In this lab, we will practice for Entropy and Decision Tree

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, precision_score, recall_score
import math
from sklearn import tree
import matplotlib.pyplot as plt
```

About this data

This dataset predicts the likelihood of becoming an astronaut or not based on the predictors age, if the person likes dogs (like_dogs) and if the person likes gravity (likes_gravity).

1. Read the data first and look at the first 5 rows. Check if there are any missing values or not (3 points)

```
#Enter your code here
```

Creating the model

2. Split the data using sklearn's train_test_split(X, y, test_size) function. This function takes in your features (X), the target variable (y), and the test_size you'd like. We will train our model on the training set and then use the test set to evaluate the model for different criterions. Use 20% of the data as test size. Use random state=5. Use all the predictors as features to predict the target variable (going to be an astronaut). (5 points)

```
#Enter your code here
```

Gini

3. Now use the gini criterion to fit the data to the training set. Continue with random state=5, with max_depth=4. Then predict on the testing set. (5 points)

```
#Enter your code here
```

4. Evaluate the accuracy, precision and recall for the model. Explain the findings. (10 points)

```
#Enter your code here
```

Explain your findings here

Entropy

5. Now use the Entropy criterion to fit the data to the training set. Continue with random state=5, with max_depth=5. Then predict on the testing set. (2 points)

```
#Enter your code here
```

6. Evaluate the accuracy, precision and recall for the model. Explain the findings. (10 points)

```
#Enter your code here
```

Explain your findings here

7. Which criteria gives the highest accuracy,precision and recall? Explain why that might be the case. (5 points)

Plot

8. **Plot the tree for both the gini and entropy function Use class_names 'No' and 'Yes'. (5 points)**

```
#Enter your code here
```

Real Life Example

9. Using the model which gives the highest accuracy: Given an individual of age 33, and that they like dogs but do not love gravity, will they become an astronaut or not? (10 points)

```
# Hint: Create a new student's data with age:33,likes_dog:1 and likes_gravity:0. Then predict using the model with highest accuracy
#Enter your code here
```

10. Find the best max_depth (from 1 to 21) in Decision Tree for entopy and Gini. If both depths are different, explain why. (10 points)

```
#Enter your code here
```

```
#Hint: Use a loop to store the depths and return the max depth
#Enter your code here
```

Explain why

Entropy and Information Gain

11. Calculate the root's entropy. Hint: look at the example in lecture slides. (10 points)

```
#Enter your code
```

Calculate the Information Gain.

12. Split based on likes_dog and going_to_be_an_astronaut. Hint: You can use a pivot table. Look at the slides in lecture (5 points).

```
#Enter your code here
#You should get a 2x2 table
```

13. Calculate the entropy for liking dogs vs not liking dogs. Then find the Information Gain. Explain your findings (20 points).

```
#Enter your code here
#Hint: Again look at how it is done in the lecture slides
```

```
#Enter your code here for the information gain
```

Explain your findings here