CS 415/515 Assignment 1 [30 points]

1. [10 points] Using information Gain as the criteria for deciding the node to split, construct a decision tree manually with the depth as 2. Using data provided in the Table below, show all the computations/steps of your work for the full credit(i.e. how do you decide which feature to use for performing the first split and the other splits). Draw the final decision tree. (You do not need to show intermediate trees.)

Day	Outlook	Temperature	Humidity	Wind	Played Tennis?
01	Sunny	Hot	High	Weak	No
02	Sunny	Hot	High	Strong	No
03	Overcast	Hot	High	Weak	Yes
04	Rain	Mild	High	Weak	Yes
05	Rain	Cool	Normal	Weak	Yes
06	Rain	Cool	Normal	Strong	No
07	Overcast	Cool	High	Strong	Yes
08	Sunny	Cool	High	Weak	No
09	Sunny	Mild	Normal	Weak	Yes
10	Rain	Hot	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

2. [8 points] Calculate the cosine similarity between two documents. Cosine similarity is defined as the dot product of the vectors divided by their magnitude.
$cosine\_similarity(A, B) = \cos\theta = \frac{A \cdot B}{||A||||B||}$
where θ is the angle between the two vectors A and B and $A \cdot B$ is the dot product of the two vectors A and B, the dot product can be calculated as $A \cdot B = \sum_{i=1}^{n} A_i B_i = [A_1 B_1 + A_2 B_2 + ... + A_n B_n]$. $A_1, A_2, ..A_n$ and $B_1, B_2, ..B_n$ are the components of the vectors A and B. $||A||$ represents the magnitude of the vectors and can be calculated as $||A|| = \sqrt{A_1^2 + A_2^2 + ...A_n^2}$.

For this assignment, you would need to first calculate the document vectors manually from the document pairs and then calculate the cosine similarity. The document pairs are:
i). *"Today is a nice day. I want to go for a walk."* and *"The man is taking his dog for a walk as the weather is nice today"*. The set of words for this problem is [today, nice, day, want, go, walk, weather, man, dog, take] Consider 'taking' to be the same word as 'take' for constructing the document vector since they are from the same root word.

ii). *"The weather is dull today"* and *"How are you doing? Are you going to the college"* Consider the set of words to be [today, weather, dull, go, college, do]. Consider 'going' to be the same word as 'go' for constructing the document vector since they are from the same root word. Similarly, consider 'doing' to be the same word as 'do' for constructing the document vector.

3. [2 + 2 points]P and Q have only binary values. P = (1, 1, 1, 1, 0, 1) and Q = (1, 0, 0, 1, 1, 0). Calculate the Simple Matching coefficient and Jaccard Coefficient.

4. [8 points] Using python, build a synthetic two-class dataset where each data point has two features x1 and x2.
Instructions
For class 1, generate 500 data points with each feature x1 and x2:

- x1 feature of class1 should follow a normal distribution with a mean of 6 and a standard deviation of 2.5.
- x2 feature of class1 should follow a normal distribution with a mean of 12 and a standard deviation of 2.5.

For class 2, generate 500 data points for each feature x1 and x2:

- x1 feature of class2 should follow a normal distribution with a mean of 14 and a standard deviation of 2.5.
- x2 feature of class2 should follow a normal distribution with a mean of 8 and a standard deviation of 2.5.

Combine these points into a single dataset, with the corresponding class labels (Class 1 labeled as 0, Class 2 labeled as 1).
Plot the data points from the two classes in different colors.
Please attach the ipynb file containing the code and the plot of the dataset.