | 85.9 | 83.2 | 211 |



Training Accuracy is the orange line and the Test Accuracy is the blue line



4.

The tree is pruned on a validation set instead of directly on the test set because that would essentially be training and adjusting the tree on the test data. Our goal is to see how well we can train the data to predict the test data correctly, and pruning using the test data would be the opposite of this. Also, the validation set is most likely going to be smaller than the test data, so it will adjust the tree quicker.

5.

In order to convert the decision tree from a classification to a ranking model in the depth and prune cases, I would focus on the leaves that aren’t all one values of labels. In these cases, to make a classification model as of now, I list the majority value as the predicted label. To make a ranking model, I would get the frequency of both values in label and then display the predicted label as the percent of that leaf being one value vs the other.