Assignment

In this assignment, we will implement and compare Naive Bayesian Linear Discriminant and Quadratic Discriminant classifiers to help predict the onset of diabetes. For the dataset, we use "diabetes" records data set from Kaggle:

`https://www.kaggle.com/datasets`
`/shantanudhakadd/diabetes-dataset-for-beginners`

Dataset Description: From the website: "This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage."

Dataset features:

1. pregnancies: number of times pregnant
2. glucose: plasma glucose concentration at 2 Hours in an oral glucose tolerance test (GTIT)
3. blood pressure: diastolic blood pressure (mm Hg)

4. skin thickness: Triceps skin fold thickness (mm)
5. insulin: 2-Hour Serum insulin (µh/ml)
6. BMI: Body mass index [weight in kg/(Height in m)]
7. diabetes pedigree: indicates the function which scores likelihood of diabetes based on family history
8. age: age of the patient
9. outcome; 0 indicates "no diabetes", 1 indicates "diabetes"

We will focus on the following subset of four features:

1. f_1: glucose
2. f_2: blood pressure
3. f_3: skin thickness
4. f_4: insulin

Question 1:

1. load the data into Pandas dataframe. Extract two dataframes with the above 4 features: df_0 for healthy patients (OUTCOME = 0) and df_1 for unhealthy patients (OUTCOME = 1)

2. for each dataset, construct the visual representations of correponding correlation matrices M_0 (from df_0) and M_1 (from df_1) and save the plots into two separate files

3. examine your correlation matrix plots visually and answer the following:

   (a) which features have the highest correlation for healthy patients?

   (b) which features have the lowest correlation for healthy patients?

   (c) which features have the highest correlation for healthy patients?

   (d) which features have the lowest correlation for unhealthy patients?

   (e) are results the same for both cases?

4. for each class and for each feature f_1, f_2, f_3, f_4, compute its mean $\mu()$ and standard deviation $\sigma()$. Round the results to 2 decimal places and summarize them in a table as shown below:

5. examine your table. Are there any obvious patterns in the distribution of each class

Question 3

class	$\mu(f_1)$	$\sigma(f_1)$	$\mu(f_2)$	$\sigma(f_2)$	$\mu(f_3)$	$\sigma(f_3)$	$\mu(f_4)$	$\sigma(f_4)$
0								
1								
all								

1. apply Logistic regression
2. apply k-NN ($k = 1, 3, 5$)
3. apply Naive-Bayesian classifier
4. apply Linear Discriminant
5. apply Quadratic Discriminant
6. compute its confusion matrices and summarize results in table below:

method	TP	FP	TN	FN	accuracy	TPR	TNR
Logistic Reg. k-NN ($k = 1$) k-NN ($k = 3$) k-NN ($k = 5$) Naive Bayesian NB Linear Discr. Quadr. Discr.							

7. examine your results and correlation matrices. Any conclusions?