The Hong Kong Polytechnic University
Department of Electrical and Electronics Engineering

Project ID: [FYP_ lpchau_20240523150449]

Machine learning model to predict the risk of diabetes

by

CHAN Hou Ting Constant

21034774D

Final Year Project - Interim Project Report 2024/2025 Sem 1

Bachelor of Science (Honours)

In

Internet and Multimedia Technologies

of

The Hong Kong Polytechnic University

Supervisor: Prof CHAU Lap Pui

Date:27-December-2024

Abstract

Diabetes has become a noteworthy social issue in the world, where more and more people have been diagnosed with diabetes in recent years. To find out the problem, machine learning is one of the approaches used to predict diabetes. This project introduces two datasets, the Pima Indian Diabetes dataset and the NHANES dataset. In addition, a machine learning framework proposed by Tasin et al. [4] is the baseline model in this model. Moreover, polynomial regression and SMOTE are applied to predict missing values and class imbalance problems. Furthermore, Random Forest and XG Boost are used for the prediction of diabetes in this project, and Random Forest and XG Boost have the highest accuracy in the Pima Indian Diabetes dataset and the NHANES dataset, respectively. In the future, exploring other key features and preprocessing methods will be the major options to get better results for this project.

Contents

List of Figures

List of Tables

1. INTRODUCTION

Diabetes has been increasingly prevalent across the world for the past several years. It is estimated that currently, in 2022, around 8.3 million people [1] exhibit symptoms of diabetes, which comprises approximately 10.4% of the world's population, which makes ignoring diabetes impossible. Modern people have a fast life that we can hardly spare time to exercise, which leads to unhealthy living habits, such as obesity, sleep deprivation, etc. There are two main categories of common diabetes: Type 1 and Type 2. The former is congenital, hereditary, etc. [2], and the latter is its own acquired poor eating habits, lack of exercise, etc. [2]. Machine learning has been applied in diabetes prediction for years to predict if a patient is likely to develop diabetes. One of the advantages of machine learning is that it can generate the corresponding prediction based on the dataset's content. It allows people to make suitable decisions based on the predictions generated by machine learning algorithms. This project explores machine learning algorithms and applies them to diabetes predictions.

1.1 Overview

1.1.1 Background

One of the most significant issues in the world is Diabetes. As mentioned earlier, many people, whether adults, youth, or children, have had diabetes in recent years. Early for diabetes, people used to do blood glucose measurements to check their blood glucose levels to see if they had diabetes or not. In the past, a blood glucose meter was used. With the evolution of science and technology, another approach that is put forth is the identification of diabetes by using machine learning algorithms. People can view the predicted output generated by machine learning algorithms to check if they are at risk for diabetes. It saves time for people who are allowed to check their health conditions without viewing the indices from the body. Moreover, it decreases the probability of misjudgment due to human factors.

1.1.2 Problem Statement

Although diabetes prediction with machine learning has been implemented recently, much research on diabetes prediction with machine learning algorithms has indicated different results. For example, Mujumdar and Vaidehi [3] indicated that Logistic Regression and Adaboost perform well with high accuracy in diabetes prediction. Tasin et al. [4] commented that XGBoost with the ADASYN approach performs well in diabetes prediction. It is difficult for people to judge whether machine learning models are the most suitable for diabetes prediction. Furthermore, much research has been done on diabetes prediction using machine learning algorithms and different datasets (NHANES and Pima Indian Diabetes) and data preprocessing methods. Making the same topic on diabetes prediction with machine learning algorithms has caused different results. Therefore, this project is a good research topic to work on. The results will be determined by collecting data from different sources, implementing the methodology, and comparing different methodologies.

2. Dataset

2.1 Pima Indian Diabetes Dataset

Pima Indian Diabetes Dataset	
Features	Pregnancies
	Glucose **(2 hours in an oral glucose tolerance test (mg/dL))**
	BloodPressure **(Diastolic blood pressure (mm Hg))**
	SkinThickness
	Insulin **(2-Hour Serum insulin (µh/ml))**
	BMI
	DiabetesPedigreeFunction
	Age
Target	Outcome

Table 1: Information on Pima Indian Diabetes Dataset

The Pima Indian Diabetes Database provided information about the patients who have Diabetes or not. The dataset source comes from the National Institute of Diabetes and Digestive and Kidney Diseases [5]. A total of 768 patients were recorded in the Pima Indian Diabetes Database, which are Pima Indians that are at least 21 years old females.

A total of 9 variables were listed in the dataset, which included eight features and one target variable. Here is the explanation of these variables:

Pregnancies: It means the number of times pregnant

Glucose (Blood Sugar): It is a group of carbohydrates [6] that provides energy for the body, and mg/dL is the measuring unit of glucose. If the glucose is lower than 140 mg/dL, it is considered normal [7].

BloodPressure: It means heart beats and pumps blood into the arteries [8]. Lack of exercise and obesity would result in Higher blood pressure, and it would cause health risks such as headache and dizziness.

SkinThickness: It estimates the body fat on thighs and limbs.

Insulin: It helps regulate blood sugar levels and is important for energy production and storage.

BMI: It measures body fat based on Height and Weight. 18.5 to 23 is considered a healthy weight and a normal body level.

$$BMI = \frac{Weight}{Height^2}$$

DiabetesPedigreeFunction: It is a function that scores the probability of Diabetes based on Family history.

Age: The age of all patients is at least 21 years old.

Outcome: A variable that diagnosed Diabetes or not.

2.2 2013-2014 NHANES Dataset

The National Health and Nutrition Examination Survey (NHANES) is a project that the National Center for Health Statistics implemented. This project aims to collect data from American adults and children through interviews and body checks. NHANES collected dietary intake, physical examinations, and laboratory tests. Also, this project uses population-based sampling that includes the entire American population. This dataset is available for open access and widely used for health research and public health initiatives. Here is the abstract of the dataset:

NHANES Dataset	**Features**
Demographic	SEQN (ID of interviewee) RIAGENDR (Gender) RIDAGEYR (Age)
Diet	DR1DAY (Intake day of the week) DR1TKCAL (Energy (kcal) take in 1 day)
Examination	BMXBMI (BMI) BPXDI1 (Blood Pressure)
Labs	LBXGLT (Glucose) LBXIN (Insulin)
Questionnaire	DIQ010 (Diabetes_Diagnosis) ALQ120Q (alcoholic drinks taken per day/ months)

Table 2: Abstract of NHANES Dataset

NHANES Dataset is divided into five parts, which are demographic, diet, examination, labs and questionnaire.

Demographic: it means the characteristics of a population, which include gender, age and marital status, etc.

Diet: it means the dietary intake information collected from the interviewees. Nutrient information like Energy taken, Vitamins, fats and carbohydrates are recorded in the database.

Examination: it means the physical examinations and medical tests conducted on the interviewees, such as BMI and blood pressure.

Labs: it means the laboratory tests performed on biological samples collected from the interviewees, such as glucose levels and Insulin.

Questionnaire: it means the self-reported information collected from the interviewees through structured interviews and surveys. It covers the topics that related to health and lifestyle like physical activity and health conditions. The details of the data processing would be explained in the following section.

3. CURRENT PROGRESS

For this project, machine learning is a suitable approach to diabetes prediction because diabetes prediction belongs to a classification task that determines whether or not the patients are diagnosed with diabetes. The process includes data preprocessing, feature selection, training, testing, and performance evaluation. All the work is done on Jupyter Notebook, a web-based application that provides an interactive computing notebook environment to describe the data analysis.

Tasin et al. [4] proposed a machine-learning framework that acquired 81% accuracy using XGBoost. The preprocessing methods are extreme gradient boosting techniques for filling the missing value; SMOTE and ADASYN are applied to address the class imbalance issue. In addition, it collected the samples from 203 people called RTML, which is used for filling "Insulin" and is the merged dataset. In this project, it will be the baseline model for the reference.

3.1 Data Preprocessing on Pima Indian Diabetes dataset:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
...	...	...	...	...	...	...	...	...	...
763	10	101	76	48	180	32.9	0.171	63	0
764	2	122	70	27	0	36.8	0.340	27	0
765	5	121	72	23	112	26.2	0.245	30	0
766	1	126	60	0	0	30.1	0.349	47	1
767	1	93	70	31	0	30.4	0.315	23	0

Figure 1: Pima Indian Diabetes Dataset (raw data)

The dataset has 768 rows (participants) with nine columns (features) before preprocessing. As shown in Figure 1, there are some values of zero in columns "SkinThickness" and "Insulin." There are no null values on the dataset, so filling in zero is unnecessary.

```
No. of rows with Glucose value 0 is: 5
No. of rows with BloodPressure 0 is: 35
No. of rows with SkinThikness 0 is: 227
No. of rows with Insulin 0 is: 374
No. of rows with BMI 0 is: 11
No. of rows with DiabetesPedigreeFunction 0 is: 0
No. of rows with age 0 is: 0
```

Figure 2: Check missing value on Pima Indian Diabetes Dataset

In Figure 2, many missing values existed in columns “SkinThickness” and “Insulin,” which count for about 30% and 50% of the dataset. Other columns “Glucose”, “BloodPressure” and “BMI” with value 0 will be filled by their mean as they are only a tiny minority of the whole dataset. To identify the rows where column “SkinThickness” is zero, variables “zero_SkinThickness_rows” and “non_zero_SkinThickness_rows” are defined to find the rows where it is zero and non-zero, respectively. To predict column “SkinThickness” become more reliable, columns “Glucose”, “BloodPressure” and “BMI” and “Age” are used to assist the prediction of column “SkinThickness”.

Polynomial Regression is applied to predict the missing value of column “SkinThickness”. Aditya Shastry et al. [9] applied polynomial regression to predict the missing value in data preprocessing and it was helpful to improve the model performance. First, the degree of the polynomial features is set to 2 and a bias column is not included in the polynomial features. Next, fit_transform() is applied to find the metrics of overall statistical properties (mean, standard deviation). Then, linear regression will be applied to train and predict the column “SkinThickness.”

```
26.47582279 27.06895818 26.58616826 32.09068795 37.04790656 37.41836047
30.74114278 14.05473733 31.06739863 36.93201267 26.33889225 26.27576328
28.85545217 37.70822675 28.43725733 31.32166556 20.98311832 31.54529092
36.17683759 38.4572848  29.16022526 39.13384227 28.46950422 40.36377076
10.455307   24.3901921  36.58915804 20.29989632 17.06743003 35.65999303
15.79472548 29.97322915 28.50144058 32.03517376 28.84767333 26.61549741
29.7697212  25.47323778 30.9171333  28.22602585 19.61014053 31.0243439
33.67315478 18.55973926 43.864927   28.6461028  27.1853426  28.80630537
27.93004012 28.47381263 41.21608174 36.13195024 18.8574458  22.2094413
25.89837917 15.85548421 28.78768946 45.48935861 17.81073123 36.73785708
32.41109423 35.90011006 29.70564443 22.4956126  32.37670799 39.70957316
32.09468255 39.58194382 11.95764468 19.26875902 29.89047941 30.35016166
24.41601562 26.66387924 29.99512954 32.99186572 18.31455787 23.9379662
20.01692718 33.18212697 24.27109452 34.88112661 28.82448483 25.08970059
27.44463308 23.17751193 29.68862076 44.42982603 17.41232504 27.51919632
24.63117663 33.77846215 41.06090853 20.71423519 32.32334077 29.4921595
37.50514086 31.77356749 34.80198969 18.08583548 37.98060149 28.28092813
38.29621309 28.76100545 32.04996562 37.26441071 24.17380535 43.3156894
41.28549284 30.38988279 15.10800553 29.01291046 25.60429482 32.49467649
27.68756409 15.86091667 24.19155664 11.39158549 28.90010375 16.38643764
21.83295302 38.82418501 26.74004242 33.41021014 27.85306568 24.64585876
20.39866422 30.50254236 15.80392531 19.40942748 28.02691584 27.34886716
24.56315557 40.40274073 28.15544757 28.51122052 28.83337287 22.31340825
29.8797612  24.45736653 23.63818676 22.11599122 34.40430959 30.7782215
29.13422408 33.28405634 27.30628908 20.7511649  40.46576164 26.6317657
29.244422   26.22874282 21.56311465 24.87806805 26.90024923 27.18631769
32.79874951 29.05364099 29.92133841 17.16918436 38.67143509 24.72632492
35.71025504 33.96004712 19.96320071 15.14872828 32.5398733  39.10354322
35.27472861 19.11946219 26.52109868 21.63239896 24.97782779 28.83886627
37.21306742 26.18941697 18.9885846  28.64230667 32.67634692 24.68002372
20.85381047 22.11411607 29.01616249 27.2877607  30.7721253  28.59112084
24.61658024 22.19830242 32.05757441 26.87561575 31.53801299 28.31480962
23.23365783 31.34689188 31.75730295 29.19207126 31.37331418 16.81864588
21.00166977 34.76140175 17.82868067 19.86979042 39.99551768 36.4639378
29.61154128 31.93897225 26.52818682 28.97638498 17.14591889 24.65871423
24.44647306 21.68829773 36.55421745 29.56861486 25.14486402 27.11004852
```

Figure 3: Fill the missing values on “SkinThickness”

In Figure 3, the predicted values of the column “SkinThickness” are based on the input features (the columns “Glucose,” “Blood Pressure,” “Age,” and “BMI”). The reason for not using the column “Insulin” as the input feature for the prediction of the column “SkinThickness” is that “Insulin” has lots of missing values, and it affected the predicted result that some of the predictions would generate negative values. Therefore, the approach that uses the column without missing value as the input feature is appropriate for predicting the features with the missing values.

```
129.28906206 132.46957463 258.28813089 177.38795949  96.45367641
107.15082068 122.71325675 279.55373151 134.92541329 209.65666514
121.10756281 246.32067016  13.22145175 124.89023428 176.59423992
106.85151155  72.93094884 172.45820912  54.55041862 222.26910097
321.08993643 211.54268217  30.84999247 123.33134807 117.60320988
 42.93714842 199.38381574  75.21851706 173.06178001   6.83659169
151.89329651  89.01488034 105.79094149 150.0768613  156.06647487
 61.9506515  142.76358205  45.18103432  99.68135723 186.18962807
106.70816578 119.03904252  51.59559684  89.24212629 184.23700985
101.38520809 101.86383679  23.85613086 201.73075084  85.38790396
252.56373827 198.3217821   99.29282368  68.97246353  31.63368303
 50.29374465 174.52737105 154.23126618  54.90230679  90.93493162
136.70740769 231.59161195 140.19621264 159.82625942 168.45412278
 97.86367613 153.28600713 130.49762218 135.59523445 139.86673422
118.17481334   2.42763621 257.809247    76.68095348 105.54796577
297.71673455 213.42137506 185.7572556  102.43946261 146.66050641
214.20671588 140.23054657 125.74789808  95.02086852  71.04587977
 80.57392853 205.45587329 166.49840075  39.17457232  40.26280321
163.01778089 307.91277009  93.965793   227.17728473 171.91859873
117.44173793  94.33366586 134.24063272 199.95891134 112.00502463
 98.58502891 169.00736236 252.28393843  58.10900728 281.86303464
207.61137959  72.93967143 148.69878614 214.36851076 114.44932332
112.76663508 284.3913629  237.19556331 157.13287406 290.72316574
290.22405825 191.54970336  34.7480137   80.54515338 176.59868268
207.41042521 157.94271491 121.91153235 162.85246038 137.25906843
 76.10821894 144.67721479 131.4173834  124.47565891 181.48924286
 92.48172466 300.3060889  162.86511483 200.79543648 183.52239676
 95.66814381 176.23751998 119.13871471  96.38534722 164.99158789
107.16084564 162.44605529 218.42066005 181.05525179 123.83972803
374.24774876 167.41612078 243.44251523 156.92183283 122.22958173
 57.75922888 133.6581394  263.67500111 232.26658579 129.70813453
140.9271225  147.80923032 279.50484115 125.40389565 290.95477848
 81.47386103 151.79123822 135.36132178 258.08740016 166.49187986
142.37306203 171.91472952  99.66294411 162.810847    92.39089078
163.45552386  72.94705585  98.6608188  181.27839517 201.46589294
282.81529124 117.83945908 279.96754823 133.99689679  65.18542071
```

Figure 4: Fill the missing values on “Insulin”

```
No. of rows with Glucose value 0 is: 0
No. of rows with BloodPressure 0 is: 0
No. of rows with SkinThikness 0 is: 0
No. of rows with Insulin 0 is: 0
No. of rows with BMI 0 is: 0
No. of rows with DiabetesPedigreeFunction 0 is: 0
No. of rows with age 0 is: 0
```

Figure 5: Check if there have missing value (0) or not

Figure 4 uses columns “Glucose,” “Blood Pressure,” “Age,” “BMI,” and “SkinThickness” as the input features for the prediction of “Insulin,” which are the missing value. As shown in Figure 4, polynomial regression did not generate negative values after the input features were used. After the polynomial regression, replace all the predicted values with all the missing values. As shown in Figure 5, there is no

missing value in each column, which means all the predicted values are successfully replaced.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148.0	72.0	35.000000	220.061558	33.6	0.627	50	1
1	1	85.0	66.0	29.000000	53.940421	26.6	0.351	31	0
2	8	183.0	64.0	19.433105	229.999034	23.3	0.672	32	1
3	1	89.0	66.0	23.000000	94.000000	28.1	0.167	21	0
4	0	137.0	40.0	35.000000	168.000000	43.1	2.288	33	1
...	...	...	...	...	...	...	...	...	...
763	10	101.0	76.0	48.000000	180.000000	32.9	0.171	63	0
764	2	122.0	70.0	27.000000	163.653538	36.8	0.340	27	0
765	5	121.0	72.0	23.000000	112.000000	26.2	0.245	30	0
766	1	126.0	60.0	30.256833	199.507326	30.1	0.349	47	1
767	1	93.0	70.0	31.000000	85.680421	30.4	0.315	23	0

Figure 6: All the missing values are replaced by the predicted values

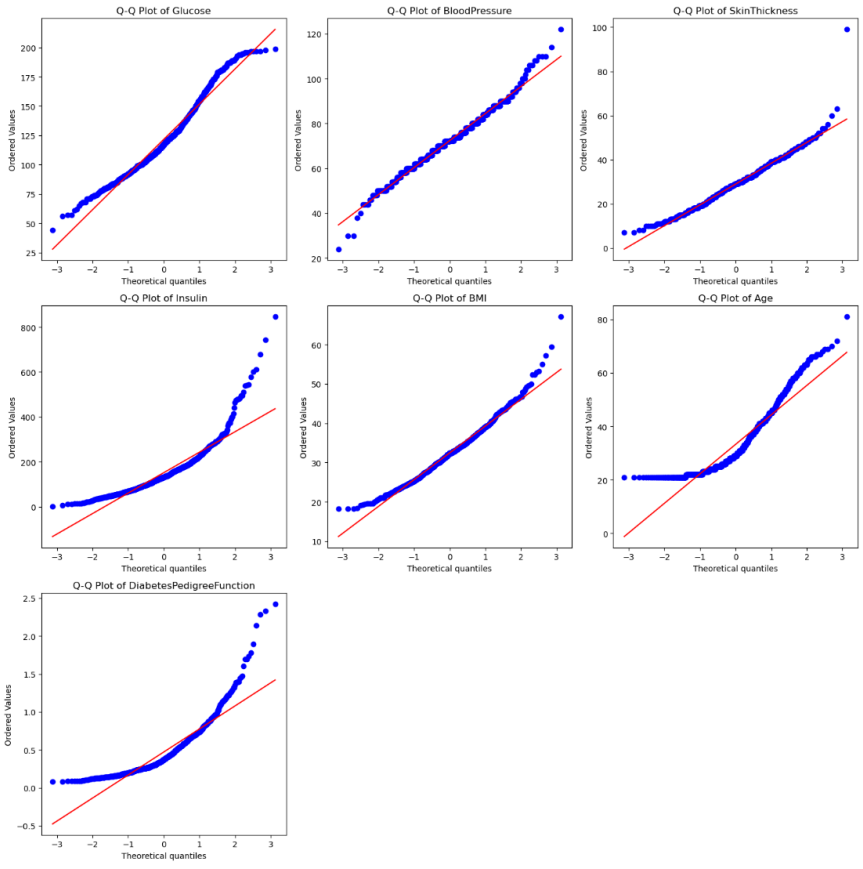

Figure 7: Quantile-Quantile Plot of all features (Pima Indian Diabetes Dataset)

As shown in Figure 7, there are the normal distribution of all features which are displayed in Q-Q plot form. Spots in features “Glucose,” “Blood Pressure,”

"SkinThickness," and "BMI" roughly follow the straight red line on the plot. Also, the spots on the features "Insulin", "Age," and "DiabetesPedigreeFunction" are deviated from the red straight line. For example, most of the spots in the feature "Age" are concentrated on 20, meaning most participants are about 21 years old.

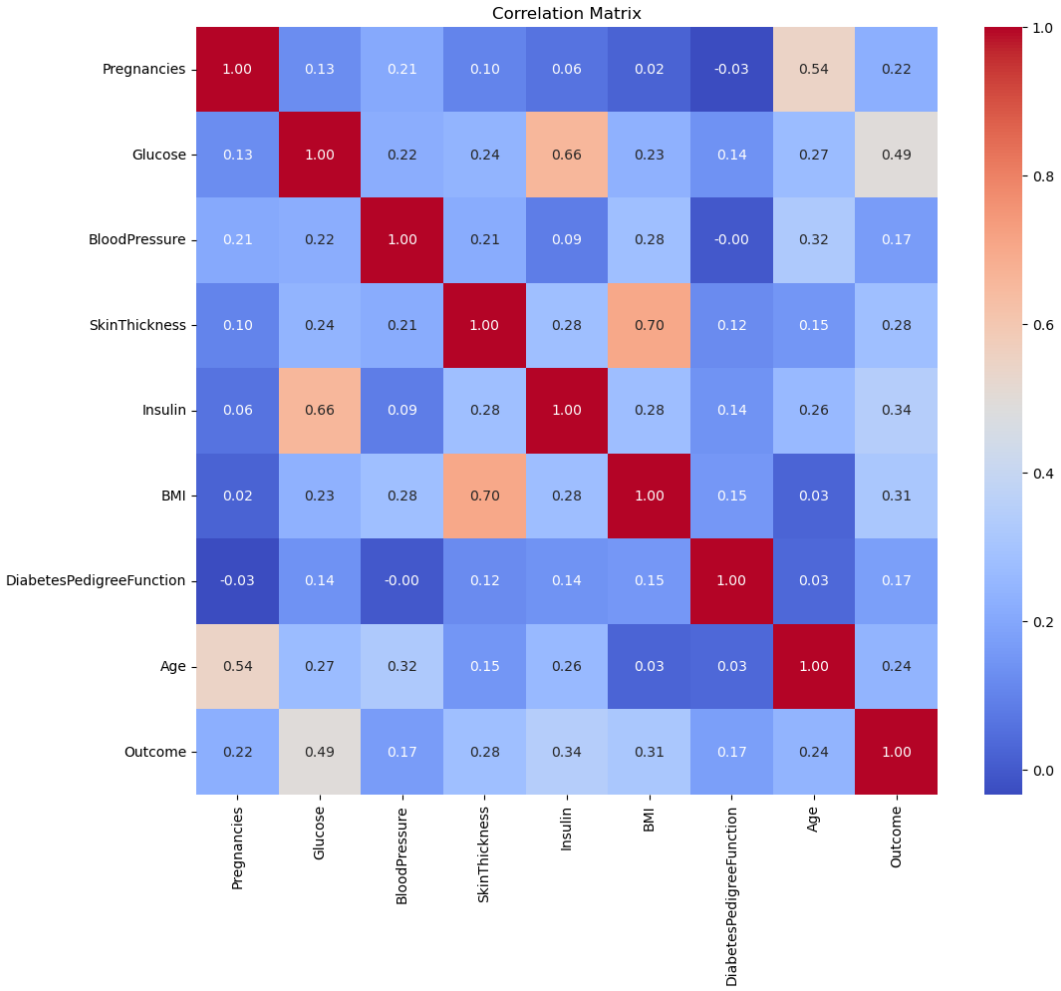

Figure 8: Correlation matrix of all features and outcome (Pima Indian Diabetes Dataset)

Figure 8 shows the correlation coefficient between all the features and the outcome. Based on the ranking in descending order, the relationship between the features and the outcome are "Glucose", "Insulin", "BMI", "SkinThickness", "Age", "Pregnancies" and "BloodPressure", which the highest score and lowest score are "Glucose" and "BloodPressure" respectively. Cleveland Clinic [10] reported that glucose level and diabetes are strongly correlated. According to ranking, features "Glucose", "Insulin", "BMI" and "SkinThickness" are selected as the key features for the prediction of diabetes.

After defining the key features and “Outcome” as a class, the data will split into 80:20, 80% for training, and the rest for testing. Next, feature scaling is applied to split data. Standardization is used in the feature scaling to normalize the data. In the standardization process, the scaler will be defined and fit into the training data. Then, the testing data will be transformed to finish the feature scaling.

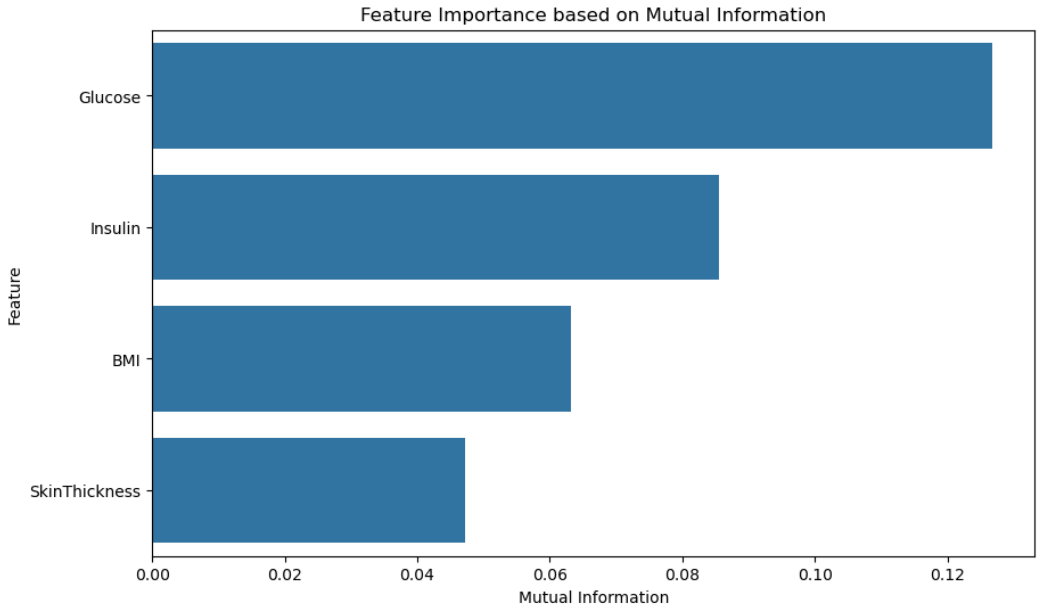

Figure 9: Feature Importance (Pima Indian Diabetes dataset)

```
Outcome
0    401
1    401
Name: count, dtype: int64
```

Figure 10: Balanced data (Pima Indian Diabetes Dataset)

As shown in Figure 9, “Glucose” has the highest rank in feature importance after standardization. Furthermore, the features rank the same as the correlation matrix. Then, SMOTE is applied to process imbalance data, which can significantly improve model performance, especially for weak learners [11]. In Figure 10, all the data are balanced at 50:50 after applying SMOTE.

3.2 Data Preprocessing on NHANES dataset:

Since the NHANES dataset has five raw data sets, it is necessary to preprocess the data to ensure that it is readable and understandable to people because there are many features in each raw data set. Originally, medication (not mentioned in Table 1) was one of the raw data of the NHANES dataset, but no complete description can be found on the NCHS official website. In addition, diet does not have similar features compared to Pima Indian Diabetes dataset. Therefore, medication and diet will not be used in data preprocessing.

	SEQN	SDDSRVYR	RIDSTATR	RIAGENDR	RIDAGEYR	RIDAGEMN	RIDRETH1	RIDRETH3	RIDEXMON	RIDEXAGM	...	DMDHREDU	DMDHRMAR
0	73557	8	2	1	69	NaN	4	4	1.0	NaN	...	3.0	4.0
1	73558	8	2	1	54	NaN	3	3	1.0	NaN	...	3.0	1.0
2	73559	8	2	1	72	NaN	3	3	2.0	NaN	...	4.0	1.0
3	73560	8	2	1	9	NaN	3	3	1.0	119.0	...	3.0	1.0
4	73561	8	2	2	73	NaN	3	3	1.0	NaN	...	5.0	1.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
10170	83727	8	2	1	26	NaN	2	2	2.0	NaN	...	3.0	NaN
10171	83728	8	2	2	2	24.0	1	1	2.0	24.0	...	3.0	1.0
10172	83729	8	2	2	42	NaN	4	4	2.0	NaN	...	5.0	3.0
10173	83730	8	2	1	7	NaN	2	2	1.0	84.0	...	4.0	1.0
10174	83731	8	2	1	11	NaN	5	6	1.0	140.0	...	5.0	1.0

Figure 11: Information about demographic (NHANES dataset)

In Figure 11, each feature is complex to read for people as they are named in terms. To make the data more readable and understandable features similar to the Pima Indian Diabetes dataset are selected and relabeled. For example, “SEQN”, “RIAGENDR” and “RIDAGEYR” are relabeled as “ID”, “Gender” and “Age” respectively. After the relabeling, all the data are merged on the "ID” and uses a left join. Then, category mapping is created for gender and split into Male and Female, and categorical variables are converted into dummy variables.

	ID	Age	BMI	BloodPressure	Glucose	Insulin	Outcome	Gender_Female	Gender_Male
0	73557	69	26.7	72.0	NaN	NaN	1.0	False	True
1	73558	54	28.6	62.0	NaN	NaN	1.0	False	True
2	73559	72	28.9	90.0	NaN	5.83	1.0	False	True
3	73560	9	17.1	38.0	NaN	NaN	2.0	False	True
4	73561	73	19.7	86.0	NaN	6.12	2.0	True	False
...	...	...	...	...	...	...	...	...	...
10170	83727	26	24.5	68.0	108.0	3.76	2.0	False	True
10171	83728	2	15.9	NaN	NaN	NaN	2.0	True	False
10172	83729	42	34.0	82.0	NaN	NaN	2.0	True	False
10173	83730	7	16.1	NaN	NaN	NaN	2.0	False	True
10174	83731	11	19.3	68.0	NaN	NaN	2.0	False	True

Figure 12: Merged dataset (NHANES dataset)

```
No. of rows with BMI 0 is: 1120
No. of rows with BloodPressure 0 is: 3086
No. of rows with Glucose value 0 is: 7830
No. of rows with Insulin 0 is: 7082
```

Figure 13: Total number of each row with value 0

In Figure 12, the merged dataset collects the key features and is renamed. Despite the conversation, many null values are still on the merged dataset. Therefore, fill the null values with zero in the entire dataset and apply polynomial regression to predict these values. In this situation, it is not suitable to use means or median to replace the null values as there are many null values in Figure 13, and this will negatively influence the prediction of diabetes if applied.

	ID	Age	BMI	BloodPressure	Glucose	Insulin	Outcome	Gender_Female	Gender_Male
0	73557	69	26.7	72.000000	140.142966	11.690010	1.0	False	True
1	73558	54	28.6	62.000000	120.565036	11.450715	1.0	False	True
2	73559	72	28.9	90.000000	137.576165	5.830000	1.0	False	True
3	73560	9	17.1	38.000000	91.461581	11.120771	2.0	False	True
4	73561	73	19.7	86.000000	137.916200	6.120000	2.0	True	False
...	...	...	...	...	...	...	...	...	...
10170	83727	26	24.5	68.000000	108.000000	3.760000	2.0	False	True
10171	83728	2	15.9	44.831670	95.956527	13.899938	2.0	True	False
10172	83729	42	34.0	82.000000	124.156978	16.658255	2.0	True	False
10173	83730	7	16.1	50.015263	94.608537	12.006523	2.0	False	True
10174	83731	11	19.3	68.000000	98.791896	11.602714	2.0	False	True

Figure 14: Merged dataset after Polynomial Regression applied (NHANES dataset)

After the polynomial regression applied, the predicted values replace the null values in the merged dataset and it looks understandable. In column “Outcome”, there are five values (1 for Yes, 2 for No, 3 for Borderline, 7 for Refused, and 9 for Do not know). Values 1 and 2 are retained, and the rest are removed because of their unclear result. Furthermore, the proportion of “Outcome” larger than two in the dataset is 2%; consequently, removing these will not significantly impact the prediction of diabetes.

In addition, it is counterintuitive to use values 1 and 2 to represent having diabetes and not having diabetes, especially since the value 2 will mistake people for invalid. Accordingly, it is necessary to use 1 and 0 to represent having diabetes and no diabetes instead of 1 and 2.

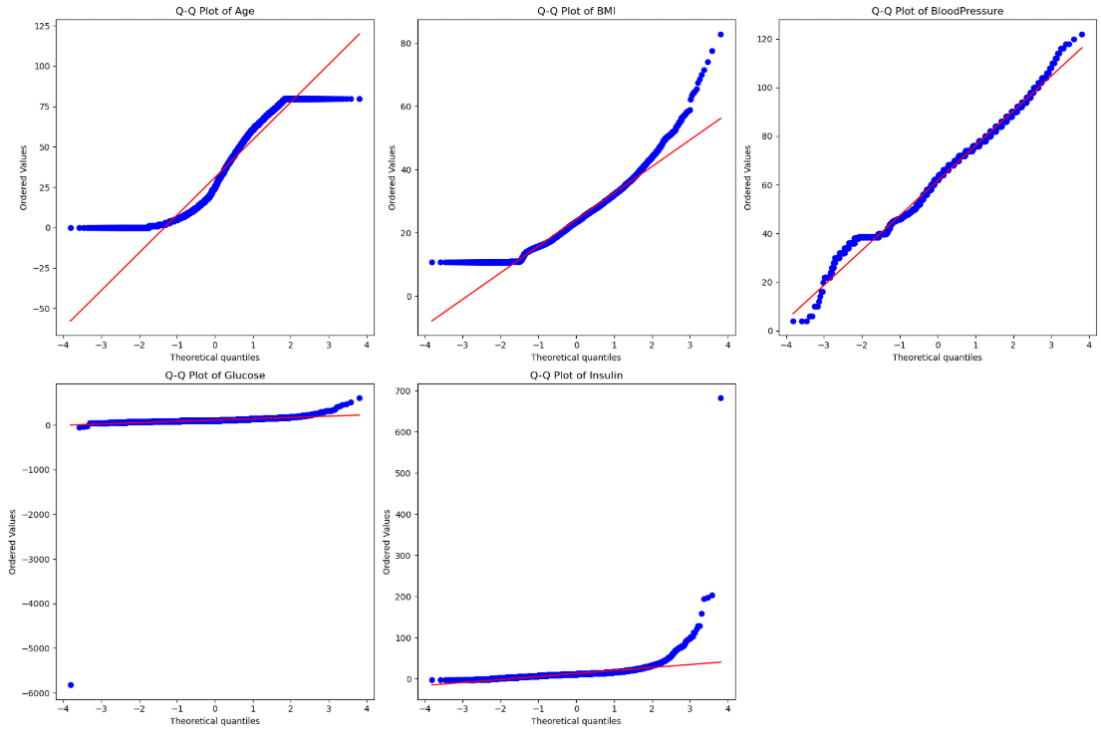

Figure 15: Quantile-Quantile Plot of all features (NHANES dataset)

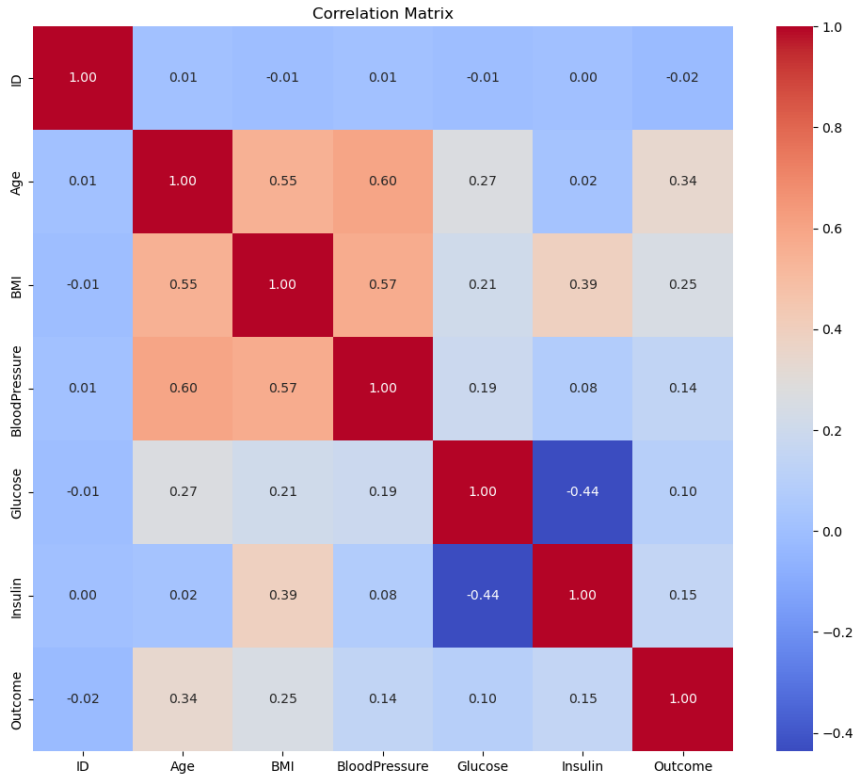

Figure 16: Correlation matrix of all features and outcome (NHANES dataset)

In Figure 15, “Blood pressure,” “Glucose,” and “Insulin” are the normal distribution, and most of the spots on these features follow the red straight line. Additionally, “Age” looks like an inverse z shape. Most spots deviate from two ranges and can be considered a bimodal distribution.

In Figure 16, put the relationship between features and “Outcome” in descending order are "Age", "BMI", "Insulin", "BloodPressure", "Glucose" and "ID". “ID” will not be used as the input feature because it is a negative for “Outcome”. In addition, “ID” is used to assign the numeric labels to each entry, which means it does not have a meaningful relationship with “Outcome”. Moreover, the result of the correlation matrix is different from the Pima Indian Diabetes dataset in that “Age” is the highest score in the NHANE dataset, but "Glucose" is the highest score in the Pima Indian Diabetes dataset.

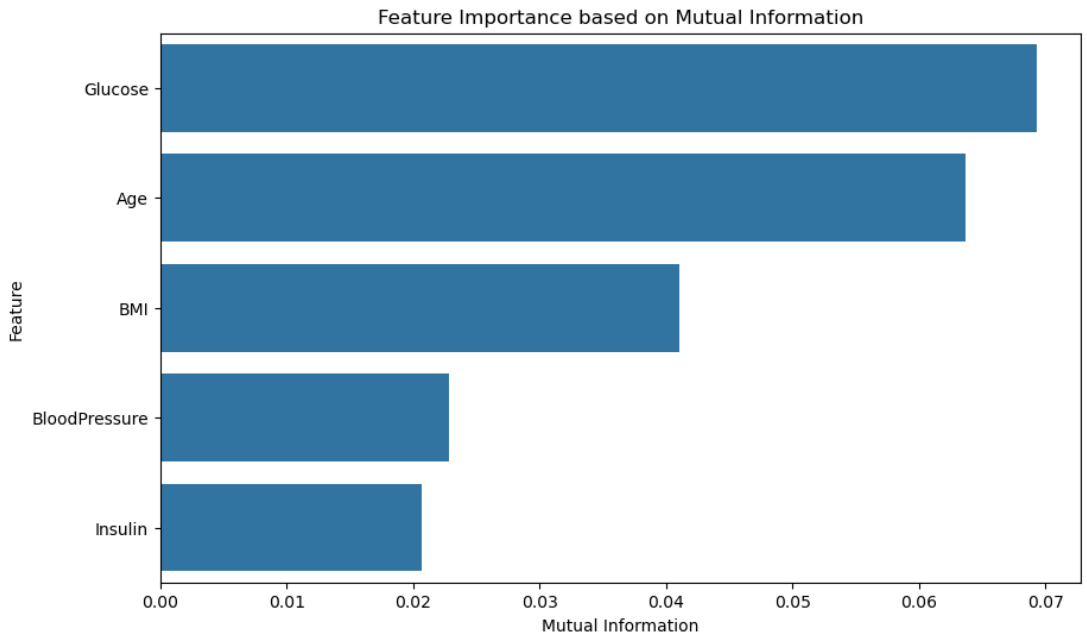

Figure 17: Feature Importance (NHANES dataset)

In Figure 17, “Glucose” has the highest feature importance after standardization, the same rank as the Pima Indian Diabetes dataset. Also, it is worthy to note that “Insulin” is at the bottom of the feature importance. In the preprocessing, “Glucose”, “Age” , “BMI”, “BloodPressure” and “Insulin” are selected as input features in order to make a similar environment to make the comparison.

3.3 Preliminary Result

```
[[95 23]
 [22 36]]
Accuracy Score 0.7443181818181818
              precision    recall  f1-score   support

           0       0.81      0.81      0.81       118
           1       0.61      0.62      0.62        58

    accuracy                           0.74       176
   macro avg       0.71      0.71      0.71       176
weighted avg       0.75      0.74      0.74       176
```

Figure 18: Reproduce Result (XGB+ADASYN)

```
[[98 20]
 [15 43]]
Accuracy Score 0.8011363636363636
              precision    recall  f1-score   support

           0       0.87      0.83      0.85       118
           1       0.68      0.74      0.71        58

    accuracy                           0.80       176
   macro avg       0.77      0.79      0.78       176
weighted avg       0.81      0.80      0.80       176
```

Figure 19: Paper Result (XGB+ADASYN)

```
roc_auc_score for ZeroR Classifier:  0.5
roc_auc_score for Bagging DecisionTree:  0.842270601987142
roc_auc_score for KNN Classifier:  0.8194038573933372
roc_auc_score for SVM Classifier:  0.8288281706604326
roc_auc_score for Random Forest Classifier:  0.836572180011689
roc_auc_score for Naive Bays Classifier:  0.8055230859146698
roc_auc_score for Ada Boost Classifier:  0.83781414377557
roc_auc_score for XG Boost Classifier:  0.8139976621858562
roc_auc_score for Logistic Regression:  0.8381063705435419
roc_auc_score for Voting Classifier:  0.8438047925189948
roc_auc_score for DecisionTree:  0.7800263004091175
```

Figure 20: Reproduce Result (AUC)

```
roc_auc_score for ZeroR Classifier:  0.5
roc_auc_score for Bagging DecisionTree:  0.842270601987142
roc_auc_score for KNN Classifier:  0.8194038573933372
roc_auc_score for SVM Classifier:  0.8289012273524253
roc_auc_score for Random Forest Classifier:  0.8389830508474576
roc_auc_score for Naive Bays Classifier:  0.8055230859146698
roc_auc_score for Ada Boost Classifier:  0.8373758036236119
roc_auc_score for XG Boost Classifier:  0.8448275862068966
roc_auc_score for Logistic Regression:  0.8397136177673875
roc_auc_score for Voting Classifier:  0.842343658679135
roc_auc_score for DecisionTree:  0.7800263004091175
```

Figure 21: Paper Result (AUC)

Despite the same dataset and environment, it is found that the results of the XG Boost Classifier are different, as shown in the figures above.

```
[[81 19]
 [19 35]]
Accuracy Score 0.7532467532467533
              precision    recall  f1-score   support

           0       0.81      0.81      0.81       100
           1       0.65      0.65      0.65        54

    accuracy                           0.75       154
   macro avg       0.73      0.73      0.73       154
weighted avg       0.75      0.75      0.75       154
```

Figure 22: Preliminary Result (XG Boost) (Pima Indian Diabetes dataset)

```
[[83 17]
 [17 37]]
Accuracy Score 0.7792207792207793
              precision    recall  f1-score   support

           0       0.83      0.83      0.83       100
           1       0.69      0.69      0.69        54

    accuracy                           0.78       154
   macro avg       0.76      0.76      0.76       154
weighted avg       0.78      0.78      0.78       154
```

Figure 23: Preliminary Result (Random Forest) (Pima Indian Diabetes dataset)`

```
[[1655  210]
 [  67   65]]
Accuracy Score 0.8612919379068603
              precision    recall  f1-score   support

         0.0       0.96      0.89      0.92      1865
         1.0       0.24      0.49      0.32       132

    accuracy                           0.86      1997
   macro avg       0.60      0.69      0.62      1997
weighted avg       0.91      0.86      0.88      1997
```

Figure 24: Preliminary Result (XG Boost) (NHANES dataset)

```
[[1543  322]
 [  34   98]]
Accuracy Score 0.8217325988983475
              precision    recall  f1-score   support

         0.0       0.98      0.83      0.90      1865
         1.0       0.23      0.74      0.36       132

    accuracy                           0.82      1997
   macro avg       0.61      0.78      0.63      1997
weighted avg       0.93      0.82      0.86      1997
```

Figure 25: Preliminary Result (Random Forest) (NHANES dataset)

Since the referenced paper proposed a machine learning framework that used XG Boost with ADASYN and other machine learning models like random forest for the comparison in the paper, XG Boost and Random Forest are used to predict diabetes in this project. For the Pima Indian Diabetes dataset, Random Forest obtains 77.9% accuracy, and XG Boost is the highest. It has 86.1% accuracy for the NHANES dataset. Adjustments have been implemented in the Pima Indian Diabetes dataset, such as the input features change to “Glucose”, “Insulin”, “BMI”, “BloodPressure” and “Age”, which are the same as the input features in the NHANES dataset. The reason for the different results in the two datasets can be the number of entries in the NHANES dataset, which is around ten thousand.

4. CHALLENGES

Data preprocessing for the NHANES dataset is challenging as it includes five different data sources. Also, each data has lots of features whether it is useful for diabetes prediction or not. Therefore, I need to read the description of the features to ensure that I can remove the redundancy and get the key features from each data and merge them into a new data frame with filtered.

Moreover, there are many missing values with 0 and null values on the Pima Indian Diabetes dataset and NHANES dataset. For example, a number of rows with “LBXGLT” and “LBXIN” (value 0) have 7830 and 7082, respectively, which comprises approximately 77% and 70% of the whole NHANES dataset. The approach of filling value zero with the means is used to predict the missing value but it performs poorly in the q-q plot as both are not the normal distribution. Therefore, polynomial regression approach is used in this situation to predict the missing value and it is better than filling value zero with the means.

5. FUTURE WORKS

1. Try to use other preprocessing methods

Standardization and SMOTE are applied as the data preprocessing approaches in this stage. For feature scaling, methods like min-max scaling and mean normalization will be the options to transform the data to fit within a specific range or scale. For the class imbalance in datasets, techniques like ADASYN will be the possible choice to ensure that the model will not lead to poor performance in the minority class.

2. Try to select other key features for comparison

In this stage, “Glucose”, “Insulin”, “BMI”, “SkinThickness” are the selected features in the Pima Indian Diabetes dataset. The criteria for selecting the input features are based on the correlation matrix, which is lower than 0.2, to ensure that no more redundancy features as the input features affect the model performance. For the NHANES dataset, “Glucose”, “Insulin”, “BMI”, “Age” and “BloodPressure” are the input features in the dataset. In the future, it is possible to explore more features as input features and make comparisons to review which performs better under the same preprocessing method.

References

[1] World Health Organization, Diabetes. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/diabetes [Accessed Dec. 09, 2024].

[2] Smart Patient, Diabetes Mellitus. [Online]. Available: https://www.smartpatient.ha.org.hk/en/smart-patient-web/disease-management/disease-information/disease/DiabetesMellitus [Accessed Dec. 09, 2024].

[3] A. Mujumdar and V. Vaidehi, "Diabetes Prediction using Machine Learning Algorithms," in 2ND INTERNATIONAL CONFERENCE ON RECENT TRENDS IN ADVANCED COMPUTING ICRTAC -DISRUP - TIV INNOVATION , 2019, AMSTERDAM: Elsevier B.V, 2019, pp. 292–299. doi: 10.1016/j.procs.2020.01.047 [Accessed Dec. 09, 2024]

[4] I. Tasin, T. U. Nabil, S. Islam, and R. Khan, "Diabetes prediction using machine learning and explainable AI techniques," *Healthcare technology letters*, vol. 10, no. 1–2, pp. 1–10, 2023, doi: 10.1049/htl2.12039 [Accessed Dec. 09, 2024]

[5] UCI Machine Learning and Kaggle Team**,** *Pima Indians Diabetes Database,* 2016**.** [Online]**.** Available**:** https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database [Accessed Dec. 09, 2024]**.**

[6] Healthline **,** *Everything You Need to Know About Glucose***,** 2024**.** [Online]**.** Available**:** https://www.healthline.com/health/glucose [Accessed Dec. 09, 2024]**.**

[7] E. Eyth, H. Basit and C.J. Swift, "Glucose Tolerance Test, "in StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing, 2024. Available: https://www.ncbi.nlm.nih.gov/books/NBK532915/#__NBK532915_dtls__ [Accessed Dec. 09, 2024]

[8] Cleveland Clinic, Blood Pressure. [Online]. Available: https://my.clevelandclinic.org/health/diagnostics/17649-blood-pressure [Accessed Dec. 09, 2024].

[9] K. Aditya Shastry *et al.*, "Regression Based Data Pre-processing Technique for Predicting Missing Values," in *Emerging Research in Computing, Information, Communication and Applications*, Singapore: Springer Singapore Pte. Limited, 2021, pp. 95–102. doi: 10.1007/978-981-16-1338-8_9

[10] Cleveland Clinic, Blood Glucose (Sugar) Test. [Online]. Available: https://my.clevelandclinic.org/health/diagnostics/12363-blood-glucose-test [Accessed Dec. 11, 2024].

[11] Train In Data, SMOTE in Python: A guide to balanced datasets. [Online]. Available: https://www.blog.traininindata.com/smote-in-python-a-guide-to-balanced-datasets/ [Accessed Dec. 11, 2024].

Appendices

Appendix 1: Pima Indian Diabetes dataset

https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

Appendix 2: 2013-2014 NHANES dataset

https://www.kaggle.com/datasets/cdc/national-health-and-nutrition-examination-survey/data?select=diet.csv