Raelyn Mendoza
Melinda Vigh
CS 422 Data Mining- Assignment 1

Question 1:
For the assignment, we used Visual Studio Code rather than Google Colab. Thus, we did not alter the code in question 1 to save to a drive.

Question 2:

(1) Please write a few lines of code (as you see fit) to replace all the non-mammals such as birds, fishes, amphibians and reptiles with 'non-mammal' in the column named 'Class'.
Hint: you may search for an answer such as 'how to replace a certain value with a different value in a column in a pandas dataframe'. You are going to need to develop the ability to phrase a question like this.

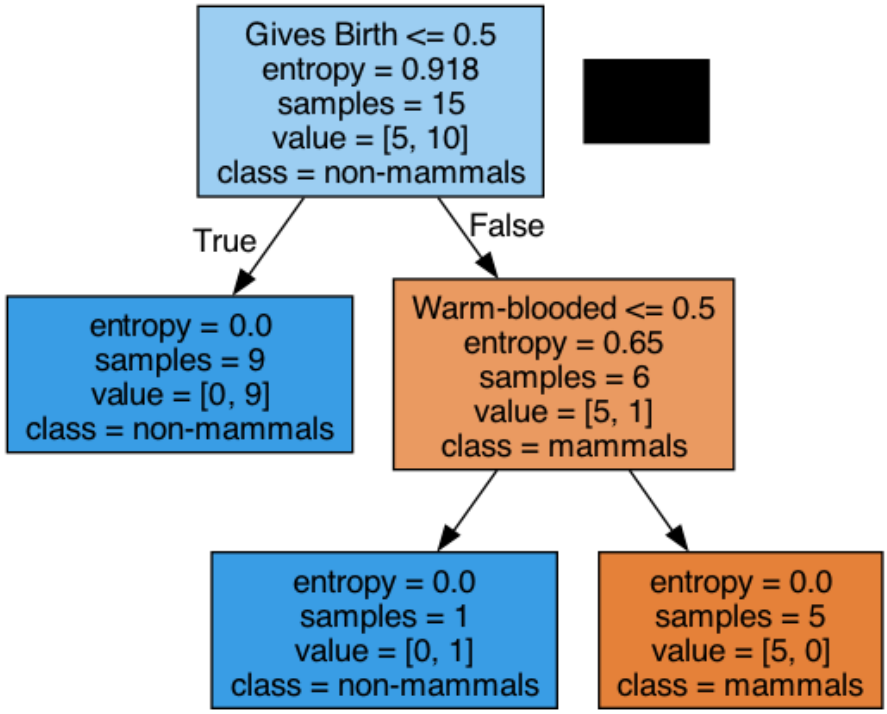
Our Code:

```
data = data.replace(['birds', 'fishes', 'amphibians', 'reptiles'], 'non-mammal')
```

(2) please provide an explanation for why it may be a good idea to congregate different animals into one single category such as 'non-mammals' here?

It is a good idea to congregate different animals into one single category such as 'non mammals' or 'mammals' because we are able to use less features for the splits. Having less features generates a smaller decision tree with a more accurate classification and avoids overfitting when training the model.

Question 3:
If you do everything right, you should have no problem running the following code and getting a tree chart. Please include the tree chart in your report. It is very important for your answer to the following question.

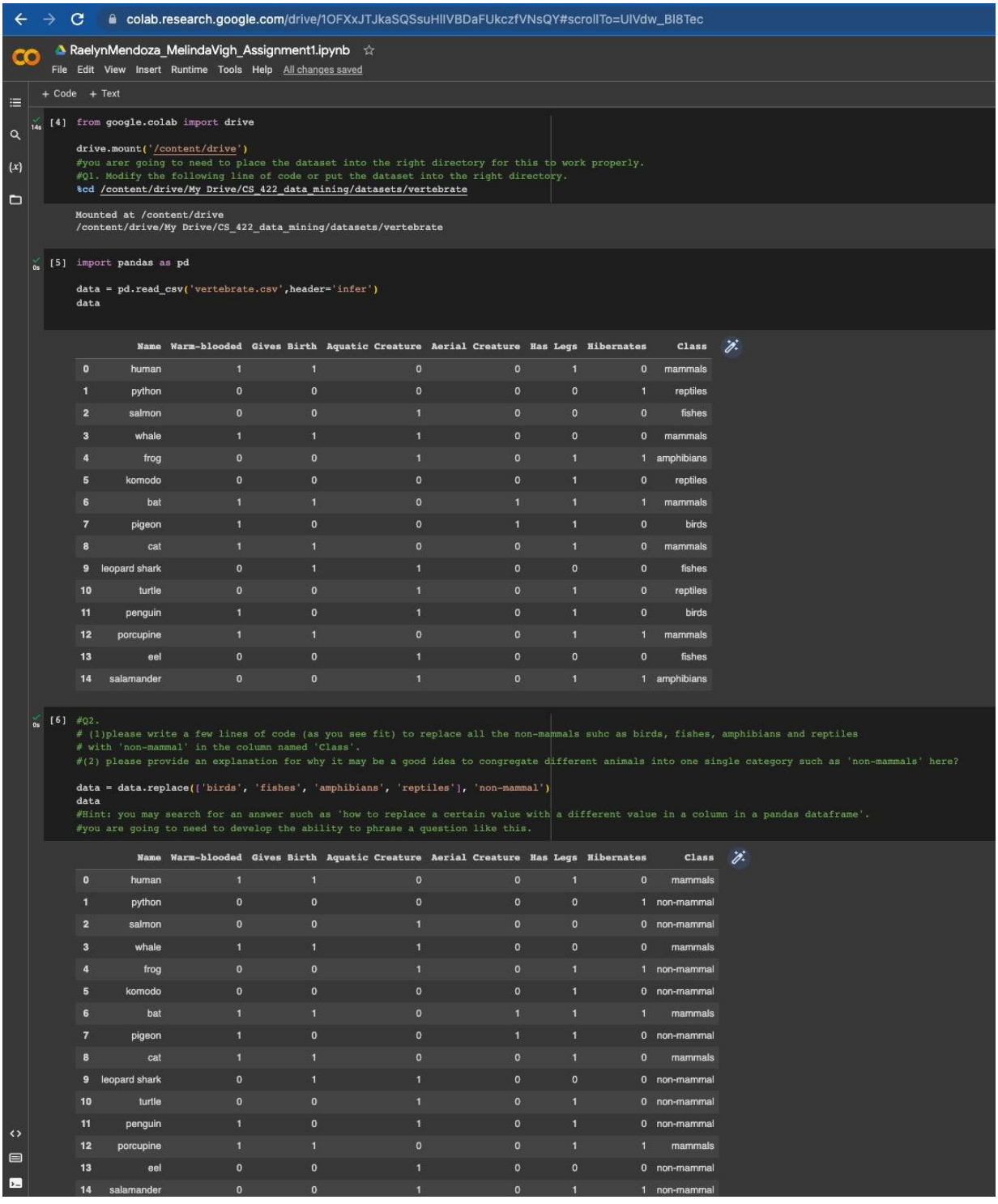


Question 4:

> Interpret the tree chart. We spent a good 10 to 15 minutes talking about how to 'read' a tree in the class. Here, it is a tree that has only a few levels. Explain how the decision tree model seems to do a good job in classifying animal into mammals and non-mammals. In other words, do you think 'warm-blooded', 'gives-birth' and so forth are useful features for us to classify animals. You need to explain the decision process until you reach all the 'leaf nodes'. A leaf node is a node that doesn't have child node. Well, you've taken CS 207.

Starting with the feature 'Gives-birth', it accurately separates the animals that have live births, which is a characteristic of mammals, from those that lays eggs. Here if ‘Gives-birth’ is less than or equal to 0.5, the animal does not give birth is true and is added to the left child node. If it is false, the animal is added the right child node. Once completed, this gives us our first leaf node with 9 non-mammals as the left child node and 6 possible mammals as the right child node. For the right child node, we use the feature 'Warm-blooded', if 'Warm-blooded' is less than or equal to 0.5, the animal does is not warm-blooded is true and is added to the left child node. If it is false, then the animal is added to the right child node. Once completed, this gives us our second leaf node with 1 non-mammals as the left child node and our third and last leaf node with 5 mammals as the right child node. Viewing the leaf nodes, we can compare to the list of mammals and non-mammals and see that the splits using the features Warm-Blooded and Gives-Birth accurately classified the animals, giving us a total of 5 mammals and 10 non-mammals.

To further verify our work, we also ran the code through Google Colab. The generated decision tree matched what we generated in Visual Studio Code. Below are the attached photos from Google Colab.



	Name	Warm-blooded	Gives Birth	Aquatic Creature	Aerial Creature	Has Legs	Hibernates	Class
0	human	1	1	0	0	1	0	mammals
1	python	0	0	0	0	0	1	reptiles
2	salmon	0	0	1	0	0	0	fishes
3	whale	1	1	1	0	0	0	mammals
4	frog	0	0	1	0	1	1	amphibians
5	komodo	0	0	0	0	1	0	reptiles
6	bat	1	1	0	1	1	1	mammals
7	pigeon	1	0	0	1	1	0	birds
8	cat	1	1	0	0	1	0	mammals
9	leopard shark	0	1	1	0	0	0	fishes
10	turtle	0	0	1	0	1	0	reptiles
11	penguin	1	0	1	0	1	0	birds
12	porcupine	1	1	0	0	1	1	mammals
13	eel	0	0	1	0	0	0	fishes
14	salamander	0	0	1	0	1	1	amphibians

	Name	Warm-blooded	Gives Birth	Aquatic Creature	Aerial Creature	Has Legs	Hibernates	Class
0	human	1	1	0	0	1	0	mammals
1	python	0	0	0	0	0	1	non-mammal
2	salmon	0	0	1	0	0	0	non-mammal
3	whale	1	1	1	0	0	0	mammals
4	frog	0	0	1	0	1	1	non-mammal
5	komodo	0	0	0	0	1	0	non-mammal
6	bat	1	1	0	1	1	1	mammals
7	pigeon	1	0	0	1	1	0	non-mammal
8	cat	1	1	0	0	1	0	mammals
9	leopard shark	0	1	1	0	0	0	non-mammal
10	turtle	0	0	1	0	1	0	non-mammal
11	penguin	1	0	1	0	1	0	non-mammal
12	porcupine	1	1	0	0	1	1	mammals
13	eel	0	0	1	0	0	0	non-mammal
14	salamander	0	0	1	0	1	1	non-mammal

```
[7] from sklearn import tree

    Y = data['Class']
    X = data.drop(['Name','Class'],axis=1)


    clf = tree.DecisionTreeClassifier(criterion='entropy',max_depth=3)
    clf = clf.fit(X, Y)
```

```
import pydotplus
from IPython.display import Image

dot_data = tree.export_graphviz(clf, feature_names=X.columns, class_names=['mammals','non-mammals'], filled=True,
                                out_file=None)
graph = pydotplus.graph_from_dot_data(dot_data)
Image(graph.create_png())
```