Part 1

1.

```
print("Gini of data is ", gini(data))
```

```
Gini of data is  0.4628099173553719
```

```
print("Entropy of data is ", entropy(data))
```

```
Entropy of data is  0.9456603046006401
```

2-1.

```
clf_depth3 = DecisionTree('gini', 3,
                          x_train, y_train
                          ,x_test, y_test)
clf_depth10 = DecisionTree('gini', 10,
                          x_train, y_train
                          ,x_test, y_test)
print(clf_depth3.accuracy, clf_depth10.accuracy)
```

```
0.78 0.69
```

Depth 10 suffers from overfitting.

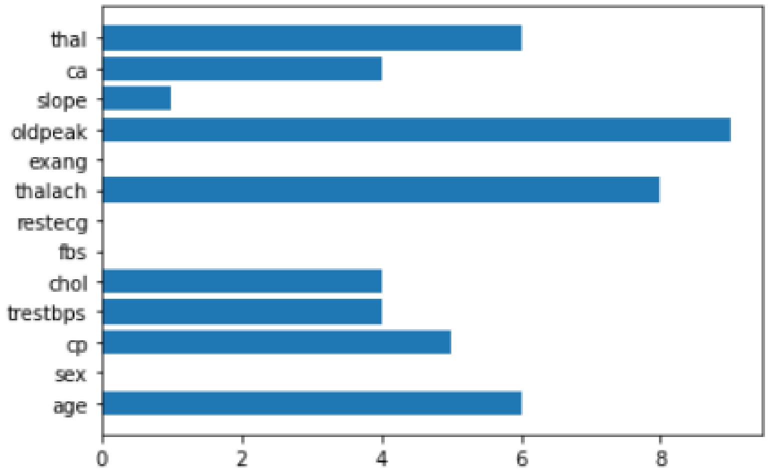
2-2.

```
clf_gini = DecisionTree('gini', 3,
                        x_train, y_train,
                        x_test, y_test)
clf_entropy = DecisionTree('entropy', 3,
                        x_train, y_train,
                        x_test, y_test)
print(clf_gini.accuracy, clf_entropy.accuracy)
```

```
0.78 0.76
```

3.

```python
plt.barh(range(len(feature_used)),
                    feature_used,
                    tick_label=list(dictionary.keys()))
plt.show()
```



4.

```python
es_ten  =  AdaBoost(10)
es_hun  =  AdaBoost(100)
```

Question 4.1

Show the accuracy score of test data by `n_estimators=10` and `n_estimators=100`, respectively.

```python
print(es_ten.accuracy,  es_hun.accuracy)
```

```
0.78 0.78
```

5-1.

```python
clf_10tree  =  RandomForest(10,  np.sqrt(x_train.shape[1]),  True,  'gini',
                    None  ,x_train,  y_train,  x_test,  y_test)
clf_100tree  =  RandomForest(100,  np.sqrt(x_train.shape[1]),  True,  'gini',
                    None  ,x_train,  y_train,  x_test,  y_test)
```

```python
print(clf_10tree.accuracy,  clf_100tree.accuracy)
```

```
0.8 0.74
```

5-2.

```
[26]
clf_random_features = RandomForest(10, np.sqrt(x_train.shape[1]), True, 'gini',
                      None ,x_train, y_train, x_test, y_test)
clf_all_features = RandomForest(10, x_train.shape[1], True, 'gini',
                      None ,x_train, y_train, x_test, y_test)
```

- Note: Use majority votes to get the final prediction, you may get slightly different results when re-building the random forest model

```
print(clf_random_features.accuracy, clf_all_features.accuracy)
```

```
0.82 0.77
```

6.

```
best_result = RandomForest(20, np.sqrt(x_train.shape[1]), True, 'entropy',
              3 ,x_train, y_train, x_test, y_test)
print(best_result.accuracy)
```

```
0.86
```

Part 2

1. A: $(300, 100) \rightarrow$ class 1

$(100, 300) \rightarrow$ class 2

misclassification: $\frac{100}{400} \times \frac{400}{800} + \frac{100}{400} \times \frac{400}{800} = \frac{1}{4}$

cross-entropy: $-\frac{1}{2}\left(\frac{1}{4}\log_2\frac{1}{4} + \frac{3}{4}\log_2\frac{3}{4}\right) - \frac{1}{2}\left(\frac{3}{4}\log_2\frac{3}{4} + \frac{1}{4}\log_2\frac{1}{4}\right) \fallingdotseq 0.811$

Gini-index: $\frac{1}{2}\left(1 - \frac{1}{16} - \frac{9}{16}\right) + \frac{1}{2}\left(1 - \frac{1}{16} - \frac{9}{16}\right) = \frac{3}{8}$

B: $(200, 400) \rightarrow$ class 2

$(200, 0) \rightarrow$ class 1

misclassification: $\frac{200}{600} \times \frac{600}{800} + \frac{0}{200} \times \frac{200}{800} = \frac{1}{4}$

cross-entropy: $-\frac{3}{4}\left(\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3}\right) - \frac{1}{4}\left(1 \cdot \log_2 1\right) \fallingdotseq 0.689$

Gini-index: $\frac{3}{4}\left(1 - \frac{1}{9} - \frac{4}{9}\right) + \frac{1}{4}(1-1) = \frac{1}{3}$

$0.689 < 0.811$ and $\frac{1}{3} < \frac{3}{8}$ => both criteria are lower

#

2. $E_{x,t}[e^{-ty(x)}] = \sum_t \int e^{-ty(x)} p(t|x)\, p(x)\, dx$

$$= \int \left[e^{-y(x)} p(t=1|x) + e^{y(x)} p(t=-1|x) \right] p(x)\, dx$$

Let $e^{y(x)} = \mu$, $p(t=1|x) = \varepsilon \Rightarrow p(t=-1|x) = 1-\varepsilon$

To minimize the function, $\frac{\varepsilon}{\mu} + \mu(1-\varepsilon)$ should be minimal

$$\Rightarrow \frac{d\left[\frac{\varepsilon}{\mu} + \mu(1-\varepsilon)\right]}{d\mu} = -\varepsilon \cdot \mu^{-2} + (1-\varepsilon) = 0$$

$$\Rightarrow \mu = \left(\frac{\varepsilon}{1-\varepsilon}\right)^{\frac{1}{2}}$$

$$\Rightarrow y(x) = \ln\left(\frac{\varepsilon}{1-\varepsilon}\right)^{\frac{1}{2}}$$

$$= \frac{1}{2} \ln\left(\frac{\varepsilon}{1-\varepsilon}\right)$$

$$= \frac{1}{2} \ln \frac{p(t=1|x)}{p(t=-1|x)}$$

#