Jon Rippe  
CSCE A415  
HW #1.2

## Problem 2: Large Data Set / Jupyter Notebook

### Entries = 10,000 Classes = 3 Attributes = 17

### Examine and Analyze Data:

To get an idea of a possible “best fit”, 1,530 decision trees were systematically created and compared with varying tree depths and test/train split ratios. Results below. The test/train splits were all stratified to promote consistency.

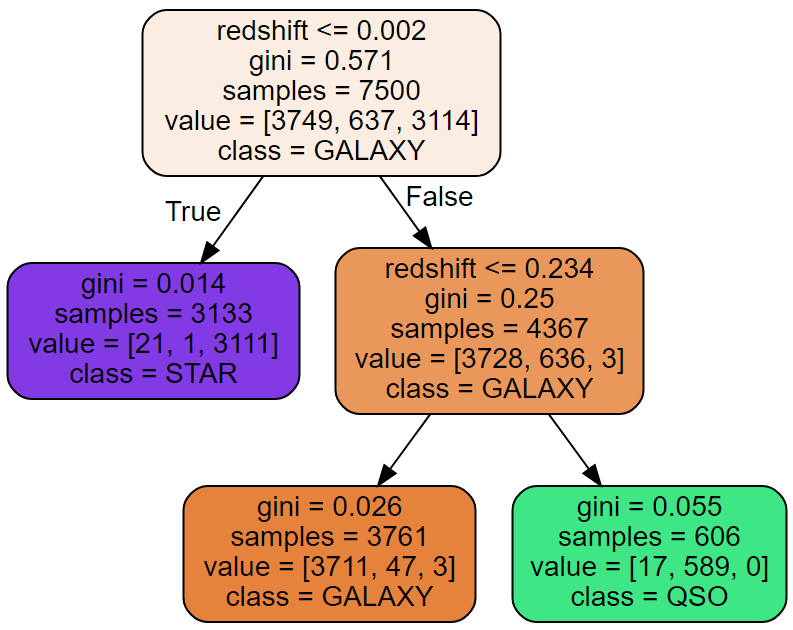
|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Test Data Ratio | | | | | | | | |  |
|  | **Tree Fit** | **0.1** | **0.2** | **0.3** | **0.4** | **0.5** | **0.6** | **0.7** | **0.8** | **0.9** | **Mean** |
| Tree Depth | **1** | 91.2% | 91.2% | 91.2% | 91.2% | 91.2% | 91.2% | 91.2% | 91.2% | 91.2% | **91.2%** |
| **2** | 98.9% | 98.7% | 98.8% | 98.8% | 98.8% | 98.8% | 98.8% | 98.7% | 98.7% | **98.8%** |
| **3** | 98.7% | 98.8% | 98.9% | 98.8% | 98.7% | 98.8% | 98.6% | 98.5% | 98.5% | **98.7%** |
| **4** | 98.8% | 98.8% | 98.8% | 98.7% | 98.7% | 98.6% | 98.6% | 98.5% | 98.3% | **98.7%** |
| **5** | 98.6% | 98.7% | 98.8% | 98.8% | 98.7% | 98.6% | 98.6% | 98.5% | 98.1% | **98.7%** |
| **6** | 98.9% | 98.9% | 98.8% | 98.7% | 98.5% | 98.5% | 98.6% | 98.4% | 98.2% | **98.7%** |
| **7** | 99.0% | 98.9% | 98.8% | 98.7% | 98.7% | 98.5% | 98.4% | 98.3% | 98.1% | **98.7%** |
| **8** | 98.7% | 98.8% | 98.8% | 98.6% | 98.7% | 98.6% | 98.3% | 98.2% | 98.1% | **98.6%** |
| **9** | 98.8% | 98.9% | 98.8% | 98.6% | 98.6% | 98.6% | 98.3% | 98.0% | 97.9% | **98.6%** |
| **10** | 98.8% | 98.8% | 98.7% | 98.5% | 98.5% | 98.4% | 98.4% | 98.2% | 98.1% | **98.5%** |
| **11** | 98.7% | 98.7% | 98.7% | 98.6% | 98.4% | 98.3% | 98.3% | 98.2% | 98.0% | **98.5%** |
| **12** | 98.9% | 98.5% | 98.5% | 98.6% | 98.5% | 98.3% | 98.1% | 98.2% | 98.0% | **98.4%** |
| **13** | 98.5% | 98.5% | 98.7% | 98.4% | 98.4% | 98.3% | 98.2% | 98.2% | 98.0% | **98.4%** |
| **14** | 98.6% | 98.5% | 98.5% | 98.5% | 98.4% | 98.3% | 98.2% | 97.9% | 98.2% | **98.4%** |
| **15** | 98.6% | 98.5% | 98.5% | 98.5% | 98.3% | 98.3% | 98.2% | 98.2% | 98.2% | **98.4%** |
| **16** | 98.5% | 98.5% | 98.5% | 98.5% | 98.3% | 98.3% | 98.1% | 98.1% | 97.8% | **98.4%** |
| **17** | 98.5% | 98.4% | 98.4% | 98.6% | 98.2% | 98.2% | 98.2% | 98.0% | 97.9% | **98.3%** |
|  | **Mean** | **98.3%** | **98.3%** | **98.3%** | **98.2%** | **98.1%** | **98.0%** | **97.9%** | **97.8%** | **97.7%** |  |

A maximum tree depth of 2 yields the most accurate decision tree (explanation as to why coming later). Additionally, tree fit appears to drop off when the test/train ratio exceeds 30%. Based on these results, a tree depth of 2 was chosen and the test/train ratio was set to the Scikit default of 25%.

### Create the Final Tree

Tree Depth = 2  
Test/Train Ratio = 0.25  
Tree Fit (Score) ≈ 98.8% - 99.0%

The biggest factor in categorizing the data into classes is an entry’s *redshift* value. Therefore, a maximum tree depth of 2 gave the best fit. Anything greater causes overfitting and ultimately reduces the tree’s accuracy.

Below is an auto-generated, two-deep, binary decision tree and a simplified, non-binary version.