Intro

過去多年以來，糖尿病在全球變得越來越普遍， 截至2022年，患有糖尿病症狀的人口已經多達八百三十萬人 [1]，大約佔全球人口的10.4%。這使得糖尿病不容忽視。現代人生活節奏急促，難以騰出時間，保持身體鍛鍊，這導致其生活習慣走向不健康的趨勢，例如肥胖和睡眠不足等等。常見的糖尿病主要分為兩種，分別是一型和二型。前者受先天因素影響，例如家族遺傳 [2]，後者受後天因素影響，例如不良的飲食習慣和缺乏運動 [2]。多年來，機器學習一直被應用於糖尿病預測，預測病人患有糖尿病的可能性。機器學習的優勢在於可以因應資料集的內容，產出相應的預測結果，讓人們能夠因應預測結果，做出適當的判斷。本次專案旨在透過探索機器學習模型並重點應用於預測糖尿病。

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 2 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. The one of advantages of machine learning is it can generate the corresponding prediction based on the content of dataset. It allows people to make a suitable decision-making based on the prediction that generated by machine learning algorithms. This project aims to explore the machine learning algorithms and apply them on diabetes predictions.

Background

糖尿病一直都是值得關注的問題。以往人們會透過血糖機，進行血糖測試並檢測血液中的血葡萄糖含量，從而判斷是否患有糖尿病。隨着科技發展，並引入機器學習進行糖尿病預測，讓人們能夠從機器學習模型的預測結果當中，知道自己是否有糖尿病的風險，這有助人們不必從多種多樣的身體數據中逐一判斷，降低了因人為因素而造成誤判的機率。

~~Diabetes is one of the major concerns in the world. As stated above, many people have got diabetes in recent years. In the past, people used blood glucose meter to do the blood glucose measurement for checking the blood glucose to see if they get diabetes or not. With the advance of science and technology, there is another way that machine learning algorithms are introduced for doing the diabetes predictions. People can view the predicted output generated by machine learning algorithms to check if they are at risk for diabetes or not. It saves the time for people that they are allowed to check their health conditions without one by one view the indices from the body. Moreover, it decreases the probability of misjudgment due to human factors.~~

One of the greatest issues in the world is Diabetes. As mentioned earlier, many people whether adults, youth or children have got diabetes in recent years. Start early for diabetes people used to do the blood glucose measurement for checking the blood glucose to see if they get diabetes or not in the past blood glucose meter has been 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 or not. It saves the time for people that they are allowed to check their health conditions without one by one view the indices from the body. Moreover, it decreases the probability of misjudgment due to human factors.

Problem Statement (Total 407 [3:36am 20/12])

儘管機器學習應用於糖尿病預測已經行之有年，但各種對糖尿病預測的機器學習研究都指向不同的結果，例如(paper1)認為xx模型最好 [3]，(paper2)則認為yyy模型有助預測糖尿病的風險 [4]。這對於人們來說，難以判斷何種模型是較為優勢並適用於糖尿病預測，因為上述研究實驗都分別採用了不同的資料集以及預處理方法，導致對同一議題的研究會出現不同的結果。因此，該專案是一個值得研究的項目，透過資料搜集和研究對比，發現所得，靈活運用大學所學知識並應用於這個專案。

~~Although diabetes prediction with machine learning have already implemented in recent years, many researches done on diabetes prediction with machine learning algorithms are 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. [3] commented XGBoost with ADASYN approach performs well in diabetes prediction. It is difficult for people to judge whether the machine learning models are the most suitable for diabetes prediction. Furthermore, lots of research done on diabetes prediction with machine learning algorithms are used different dataset (NHANES and Pima Indian Diabetes) and data preprocessing methods and make the same topic on diabetes prediction with machine learning algorithms caused different results. Therefore, this project is a good research topic to work on it that find the results through collecting the data from different sources, implement the methodology and make a comparison of different methodologies.~~

Although diabetes prediction with machine learning has been implemented in recent years, 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. [3] commented that XGBoost with the ADASYN approach performs well in diabetes prediction. It is difficult for people to judge whether the 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.

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 |

The Pima Indian Diabetes Database provided information about the patients who have Diabetes or not. The original source of the dataset comes from the National Institute of Diabetes and Digestive and Kidney Diseases [6]. 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 [7] 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 [8].

BloodPressure: It means heart beats and pumps blood into the arteries [9]. 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.

**NHANES Dataset**

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

|  |  |
| --- | --- |
| NHANES Dataset | |
| 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 1: 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 included 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.

Methodology:

Machine learning is a suitable approach to do diabetes prediction because the diabetes prediction belongs to classification task that determines whether the patients diagnosed with having diabetes. The process is included data preprocessing, feature selection, training, testing and the performance evaluation. All the works are done on Jupyter Notebook which is a web based application that provides interactive computing notebook environment to describe the data analysis.

Challenges:

~~For the NHANES dataset, it is the challenging task of data preprocessing 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 lots of missing value with 0 and null value on the Pima Indian Diabetes dataset and NHANES dataset. For example, 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. I used the approach that filling value zero with the means but it performs poor in the q-q plot as both are not the normal distribution. Therefore, I used polynomial regression to predict the missing value and it is better than filling value zero with the means.~~

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. I used the approach of filling value zero with the means, but it performs poorly in the q-q plot as both are not the normal distribution. Therefore, I used polynomial regression to predict the missing value and it is better than filling value zero with the means.

CURRENT PROGRESS: (around 900-1000)

Data Preprocessing on Pima Indian Diabetes dataset:

一張含有 文字, 螢幕擷取畫面, 數字, 平行 的圖片

自動產生的描述 Figure 1: Pima Indian Diabetes Dataset (raw data)

There are 768 rows (interviewees) with 9 columns (features) in the dataset before preprocessing. As shown in the figure 1, there have some value zero in columns “SkinThickness” and “Insulin”. No null values on the dataset so there is not necessary to fill zero.

一張含有 文字, 字型, 白色, 收據 的圖片

自動產生的描述

Figure 2: Check missing value on Pima Indian Diabetes Dataset

In figure 2, lots of missing value existed in columns “SkinThickness” and “Insulin” that they count for about 30% and 50% of the whole 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 is zero and non-zero respectively. To make the prediction on column “SkinThickness” become more reliable, columns “Glucose”, “BloodPressure” and “BMI” and “Age” are used to assist the prediction of column “SkinThickness”.

========================(start below here)

Polynomial Regression is applied to predict the missing value of column “SkinThickness”. 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.”

一張含有 文字, 紙張, 文件 的圖片

自動產生的描述

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.

一張含有 文字, 文件, 紙張 的圖片

自動產生的描述

Figure 4: Fill the missing values on “Insulin”

一張含有 文字, 螢幕擷取畫面, 字型 的圖片

自動產生的描述

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 that all the predicted values are replaced successfully.

一張含有 文字, 螢幕擷取畫面, 數字, 字型 的圖片

自動產生的描述

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

一張含有 文字, 圖表, 行, 繪圖 的圖片

自動產生的描述

Figure 7: Quantile-Quantile Plot of all features

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”, “BloodPressure”, “SkinThickness” and “BMI” are roughly following the red straight line on the plot. Alos, the spots on the features “Insulin”, “Age” and “DiabetesPedigreeFunction” are deviated from the red straight line. For example, most of the spots in feature “Age” are concentrated in 20, which means most of participants are about 21 years old.

一張含有 文字, 螢幕擷取畫面, 正方形, Rectangle 的圖片

自動產生的描述

Figure 8: Correlation matrix of all features and target variable (outcome)

~~Figure 8 shows the correlation coefficient between all the features and the outcome. The ranking of relationship between the features and the outcome in descending order 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 the higher glucose level, the higher risk of be diagnosed with diabetes. According to the ranking, “Glucose”, “Insulin”, “BMI” and “SkinThickness” are selected as the key features for the prediction of diabetes.~~

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 process of Standardization, the scaler will be defined and fit into the training data, then transform the testing data to finish the feature scaling.

一張含有 螢幕擷取畫面, 文字, 圖表, 行 的圖片

自動產生的描述

Figure 9: Feature Importance

一張含有 文字, 字型, 螢幕擷取畫面, 白色 的圖片

自動產生的描述

Figure 10: Feature Importance

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

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.

Data Preprocessing on NHANES dataset:

~~Since the NHANES dataset has five raw data, it is necessary to do the data preprocessing to ensure that the data is readable and understandable to people because there are lots of features in each raw data. Originally, medication (not mentioned in Table 1) is one of the raw data of the NHANES dataset but there is no complete description that can be found on 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.~~

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.

一張含有 文字, 螢幕擷取畫面, 數字, 字型 的圖片

自動產生的描述

Figure 11: Information about demographic (NHANES dataset)

~~In Figure 11, each feature is hard to read for people as they are named in terms. To make the data become more readable and understandable, features which are similar to 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 convert categorical variables to dummy variables.~~

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.

一張含有 文字, 螢幕擷取畫面, 數字, 字型 的圖片

自動產生的描述

Figure 12: Merged dataset (NHANES dataset)

一張含有 文字, 字型, 白色, 代數 的圖片

自動產生的描述

Figure 13: Total number of each row with value 0

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

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.

一張含有 文字, 螢幕擷取畫面, 字型, 數字 的圖片

自動產生的描述

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, 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% and consequently remove these will not have an significant impact on the prediction of diabetes.~~

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 a counterintuitive way to use the values 1 and 2 to represent having diabetes and no diabetes, especially the 2 will mistake people for invalid. Accordingly, it is necessary to use 1 and 0 to represent having diabetes and no diabetes instead 1 and 2.~~

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.

一張含有 文字, 圖表, 行, 繪圖 的圖片

自動產生的描述

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

一張含有 文字, 螢幕擷取畫面, 正方形, Rectangle 的圖片

自動產生的描述

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

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

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” are used to assign the numeric labels to each individual entry, which means it does not have a meaningful relationship with “Outcome”.~~

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.

一張含有 螢幕擷取畫面, 文字, 行, 圖表 的圖片

自動產生的描述

Figure 17: Feature Importance (NHANES dataset)

~~In Figure 17, “Glucose” is the highest rank in the feature importance after standardization, which is the same rank as 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 feature in order to make a similar environment to do the comparison.~~

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.

===================

# FUTURE WORKS (200)

1. (Comparison: use other preprocessing method like ADYSN? To vs SMOTE)

Try to use other preprocessing methods

~~In this stage, standardization and SMOTE are applied as the approaches of the data preprocessing. For the feature scaling, other methods like min-max scaling and mean normalization will be the possible 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 on the minority class.~~

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. (Select another key features for comparison on model performance)

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 are lower than 0.2 to ensure that no more redundancy features as the input features to 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 the input features and do the comparison to review which perform better under the same preprocessing method.~~

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.

3. (Maybe try another model like NN?)

Try to use another model for the comparison

In this stage,

==============

~~The baseline model referenced in this project is proposed by Tasin et al [4].~~

~~Tasin et al. [4] proposed a machine learning framework that acquired 81% accuracy by using XGBoost. The preprocessing methods are extreme gradient boosting technique for filling the missing value, SMOTE and ADASYN are applied for the class imbalance issue. In this project, it will be the baseline model for the reference.~~

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.

一張含有 文字, 螢幕擷取畫面, 字型, 數字 的圖片

自動產生的描述 一張含有 文字, 螢幕擷取畫面, 字型, 數字 的圖片

自動產生的描述

Figure 18: Reproduce Result

Figure 19: Paper Result

一張含有 文字, 螢幕擷取畫面, 字型, 數字 的圖片

自動產生的描述

Figure 20: Reproduce Result

一張含有 文字, 螢幕擷取畫面, 字型, 文件 的圖片

自動產生的描述

Figure 21: Paper Result

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.

一張含有 文字, 螢幕擷取畫面, 字型, 數字 的圖片

自動產生的描述一張含有 文字, 螢幕擷取畫面, 字型, 數字 的圖片

自動產生的描述

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

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

一張含有 文字, 螢幕擷取畫面, 字型, 數字 的圖片

自動產生的描述一張含有 文字, 螢幕擷取畫面, 字型, 數字 的圖片

自動產生的描述

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

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

~~Since the referenced paper proposed a machine learning framework that used XG Boost with ADASYN and used other machine learning models like random forest for the comparison in the paper, XG Boost and Random Forest are used to do the prediction of diabetes in this project. For the Pima Indian Diabetes dataset, Random Forest obtains 77.9% accuracy and XG Boost is the highest that it has 86.1% accuracy for the NHANES dataset. Adjustments have 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 NHANES dataset. The reason for the different result in two datasets can be the number of the entries that NHANES dataset has around ten thousand entries.~~

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.

Conclusion

Abstract

~~Diabetes has become a noteworthy social issue in the world where more and more people are diagnosed with diabetes in recent years. To find out the problem, machine learning is one of the approaches that is used to predict diabetes. This project introduces two datasets which are Pima Indian Diabetes dataset and 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 do the prediction of missing value and class imbalance problem. Furthermore, Random Forest and XG Boost are used to the prediction of diabetes in this project, and Random Forest and XG Boost are the highest accuracy in the Pima Indian Diabetes dataset and the NHANES dataset respectively. In the future, exploring other key features and preprocessing methods are the major options to get better results in this project.~~

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.

(講舊時vs現今人們的飲食習慣-> 多油多糖-> 糖尿病upup)

~~現時，糖尿病目前並未有能夠根治的方法~~

~~本次專案旨在透過探索機器學習並開發用於預測糖尿病的機器學習模型~~

Reference (Latest)

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].

References (Old)

1. Brother LeVon X Community Reporting, We can take control of our diabetes or diabetes can control us, 2023. [Online]. Available: https://www.brotherlevonxcommunityreporting.com/post/we-can-take-control-of-our-diabetes-or-diabetes-can-control-us [Accessed Sept. 18, 2024].
2. International Diabetes Federation, Diabetes facts & figures, 2022?. [Online]. Available: https://idf.org/about-diabetes/diabetes-facts-figures/ [Accessed Sept. 18, 2024].
3. International Diabetes Federation, Diabetes facts & figures, 2022?. [Photograph]. Available: https://idf.org/about-diabetes/diabetes-facts-figures/ [Accessed Sept. 18, 2024].
4. Y. Qin et al., “Machine Learning Models for Data-Driven Prediction of Diabetes by Lifestyle Type,” Int. J. of Environmental Research and Public Health, vol 19, no. 22, pp. 15027-, Nov. 2022. [Online]. doi: 10.3390/ijerph192215027 [Accessed Sept. 18, 2024].
5. R. G. Mertig, Diabetes. New York: Demos Health, 2011.
6. 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 Sept. 18, 2024]
7. SHAP, An introduction to explainable AI with Shapley values, 2018?. [Online]. Available: https://shap.readthedocs.io/en/latest/example\_notebooks/overviews/An%20introduction%20to%20explainable%20AI%20with%20Shapley%20values.html [Accessed Sept. 19, 2024].