**Project ID: [FYP\_ lpchau\_20240523150449]**

**Machine learning model to predict the risk of diabetes**

**by**

**CHAN Hou Ting Constant**

**21034774D**

**Final Year Project - Final Report 2024/2025 Sem 2**

**Bachelor of Science (Honours)**

In

**Internet and Multimedia Technologies**

**of**

**The Hong Kong Polytechnic University**

Supervisor: Prof CHAU Lap Pui Date:3-April-2025

Abstract

Diabetes has become a significant health problem in the world, which has been increasingly prevalent across the world for the past several years. Machine learning is one of the methods to predict Diabetes. This final year project studied two datasets, the Pima Indian Diabetes dataset and the 2013-2014 NHANES dataset, which will be used in the project. In addition, the machine learning framework proposed by Tasin et al. [23] and Qin et al. [26] are the baselines for the reference. The preprocessing methods Polynomial Regression and the imputation of missing values by mean are used to fill the missing values. Standardization is also applied to normalize the input features. In the experiment, XGBoost, CatBoost, and Random Forest are applied in the in experiment, in which XGBoost achieved 82.46% accuracy with an AUC of 0.86 and 93.94% accuracy with an AUC of 0.91 in the Pima Indian Diabetes dataset and the 2013-2014 NHANES dataset respectively.

Contents

[Abstract i](#_Toc194537820)

[Contents ii](#_Toc194537821)

[1. Introduction 3](#_Toc194537822)

[2. Related Work 4](#_Toc194537823)

[3. Methodology 5](#_Toc194537824)

[3.1 Random Forest 5](#_Toc194537825)

[3.2 XGBoost 5](#_Toc194537826)

[3.3 CatBoost 5](#_Toc194537827)

[3.4 Proposed Framework 6](#_Toc194537828)

[4. Dataset Study 8](#_Toc194537829)

[4.1 Pima Indian Diabetes Dataset 8](#_Toc194537830)

[4.2 2013-2014 NHANES Dataset 11](#_Toc194537831)

[5. Experiment 14](#_Toc194537832)

[5.1 Data Preprocessing on Pima Indian Diabetes dataset 14](#_Toc194537833)

[5.2 Data Preprocessing on 2013-2014 NHANES dataset 20](#_Toc194537834)

[6. Result 24](#_Toc194537835)

[6.1 Results on Pima Indian Diabetes dataset 24](#_Toc194537836)

[6.2 Results on 2013-2014 NHANES dataset 29](#_Toc194537837)

[7. Discussion 33](#_Toc194537838)

[8. Conclusion 35](#_Toc194537839)

[References 36](#_Toc194537840)

List of Figures

Figure 1. Diabetes around the world in 2021 & prediction in 2030 & 20453

Figure 2. Proposed Framework 6

Figure 3. Information of Pima Indian Diabetes dataset14

Figure 4. Missing value count on the Pima Indian Diabetes dataset14

Figure 5. Correlation Matrix of Pima Indian Diabetes dataset 15

Figure 6. The predicted values on “SkinThickness” 16

Figure 7. Pima Indian Diabetes dataset (Predicted value filled)17

Figure 8. Q-Q Plot of each feature in the Pima Indian Diabetes dataset (after the

Polynomial Regression)18

Figure 9. Information of Raw data Demographic 20

Figure 10. Information of Merged data (Not preprocessed) 21

Figure 11. Merged data (Processed)21

Figure 12. Correlation Matrix of the 2013-2014 NHANES dataset22

Figure 13. ROC Curve of the Pima Indian Diabetes dataset28

Figure 14. ROC Curve of the 2013-2014 NHANES dataset32

List of Tables

Table 1. Abstract of Pima Indian Diabetes Dataset 8

Table 2. Abstract of NHANES Dataset 11

Table 3. Settings on Polynomial features 15

Table 4. Feature(s) selected to predict “SkinThickness” 15

Table 5. Feature(s) selected to predict “Insulin” 15

Table 6. The combination of key features in the Pima Indian Diabetes dataset 19

Table 7. Abstract of Selected Raw data and Features 20

Table 8. The combination of key features in the 2013-2014 NHANES dataset 23

Table 9. Hyperparameter tuning on XGBoost – scale\_pos\_weight (Pima Indian

Diabetes dataset) 24

Table 10. Hyperparameter tuning on XGBoost – max\_depth (Pima Indian Diabetes

dataset) 24

Table 11. Hyperparameter tuning on XGBoost – learning\_rate (Pima Indian Diabetes

dataset) 25

Table 12. A comparison between XGBoost in different preprocessing techniques

(Pima Indian Diabetes dataset) 26

Table 13. Abstract of Baseline VS Proposed framework (Pima Indian Diabetes

dataset) 27

Table 14. Hyperparameter tuning on CatBoost (2013-2014 NHANES dataset) 29

Table 15. A comparison between CatBoost in different preprocessing techniques

(2013-2014 NHANES dataset) 30

Table 16. Baseline VS Proposed framework (2013-2014 NHANES dataset) 31

Table 17. A comparison between XGBoost in different combinations of key

features (Pima Indian Diabetes dataset) 33

Table 18. A comparison between XGBoost in different combinations of key

features (2013-2014 NHANES dataset) 33

# Introduction

Diabetes is a chronic disease in which the human body cannot utilize insulin generated by the body or produce enough insulin [16]. In other words, it affects the body's function and causes the body function to not work properly because of disordered metabolism.

一張含有 文字, 地圖, 地圖集 的圖片

自動產生的描述

Figure 1: Diabetes around the world in 2021 & prediction in 2030 & 2045 [17]

Nowadays, Diabetes becomes a significant health issue in the world, which has been increasingly prevalent across the world for the past several years. In Fig 1, it shows the prediction that the number of patients with Diabetes will increase in 2030 and 2045, of which Africa will increase to 134 percent. In addition, the World Health Organization [18] estimated that 830 million people exhibit symptoms of Diabetes, which comprises approximately 10% of the world's population. In addition, IDF Diabetes Atlas [19] found that 537 million adults were having Diabetes and 6.7 million people died because of Diabetes in 2021.

In view of the growing popularity of technologies, machine learning has become one of the approaches used to predict if a person is likely to develop Diabetes. The main advantage of machine learning is that it can generate the corresponding prediction based on datasets and preprocessing, which enables experts to make decisions based on the prediction. Two famous Diabetes datasets were used for the research: the Pima Indian Diabetes dataset [20] and the NHANES dataset [21]. All these datasets come from people who truly lived in the world.

# Related Work

Before the widespread of Gradient Boosting, there were studies based on traditional machine learning algorithms to do the classification task. For example, Abdillah and Suwarno [22] used a Support Vector Machine to predict Diabetes in 2016.

This final year project is based on two research papers that used the Pima Indian Diabetes dataset and the 2013-2014 NHANES dataset, respectively. They proposed two different approaches to the classification.

Tasin et al. [23] proposed a method based on eXtreme Gradient Boosting (XGBoost) on the Pima Indian Diabetes dataset. XGBoost is a robust machine-learning algorithm good at classification and regression tasks. In the preprocessing, they used ADASYN to process the class imbalance problem. In addition, they merged the extra dataset called the RTML private dataset with the Pima Indian Diabetes dataset. They replaced the missing values of features besides Insulin and Age with the mean value. Moreover, they used Polynomial Regression, a form of regression analysis, to model the relationship between input variables (x) and output variables (y) by fitting a polynomial equation to the data [24] to predict the missing value of Insulin. It is noted that some research on Diabetes prediction had also used Polynomial Regression to predict the missing values and brought gains to the experiment results. For example, Aditya Shastry et al. [25] applied polynomial regression to predict the missing value of the features in data preprocessing, which improved model performance. This method gave them 81% accuracy and an AUC of .84.

Qin et al. [26] proposed a method based on CatBoost on the 2013-2014 NHANES dataset. In the preprocessing, they used ADASYN and One-Hot encoding to solve the class imbalance and categorical features, respectively. Also, they compared five machine learning models, which were XGBoost, CatBoost, Support Vector Machine. Random Forest and Logistic Regression, of which CatBoost performs well with 82.1% accuracy and an AUC of 0.83.

# Methodology

## Random Forest

Random Forest is an ensemble machine learning algorithm combining multiple Decision trees for classification and regression tasks. Compared with the Decision Tree algorithm, Random Forest can provide better performance that improves the model's accuracy, especially for complex datasets. For example, Chen et al. [27] commented that Random Forest was the most effective algorithm compared to Naïve Bayes Tree and Alternating Decision Tree, which provided higher accuracy and can handle complex datasets like historical flood data and geographical factors.

## XGBoost

XGBoost is an ensemble machine learning algorithm based on the gradient boosting framework and decision tree to process the tasks like classification and regression [28]. In other words, it is an optimized gradient boosting framework and is widely used in competitions like Kaggle. The main feature of XGBoost is that it provides higher performance and speed to handle large datasets because of parallel processing. Also, it can learn and handle the missing values during the training to save time for preprocessing, which is suitable for situations in which datasets have many missing values.

## CatBoost

CatBoost is a machine learning algorithm based on a gradient boosting framework and decision tree, similar to XGBoost. Compared with XGBoost, CatBoost is better at handling categorical features because it uses ordered encoding to encode categorical variables [29], which can handle overfitting and improve the model performance. In addition, Hancock and Khoshgoftaar [30] commented that CatBoost is good at processing categorical and heterogeneous data. Therefore, it is suitable for datasets with categorical features.

## Proposed Framework

Figure 2: Proposed Framework

As shown in Fig. 2, the Pima Indian Diabetes dataset and the 2013-2014 NHANES dataset are the target datasets in the proposed framework. Before splitting the data for training and testing, it is necessary to preprocess the datasets because there are raw data in the datasets that have not been processed. Also, most of the datasets have null values or missing values, which negatively affect the overall result, such as generating abnormal model performance, leading to biased parameter estimates and incorrect conclusions on the experiment result [31]. Therefore, filling in the missing values is important to prevent the situations above.

In “Missing Value”, polynomial regression and the imputation of missing values by mean are used to fill the missing values in the Pima Indian Diabetes dataset and the 2013-2014 NHANES dataset. In “Feature Importance,” a correlation matrix is created to view the relationship between features and output and select the key features as input.

In “Split data”, the data will be split into an 80:20 ratio, in which 80 is the training set, and 20 is the testing set. In “Feature Scaling,” Feature Scaling can ensure fair treatment of all features [32], in which Standardization will be applied in normalizing the input features. After the training and testing in “Train” and “Test,” the evaluation will be conducted to review the results and make adjustments if the results do not meet the expectation.

# Dataset Study

## Pima Indian Diabetes Dataset

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

Table 1: Abstract of Pima Indian Diabetes Dataset

Nine variables are provided in the Pima Indian Diabetes Dataset. They are Pregnancies, Glucose, Blood Pressure, Insulin, BMI, DiabetesPedigreeFunction, Age and Outcome. In the experiment, Outcome will be defined as a target variable because its variable indicates a person who is diagnosed with having Diabetes or not directly (“1” means having Diabetes and “0” means not having Diabetes), and the rest are classified into Features because they are the risk factors of having Diabetes or not. Here is the information about each feature below:

Pregnancies: it means the number of times a female is pregnant or not, and the median value of Pregnancies is 3. There was the study explored the relationship between pregnancy and Diabetes. Buchanan et al. [4] commented that there is a higher risk of having Diabetes if a female has Gestational diabetes mellitus (GDM) during pregnancy.

Glucose (Blood Sugar): it is a group of carbohydrates [5] that provides energy for the human body that is important to body function. Eyth et al. [6] commented that it is considered a moderate level if the glucose level is lower than 140 mg/dL in the oral glucose tolerance test. The median value of Glucose is 117, which means that the typical glucose level for the participants in the dataset is relatively moderate. Generally, Glucose is directly related to Diabetes as the body cannot regulate blood glucose levels generally if a person is diagnosed with having Diabetes. For example, the Mayo Clinic [7] noted that Diabetes can cause excessive Glucose in the blood, which cannot be properly regulated and results in serious health issues.

BloodPressure: it means the heart beats and pumps blood into the arteries [8] that deliver oxygen and nutrients into the body. It is considered moderate if systolic and diastolic pressure are less than 120 and 80 [9]. In the dataset, BloodPressure used diastolic blood pressure as the unit to measure blood pressure. The median value of BloodPressure is 72, which means that the typical blood pressure level for the participants in the dataset is relatively normal. Several organizations explored the relationship between blood pressure and Diabetes. For example, the probability of getting high pressure is higher than a person without Diabetes if a person has Diabetes [10].

SkinThickness: In general, it measures body fat on thighs and limbs. In the dataset, triceps (the back of the upper arm) was used to estimate the body fat. The average triceps skin thickness of human and women are 18.7 ± 8.5 mm and 23.6 ± 7.5 mm respectively [11]. It is noteworthy that all the participants in the dataset are females and the median value of SkinThickness before preprocessing is 23 mm, which is similar to the measurement on triceps skin thickness above. In addition, Särnblad et al [12] found that the skin thickness with a person with Diabetes is higher than a health person, even they have a similar BMI.

Insulin: it regulates blood sugar levels and is important for energy production and storage. The average insulin level (2-hour Serum insulin (µh/ml)) of a human is between 16 and 166 [13], and the median value of Insulin before preprocessing is 30.5. Moreover, Insulin is related to glucose and Diabetes. Better Health Channel [14] states that Diabetes will cause the body not to have enough Insulin and result in high blood glucose levels.

BMI: it measures body fat based on Height and Weight. The median BMI in the dataset is 32, which is considered obesity, and 18.5 to 24.9 is also considered a healthy body [15].

DiabetesPedigreeFunction: it is a function that scores the probability of Diabetes based on Family history. Smith et al. [16] defined that the value of DiabetesPedigreeFunction would be higher if more people had Diabetes in the family and vice versa.

Age: the age of all patients is at least 21 years old, and the median age is 29.

Outcome: it is a target variable that diagnosed Diabetes or not.

## 2013-2014 NHANES Dataset

The National Health and Nutrition Examination Survey (NHANES)

|  |  |
| --- | --- |
| NHANES Dataset (Raw data) | **Features (selected)** |
| **Demographic** | SEQN (ID of participant)  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

Five raw data sets are in the 2013-2014 NHANES Dataset: Demographics, Diet, Examination, Labs and Questionnaire.

Demographic is the characteristics of the participants, which describe personal information and social status. For example, it collects income, education, and marital status. Generally, these variables are used to analyze the relationship between health and socioeconomic factors to help researchers understand the health disparity between populations. Three variables in demographics are captured as the selected features in data preprocessing, which are SEQN (ID of the participant), RIAGENDR (Gender of the participant), and RIDAGEYR (Age of the participant). The reason that these 3 variables are the selected features in the experiment is that they are similar to the features provided in the Pima Indian Diabetes Dataset to make a comparison of two different datasets (Pima Indian Diabetes Dataset and 2013-2014 NHANES Dataset)' model performances. This approach will be applied to other raw data sets in the following experiment to make the results more comparable.

Diet collects the participants’ intake of food and drink. The purpose of gathering these attributes is to evaluate the nutrition and dietary habits of the participants to understand the relationship between Diet and physical disease (e.g., Obesity, Diabetes, Cardiovascular disease). Two variables in Diet are captured as the selected features in the data preprocessing, which are DR1DAY (Intake day of the week) and DR1TKCAL (Energy (kcal) taken in 1 day). These two variables are the selected features in the experiment because calorie intake is a key factor in preventing and managing type 2 diabetes [1].

Examination is the physical and physiological measurements. The purpose of collecting these measurements is to reflect the health data about the participants and enable the researchers to explore the relationship between the measurements and other NHANES raw data sets (e.g., Diet and Labs). In the experiment, BMXBMI (BMI) and BPXDI1 (Blood Pressure) are the selected features in the data preprocessing, as these selected features are the same as those provided in the Pima Indian Diabetes dataset.

Labs means the laboratory tests performed on biological samples collected from the participants. Collecting these biological samples aims to provide objective and quantifiable measures of health and nutrition to enable the researchers to understand participants’ nutritional status and health conditions. LBXGLT (Glucose) and LBXIN (Insulin) were selected features in the experiment. They are used in data preprocessing since these selected features are the same as those provided in the Pima Indian Diabetes dataset.

Questionnaire collects information from a series of interviews and questionnaires. The questions are related to family or individual-level information [2]. The questionnaire aims to collect information (e.g. health behaviors, medical history, and lifestyle choices) from participants to provide a critical context for understanding health outcomes and risk factors. In addition, it helps researchers to study risk factors for diseases such as Diabetes and Hypertension. In the experiment, DIQ010 (Diabetes Diagnosis) and ALQ120Q (alcoholic drinks taken per day/ months) will be taken as the selected features in the data preprocessing as DIQ010 indicates whether a participant is diagnosed with Diabetes or not (which is similar to "Outcome" in Pima Indian Diabetes dataset). Also, there was a study that explored the relationship between alcohol consumption and Diabetes. For example, Koppes et al. [3] commented that moderate alcohol consumption could reduce the risk of getting type 2 Diabetes.

# Experiment

## Data Preprocessing on Pima Indian Diabetes dataset

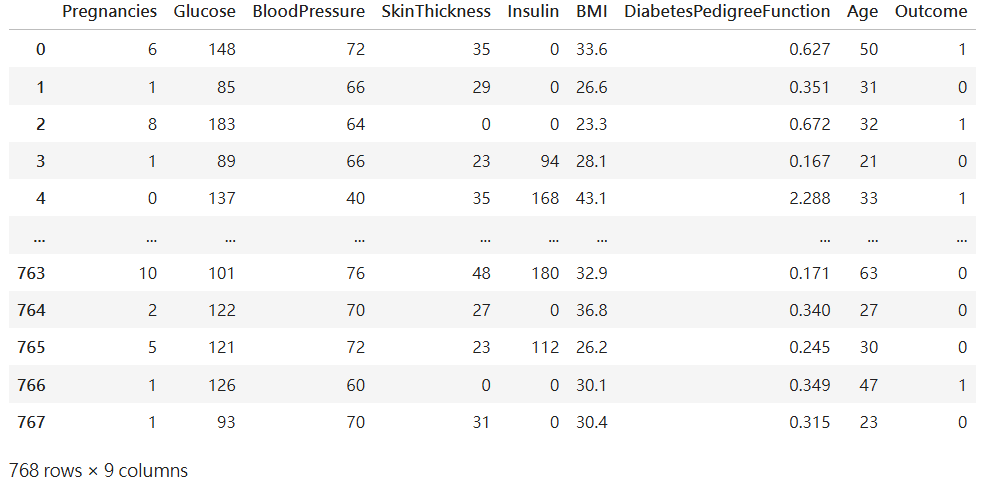


Figure 3: Information of Pima Indian Diabetes dataset

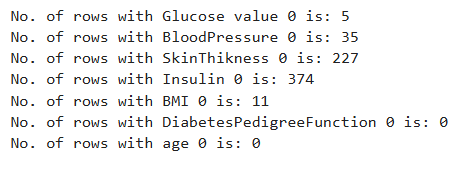


Figure 4: Missing value count on the Pima Indian Diabetes dataset

In Fig. 3, 768 entries were collected in the dataset, including eight features and the target variable. In addition, some features like “Glucose” and “BloodPressure” exist the missing values in Fig.4, in which “SkinThickness” and “Insulin” have 227 (approximately 30% of total entries in the dataset) and 374 (approximately 49% of total entries in the dataset) missing value respectively. There is no null value on the dataset. To fill in the missing values, the features “Glucose”, “Blood Pressure,” and “BMI” will be filled by their mean value as they are a minority of missing values in the dataset. For the features “SkinThickness” and “Insulin,” Polynomial Regression is applied to predict the missing values, and the predicted values will replace them.

|  |  |
| --- | --- |
|  | **Parameter settings** |
| degree | 2 |
| include\_bias | False |

Table 3: Settings on Polynomial features

|  |  |  |  |
| --- | --- | --- | --- |
| Feature(s) selected to predict “SkinThickness” | | | |
| Glucose | BMI | Age | BloodPressure |

Table 4: Feature(s) selected to predict “SkinThickness”

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Feature(s) selected to predict “Insulin” | | | | |
| Glucose | BMI | Age | BloodPressure | SkinThickness |

Table 5: Feature(s) selected to predict “Insulin”

一張含有 文字, 螢幕擷取畫面, 正方形, Rectangle 的圖片

AI 產生的內容可能不正確。

Figure 5: Correlation Matrix of Pima Indian Diabetes dataset

一張含有 文字, 紙張, 文件 的圖片

自動產生的描述

Figure 6: The predicted values on “SkinThickness”

Before the Polynomial Regression, it is necessary to identify the features for predicting “SkinThickness” and “Insulin”. In Table 4 and Table 5, the common features are “Glucose”, “BMI”, “Age” and “BloodPressure”. The reason for choosing four features is based on the Correlation Matrix, as shown in Fig. 5. In the prediction of  “Insulin,” the reason for selecting “BloodPressure” and “SkinThickness” instead of “DiabetesPedigreeFunction” is to test the performance under the same four common features. In addition, two research papers mentioned in the Related Work section did not use the predicted values by using Polynomial Regression in another prediction of the features that exist missing values. The experiment combines linear regression and polynomial features to form the Polynomial Regression. The degree of the polynomial features is set to 2, and a bias is not included in the polynomial features, as shown in Table 3. Next, fit\_transform is applied to the non-zero values features to find the metrics like mean value and standard deviation in overall statistics. Then, the non-zero values are normalized by transforming them afterward. Then, the values will be predicted based on the features of non-zero values in Fig. 6 and replace zero values with predicted values.

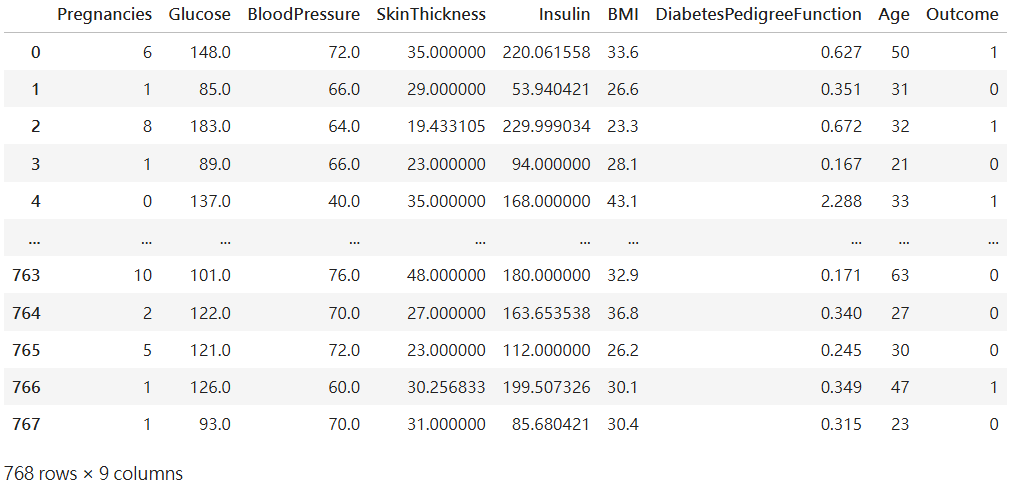


Figure 7: Pima Indian Diabetes dataset (Predicted value filled)

Fig.7 shows that all the predicted values were inserted into the corresponding features when the Polynomial Regression was done. Also, these predicted values were inserted based on the values of other features instead of randomly inserted the predicted values, which preserved the characteristic of the dataset that these values were learned from other features to make the prediction and predicted with a rational value. Therefore, they would not harm the experiment environment and make the results more comparable.

一張含有 文字, 圖表, 行, 繪圖 的圖片

AI 產生的內容可能不正確。Figure 8: Q-Q Plot of each feature in the Pima Indian Diabetes dataset (after the Polynomial Regression)

Q-Q Plot (Quantile-Quantile Plot) reflects the distribution of the features, whether they are normal or not. In Fig.8, features “Glucose,” “BloodPressure,” “SkinThickness,” and “BMI” were displayed in a normal distribution form as they roughly followed the red slope on the plot. On the other hand, the q-q plots of features “Insulin,” “Age,” and “DiabetesPedigreeFunction” were displayed in a curved pattern, which means they are not the normal distribution. For example, most of the spots in the feature “Age” are approximately concentrated on those above 20, which is in line with the description of the dataset that the participants are at least 21 years old.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **The combinations of Key Features** | | | | | |
| **Key Features** | Glucose | Insulin | BMI | Age |  |
| Glucose | Insulin | BMI | SkinThickness |  |
| Glucose | Insulin | BMI | Age | BloodPressure |
| Glucose | Insulin | BMI | Age | SkinThickness |

Table 6: The combination of key features in the Pima Indian Diabetes dataset

To define the key features and target variable in the prediction, choosing the features as input is important in the experiment. Four combinations of key features were tested in the experiment to compare the different key features in the result, as shown in Table 6. Most of the key features are selected in order of the score in Correlation Matrix besides “Pregnancies” and “DiabetesPedigreeFunction”. The reason there are no “Pregnancies” and “DiabetesPedigreeFunction” in the combinations is that “Pregnancies” is only related to females, and it will affect the experiment result. Therefore, it will affect the prediction of Diabetes in a male if “Pregnancies” counted as a key feature. In addition, the score of “BloodPressure” and “DiabetesPedigreeFunction” are the same, but “BloodPressure” is directly related to the body, especially Diabetes, which is mentioned in the section Dataset – Pima Indian Diabetes Dataset. On the other hand, “DiabetesPedigreeFunction” is a mathematical function that represents the probability of Diabetes based on Family history, and it is hard to directly link to the human body. Therefore, “Pregnancies” and “DiabetesPedigreeFunction” are not selected as the prediction's key features.

## Data Preprocessing on 2013-2014 NHANES dataset

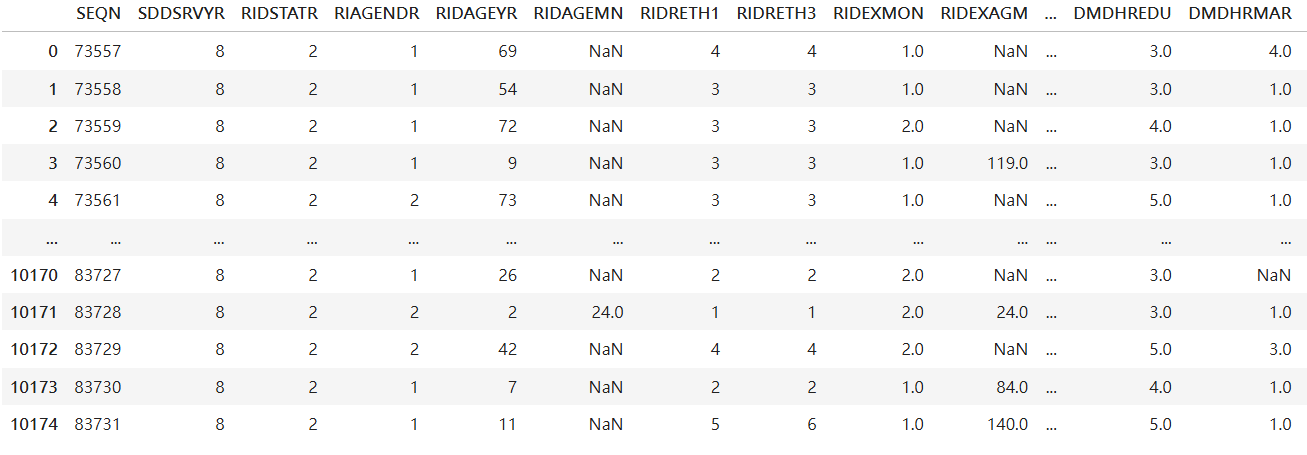


Figure 9: Information of Raw data Demographic

|  |  |  |
| --- | --- | --- |
| NHANES Dataset (Selected Raw data) | **Features in code** | **Renamed Features** |
| **Demographic** | SEQN (ID of participant)  RIAGENDR (Gender)  RIDAGEYR (Age) | ID  Gender  Age |
| **Examination** | SEQN (ID of participant)  BMXBMI (BMI)  BPXDI1 (Blood Pressure) | ID  BMI  BloodPressure |
| **Labs** | SEQN (ID of participant)  LBXGLT (Glucose)  LBXIN (Insulin) | ID  Glucose  Insulin |
| **Questionnaire** | SEQN (ID of participant)  DIQ010 (Diabetes Diagnosis) | ID  Outcome |

Table 7: Abstract of Selected Raw data and Features

In the experiment, five raw data will be merged into one. In Fig.9, there are too many variables in each raw data. It is necessary to select relevant variables from each raw data to make the dataset readable to people. The features were selected as shown in Table 2. In the 2013-2014 NHANES dataset, all the features were written in code, which was not readable to the public. Therefore, the key features are renamed, as shown in Table 7. Also, to create a similar environment to the Pima Indian Diabetes dataset and make the results more comparable, the raw data Diet and ALQ120Q in the questionnaire were removed from the dataset as there were no features that matched the Pima Indian Diabetes dataset.

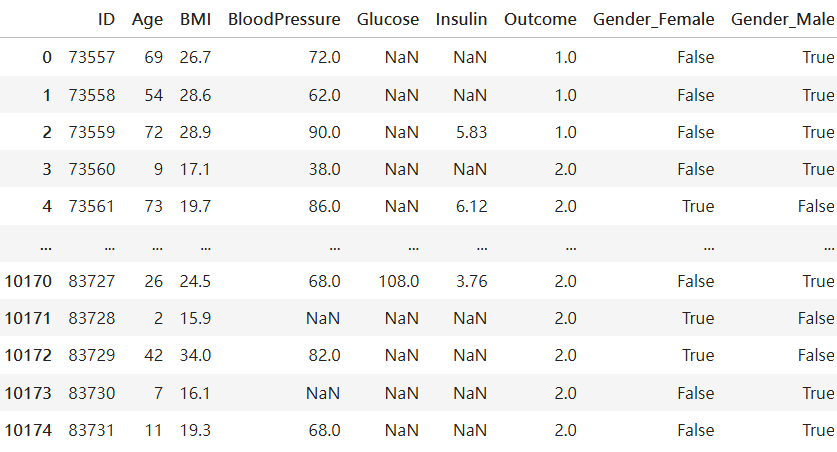


Figure 10: Information of Merged data (Not preprocessed)

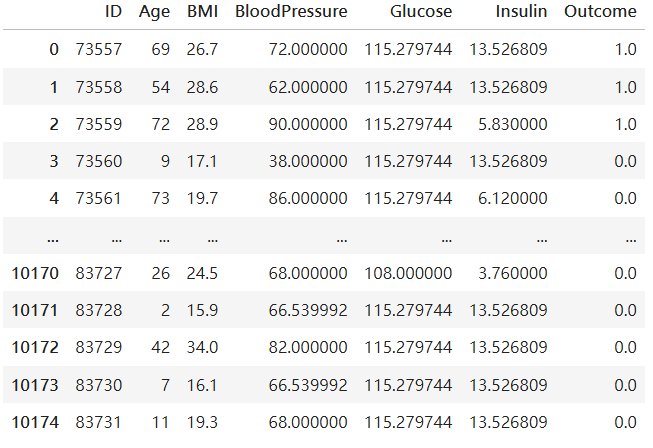


Figure 11: Merged data (Processed)

To merge the raw data into one, ID is the common feature in each raw data, and ID is used to combine all the raw data. Despite the merged data, null and missing values exist in Fig.10. In this situation, the mean fills all the null and missing values in each feature. Also, there are two types of values in Outcome, which 1.0 is “Diagnosed with Diabetes” and 2.0 is “No Diabetes”. The values in Outcome are renamed 1 and 0, respectively, similar to the Pima Indian Diabetes dataset. Moreover, some work is done on the merged data in Fig. 11, such as removing the columns “Gender\_Female” and “Gender\_Male” and removing the rows where “Glucose” and “Insulin” are less than 0.

一張含有 螢幕擷取畫面, 文字, 正方形, Rectangle 的圖片

AI 產生的內容可能不正確。

Figure 12: Correlation Matrix of the 2013-2014 NHANES dataset

In Fig.12, the order of the scores from highest to lowest in the correlation matrix on “Outcome” are “Age,” “BMI,” “Insulin,” “Blood pressure,” and “Glucose. “ID” is not considered a feature because “ID” is only used to join the raw data.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **The combinations of Key Features** | | | | | |
| **Key Features** | Glucose | Insulin | BMI | Age |  |
| Glucose | Insulin | BMI | Age | BloodPressure |
|  | Insulin | BMI | Age |  |

Table 8: The combination of key features in the 2013-2014 NHANES dataset

The experiment tested three combinations of key features to compare the key features in the result, as shown in Table 8. The key features are selected in order of the score in the Correlation Matrix besides “ID.” For the comparison, the combination of “Glucose,” “Insulin,” “BMI,” “Age,” and “Blood Pressure” is added to test the performance of adding all features as input.

# Result

## Results on Pima Indian Diabetes dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Gamma | Learning\_rate | Max\_depth | N\_estimators | Scale\_pos\_weight | Accuracy (%) |
| 0 | 0.2 | 8 | 100 | **1** | **81.17%** |
| 2 | 77.2% |
| 3 | 78.57% |

Table 9: Hyperparameter tuning on XGBoost – scale\_pos\_weight (Pima Indian Diabetes dataset)

In Table 9, it shows the adjustment on scale\_pos\_weight on XGBoost. In general, Scale\_pos\_weight can control the balance of positive and negative weights to solve the class imbalance problem in machine learning. In this situation, 1 is suitable for the model because it performs well with an accuracy of 81.17%.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Gamma | Learning\_rate | Max\_depth | N\_estimators | Scale\_pos\_weight | Accuracy (%) |
| 1 | 0.2 | 15 | 100 | 1 | 79.87% |
| 9 | 81.17% |
| **8** | **81.82%** |
| 7 | 80.52% |
| 3 | 78.57% |

Table 10: Hyperparameter tuning on XGBoost – max\_depth (Pima Indian Diabetes dataset)

Table 10 shows the adjustment of max\_depth on XGBoost. Max\_depth means the maximum tree depth. In this situation, setting the max depth to 8 performs well with an accuracy of 81.82%.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Gamma | Learning\_rate | Max\_depth | N\_estimators | Scale\_pos\_weight | Accuracy (%) |
| 1 | 0.05 | 8 | 100 | 1 | 80.52% |
| **0.1** | **82.47%**  **(AUC=0.86)** |
| 0.15 | 82.47% (AUC=0.85) |

Table 11: Hyperparameter tuning on XGBoost – learning\_rate (Pima Indian Diabetes dataset)

Table 11 shows the adjustment of the learning\_rate on XGBoost. The size of the learning rate can prevent the model from being overfitting or underfitting. In the experiment, the learning rate of 0.1 is better than other values, and the accuracy is 82.47%, which is a noticeable improvement in the model performance compared to Table 8 and 9. Although the learning rate of 0.15 has the same accuracy as the learning rate of 0.1, the AUC is decreased compared to the learning rate of 0.1.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Gamma** | **Learning\_rate** | **Max\_depth** | **N\_estimators** | **Scale\_pos\_weight** |
| 1 | 0.1 | 8 | 100 | 1 |
| **Selected Features** | | | | |
| Glucose | Insulin | BMI | SkinThickness |  |
|  | **Accuracy**  **(%)** | **Precision**  **(%)** | **F1 Score**  **(%)** | **AUC**  **(0–1)** |
| XGBoost (Filled by Mean only) | 80.52%  (+1.94%) | 80%  (+2%) | 80%  (+2%) | 0.85  (+2) |
| Proposed framework (XGBoost, Polynomial Regression & Filled by Mean) | **82.46%** | **82%** | **82%** | **0.86** |
| XGBoost (Filled by Mean, SMOTE) | 75.97%  (+6.49%) | 76%  (+6%) | 76%  (+6%) | 0.84  (+0.02) |

Table 12: A comparison between XGBoost in different preprocessing techniques (Pima Indian Diabetes dataset)

Table 12 shows the performance comparison on XGBoost by using different preprocessing methods. The hyperparameters of XGBoost in different preprocessing techniques are the same. From the result, the proposed framework that uses the Polynomial Regression and means to fill the missing values without SMOTE gains the highest accuracy with 82.46% and an AUC of 0.86. In addition, the preprocessing method, which used the imputation of missing values by mean, was used to fill in the missing values, and SMOTE only achieved 75.97% accuracy with an AUC of 0.84, which is the worst performance overall.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Accuracy**  **(%)** | **Precision**  **(%)** | **F1 Score**  **(%)** | **AUC**  **(0–1)** |
| Baseline (XGBoost, Tasin et al. [23]) | 81%  (+1.46%) | 81%  (+1.46%) | 81%  (+1.46%) | 0.84  (+0.02) |
| Proposed framework (XGBoost, Polynomial Regression & Filled by Mean) | **82.46%** | **82%** | **82%** | **0.86** |
| Proposed framework (Random Forest) | 80.52%  (+2.06%) | 80%  (+2%) | 80%  (+2%) | 0.85  (+0.01) |

Table 13: Abstract of Baseline VS Proposed framework (Pima Indian Diabetes dataset)

一張含有 文字, 行, 圖表, 螢幕擷取畫面 的圖片

AI 產生的內容可能不正確。

Figure 13: ROC Curve of the Pima Indian Diabetes dataset

Table 13 shows the results of the baseline model and the proposed framework. In comparison with the baseline model, XGBoost with the proposed framework gets a satisfying result as it has 82.46% accuracy with an AUC of 0.86, as shown in Fig.13, which is better than the baseline model proposed by Tasin et al. [23] and Random Forest with the proposed framework.

## Results on 2013-2014 NHANES dataset

|  |  |  |
| --- | --- | --- |
| Iteration | Learning\_rate | Accuracy (%) |
| 500 | 0.1 | 93.34% |
| 0.15 | 93.04% |
| 0.2 | 92.24% |
| 1000 | 0.1 | 92.74% |
| 0.15 | 92.34% |
| 0.2 | 92.29% |
| 2000 | 0.1 | 92.59% |
| 0.15 | 92.04% |
| 0.2 | 92.14% |

Table 14: Hyperparameter tuning on CatBoost (2013-2014 NHANES dataset)

Since the baseline model proposed by Qin et al. [26] used CatBoost as the training model in the prediction of Diabetes in the 2013-2014 NHANES dataset, CatBoost was applied in the experiment for the comparison, and the highest accuracy succeeded when the iteration was equal to 500 with the learning rate of 0.1.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Iteration** | **Learning\_rate** |  |  |  |
| 500 | 0.1 |  |  |  |
| **Selected Features** | | | | |
| Glucose | Insulin | BMI | Age | BloodPressure |
|  | **Accuracy**  **(%)** | **Precision**  **(%)** | **F1 Score**  **(%)** | **AUC**  **(0–1)** |
| CatBoost (Polynomial Regression) | 91.58%  (+1.76%) | 89%  (+3%) | 90%  (+3%) | 0.90  (+0) |
| Proposed framework (CatBoost, Filled by Mean) | **93.34%** | **92%** | **93%** | **0.90** |
| CatBoost (Polynomial Regression, SMOTE) | 91.58%  (+1.76%) | 89%  (+3%) | 90%  (+3%) | 0.90  (+0) |

Table 15: A comparison between CatBoost in different preprocessing techniques (2013-2014 NHANES dataset)

Table 15 shows the performance comparison of CatBoost by using different preprocessing methods. The hyperparameters of CatBoost in different preprocessing techniques are the same. As the table shows, the proposed framework that uses the imputation of missing values by mean is used to fill the missing values, gaining the highest accuracy with 93.34% and an AUC of 0.9. Also, the same results were achieved using polynomial regression and SMOTE, and polynomial regression only.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Accuracy (%)** | **Precision(%)** | **F1 Score (%)** | **AUC (0–1)** |
| Baseline  (CatBoost, Qin et al. [26]) | 82%  (+11.94%) | 82%  (+11%) | 82%  (+11%) | 0.83  (+0.08) |
| Proposed framework  (CatBoost, Filled by Mean) | 93.34%  (+0.6%) | 93  (+0%) | 93  (+0%) | 0.90  (+0.01) |
| Proposed framework  (XGBoost, Fill by Mean) | **93.94%** | **93%** | **93%** | **0.91** |
| Proposed framework (Random Forest, Filled by Mean) | 93.19%  (+0.75%) | 92%  (+1%) | 92%  (+1%) | 0.88  (+0.03) |

Table 16: Baseline VS Proposed framework (2013-2014 NHANES dataset)

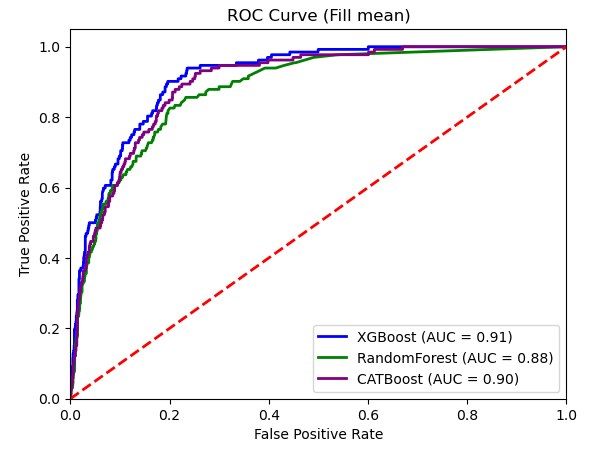
 Figure 14: ROC Curve of the 2013-2014 NHANES dataset

Table 16 shows the results of the baseline model and the proposed framework. In the experiment, the XGBoost hyperparameter is the same as the Pima Indian Diabetes dataset in Table 11. From Table 15, XGBoost achieved the highest accuracy of 93.94%, with an AUC of 0.91 (as shown in Fig.14), which is better than the baseline and other models under the same proposed framework.

# Discussion

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Performance of the combinations of Key Features (Proposed Framework, XGBoost)** | | | | | | **Accuracy%** |
| **Key Features** | Glucose | Insulin | BMI | Age |  | 80.52% |
| Glucose | Insulin | BMI | SkinThickness |  | 82.46% |
| Glucose | Insulin | BMI | Age | BloodPressure | 77.27% |
| Glucose | Insulin | BMI | Age | SkinThickness | 78.57% |

Table 18: A comparison between XGBoost in different combinations of key features (Pima Indian Diabetes dataset)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Performance of the combinations of Key Features (Proposed Framework, XGBoost)** | | | | | | **Accuracy%** |
| **Key Features** | Glucose | Insulin | BMI | Age |  | 93.44% |
| Glucose | Insulin | BMI | Age | BloodPressure | 93.94% |
|  | Insulin | BMI | Age |  | 93.08% |

Table 18: A comparison between XGBoost in different combinations of key features (2013-2014 NHANES dataset)

The above results show that the proposed framework can achieve better performance, which is better than the baseline. Also, from the above results, it can be found that the performance without SMOTE is higher than the SMOTE applied. The reason for the difference between the ensemble learning approach and the class imbalance method (e.g., SMOTE) is that the ensemble learning method has a built-in hyperparameter for controlling the balance of positive and negative weights to process the class imbalance problem in machine learning (e.g., scale\_pos\_weight” in XGBoost). Therefore, using the class imbalance technique (e.g., SMOTE) is another possible option if the model does not provide the hyperparameters to control the balance of positive and negative weights to process the class imbalance.

In Table 17 and Table 18, the combinations of key features (Glucose, Insulin, BMI and SkinThickness) in the Pima Indian Diabetes dataset and (Glucose, Insulin, BMI, Age and BloodPressure) in the 2013-2014 NHANES dataset both achieved satisfactory performance, in which improved the performance by 3-4% by removing “Age” in the combination of key features in the Pima Indian Diabetes dataset. Also, the common features of these datasets are Glucose, Insulin, and BMI, which can be considered the key factors in predicting Diabetes.

# Conclusion

In this final year project, XGBoost, CatBoost, and Random Forest were applied to predict Diabetes in the Pima Indian Diabetes dataset and the 2013-2014 NHANES dataset. Also, polynomial regression and the imputation of missing values by mean were the preprocessing techniques used to fill in the missing values in these datasets. In addition, two baselines were used as the benchmark for the comparison and the reference in two different datasets, which were proposed by Tasin et al. [23] and Qin et al. [26] in the Pima Indian Diabetes dataset and 2013-2014 NHANES dataset respectively. For the results, XGBoost achieved the highest accuracy in both datasets, in which the ideal combinations of key features are (Glucose, Insulin, BMI, and SkinThickness) in the Pima Indian Diabetes dataset and (Glucose, Insulin, BMI, Age, and blood pressure) in the 2013-2014 NHANES dataset. Moreover, applying the class imbalance technique, if using the ensemble learning model, is unnecessary as it would decrease the model performance.

Despite the results meeting the expectations, there have been improvements in the experiment, such as adding a neural network to compare the performance between the neural network and the ensemble learning method in different datasets and the hyperparameters tuning on Random Forest.

References

[1] N. G. Forouhi, A. Misra, V. Mohan, R. Taylor, and W. Yancy, “Dietary and nutritional approaches for prevention and management of type 2 diabetes,” BMJ (Online), vol. 361, pp. k2234–k2234, 2018, doi: 10.1136/bmj.k2234 [Accessed Mar. 12, 2025]

[2] National Center for Health Statistics, NHANES 2013-2014 Questionnaire Data Overview. [Online]. Available: https://my.clevelandclinic.org/health/diagnostics/17649-blood-pressure [Accessed Mar. 12, 2024]

[3] L. L. J. Koppes, J. M. Dekker, H. F. J. Hendriks, L. M. Bouter, and R. J. Heine, “Moderate Alcohol Consumption Lowers the Risk of Type 2 Diabetes: A meta-analysis of prospective observational studies,” Diabetes care, vol. 28, no. 3, pp. 719–725, 2005, doi: 10.2337/diacare.28.3.719 [Accessed Mar. 12, 2025]

[4] T. A. Buchanan, A. H. Xiang, and K. A. Page, “Gestational diabetes mellitus: risks and management during and after pregnancy,” *Nature reviews. Endocrinology*, vol. 8, no. 11, pp. 639–649, 2012, doi: 10.1038/nrendo.2012.96 [Accessed Mar. 14, 2025]

[5] Healthline , Everything You Need to Know About Glucose, 2024. [Online]. Available: https://www.healthline.com/health/glucose [Accessed Mar. 14, 2025].

[6] 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 Mar. 14, 2025]

[7] Mayo Clinic, Diabetes, 2024. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/diabetes/symptoms-causes/syc-20371444 [Accessed Mar. 14, 2025].

[8] Cleveland Clinic, Blood Pressure, 2022. [Online]. Available: https://my.clevelandclinic.org/health/diagnostics/17649-blood-pressure [Accessed Mar. 15, 2025].

[9] National Heart, Lung, and Blood Institute, What Is High Blood Pressure?, 2024. [Online]. Available: https://www.nhlbi.nih.gov/health/high-blood-pressure [Accessed Mar. 15, 2025].

[10] Johns Hopkins Medicine, Diabetes and High Blood Pressure, n.d. [Online]. Available: https://www.hopkinsmedicine.org/health/conditions-and-diseases/diabetes/diabetes-and-high-blood-pressure [Accessed Mar. 15, 2025].

[11] W. Li *et al.*, “Associations Between Adult Triceps Skinfold Thickness and All-Cause, Cardiovascular and Cerebrovascular Mortality in NHANES 1999–2010: A Retrospective National Study,” *Frontiers in cardiovascular medicine*, vol. 9, pp. 858994–858994, 2022, doi: 10.3389/fcvm.2022.858994 [Accessed Mar. 17, 2025].

[12] S. Särnblad, A. Magnuson, U. Ekelund, and J. Åman, “Body fat measurement in adolescent girls with type 1 diabetes: a comparison of skinfold equations against dual-energy X-ray absorptiometry,” *Acta Paediatrica*, vol. 105, no. 10, pp. 1211–1215, 2016, doi: 10.1111/apa.13366 [Accessed Mar. 17, 2025].

[13] Medscape, Insulin, 2024. [Online]. Available: https://emedicine.medscape.com/article/2089224-overview#a1 [Accessed Mar. 17, 2025].

[14] Better Health Channel, Diabetes and insulin, 2021. [Online]. Available: https://www.betterhealth.vic.gov.au/health/conditionsandtreatments/diabetes-and-insulin#type-2-diabetes [Accessed Mar. 17, 2025].

[14] Cleveland Clinic, Body Mass Index (BMI), 2022. [Online]. Available: https://my.clevelandclinic.org/health/articles/9464-body-mass-index-bmi [Accessed Mar. 17, 2025].

[15] J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, and R. S. Johannes, “Using the ADAP learning algorithm to forecast the onset of diabetes mellitus,” in *Proceedings - Symposium on Computer Application in Medical Care*, 1988, pp. 261–265 [Accessed Mar. 17, 2025].

[16] Centre for Health Protection, Department of Health, The Government of the Hong Kong Special Administrative Region, What is Diabetes?, 2022. [Online]. Available: https://www.chp.gov.hk/en/features/103650.html [Accessed Mar. 18, 2025].

[17] International Diabetes Federation, Facts & figures, 2021?. [Photograph]. Available: https://idf.org/about-diabetes/diabetes-facts-figures/ [Accessed Mar. 18, 2025].

[18] World Health Organization, Diabetes, 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/diabetes [Accessed Mar. 18, 2025].

[19] IDF Diabetes Atlas, Diabetes around the world in 2021, 2021?. [Online]. Available: https://diabetesatlas.org/ [Accessed Mar. 18, 2025].

[20] Kaggle, Pima Indians Diabetes Database, 2016. [Online]. Available: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database [Accessed Mar. 18, 2025].

[21] National Center for Health Statistics, Center for Disease Control and Prevention, NHANES Questionnaires, Datasets, and Related Documentation, n.d. [Online]. Available: https://wwwn.cdc.gov/nchs/nhanes/ [Accessed Mar. 18, 2025].

[22] Abdul Azis Abdillah and Suwarno Suwarno, “Diagnosis of Diabetes Using Support Vector Machines with Radial Basis Function Kernels,” *International Journal of Technology*, vol. 7, no. 5, pp. 849–858, 2016, doi: 10.14716/ijtech.v7i5.1893 [Accessed Mar. 19, 2025].

[23] 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 Mar. 19, 2025].

[24] ScienceDirect, Polynomial Regression, n.d. [Online]. Available: https://www.sciencedirect.com/topics/computer-science/polynomial-regression [Accessed Mar. 18, 2025].

[25] 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 [Accessed Mar. 19, 2025].

[26] Y. Qin et al., “Machine Learning Models for Data-Driven Prediction of Diabetes by Lifestyle Type,” International journal of environmental research and public health, vol. 19, no. 22, pp. 15027-, 2022, doi: 10.3390/ijerph192215027 [Accessed Mar. 19, 2025].

[27] W. Chen et al., “Modeling flood susceptibility using data-driven approaches of naïve Bayes tree, alternating decision tree, and random forest methods,” The Science of the total environment, vol. 701, pp. 134979–134979, 2020, doi: 10.1016/j.scitotenv.2019.134979 [Accessed Mar. 20, 2025].

[28] DMLC XGBoost, XGBoost Documentation, 2022?. [Online]. Available: https://xgboost.readthedocs.io/en/stable/ [Accessed Mar. 20, 2025].

[29] Esri, How CatBoost algorithm works, n.d. [Online]. Available: https://pro.arcgis.com/en/pro-app/latest/tool-reference/geoai/how-catboost-works.htm [Accessed Mar. 20, 2025].

[30] J. T. Hancock and T. M. Khoshgoftaar, “CatBoost for big data: an interdisciplinary review,” Journal of big data, vol. 7, no. 1, pp. 94–94, 2020, doi: 10.1186/s40537-020-00369-8 [Accessed Mar. 20, 2025].

[31] P. D. Allison, *Missing data*. Thousand Oaks, Calif: Sage Publications, 2002. [ebook]. Available: https://loyola.sites.oasis.unc.edu/soci709/cdocs/allison.pdf [Accessed Mar. 21, 2025].

[32] Medium, What is Feature Scaling and Why Does Machine Learning Need It?, 2023. [Online]. Available: https://medium.com/@shivanipickl/what-is-feature-scaling-and-why-does-machine-learning-need-it-104eedebb1c9 [Accessed Mar. 21, 2025].