CENG484 Data Mining Assignment2

Task 1

a) What will be the output for this input? **Input:** 4228.png

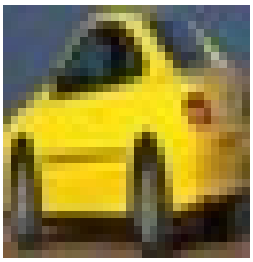
Most similar three images with similarity values: 4064.png has 0.859885 cos similarity
4162.png has 0.84638953 cos similarity
4766.png has 0.83251476 cos similarity

Input

```
Enter the filename of the image: 4228.png

Most similar three images with similarity values:

4766.png 4162.png 4064.png
0.83251476 0.84638953 0.859885

Process finished with exit code 0
```

Output of Task_1.py

0.859885



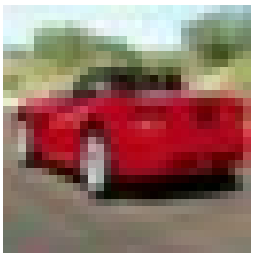
0.84638953

0.83251476

b) What will be the output for this input? **Input:** 3861.png

Most similar three images with similarity values: 3952.png has 0.8567403 cos similarity
4946.png has 0.85013604 cos similarity
3819.png has 0.8490281 cos similarity

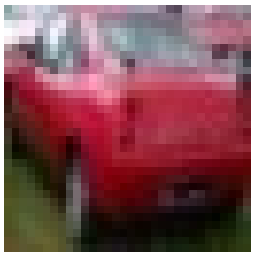
Input



```
Enter the filename of the image: 3861.png

Most similar three images with similarity values:

3819.png 3952.png 4946.png
0.8490281 0.8567403 0.85013604

Process finished with exit code 0
```

Output of Task_1.py

0.8567403



0.85013604

0.8490281

Task 2

a) Compute the **E** and **GI** for the overall collection of training examples.

Entropy = 1.0
Gini Index = 0.5

b) Compute the **E** and **GI** for the **age** attribute.

Entropy = 0.5916727785823275
Gini Index = 0.24489795918367352

c) Compute the **E** and **GI** for the **cp** attribute.

Entropy = 0.37123232664087563
Gini Index = 0.1326530612244898

d) Compute the **E** and **GI** for the **trestbps** attribute.

Entropy = 0.9007930640987631
Gini Index = 0.4331550802139037

e) Which attribute is better according to calculations?

We should assess the goodness of an attributes according to their entropy and gini index values. Entropy is a measure of randomness so the higher the entropy, the harder to summarize the dataset. Gini index measure the impurity of splitted data so the smaller degree impurity, the more skewed the class distribution. Based on those information, we need to choose the lowest entropy and gini index values. **Cp** attribute holds this condition so **cp** is better according to calculations.

f) Which attribute can be chosen as the root? Explain why.

The default root node is calculated in the part a and the entropy value is 1.0 and the gini index value is 0.5. Any attribute that makes this default root better can be chosen as root. Therefore, according to the calculations made in part b, part c and part d, the attributes: age, cp and trestbps lowers the entropy and gini index and anyone of those attributes**(age, cp, trestbps) can be good candidates for the root**. However, if choosing a better attribute to be root is desired, then cp attribute can be chosen because of the reasons explained in part e.

```
Computation of E and GI for overall collection of training examples

Entropy =  1.0
Gini Index =  0.5

Computation of E and GI for age

Entropy =  0.5916727785823275
Gini Index =  0.24489795918367352

Computation of E and GI for cp

Entropy =  0.37123232664087563
Gini Index =  0.1326530612244898

Computation of E and GI for trestbps

Entropy =  0.9007930640987631
Gini Index =  0.4331550802139037

Process finished with exit code 0
```

Output of Task_2.py