Department: Mathematics and Computer Science
Course Name: Data Mining
Course Code: Stat 412,404
Professor Name: Dr. Laila M. Fatehy
Level: Four

جامعة الإسكندرية
ALEXANDRIA UNIVERSITY
كلية العلوم

Assignment (1)
Decision Trees & KNN Implementation

1) Giving the following Data ("Weather_New")

outlook	temperature	humidity	windy	play
sunny	hot	high	f	no
sunny	hot	high	t	no
overcast	hot	high	f	yes
rainy	mild	high	f	yes
rainy	cool	normal	f	yes
rainy	cool	normal	t	no
overcast	cool	normal	t	no
sunny	mild	high	f	no
sunny	cool	normal	f	yes
rainy	mild	normal	f	yes
sunny	mild	normal	t	no
overcast	mild	high	t	yes
overcast	mild	normal	f	yes
rainy	mild	high	t	no
rainy	mild	high	f	yes
rainy	cool	normal	f	yes
rainy	cool	normal	t	no
overcast	mild	normal	t	yes
sunny	mild	normal	f	no
sunny	cool	normal	f	no
rainy	mild	normal	f	yes
sunny	mild	normal	t	yes
overcast	mild	high	t	yes
overcast	hot	high	f	yes
rainy	cool	high	t	no
sunny	normal	high	t	no
overcast	normal	high	f	yes
rainy	cool	high	f	yes
sunny	hot	high	f	yes
rainy	mild	high	f	yes
sunny	cool	normal	f	yes
rainy	cool	normal	t	no
sunny	cool	high	t	no
overcast	mild	high	f	no
overcast	cool	normal	f	yes

a) In this dataset, there are 5 categorical attributes outlook, temperature, humidity, windy, and play you must at first load our dataset "Weather_New.csv"

b) We are interested in building a system which enable us to decide whether or not to play the game on the basis of weather conditions .

c) Find the confusion matrix and calculate the Accuracy

d) Calculate the Precision, F1 score, Error rate and Sensitivity

e) Plot the ROC Graph (receiver operating characteristic curve)

2) Giving the following Data ("Iris_Dataset")

a) In this dataset, there are 5 attributes SepalLengthCm, SepalWidthCm, PetalLengthCm, PetalWidthCm, and Species of 50 samples of three species of (Iris setosa, Iris virginica and Iris versicolor) Four features were measured from each sample: the length and the width of the sepals and petals, in centimeters These measures were used to create a KNN model to classify the species.
you must at first load our dataset "iris.csv"

b) Find the confusion matrix and calculate the Accuracy

c) Calculate the Precision, F1 score, Error rate and Sensitivity

d) Plot the ROC Graph (receiver operating characteristic curve)

What do you need to install?

You will need an installation of Python and also Anaconda from the following links

- Python
- Anaconda

You will need an installation of R and also RStudio from the following links

- R
- RStudio

Due Date (Submission Deadline)

April 13th at 11:59 PM

Assignment Discussion

April 16th Before and after Data Mining Section on the same day

Assignment instructions

submit your Assignment to(mahmoud_samy@alexu.edu.eg). Get organized before you begin and you must submit it with your Full name and Academic ID before Deadline.

Avoid plagiarism !

Plagiarism is any act claiming or implying *another person's work is your own*, the below are some coding-specific examples of what constitutes plagiarism

Examples of plagiarism:

- **Copying someone's code exactly, in whole or part.**
- **Combining code copied from multiple sources.**
- **Copying someone's code, or part, and making changes (e.g. changing variable or function names, comments, order of function definition).**

If one of the previous points **Verified in your project**, your Grade will be **Zero**

Good Luck
M. Samy