Final Exam Computer Programming for Data Science and Artificial Intelligence

Note: Please save the file as FinalExam_<student_id>_<student_name>/ Also put the link of your deployed site in teal / Please Upload answers to 5a and 5b in Word (The file should have your <student_id>_<student_name>)

Reminder:

- You're only allowed to code in Visual Studio Code
- You're not allowed to use Github Co-pilot for AI extensions such as these
- You're not allowed to reference any notes/websites except your A4 cheat sheet
- Please take note of the file name and please remember to put the link in teal

---

Description: Given a dataset, develop a classification model to predict the chance of diabetes. Show all your work and justify your answers.

Tasks

1. Load the dataset (Link:) (1 mark)
2. EDA - (5 marks)
   a. Analyze the numerical distribution of data and identify the null values (1 mark)
   b. Plot the distribution of the target (1 mark)
   c. Plot the distribution of continuous features (1 mark)
   d. Analyze the EDA results you have concluded in short. Investigate Class Imbalance, outliers, and distribution of data (2 marks)
3. Modeling with Pipeline - (24 marks)
   a. Encoding, Scalerisation, Splitting in the pipeline - (1+1+2 marks)
   b. Classification Model (Logistic Regression and perform grid search and CV Split to find the best params for the best model) - (5 marks)
   c. K Means Model (5 marks)
      i. Apply the K-Means model using 2 clusters (2.5 marks)
      ii. Scatterplot of the clusters found and analysis of the results (2.5 marks)
   d. Classification Model of Clusters - (10 marks).
      i. Apply classification model using Logistic Regression for 1st cluster (4 marks)
      ii. Apply classification model using Logistic Regression for 2nd cluster (4 marks)
      iii. Analysis of the resulting impact based on each feature (2 marks)
4. Analysis of the performance of Logistic Regression with and without clustering (3d and 3b) using the following evaluation metrics - (10 marks)
   a. AUC - (2 marks)
   b. Classification Report - (2 marks)

c. Confusion matrix - (2 marks)
d. Analysis and comparison between the models (List some problems with the given solutions and provide a better way to improve them by the implementation) - (1 + 3 marks)

5. Deployment of the Logistic Regression Model from 2 - (5 +5 marks)
   a. Screenshot of the website with IP visible (5 marks)
   b. Screenshot of the result with the given inputs (5 marks)

Note: Please Upload 5a and 5b in Word (The file should have your <student_id>_<student_name>) together with the link of your deployment.

<Make sure the Input/output are as follows :>
Input:

1. **Gender: Female, Age: 80, Hypertension: 0, Heart_diseae: 0, smoking_history: former, BMI: 21.97, HbA1c_level: 7, blood_glucose_level: 300**
2. **Gender: Male, Age: 51, Hypertension: 0, Heart_diseae: 0, smoking_history: never, BMI: 27.32, HbA1c_level: 4.8, blood_glucose_level: 145**

Output: Diabetes and Non-Diabetes based on the Input with the Confidence Values of each prediction