MODELS THAT EXPLAIN THEMSELVES
PROJECT 1: DECISION TREES

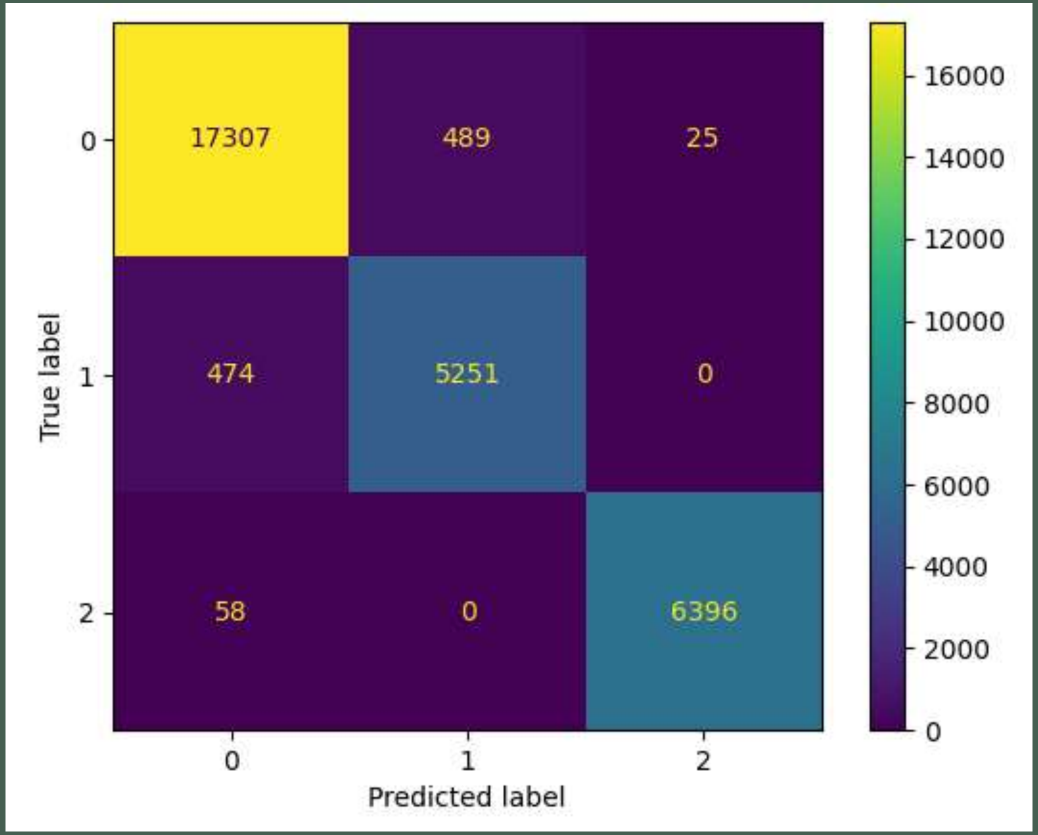
Valentina Tretti
Radhika Yadav

Dataset and task

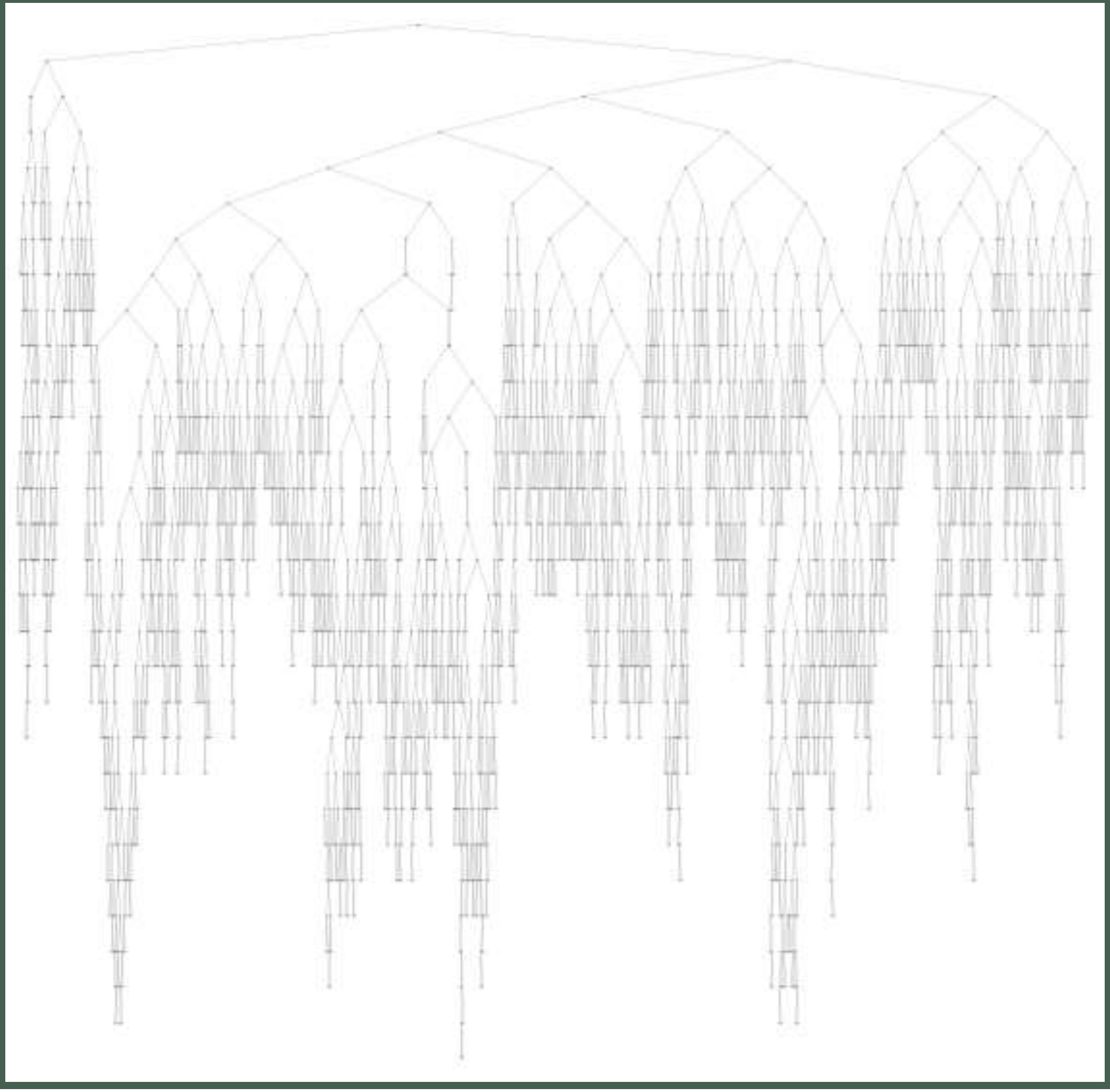
- **Name:** Stellar Classification Dataset - SDSS17
- **Authors:** Released by Sloan Digital Sky Survey under public domain
- **Size:** 100,000 rows and 18 columns
- **Task:** Stellar classification refers to the classification of stars based on certain spectral characteristics. The aim of the dataset is to classify galaxies (0), quasars (1) and stars (2) based on these.
- **Columns:** obj_ID, alpha, delta, u, g, r, i, z, run_ID, rerun_ID, cam_col, field_ID, spec_obj_ID, class, redshift, plate, MJD, fiber_ID

Accuracy and Confusion Matrix

- ❖ Accuracy: 0.9651333333333333



Non-optimized Decision Tree

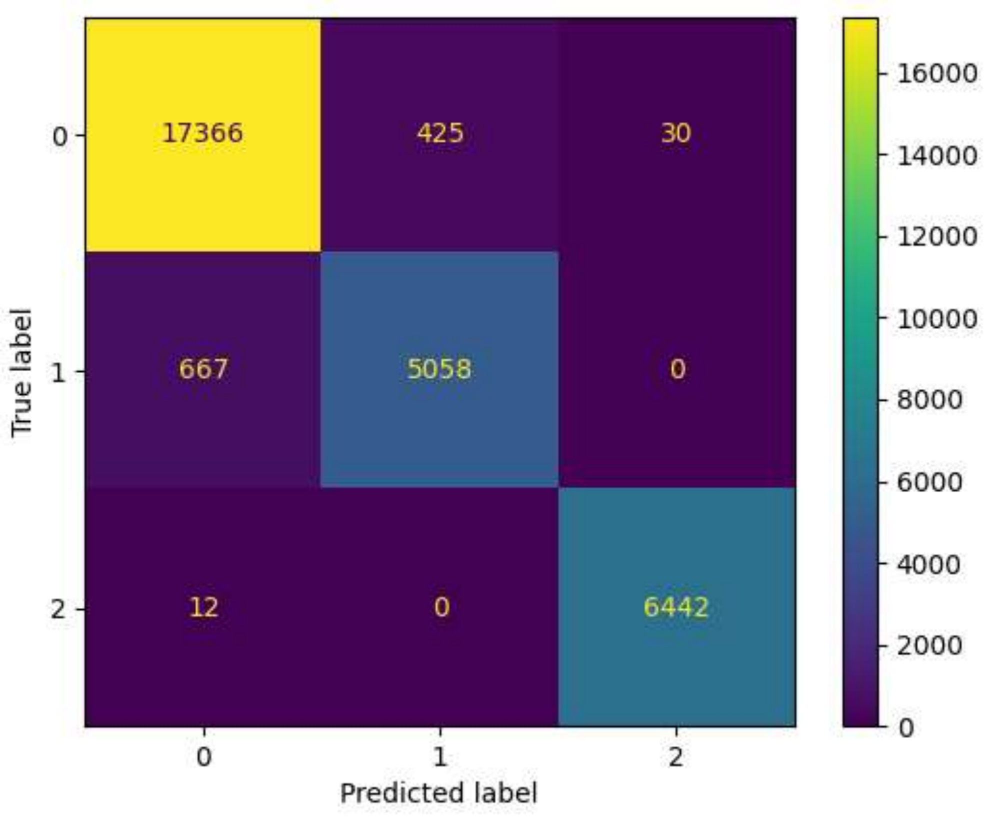


Hyperparameter Tuning

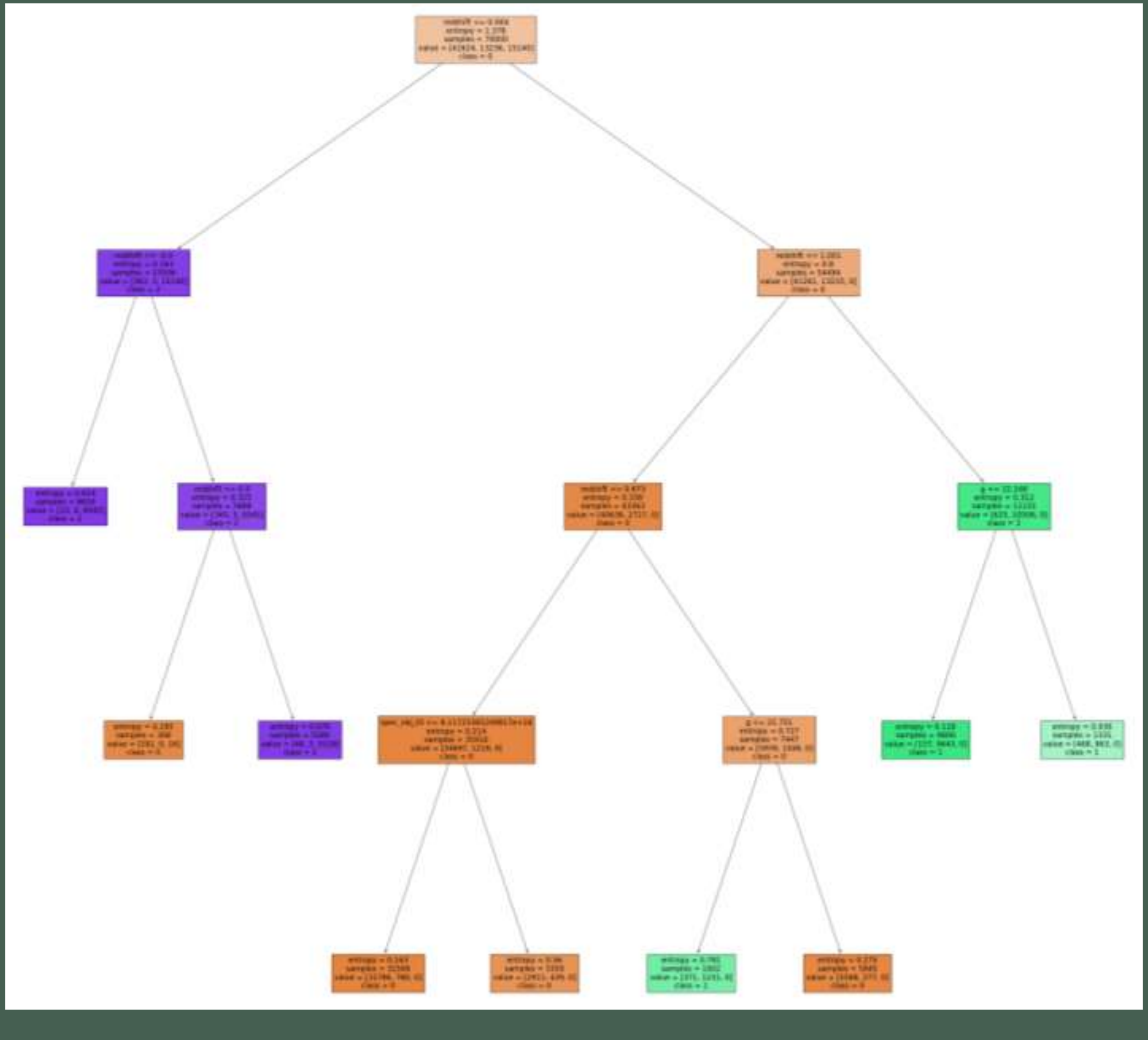
- Hyperparameters optimized:
   1. criterion: ['gini’, 'entropy’, 'log_loss’]
   2. splitter: ['best', 'random’]
   3. max_depth: np.arange(1,10)
   4. max_leaf_nodes: np.arange(2,10)
- Best hyperparameters for the model:
   1. criterion: 'entropy’
   2. splitter: best
   3. max_depth: 4
   4. max_leaf_nodes: 9

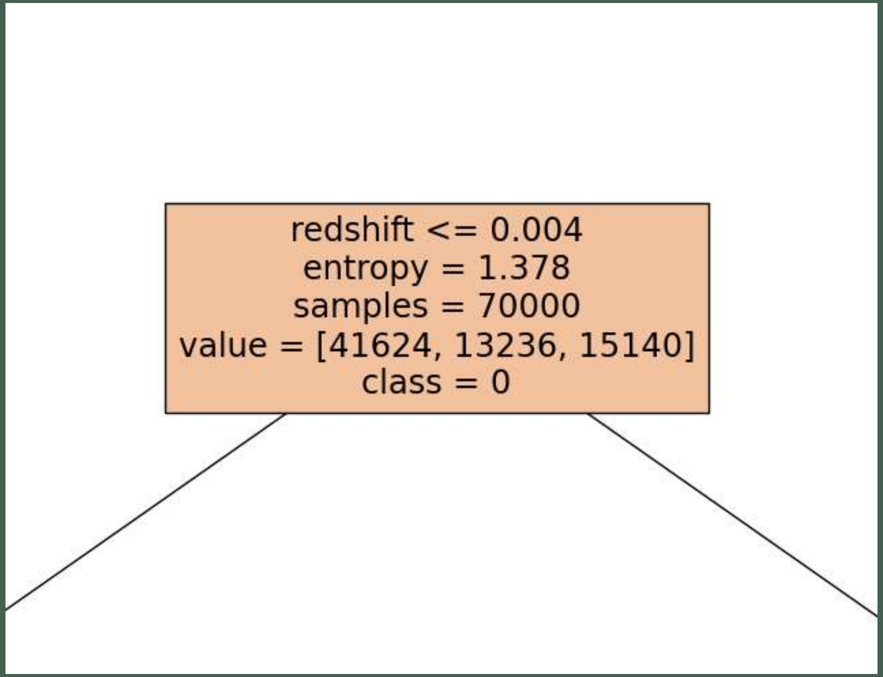
Accuracy and Confusion Matrix

- Accuracy: 0.9622



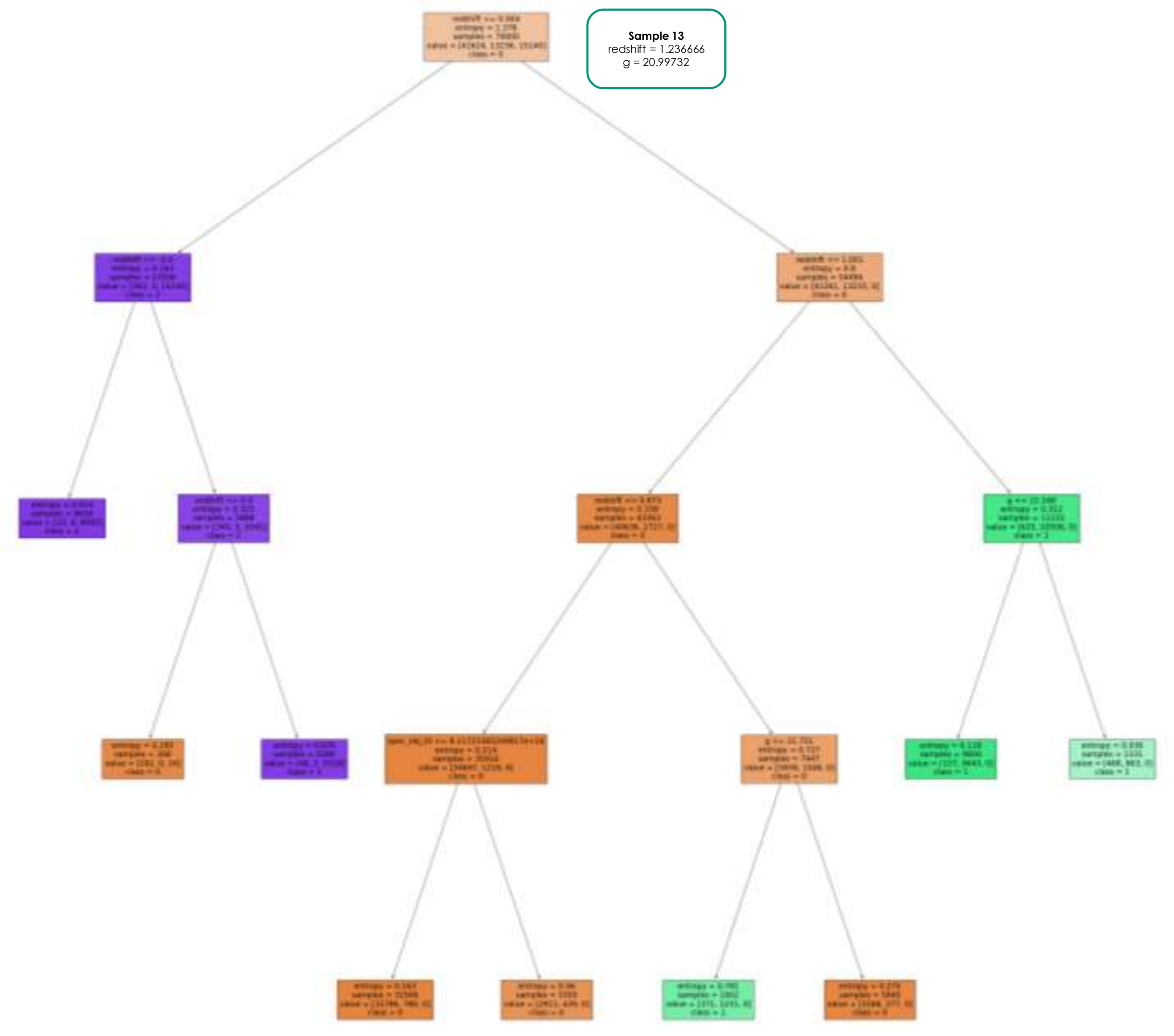
Optimized Decision Tree



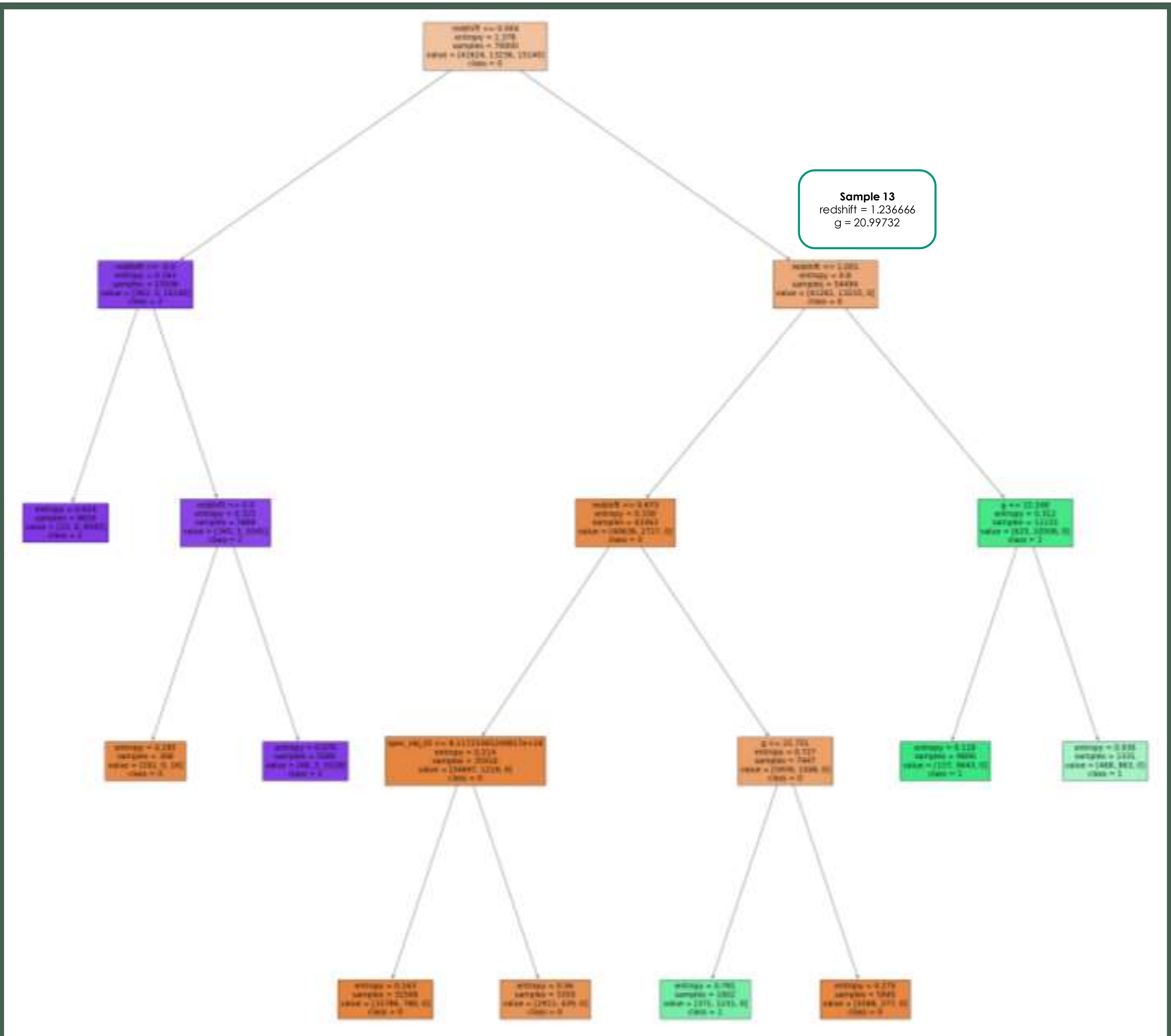


redshift <= 0.004
entropy = 1.378
samples = 70000
value = [41624, 13236, 15140]
class = 0

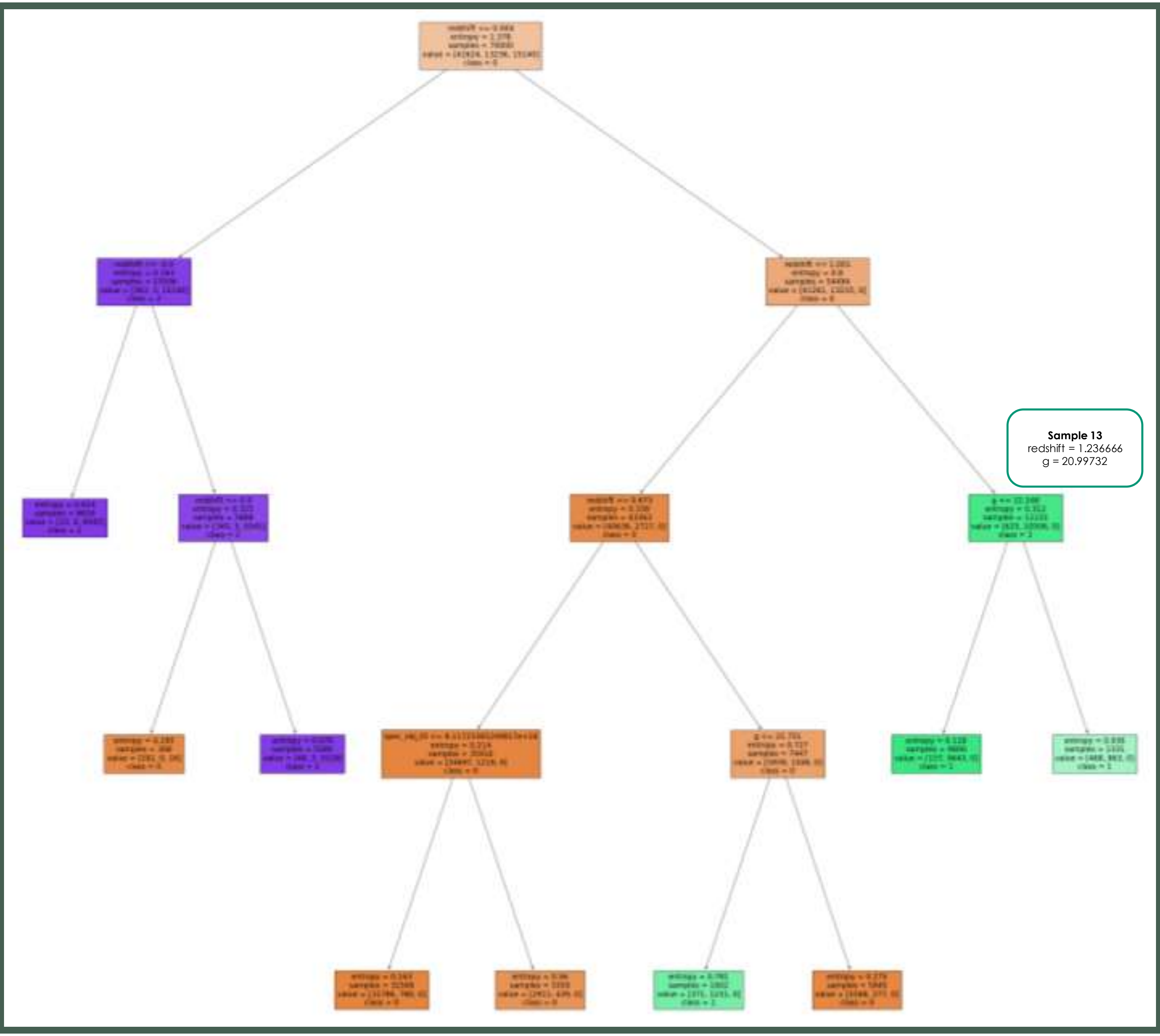
Correctly classified Example 1



Sample 13
redshift = 1.236666
g = 20.99732

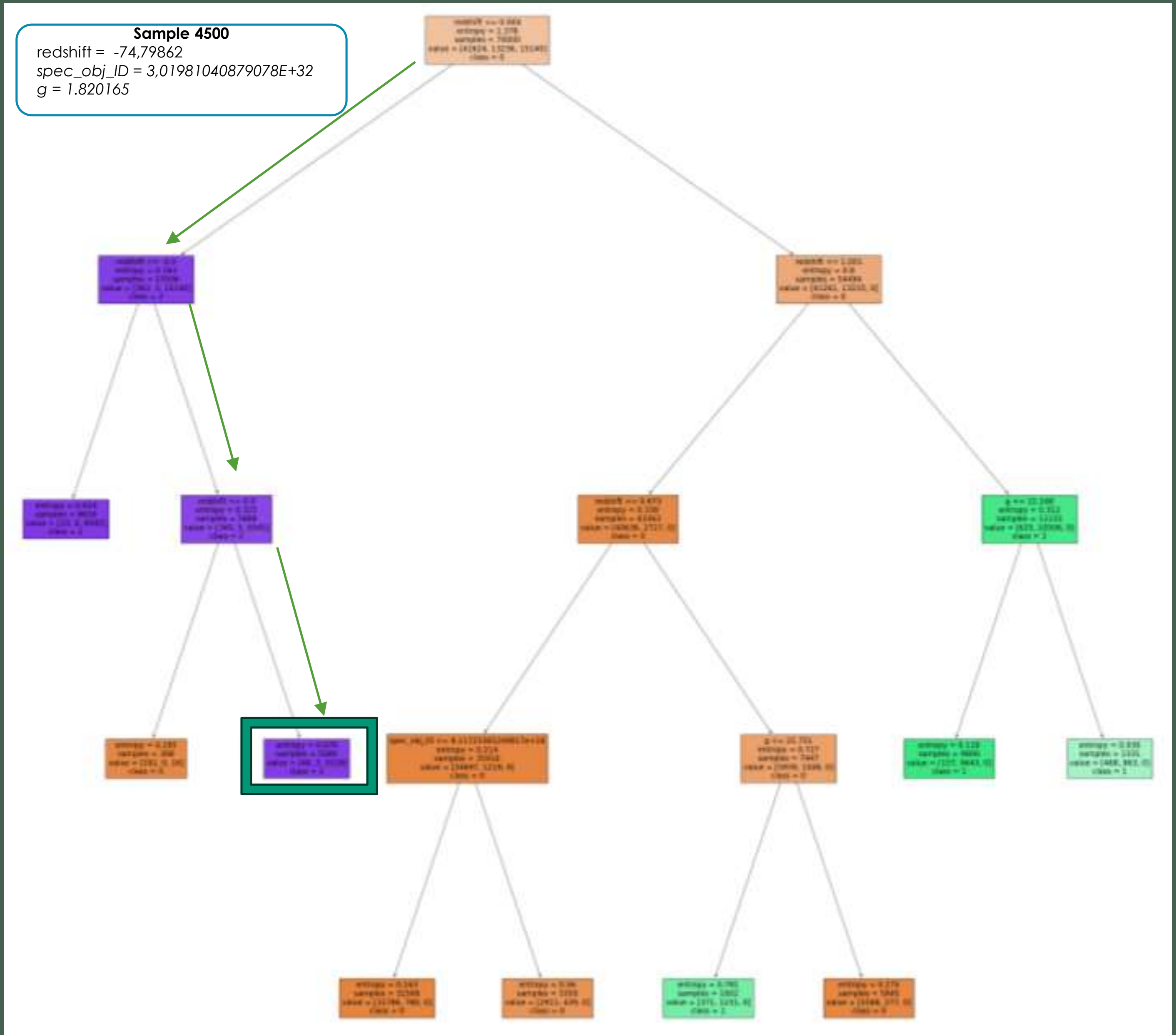


Sample 13
redshift = 1.236666
g = 20.99732



Sample 13
redshift = 1.236666
g = 20.99732

Correctly classified Example 2



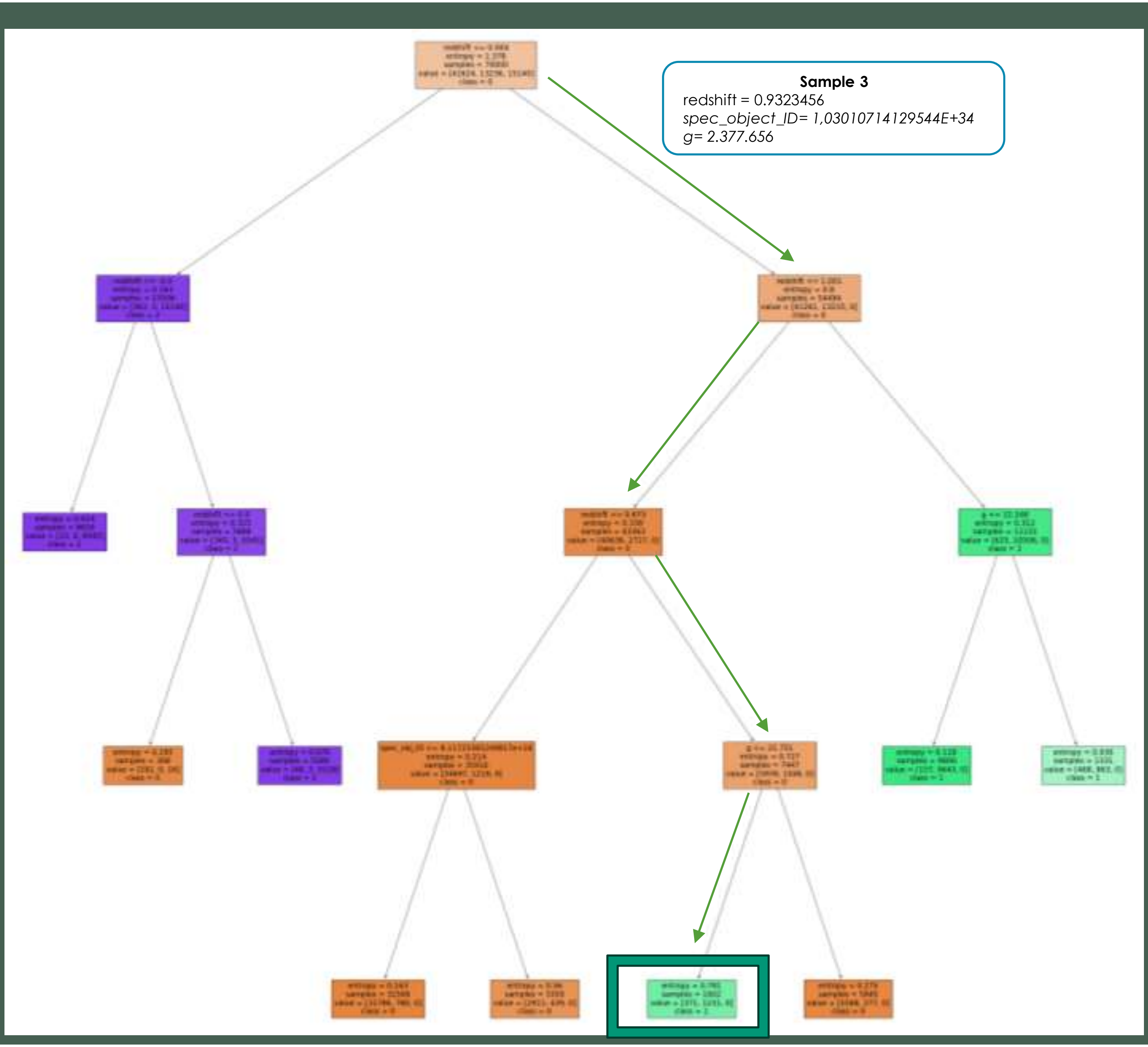
Sample 4500

redshift = -74,79862

spec_obj_ID = 3,01981040879078E+32

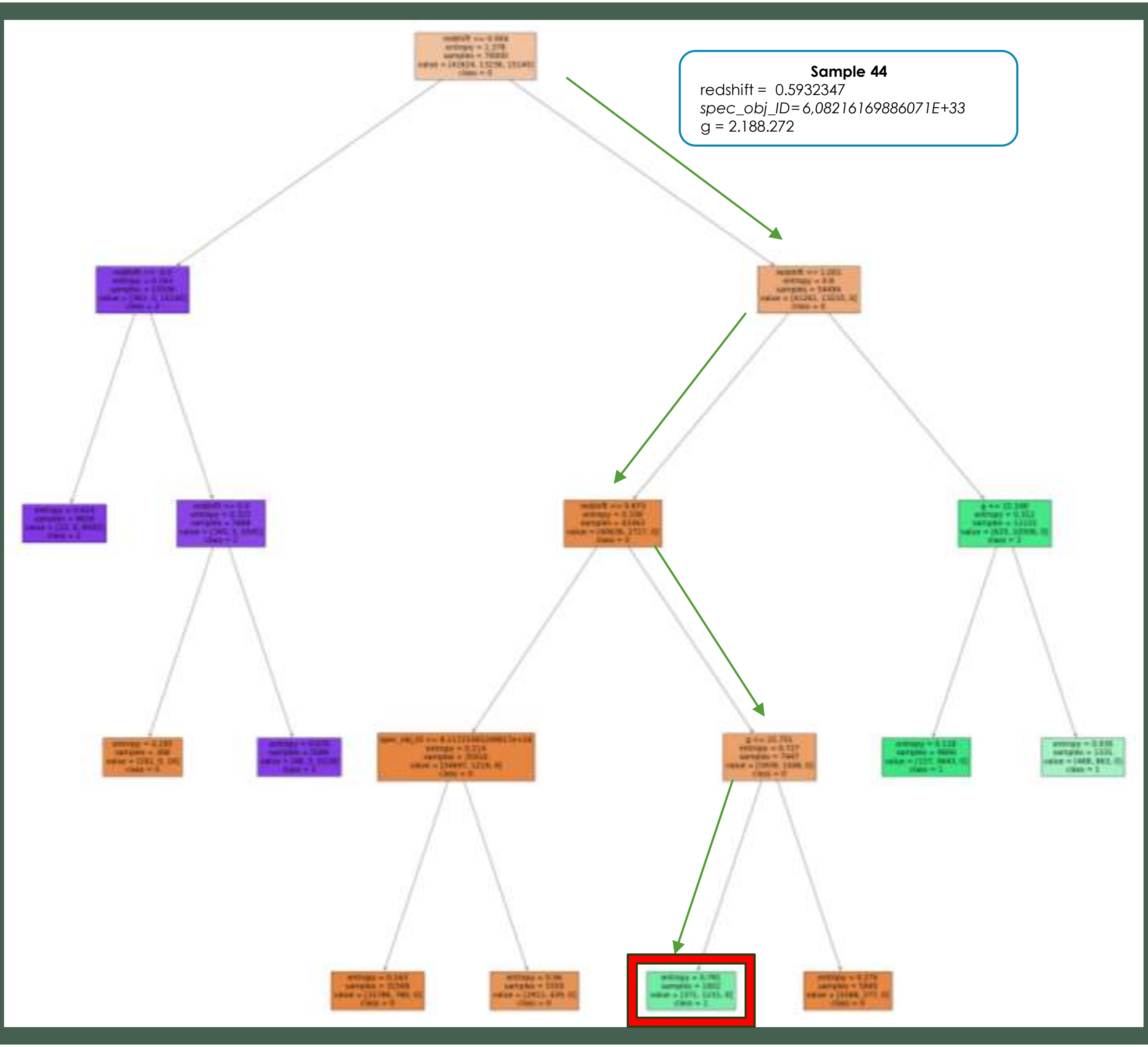
g = 1.820165

Correctly classified Example 3



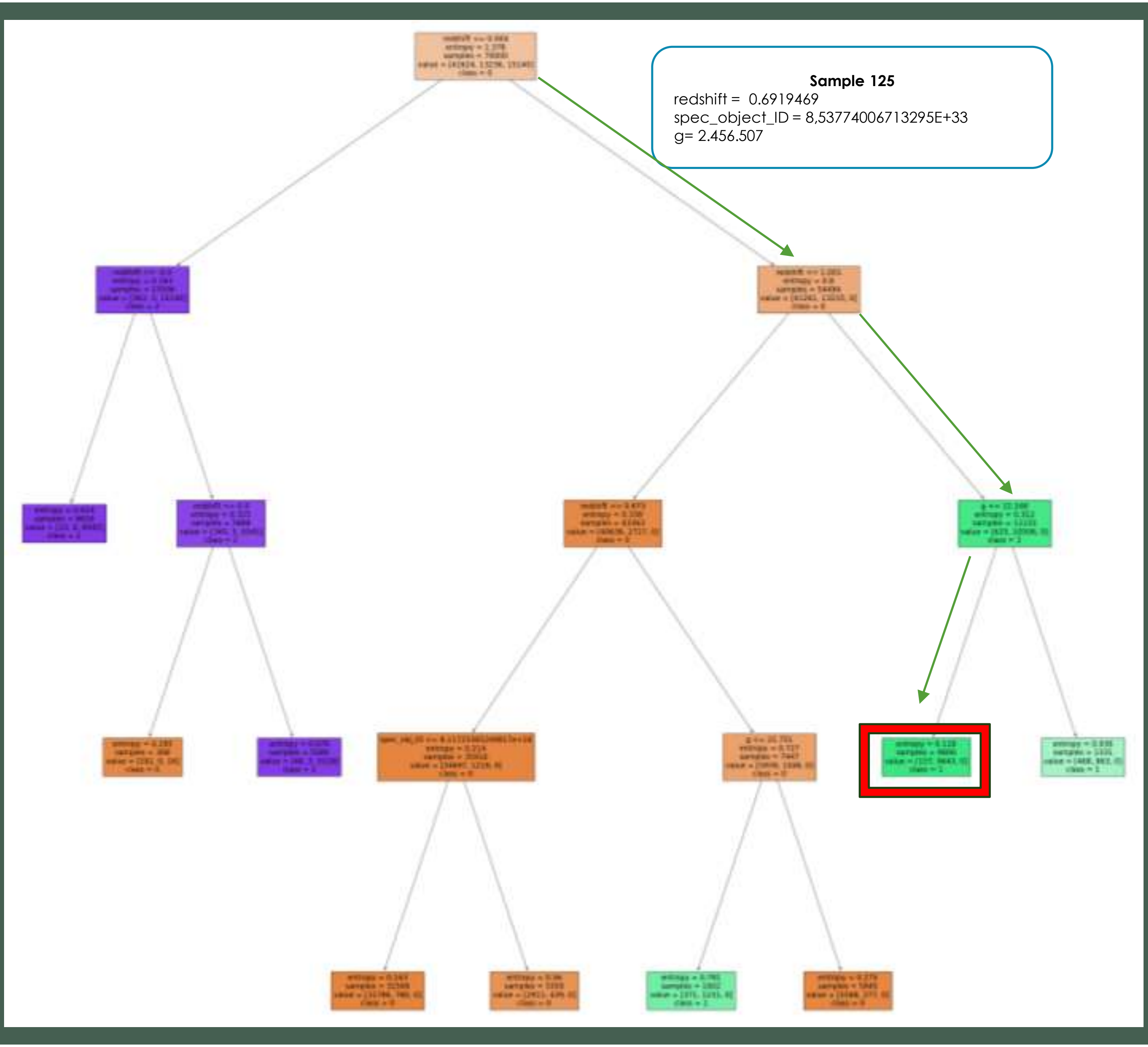
Sample 3
redshift = 0.9323456
spec_object_ID= 1,03010714129544E+34
g= 2.377.656

Incorrectly classified Example 1



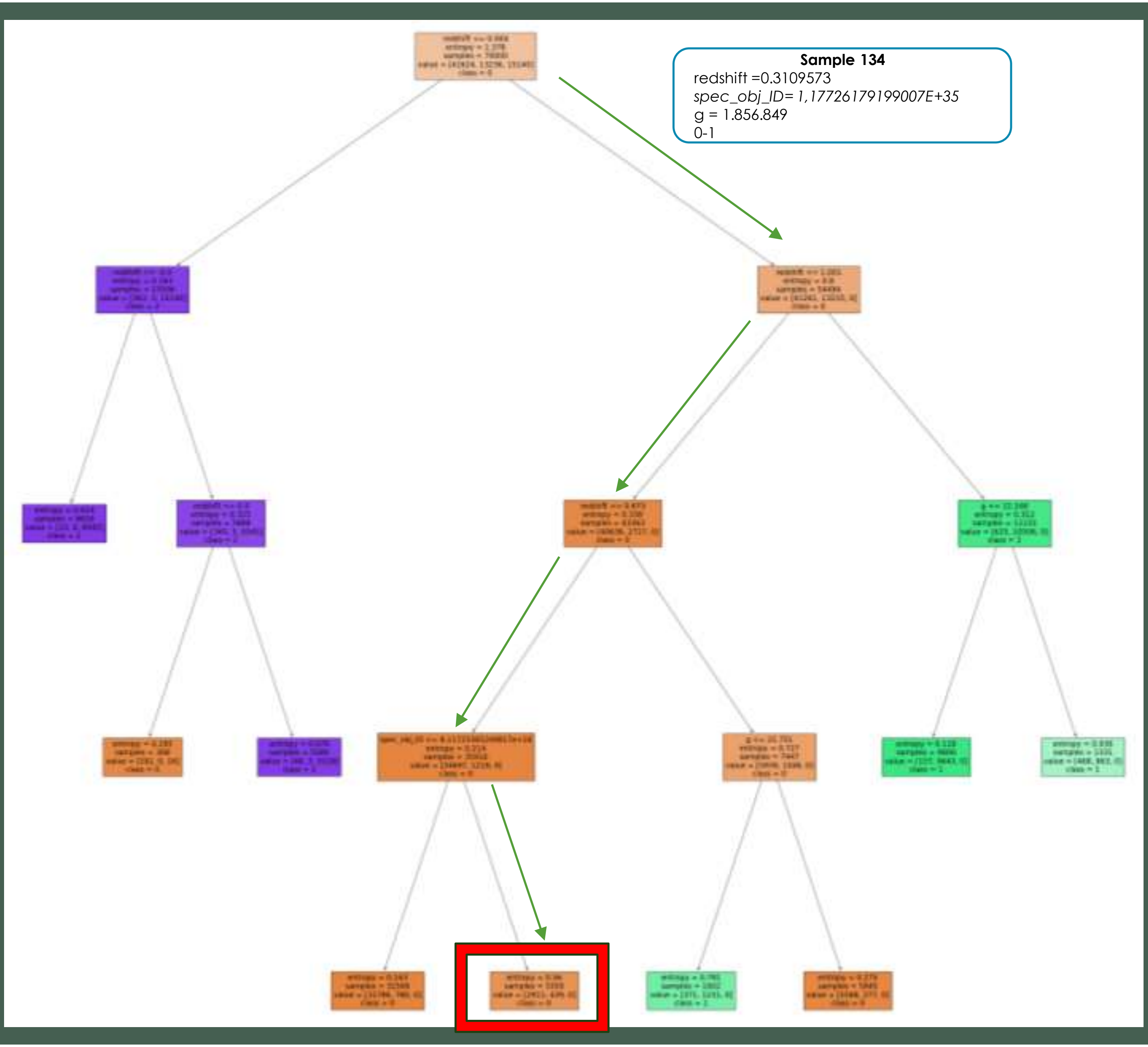
Sample 44
redshift = 0.5932347
spec_obj_ID=6,08216169886071E+33
g = 2.188.272

Incorrectly classified Example 2



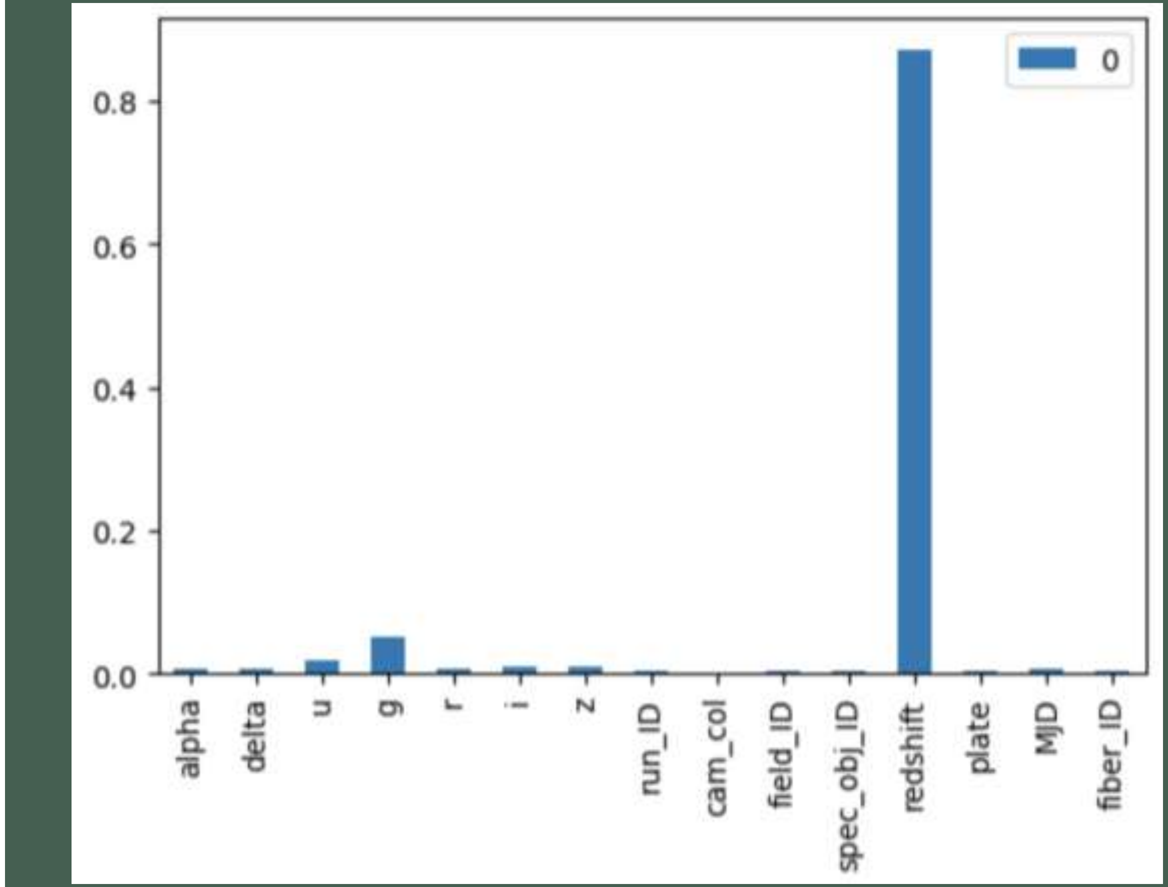
Sample 125
redshift = 0.6919469
spec_object_ID = 8,53774006713295E+33
g= 2.456.507

Incorrectly classified Example 3



Sample 134
redshift =0.3109573
spec_obj_ID= 1,17726179199007E+35
g = 1.856.849
0-1

Feature Importance

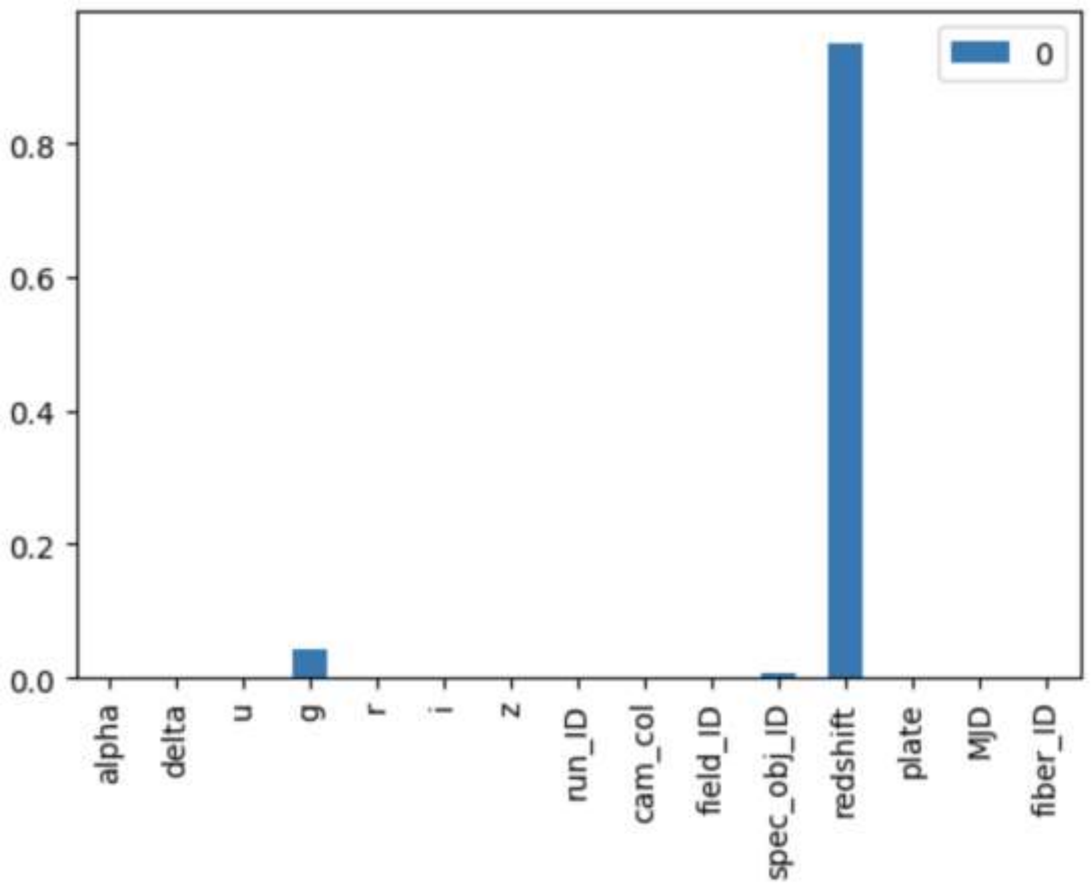


Non optimized

- ❖ Top most important features **before** optimizing:
  - ❖*u, g, r, i, z and redshift*

Feature Importance

- Top most important features **after** optimizing:
  - *g, spec_obj_ID, and redshift*



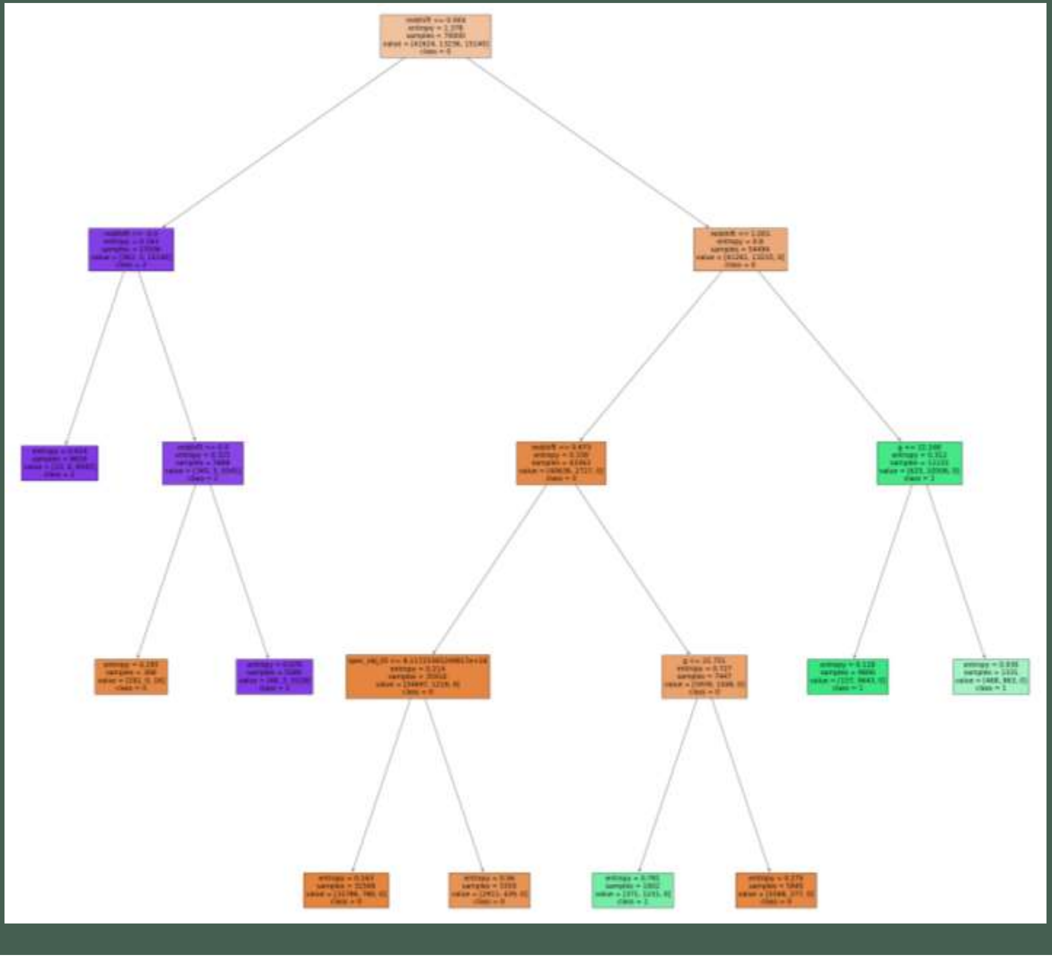
Optimized

Model's Accuracy: after removing 3 top features

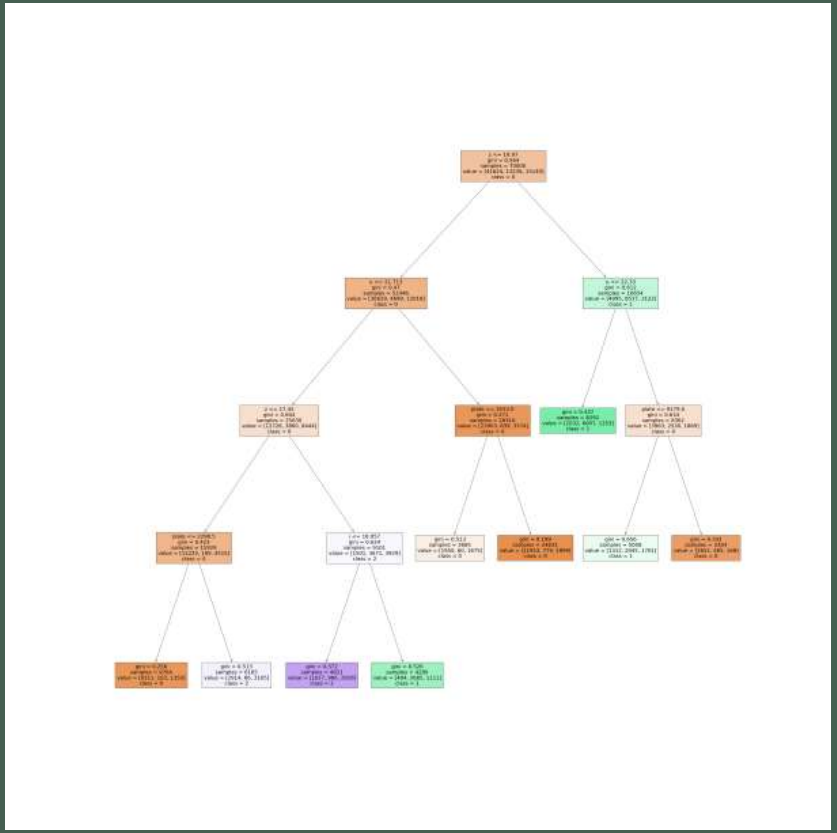
- Accuracy of non optimized model:
  - **0.795033**
  - Difference of 17,01%

- Accuracy of optimized model:
  - **0.738766**
  - Difference of 22,35%

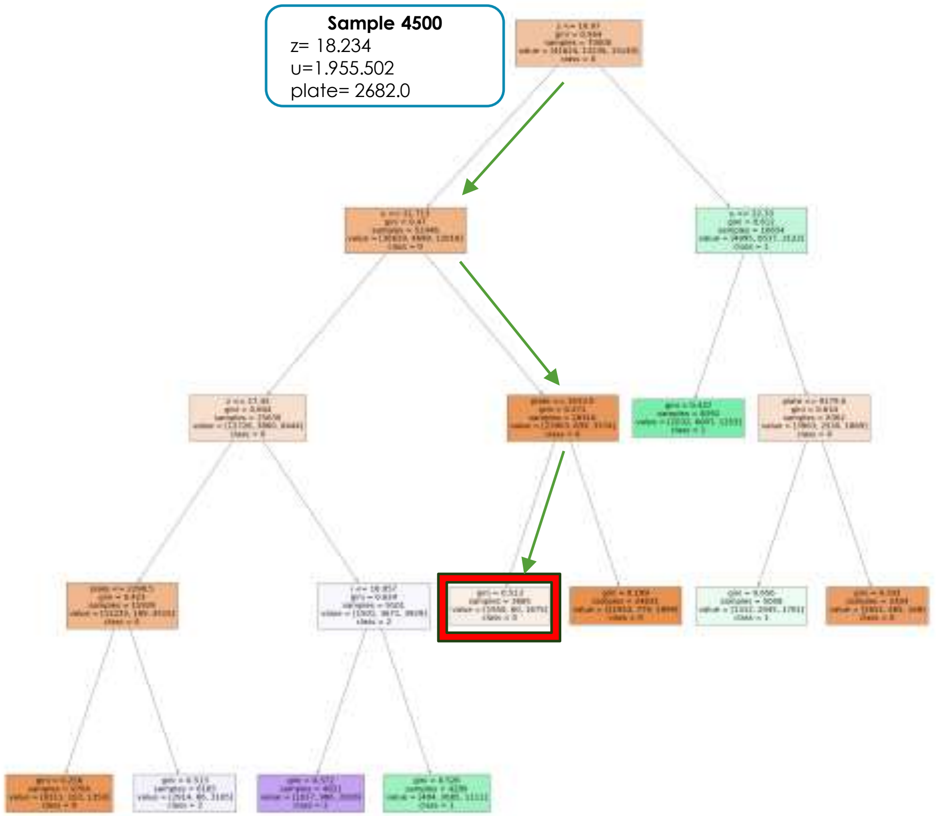
Optimized Decision Tree with 3 best features



Optimized Decision Tree without 3 best features



Incorrectly classified sample



Sample 4500
z= 18.234
u=1.955.502
plate= 2682.0