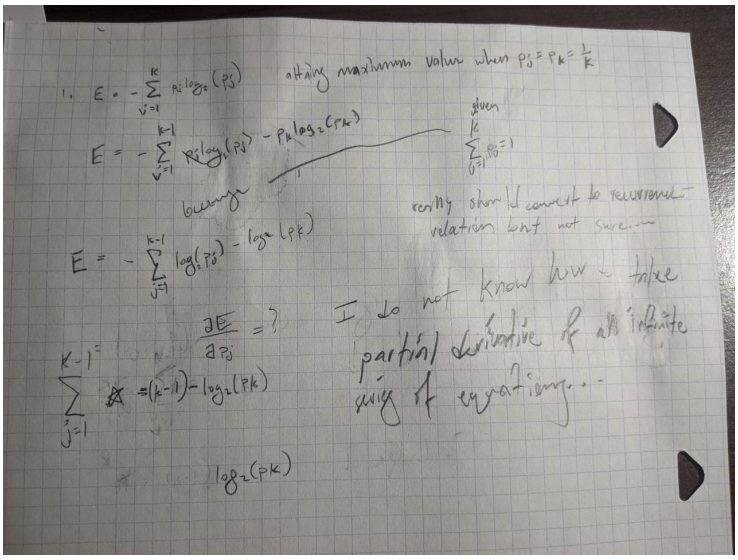
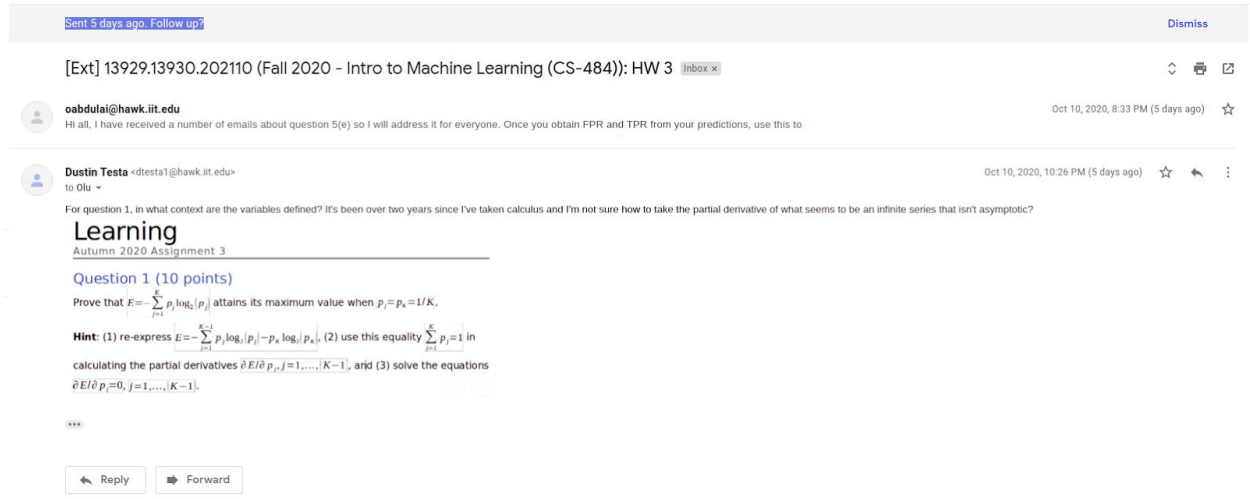
Dustin Van Tate Testa

1. Here's a partial solution using guesswork and what was given in the assignment because none of this makes sense. I'm gonna go kms now, thanks…





If you tell me that you are too busy during office hours but don't respond to emails how am I going to pass this class?

2. The probability that any item in the dataset is misclassified is defined as:

$$P(misclassification) = \sum_{i=0}^{N} P(put\ in\ wrong\ category_i) \times P(pulled\ from\ wrong\ category_i)$$

Gini Impurity is defined as the probability that a randomly selected item in the dataset is put in wrong category

$$I_G = \sum_{i=0}^{N} p_i^2$$

Dustin Van Tate Testa

Where p_i is the fraction of items having the same classification as the randomly selected one

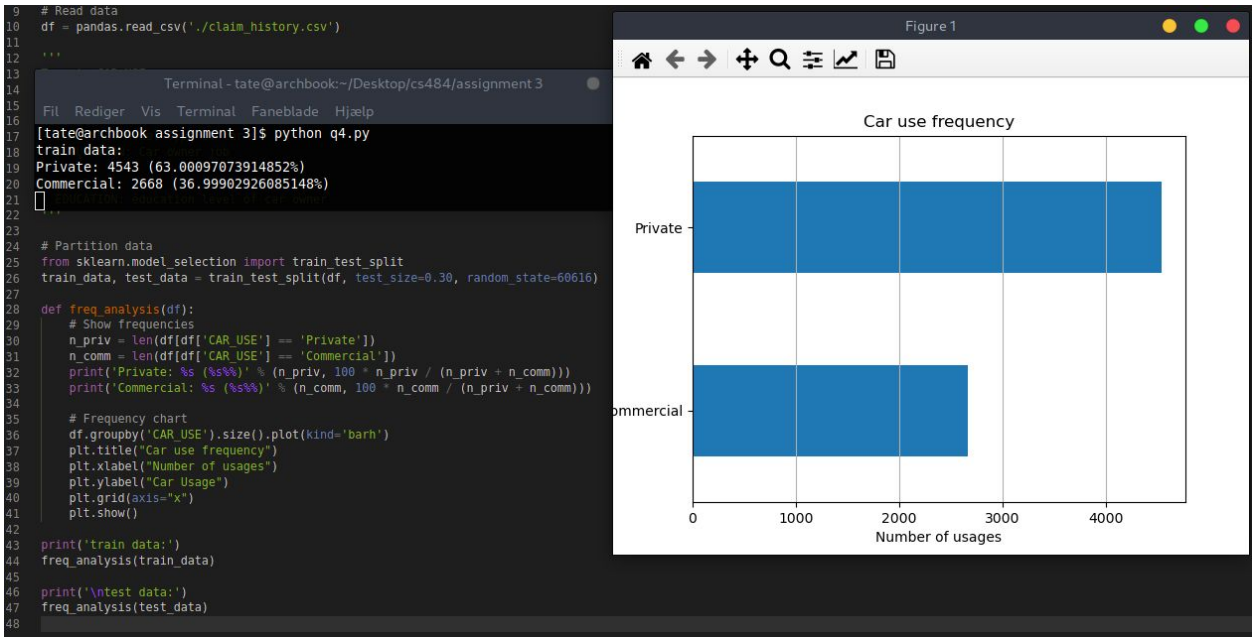
Thus because the set is entirely random P(put in wrong category) and P(pulled from wrong category) will be the same value for the entire dataset

This makes the two equivalent.

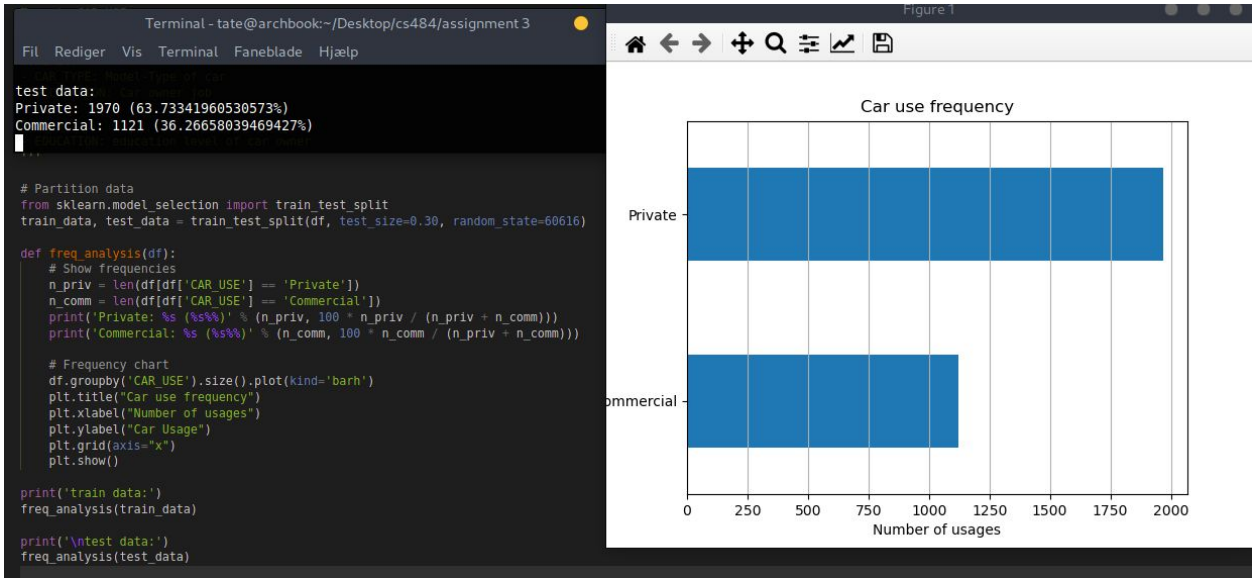
3. If the set contains only a single misclassified node, the Gini Impurity would have to be 1 as the only item to select would be misclassified, giving it a 100% chance of being as such.

4.

a. The training data consisted of 63% private vehicles and the test data consisted of 63.7% private vehicles making the proportions comparable

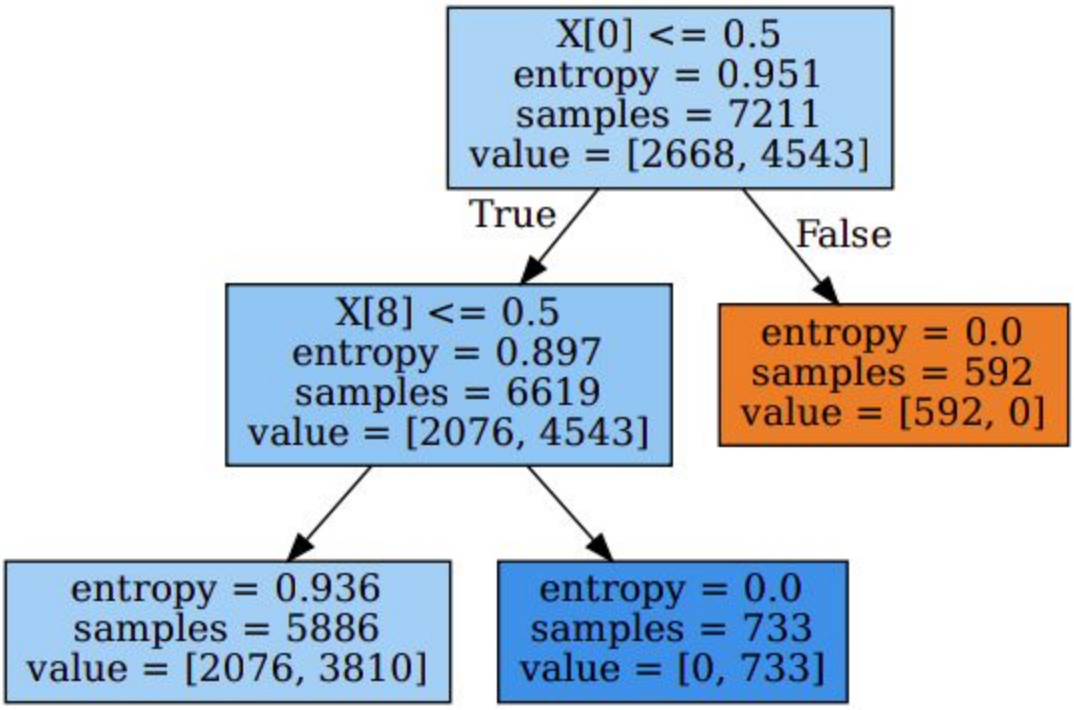


Dustin Van Tate Testa



b. Because both groups have approximately the same proportions, the probability that a private vehicle will will fall in the test group is just the ratio of the sizes. Therefore: 30/100 = 30%

5. This graph was generated using the the code in tree.py



a. The entropy of first node is 0.951
b. By enumerating over the column names for the one hot encoded training data we can see what it's checking for in each branch

i. In the root node it's checking if the vehicle is a `Panel Truck` and classifying the vehicle as commercial if so
ii. Otherwise it goes to the second branch where it checks if the driver's Occupation is 'Lawyer'
iii. These specific checks make sense as Panel Trucks are almost never used as private vehicles and Lawyers seldom drive commercial vehicles

```
print(graph)
graph.render('graph')

# print column names
for i, col in enumerate(one_hot_inputs):
    print('X[%s] = %s' % (i, col))
    if i > 10: break
```

```
[tate@archbook assignment 3]$ python tree.py
<graphviz.files.Source object at 0x7f4248dd2e50>
X[0] = CAR_TYPE_Panel Truck
X[1] = CAR_TYPE_Pickup
X[2] = CAR_TYPE_SUV
X[3] = CAR_TYPE_Sports Car
X[4] = CAR_TYPE_Van
X[5] = OCCUPATION_Clerical
X[6] = OCCUPATION_Doctor
X[7] = OCCUPATION_Home Maker
X[8] = OCCUPATION_Lawyer
X[9] = OCCUPATION_Manager
X[10] = OCCUPATION_Professional
X[11] = OCCUPATION_Student
[tate@archbook assignment 3]$
```

c. 0.897 according to the graphviz plot at top
d. Description of nodes:
   i. From the left

Rule	Count	Entropy		
Occupation != Lawyer Car != panel truck	5886	0.936		
Occupation == Lawyer Car != panel truck	733	0.0		
Car == panel truck	592	0.0		

e. https://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test
https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.kstest.html
f. Our tree only correctly classified

6.

a. The misclassification rate for the test data was found to be about 27.82%

```
# Find misclassification rate for test data
one_hot_inputs = pandas.get_dummies(
    test_data[['CAR_TYPE', 'OCCUPATION', 'EDUCATION']],
    drop_first = True)
targets = test_data[['CAR_USE']]
print('misclassification rate: %s%%' % (100 - 100 * dt.sco
```

```
X[8] = OCCUPATION_Lawyer
X[9] = OCCUPATION_Manager
X[10] = OCCUPATION_Professional
X[11] = OCCUPATION_Student


misclassification rate: 27.822711096732448%
[tate@archbook assignment 3]$
```

b.