Outline

The choose diabetes diagnosis problem as it is the fourth leading cause of death in the world and one of the most common endocrine disorders. According to studies, Type-2 diabetes kills thousands of people around the world every year and imposes huge costs on societies in the form of surgeries and other treatment programs, as well as controlling complications and disability. Therefore, predicting and early diagnosis of this disease will greatly help governments and patients.

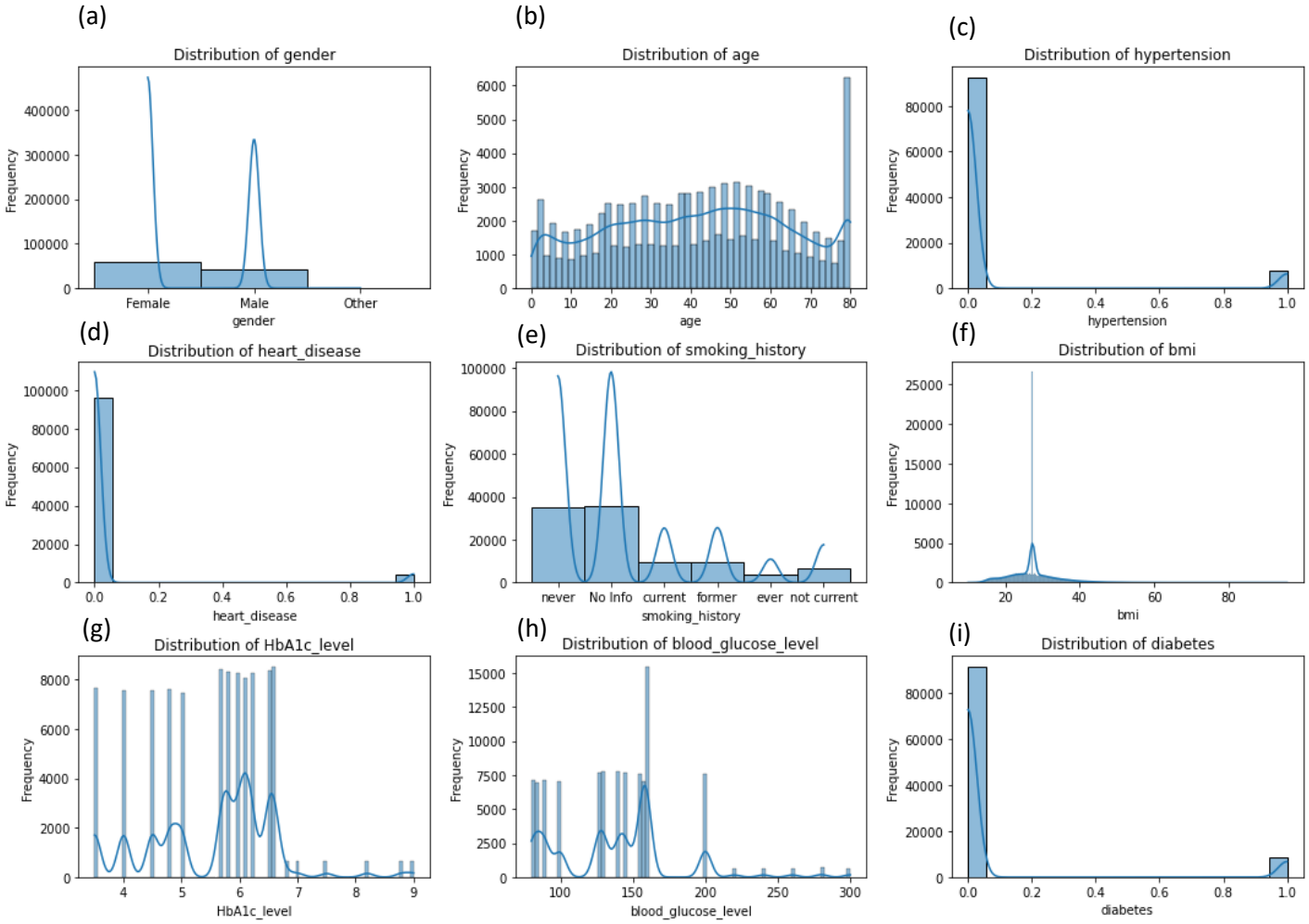


Fig.1. The frequency and distribution of features, including (a) gender, (b) age, (c) hypertension, (d) heart disease, (e) smoking history, (f) bmi, (g) HbAlc , and (h) blood glucose level.

The size of the dataset we utilized to predict early diagnosis on diabetes is 100000 instances by 9 features (attributes), including gender, age, hypertension, heart_disease, smoking_history, bmi abbreviated for body mass index, HbA1c_level, blood_glucose_level, and diabetes which is if the subject is diagnosed with diabetes or not labeled as 1 and o, respectively. Gender and smoking_history are categorical variables (see Fig. 1 (a) and Fig. 1(e)), and the rest of features are numerical. Worth mentioning, hypertension, heart_disease, and diabetes are features shown in Fig. 1(c, g, and i) taking only 0 and 1 showing a logistic distribution. We will report standard deviation of this sample with 100000 observations and ratio (%) for the features that contain non-binary values and binary values, respectively. Note that age is in the range of 0 to 80 (see Fig. 1(b)), and blood glucose level starts at 30 and ends at 300 (see Fig. 1(h)). Also, Fig. 2 is an example of the bmi table which is the ratio of weight and height divided with five classes from normal to obesity. Smoking history contains six levels and roughly 360,000 observations without any information.

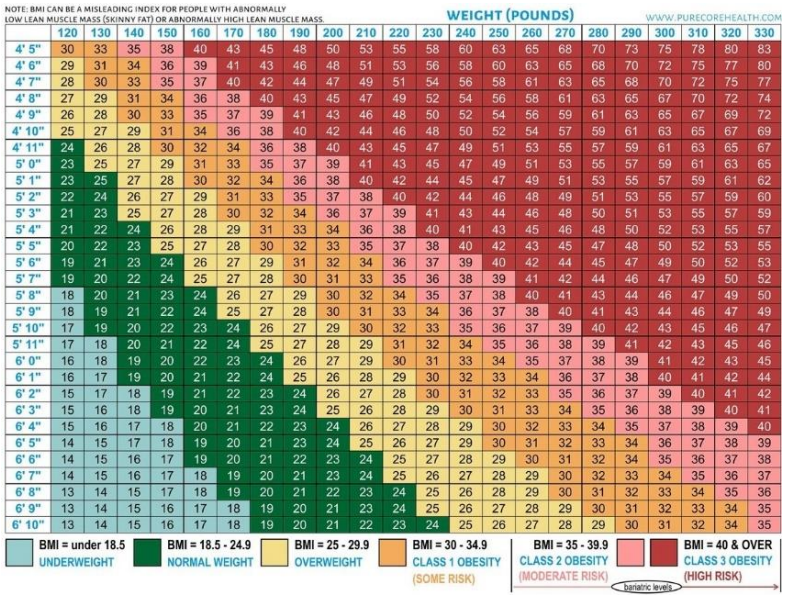
Statistical Analysis

Initially, the statistical summary of the dataset is executed to have an overview on the information, including mean, median, mode, quantiles, and a function is introduced as IQR which calculates the difference between Q1 and Q3 as the first and the third quartiles (see Eq. 1). We will use IQR to find the outliers in the distribution of each feature and Eq. (1-3) are used for this purpose. Note that datapoints which fall below Q1-1.5IQR (Eq. 2) and above Q3+1.5IQR (Eq. 3) are considered as outliers, and they are added to the summary table in the program.

$$IQR = Q3 - Q1 \quad \textbf{(1)}$$

$$\text{Outlier (low range)} = Q1 - (1.5.IQR) \quad \textbf{(2)}$$

$$\text{Outlier (higher range)} = Q3 + (1.5.IQR) \quad \textbf{(3)}$$



	120	130	140	150	160	170	180	190	200	210	220	230	240	250	260	270	280	290	300	310	320	330
4' 5"	30	33	35	38	40	43	45	48	50	53	55	58	60	63	65	68	70	73	75	78	80	83
4' 6"	29	31	34	36	39	41	43	46	48	51	53	56	58	60	63	65	68	70	72	75	77	80
4' 7"	28	30	33	35	37	40	42	44	47	49	51	54	56	58	61	63	65	68	70	72	75	77
4' 8"	27	29	31	34	36	38	40	43	45	47	49	52	54	56	58	61	63	65	67	70	72	74
4' 9"	26	28	30	33	35	37	39	41	43	46	48	50	52	54	56	59	61	63	65	67	69	72
4' 10"	25	27	29	31	34	36	38	40	42	44	46	48	50	52	54	57	59	61	63	65	67	69
4' 11"	24	26	28	30	32	34	36	38	40	43	45	47	49	51	53	55	57	59	61	63	65	67
5' 0"	23	25	27	29	31	33	35	37	39	41	43	45	47	49	51	53	55	57	59	61	63	65
5' 1"	23	25	27	28	30	32	34	36	38	40	42	44	45	47	49	51	53	55	57	59	61	62
5' 2"	22	24	26	27	29	31	33	35	37	38	40	42	44	46	48	49	51	53	55	57	59	60
5' 3"	21	23	25	27	28	30	32	34	36	37	39	41	43	44	46	48	50	51	53	55	57	59
5' 4"	21	22	24	26	28	29	31	33	34	36	38	40	41	43	45	46	48	50	52	53	55	57
5' 5"	20	22	23	25	27	28	30	32	33	35	37	38	40	42	43	45	47	48	50	52	53	55
5' 6"	19	21	23	24	26	27	29	31	32	34	36	37	39	40	42	44	45	47	49	50	52	53
5' 7"	19	20	22	24	25	27	28	30	31	33	35	36	38	39	41	42	44	46	47	49	50	52
5' 8"	18	20	21	23	24	26	27	29	30	32	34	35	37	38	40	41	43	44	46	47	49	50
5' 9"	18	19	21	22	24	25	27	28	30	31	33	34	36	37	38	40	41	43	44	46	47	49
5' 10"	17	19	20	22	23	24	26	27	29	30	32	33	35	36	37	39	40	42	43	45	46	47
5' 11"	17	18	20	21	22	24	25	27	28	29	31	32	34	35	36	38	39	41	42	43	45	46
6' 0"	16	18	19	20	22	23	24	26	27	29	30	31	33	34	35	37	38	39	41	42	43	45
6' 1"	16	17	19	20	21	22	24	25	26	28	29	30	32	33	34	36	37	38	40	41	42	44
6' 2"	15	17	18	19	21	22	23	24	26	27	28	30	31	32	33	35	36	37	39	40	41	42
6' 3"	15	16	18	19	20	21	23	24	25	26	28	29	30	31	33	34	35	36	38	39	40	41
6' 4"	15	16	17	18	20	21	22	23	24	26	27	28	29	30	32	33	34	35	37	38	39	40
6' 5"	14	15	17	18	19	20	21	23	24	25	26	27	29	30	31	32	33	34	36	37	38	39
6' 6"	14	15	16	17	19	20	21	22	23	24	25	27	28	29	30	31	32	34	35	36	37	38
6' 7"	14	15	16	17	18	19	20	21	23	24	25	26	27	28	29	30	32	33	34	35	36	37
6' 8"	13	14	15	17	18	19	20	21	22	23	24	25	26	28	29	30	31	32	33	34	35	36
6' 9"	13	14	15	16	17	18	19	20	21	23	24	25	26	27	28	29	30	31	32	33	34	35
6' 10"	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	34	35

Fig.2. The bmi table, indicating five classes; underweight, normal weight, overweight, class 1 obesity, class-2 obesity, class-3 obesity 3.

Visualization

Note that Fig.1 shows the distribution of all features in a 3 by 3 plot depicted with histograms and graph-lines. Features are assigned to the x-axis and y-axis is frequency. Note that Fig.1(a) depicts gender with three groups, including males, females, and other. The number of males is slightly higher than females, and the other group of subjects is negligible in comparison with the other two groups. It is interesting that each group of gender is normally distributed. Age is normally distributed and bmi is almost normal with a slight positive skewness. Blood glucose and red blood cells indicated as HbA1c are skewed to the left. We used different sorts of visualization for different purposes that will be discussed. We used boxplots to show min, max, median, Q1, Q3, and probable outliers. It can be seen in Fig. 4(a) that age is almost normally distributed but bmi, glucose level, and HbA1c are left skewed (see Fig. (b-d)). Moreover, we calculated the difference between Q1 and Q3 indicated as IQR to use it for finding outliers for each feature which is added to the summary table. Moreover, we extracted all outliers and saved them in an individual table as outliers.

We used Scatterplots to show the relationship between the features to look at their trend (see Fig. 5). For instance, we plotted bmi over age to see if older people may have higher bmi indicating over-/under-weight. Generally, scatterplots give us a clear visual on the outliers. Also, we thought of another hypothesis testing if people who smoke may have lower glucose than others who do not smoke regularly. But they are not highly correlated since we see a common distribution of blood glucose for people with different smoking history (see Fig. 5(g)). Note that Fig. 5(e) shows unequal variability at different levels of blood glucose over age. Fig. 5(c) depicts that the blood glucose level of subjects with HbA1c higher than 7 starts at roughly 125, and people with HbA1c lower than 6 reached 200 of blood glucose which is totally the opposite

trend compared to people with HbA1c higher than 7. Furthermore, we can see people diagnosed with or without diabetes regardless of the smoking history (see Fig. 5(h)). Finally, we illustrated the normality of age, bmi, blood glucose, and HbA1c by QQ-plot which is indicated with blue color and the red line is mean (see Fig. 6). They show lower normality when the trend of each feature gets less linear and far from the mean line. Also, we can see the variability of the observations for each feature.

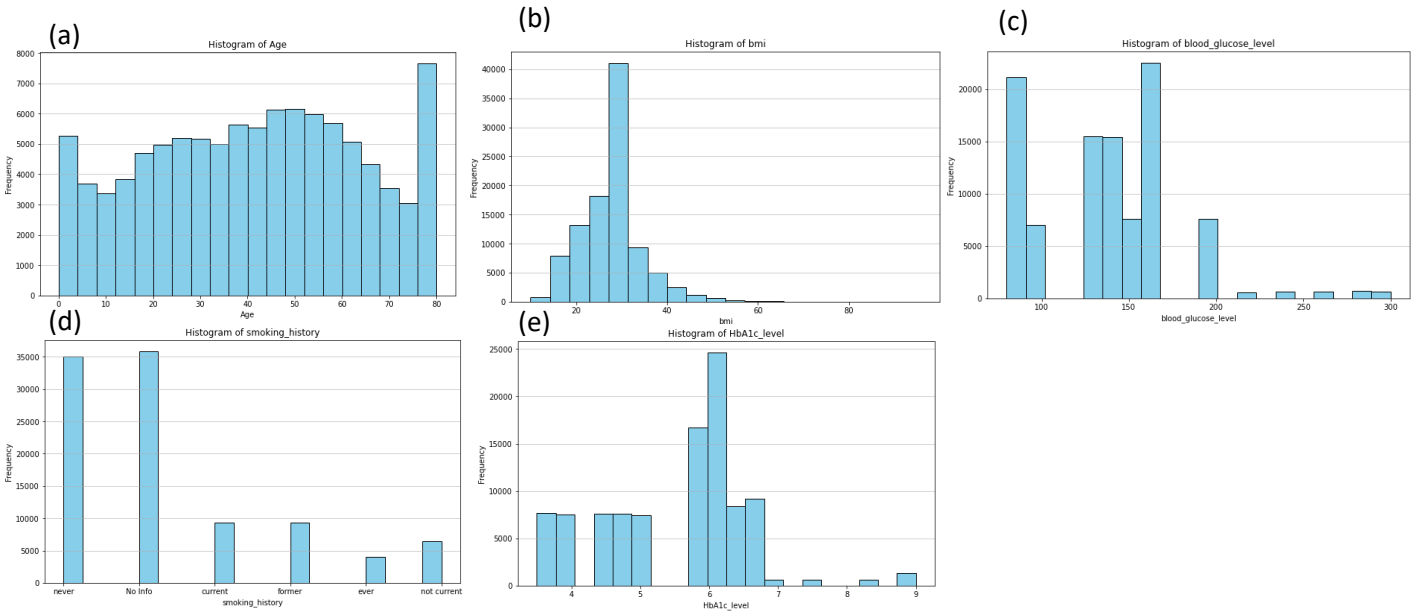


Fig.3. Histograms of (a) age, (b) bmi, (c) blood glucose level, (d) HbA1c level, and (e) smoking history.

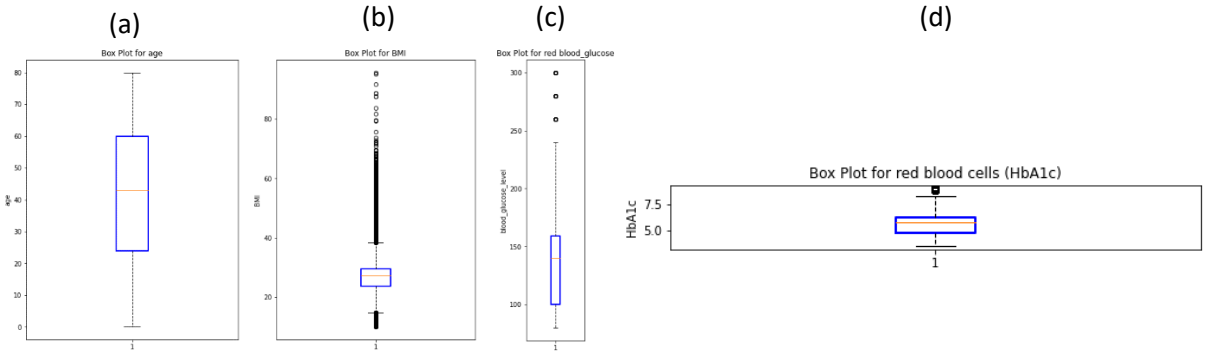


Fig.4. Boxplots of (a) age, (b) bmi, (c) blood glucose level, and (d) HbA1c level.

We calculated the frequency of each group at different levels to determine the ratio of each group over the total population of 100000. Also, we drew the Pearson's correlation table shown in Fig. 7 to look at the relationship of pair-wise features. This can be useful for feature selection or understanding which features may be inversely related to diabetes diagnosis. It can also give us the feasibility of feature reduction for the ones that might be highly correlated to decrease the process cost for which no feature is detected, and we may utilize other methods of feature engineering. Diabetes status is mostly correlated with blood glucose level and the amount of red blood cells (HbA1c) with 42% and 40%, respectively followed by 26% of correlation with Diabetes. Note that bmi and age have the highest correlation with 34%. Heart disease is least correlation with bmi with only 6% followed by correlations of heart disease and hypertension with HbA1c level, indicated as 7% and 8%, respectively.

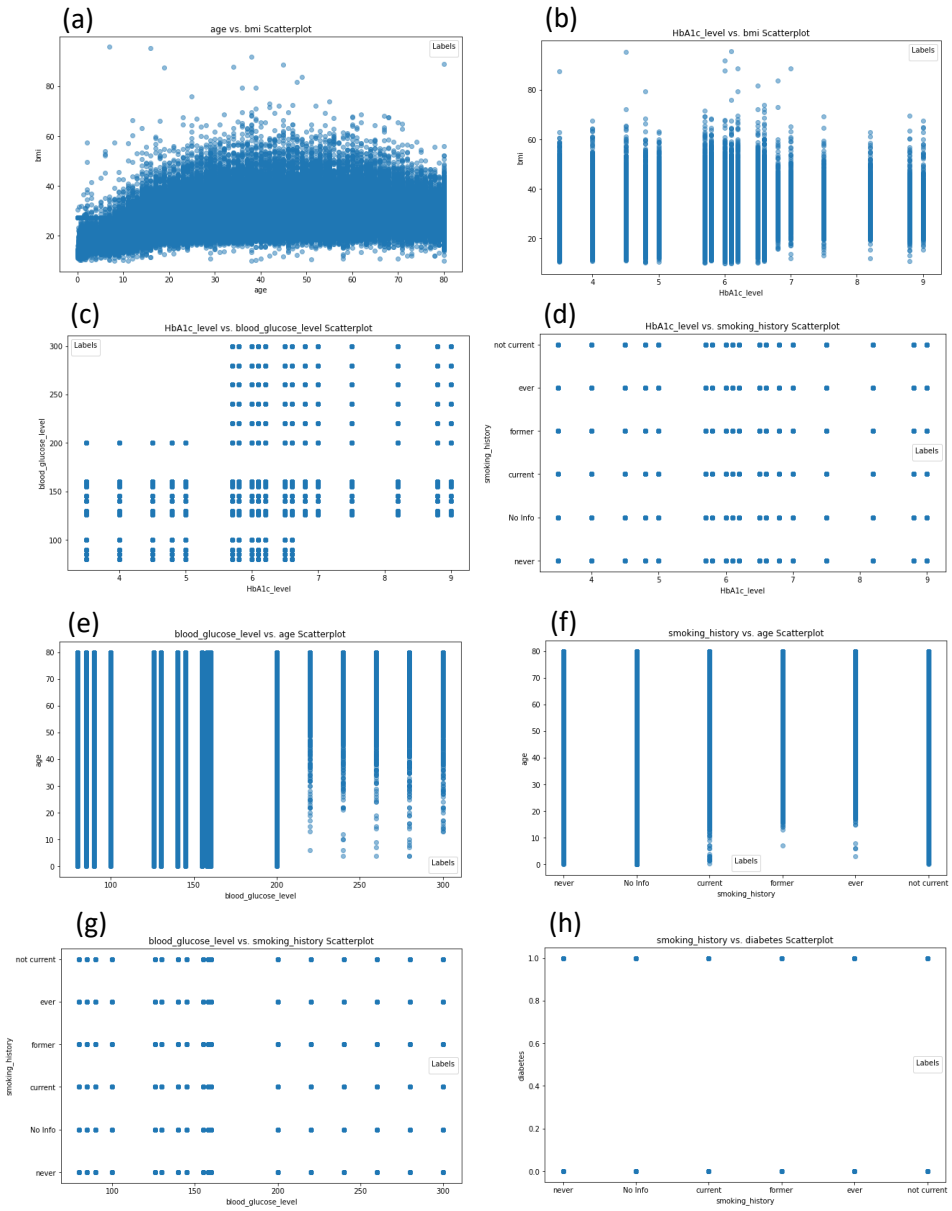


Fig.5. Scatterplots of (a) age vs. bmi, (b) bmi vs. HbA1c, (c) blood glucose vs. HbA1c, (d) smoking history vs. HbA1c, (e) age vs. blood glucose, (f) age vs. smoking history, (g) smoking history vs. blood glucose, and (h) diabetes status vs. smoking history.

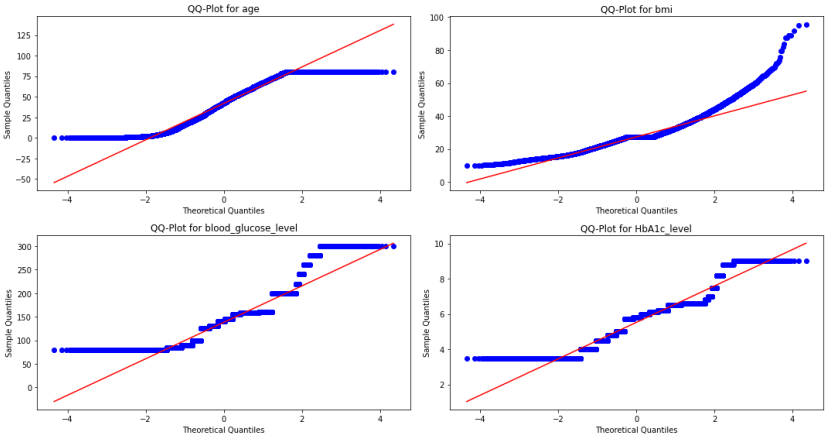


Fig.6. QQ-plot for (a) age, (b) bmi, (c) blood glucose, and (d) HbA1c.

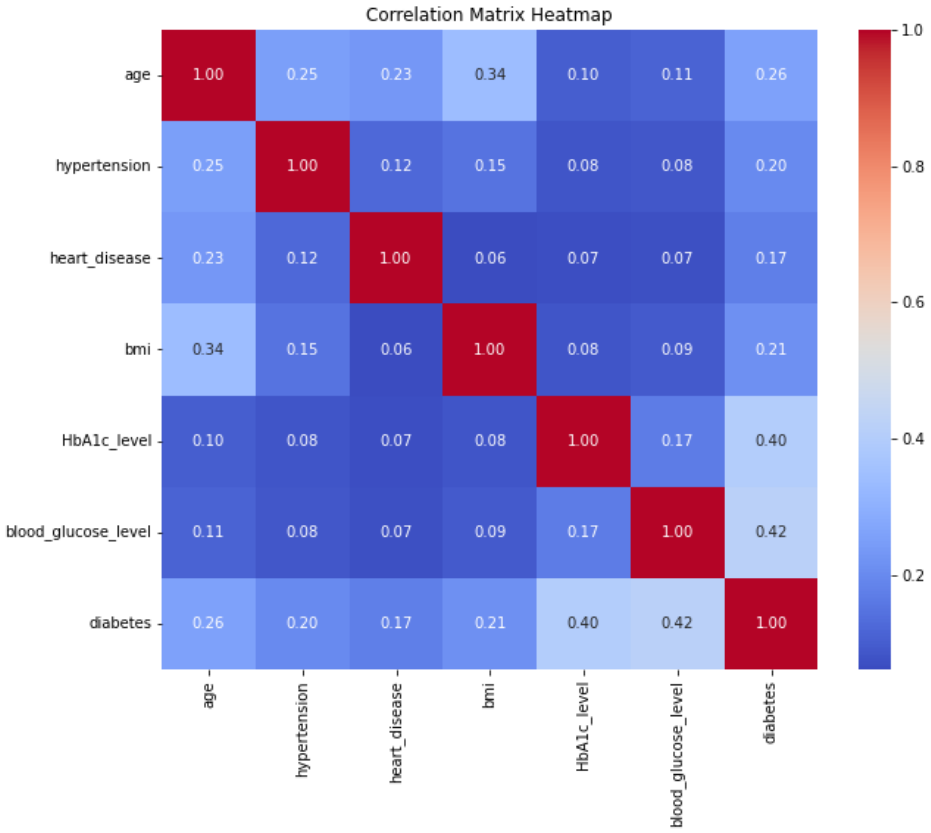


Fig.6. Pearson's correlation table for the numerical features, including, age, hypertension, heart-disease, bmi, HbA1c_level, blood glucose level, and diabetes.