Mid Term Test (15%)

Type 2 diabetes is a chronic condition that affects how the body metabolizes glucose, a vital source of energy. It is particularly prevalent in specific populations due to genetic, lifestyle, and socioeconomic factors. One such group is the Pima Indian community in Arizona, who have historically exhibited high rates of diabetes. To address this growing health concern, the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) conducted a study collecting medical and personal health data from female Pima Indian patients aged 21 and older.

The dataset includes features such as blood pressure, glucose levels, insulin levels, body mass index (BMI), age, and pregnancy history—factors commonly associated with the risk of diabetes. This rich data presents an opportunity to apply machine learning techniques to predict whether an individual is likely to be diagnosed with diabetes.

Instructions: Work individually. Answer all the questions. Explain adequately how you get the answers that can include: 1) codes or process used; and 2) print screens and/or related files that can justify your outcome. Please create a word file in .docx (or in .ipynb) to explain your answer. Convert the file to pdf format and submit to the spectrum "Mid Term" submission page.

Part 1: (10 marks)

1. Download one of the diabetes datasets in CSV format. Please refer to Appendix 1 (at the end of the document) on which dataset you should download. Load the data and preprocess whenever necessary. (2 marks)
2. Perform K-means clustering, using the first 7 columns of the dataset as your features. Find out the optimum number of clusters, and explain what the resulting clusters mean, to the best of your understanding. (2 marks)
3. Train linear regression model, using the first 7 columns of the dataset as your features, and the 8^{th} column as your target variable. Evaluate the outcome with and without data preprocessing. (3 marks)
4. Train decision tree model, using the first 7 columns of the dataset as your features, and the 9^{th} column as your target variable. Evaluate the outcome with and without data preprocessing. (3 marks)

Part 2: (5 marks)

1. Based on the outcome in Part 1 Question 3 or Question 4 (pick one of it), improve the outcome using the proposed machine learning solution of your choice. Explain your solution in detail. (5 marks)

Appendix 1: Please download respective datasets (in .csv format, e.g. Set1.csv) based on the name list below:

No	Name	Set
1.	KEVIN WONG XIN KAI	1
2.	TAN YONG SHENG	2
3.	WANG YUFENG	3
4.	VINOD KUNHI KRISHNAN ANNUKARAN	4
5.	CHANG CHI HUI	5
6.	MIAO XINYU	6
7.	RABITA BHUIYA TANAYA	7
8.	HUA SHAOJIE	8
9.	SEAN LEE	9
10.	AIDA WANIE BINTI JASNI	10
11.	DIVA ALIFTA CHANDRA	1
12.	CHONG KAH HOE	2
13.	SHARIFAH NURUL AMIRAH BINTI SYED AHMAD FAUZI	3
14.	SAN YONGLI	4
15.	LOW ZI YANG	5
16.	VIJAYKUMAR KARTHA RAMACHANDRAN	6
17.	MAWADDAH BINTI MUSTHAFA	7
18.	NURUL HAFIZAH BINTI ZAINI	8
19.	NURUL NADIA BINTI ABD RAHMAN	9
20.	NUR MAISARAH BINTI JALALULAIL	10
21.	AHMAD MARWAN BIN MURSHIDI	1
22.	TAN RIK EE	2
23.	MUHAMAD RAFIQ IQBAL BIN SAMSUDIN	3
24.	TAN JIAN LIN	4
25.	TEO KAI NING	5
26.	YANWEI ZHANG	6
27.	LI YUE XIN	7
28.	SITI NUR LIYANA BINTI ROSLAN	8
29.	TAN FOO HOU	9
30.	MUHAMMAD HAKIM BIN NASARUDDIN	10
31.	ARUN KUMAR	1
32.	LEE MIN QI	2
33.	LOW E-JIE	3
34.	LOH KE YI	4
35.	CHUA SZE YAN	5
36.	SITI HAIRUNEESHA BINTI ZAINUDDIN	6
37.	LAU WEN XI	7
38.	NUR ARIANA SOFEA BINTI BADRUL HISHAM	8
39.	SHAIK MOHAMMED MUZEEB	9
40.	VETRI A/L THANABALAN	10
41.	LUO QINGZHEN	1
42.	YEE SEE MARN	2
43.	LIM SZE GEE	3
44.	ALESSANDRO GUIDO JACQUES BROZZONI	4
45.	LISA HO YEN XIN	5
46.	CHOW KOO LI	6
47.	TOH CHU XIAN	7
48.	TAN CHEE YONG	8
49.	LAW YU XUAN	9
50.	LIJINZHAO	10

51.	NICHOLAS OOI JIAWEI	1
52.	HUSSAIN ALI KAZIM	2
53.	KARENINA KAMILA	3
54.	AREEJ ABDURAHMAN MOHAMMAD	4
55.	MOHAMMED IQRAM	5
56.	LEE XUAN YU	6
57.	LOOI XUE YING	7
58.	NUR RIDWANA BINTI MOHD RAFIX	8
59.	SITI NURAISHAH BINTI AB MANAM	9
60.	TIAN TIANCHU	10
61.	YANG JIYU	1
62.	YAO YUANWEI	2
63.	ANG ZHI YANG	3
64.	LOONG SHIH-WAI	4
65.	WANG JIAJU	5
66.	TENGKU MUHAMAD FIRDAUS MAHMOOD BIN TENGKU ZAMBRI	6
67.	AHMAD DANIEL BIN MOHD SUPANDI	7
68.	WANG KEXIN	8
69.	PAARISHA EMILIE	9
70.	KARAM ALJANADI	10
71.	MUHAMMAD TAUFIQ BIN ISMAIL	1
72.	LIM SZE CHIE	2
73.	MOHAMMAD IQBAL AFIF BIN MOHAMAD SHAHNAZ	3
74.	KHOR KEAN TENG	4
75.	TAN YEE THONG	5