**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 2**

**Bachelor of Science (Honours)**

In

**Internet and Multimedia Technologies**

**of**

**The Hong Kong Polytechnic University**

Supervisor: Prof CHAU Lap Pui Date:2-April-2025

Abstract (list what u did in the prj)

XX

Contents

[Abstract (list what u did in the prj) i](#_Toc192596862)

[Contents ii](#_Toc192596863)

[1. Introduction (background of the prj) 3](#_Toc192596864)

[1.1 Overview 3](#_Toc192596865)

[1.1.1 Motivation (why u want to do this project?(want to know how machine learning operate in diabetes prediction / NHANES dataset are rarely used in research paper (most paper rather used Pima Indian dataset/ ) 3](#_Toc192596866)

[1.1.2 Objective (what do you want to achieved in this project? (e.g. proof that xx is more better / try to used another method that is rarely used in research paper to see the result is better or not) 4](#_Toc192596867)

[2. Related Work (List how machine learning works on Diabetes prediction in tradition / List how neural network work on Diabetes prediction (Maybe can write it?) /List the method used in reference paper (baseline) ) 5](#_Toc192596868)

[3. Methodology 5](#_Toc192596869)

[3.1 Proposed Framework (My proposed framework (how it works) ; why choose that component in the proposed framework?(e.g. improve acc? / Convenient? / haven’t used in research paper? (most of paper use polynomial regression) ) 5](#_Toc192596870)

[4. Dataset 6](#_Toc192596871)

[4.1 Pima Indian Diabetes Dataset 6](#_Toc192596872)

[4.2 2013-2014 NHANES Dataset 8](#_Toc192596873)

[5. Experiment 10](#_Toc192596874)

[5.1 Data Preprocessing on Pima Indian Diabetes dataset: 10](#_Toc192596875)

[5.2 Data Preprocessing on NHANES dataset: 11](#_Toc192596876)

[6. Result (list the result [acc/auc/precision..] / list the hyperparameters used/adjusted in the experiment in TWO dataset) 12](#_Toc192596877)

[7. Discussion (what u find/observe in the experiment) 12](#_Toc192596878)

[8. Conclusion (summarize what u achieved in the prj & any improvements / future work) 12](#_Toc192596879)

[References 13](#_Toc192596880)

[Appendices 14](#_Toc192596881)

[Appendix 1: Pima Indian Diabetes dataset 14](#_Toc192596882)

[Appendix 2: 2013-2014 NHANES dataset 14](#_Toc192596883)

List of Figures

Figure 1. Pima Indian Diabetes Dataset (raw data)9

Figure 2. Check missing value on Pima Indian Diabetes Dataset 10

Figure 3. Fill the missing values on “SkinThickness”11

Figure 4. Fill the missing values on “Insulin”12

Figure 5. Check if there have missing value (0) or not 12

Figure 6. All the missing values are replaced by the predicted values 13

Figure 7. Quantile-Quantile Plot of all features (Pima Indian Diabetes Dataset)13

Figure 8. Correlation matrix of all features and outcome (Pima Indian Diabetes Dataset)14

Figure 9. Feature Importance (Pima Indian Diabetes dataset)15

Figure 10. Balanced data (Pima Indian Diabetes Dataset)15

Figure 11. Information about demographic (NHANES dataset)16

Figure 12. Merged dataset (NHANES dataset)17

Figure 13. Total number of each row with value 017

Figure 14. Merged dataset after Polynomial Regression applied (NHANES dataset)18

Figure 15. Quantile-Quantile Plot of all features (NHANES dataset)19

Figure 16. Correlation matrix of all features and outcome (NHANES dataset)19

Figure 17. Feature Importance (NHANES dataset)20

Figure 18. Reproduce Result (XGB+ADASYN)21

Figure 19. Paper Result (XGB+ADASYN)21

Figure 20. Reproduce Result (AUC)21

Figure 21. Paper Result (AUC)21

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

Figure 23. Preliminary Result (Random Forest) (Pima Indian Diabetes dataset)22

Figure 24. Preliminary Result (XG Boost) (NHANES dataset)23

Figure 25. Preliminary Result (Random Forest) (NHANES dataset)23

List of Tables

Table 1. Information on Pima Indian Diabetes Dataset 5

Table 2. Abstract of NHANES Dataset 7

# Introduction (background of the prj) (each section write 250-300)

## Overview

### Motivation (why u want to do this project?(want to know how machine learning operate in diabetes prediction / NHANES dataset are rarely used in research paper (most paper rather used Pima Indian dataset/ )

Oxx

### Objective (what do you want to achieved in this project? (e.g. proof that xx is more better / try to used another method that is rarely used in research paper to see the result is better or not)

xx

# Related Work (List how machine learning works on Diabetes prediction in tradition / List how neural network work on Diabetes prediction (Maybe can write it?) /List the method used in reference paper (baseline) )

# Methodology

## Proposed Framework (My proposed framework (how it works) ; why choose that component in the proposed framework?(e.g. improve acc? / Convenient? / haven’t used in research paper? (most of paper use polynomial regression) )

# Dataset

## 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.

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.

## 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.

# Experiment

xx

## Data Preprocessing on Pima Indian Diabetes dataset:

The

## Data Preprocessing on NHANES dataset:

Sxx

# Result (list the result [acc/auc/precision..] / list the hyperparameters used/adjusted in the experiment in TWO dataset)

# Discussion (what u find/observe in the experiment)

# Conclusion (summarize what u achieved in the prj & any improvements / future work)

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.trainindata.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