Instructions for Assignment 2

For the Second assignment, we are going to be implementing three classifiers - **Decision Tree**, **Naïve Bayes**, and **Nearest Neighbor classifier.**

- Each classifier is using its own dataset.
- Each assignment is having its own tasks.
- Each assignment is presenting a challenge.

Decision Tree Classifier:

Dataset information

Load dataset_DT.csv
The data is ordered by date (day, month)

Features Description:

- age → age
- job → type of Job
- marital → marital status
- education → highest education finished
- default → already has credit in default?
- balance → account balance
- housing → taken housing loan?
- loan → taken personal loan?
- contact → communication via...
- day → day of last contact
- month → month of last contact
- duration → duration of last contact
- campaign → number of contacts made to the client during the campaign
- pdays → number of days that passed by after the client was last contacted from a previous campaign (999 means client wasn't previously contacted)
- previous → number of contacts performed before this campaign and for this client
- poutcome → outcome of the previous marketing campaign

Target variable:

- y → has the client subscribed a term deposit?

Tasks

1. Import the libraries and load the dataset (from the csv file) [5 points]
2. Pre-process the dataset and provide a detailed explanation. [25 points]
3. Perform 2 visualizations of the features with respect to target variable with detailed explanation. [15 points]

4. Determine Six most influential attributes on target attribute (with explanation). You do not necessarily need to drop the remaining features for the Decision Tree Classifier. Your task is just to determine and show the Six most influential attributes with detailed explanation. [10 points]
5. Split your dataset 75% for training, and 25% for testing the classifier. [2 points]
6. Use gini and entropy (play around with max_depth and min_samples_leaf parameters) to measure the quality of a split. [5 points]
7. Use comments to explain your code and variable names. [3 points]
8. Calculate and print the confusion matrix (use graphics instead showing a 2D array), and the classification Report (includes: precision, recall, f1-score, and support) for both criteria. [20 points]
9. Compare the results of the classifier using gini and entropy [5 points]
10. Print the decision tree visualization with depth of 5 [5 points]
11. Include a paragraph describing the contribution made by each team member

Hints:

1. Apart from null values, the dataset consists of "unknown" (string) values in multiple columns. You need to handle them as a part of null values.
2. To compress the data, numerous columns might be combined into one feature. There might even be columns with redundant data, i.e., information from a column might also be available from another column. If there are such columns, you can drop them.
3. Categorized Data is preferable for decision trees. If needed, figure out how to convert continuous feature to categorical and implement it.
4. Unwanted data can reduce the model's accuracy.

Naïve Bayes Classifier:

Dataset information

Features Description:

- email → text data of the actual email

Target variable:

- Label → spam(1) or not spam(0)

Tasks

1. Load the dataset as pandas dataframe
2. You have textual data that you cannot feed into the model. Therefore, you need to extract features from the text (email) and transform the data.
3. Test train split, using 80% for training, rest for testing.
4. Train NB model (Gaussian) for classification, on the training data.
5. Predict on the test data.
6. Get the accuracy, plot the confusion matrix, report Accuracy Score(metrics.accuracy_score), and plot Confusion Matrix(metrics.confusion_matrix) plotted graphically (It needs to be in the final Jupyter file which you submit)
7. Create a report (Jupyter markdown) file to include concise answers to the following questions -

   a. Briefly explain your approach, any preprocessing, explain the output, any visualization for explanation, any feature extraction, in same Jupyter file (put it in markdown). Do not submit separate pdf report file (3-4 paragraphs max)
8. Include a paragraph describing the contribution made by each team member

Hints-

1. There are techniques to extract features, such as Bag of Words, n-grams, Tf-Idf, Word2Vec, CountVectorizer, and many others.
2. Know your data. Look at the data in dataset (Open the data file and see the data or use pandas to check the info).

Nearest Neighbor Classifier:

Dataset information

Load dataset_NN.csv
The data is ordered by age

Features Description:

- Pregnancies → Number of times pregnant
- Glucose → Plasma glucose concentration 2 hours in an oral glucose tolerance test
- BloodPressure → Diastolic blood pressure (mm Hg)
- SkinThickness → Triceps skin fold thickness (mm)
- Insulin → 2-Hour serum insulin (mu U/ml)
- BMI → Body mass index (weight in kg/(height in m)^2)
- DiabetesPedigreeFunction → Diabetes pedigree function
- Age → Age (years)

Target Variable:

- Outcome - Class variable (0 or 1)

Tasks:

1. Load dataset_NN.csv dataset. [2 points]
2. Data Pre-processing. [10 points]
3. Using Pearson's Correlation Coefficient find out the relation between variables using Heat Map (Draw heat maps before and after cleaning data to find differences) [5 points]
4. Scale the data and mention which scaling technique used [2 points]
5. Split your dataset 75% for training, and 25% for testing and do cross validation for the classifier. [2 points]
6. Find the best K using elbow method. [5 points]
7. Use Euclidean distance. [2 points]
8. Select three best attributes and explain why you chose them. [15 points]
9. Test the classifier with three different k values for neighbors and record the results. [15 points]
10. Plot the ROC curve for best K value. [5 points]
11. Use comments to explain your code at each step of all points. [2 points]

12. Calculate and print the confusion matrix, and the classification Report (includes: precision, recall, f1-score, and support) for all three different numbers. Plot the Error rate vs. K-value. [15 points]
13. Create a report (Jupyter markdown) file to include concise answers to the following questions:
    a. Describe the Nearest Neighbors method and why scaling is important in KNN. [5 points]
    b. Explain what your criteria was for selecting the three attributes. What other 3 attribute can you choose? Visualizations of the target variable w.r.t three most significant attributes in a 2D projection, and write your observations in 4 - 5 lines [5 points]
    c. Explain Pearson's Correlation Coefficient, write the observations from heatmaps drawn. [5 points]
    d. Interpret and compare the results. [5 points]
14. Include a paragraph describing the contribution made by each team member

Hints:

1. Dataset consists of Nan values/ null values, to pre-process the data, you simply should not replace with mean/median, instead understand the data distribution, and do data preprocessing.
2. You can use libraries: NumPy, Pandas, Scikit-learn, Matplotlib and Seaborn
3. While choosing K-values, that should be meaningful, you cannot just simply choose and do analysis. Describe why you choose only those particular K values.
4. Models' accuracy depends on the first step i.e., data preprocessing

Programming Assignment Details:

- For this assignment use Jupyter notebook.
- You can use libraries: NumPy, Pandas, Scikit-learn, Matplotlib and Seaborn
- Make sure to explain any kind of visualization
- Do not to forget to cite any external sources used by you.
- Do not rename the dataset files

Submission:

Fill your name and ID in all the Jupyter notebooks for each group member in the following format:

First Student Name and ID: ABC 1001XXXXXX
Second Student Name and ID: ABC 1001XXXXXX
Third Student Name and ID: ABC 1001XXXXXX

Name your submission files:

- yourLastName_Last4digitsofyourID_DT.ipynb
- yourLastName_Last4digitsofyourID_NB.ipynb
- yourLastName_Last4digitsofyourID_NN.ipynb

For example:

Last name	-	last 4 digits of three team members:
abc	-	1234
def	-	5678
xyz	-	3819

Name for Decision Tree file: '**abc_1234_def_5678_xyz_3819_DT.ipynb**'

NOTE: Only one of the team members will submit all the files

Expected Output: **(Each file should be submitted separately, do not compress them)**