

**CENG 484 – DATA MINING**

**Assignment 2 Report**

**Kemal Türk**

**250201073**

## Task 1)

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

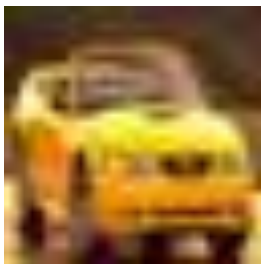
**Answer – A)**

Most similar ones to 4228.png:

4064.png value: 0.859885



4162.png value: 0.84638953



4766.png value: 0.83251476



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

**Answer – B)**

Most similar ones to 3861.png:

3952.png value: 0.8567403



4946.png value: 0.85013604



3819.png value: 0.8490281



## **Task 2)**

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

**Answer – A)**

Entropy of overall: 1.0

Gini of overall: 0.5

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

**Answer – B)**

Entropy of age: 0.5916727785823275

Gini of age 0.24489795918367352

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

**Answer – C)**

Entropy of cp: 0.37123232664087563

Gini of cp: 0.13265306122448983

**D) Compute the E and GI for the trestbps attribute**

**Answer – D)**

Entropy of trestbps: 0.9007930640987631

Gini of trestbps: 0.43315508021390364

**E) Which attribute is better according to calculations?**

**Answer – E)**

Because of having lower results on both Entropy and Gini calculations “cp” attribute is a better attribute than the others.

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

**Answer – F)**

“cp” attribute should be chosen as root because it has the lowest Entropy result which would lead us to most information gain, and according to ID3 algorithm attribute with most information gain should be the root node.