Computer Science Department
California State University Channel Islands

COMP 478 - Homework 1

Deadline: 02/21/2022, 11:59 am

1. (50 points) Write a python code to implement **KNN**.
   Note: in this question, you are NOT allowed to use Scikit-Learn package!

   (a) *Step 1*: Write a function to calculate pair-wise distance between two data points. (args: datapoint1, datapoint2, dist-fcn) "dist-fcn" should be determined by the user (can be either "Manhattan" or "Euclidean")

   (b) *Step 2*: Load iris dataset (iris-dataset-1) using the Pandas package. The test dataset includes indices: 5, 15, 25, 35, 45, 55, 65, 75, 85, 95, 105, 115, 125, 135, 145 and the rest of the data points will be your training dataset.

   (c) Step 3: Set K = 11 and use your function in Step 1 to find the K closest data points to the test samples and predict the label for each.

   (d) Step 4: Calculate the accuracy.

2. (50 points) One of the challenges in decision trees is selecting the appropriate splitting criterion. In this problem, we will learn a new splitting criterion called Gini-Index.

   (a) *Step 1*: Research about Gini-Index and briefly explain how it differs from Information Gain.

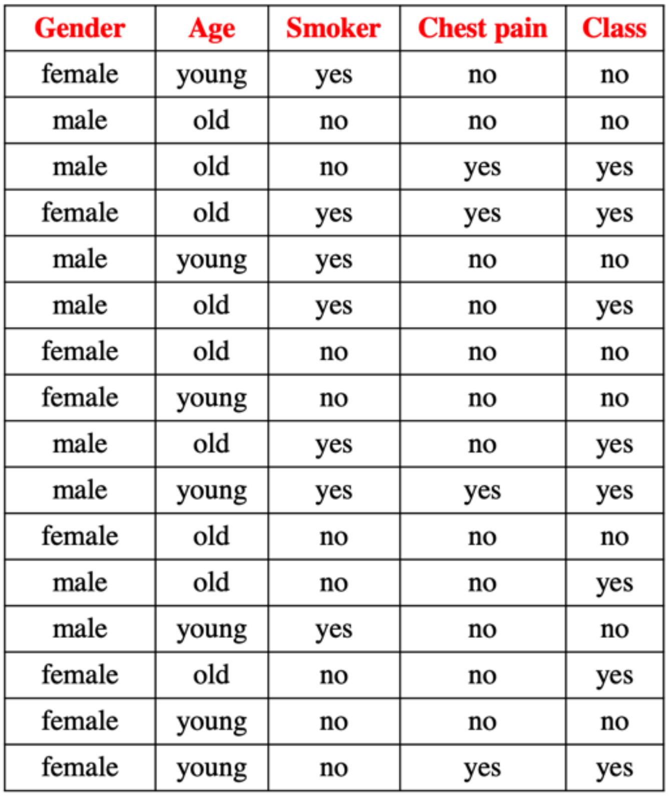
   (b) *Step 2*: A researcher has discovered a new cure for a disease, however, it can cause some side effects. To predict it's side effects, he has tested that on 16 volunteer patients. The table below presents his findings. Now, he wants to use decision trees to make a prediction for future patients using four features (gender, age, smoker, chest pain). (Note: this is just a toy problem, not a real world data!)

   - What is the decision tree for this problem, if he uses Information Gain as splitting criterion? Write the complete solution. (Note: You are only allowed to use a maximum of 3 decision nodes in your tree.)

   - What is the training accuracy for your decision tree?

   (c) *Step 3*: What if he uses Gini-Index as the splitting criterion? Write the complete solution. (Note: You are only allowed to use a maximum of 3 decision nodes in your tree.)

   - What is the training accuracy for your decision tree?



Gender	Age	Smoker	Chest pain	Class
female	young	yes	no	no
male	old	no	no	no
male	old	no	yes	yes
female	old	yes	yes	yes
male	young	yes	no	no
male	old	yes	no	yes
female	old	no	no	no
female	young	no	no	no
male	old	yes	no	yes
male	young	yes	yes	yes
female	old	no	no	no
male	old	no	no	yes
male	young	yes	no	no
female	old	no	no	yes
female	young	no	no	no
female	young	no	yes	yes