CSCI 184 HW 2 -- Part 2: Programming

Due Date: May 15, 11:59pm 2024

All assignments MUST have your name, student ID, course name/number at the beginning of your documents.

Your homework MUST be submitted via Camino with the file format and name convention as follows:

For Question Answering part, you can either write by hand or type your answers, but please ensure your submission is **pdf** file with name "**HW#_Name.pdf**".

For programming questions, please upload your code and supporting files in "**HW#_Name.zip**".

If you have any questions, please don't hesitate to contact me :)

Naive Bayes – Cancer Tumor Classification

For this part, you will focus on a cancer dataset that comprises of 569 rows and 32 columns and perform Naive Bayes Classification.

What do you need to do?

1. Load the dataset from 'cancer.csv' into a pandas DataFrame and print it along with its shape. 'diagnosis' is the target variable.
2. Print the column names and the data type of each column.
3. Plot the 'Radius Mean' VS 'Texture Mean' along with the classes represented as colors or shapes. Is the data linearly separable?
4. Perform encoding on the target variable (here label encoding will suffice).
5. Divide the data into X and Y, where X is the set of features and Y is the target variable.
6. Split the data into train and test data. Choose a split size of 70 - 30.
7. Given the nature of the data and its features, choose which Naive Bayes is the most suitable. Mention this in your report along with why you make your choice.
You may use the Naive Bayes from sklearn.
8. Once you have trained your model, evaluate the model performance by printing the performance matrix.
9. Write a report with screenshots of your results and the final results for step 8.
10.Submit your code as an .ipynb file and a pdf file reporting your findings.