STATISTICAL ANALYSIS

Group Project Report

EXPLORATORY ANALYSIS AND PREDICTION OF DIABETES

--By Group 6

Group Members:

1155202878 TAI Yuting

1155203494 LIU Xiaojia

1155208413 HONG Zhihui

1155206032 YE Ruixin

1155201562 WANG Shuonan

Research Background

Diabetes was one of the first diseases described with an Egyptian manuscript from 1500 B.C. mentioning "too great emptying of the urine." It is a group of common endocrine diseases characterized by sustained high blood sugar levels. Classic symptoms include thirst, polyuria, weight loss, and blurred vision. If left untreated, this disease can lead to various health complications, including disorders of the cardiovascular system, eye, kidney, and nerves.

Diabetes can be broadly categorized into two types. Type1 diabetes is characterized by loss of the insulin-producing beta cells of the pancreatic islets, leading to severe insulin deficiency, and can be further classified as immune-mediated or idiopathic (without known cause). This accounts for 5% to 10% of diabetes cases and is the most common in patients under 20 years old, whileType2 diabetes accounts for 95% of diabetes. Different from Type1 diabetes, Type2 starts with an abnormal insulin impedance action (abnormal and insensitive cellular response to insulin) without any pathologic problem in the pancreas itself.

In 2021, an estimated 537 million people had diabetes worldwide accounting for 10.5% of the adult population, with Type2 making up about 90% of all cases. Diabetes is responsible for 6.7 million deaths in 2021 - 1 every 5 seconds, and it caused at least USD 966 billion dollars in health expenditure – a 316% increase over the last 15 years.

However, this trend can be curbed by making people aware of the triggers of diabetes and finding appropriate interventions. We selected common factors to discover their association with diabetes.

Research Questions

1. Is there a significant relationship between gender and diabetes?
2. Are people without smoking history less likely to develop diabetes?
3. Does heart disease increase the possibility of having diabetes?
4. Which model performs better for prediction?

Dataset Description

The dataset used by this project comes from Kaggle (Mustafa, 2023) and consists of 100,000 records. During the initial stages of the project, the dataset underwent data cleaning by removing duplicates along with null values. Furthermore, entries that had gender tagged as ‘other’ and entries showing the person's smoking history as 'not current' or 'no info' were left out. Based on the message provided by the dataset creator, the labeling 'not current' and 'former' had a very similar interpretation, so instances labeled as "not current" were dropped in order to avoid potential ambiguity. The removal of entries labeled "no information" for gender and "other" for smoking history was to ensure the dataset more informative. Finally, 56,888 usable observations were maintained.

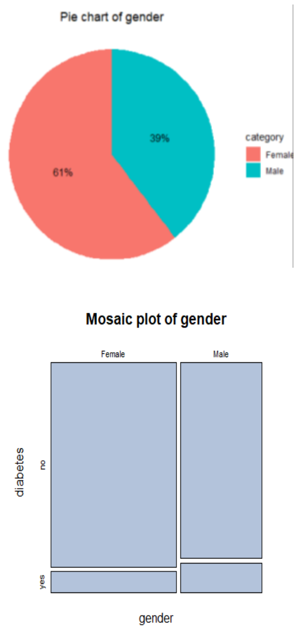
The dependent variable in the dataset is diabetes, which is a binary variable categorized as "0" (no diabetes) and "1" (diabetes). Independent variables are typical demographic and medical characteristics, including categorical variables like gender, hypertension, and smoking history, and numerical variables like age, weight, HbA1c_level and blood glucose level.

Smoking history variables were categorized into four groups, including never, current, ever and former smokers. Never indicates never smoked. Currently indicates being a smoker. Ever indicates having smoked but not currently smoking. Former indicates having a history of smoking at some point.

Although both HbA1c_level and Blood_glucose_level reflect an individual's blood glucose control, HbA1c provides a longer-term perspective, specifically the average blood glucose level over the past two to three months, and blood glucose provides a snapshot at a point in time.

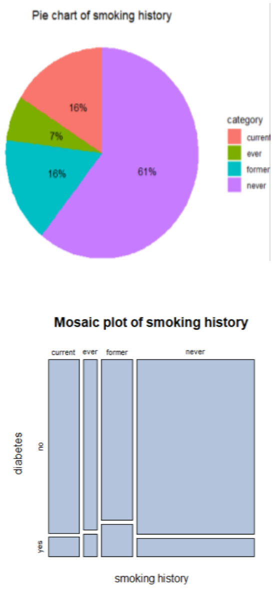
Visualization

In this part, we plot various kinds of graphs to explore the relationship between the independent variables we are interested in and the dependent variable. Because both of them are categorical variables, we plot mosaic plot and pie chart on them.



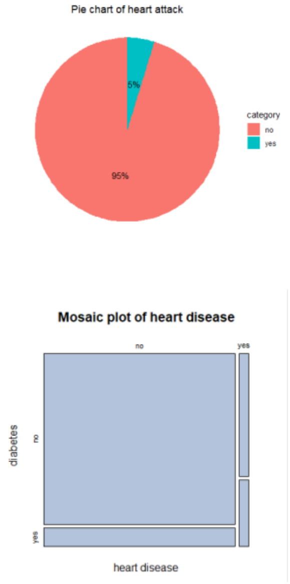
Graph 4-1

In the mosaic graph, we can see that the proportion of men among diabetic patients is higher than that among people without diabetes, which means that men are more likely to develop diabetes than women.



Graph 4-2

In the mosaic graph, we can see that the proportion of smokers among diabetic patients is higher than that among people without diabetes, which means that smoker more tend to develop diabetes.

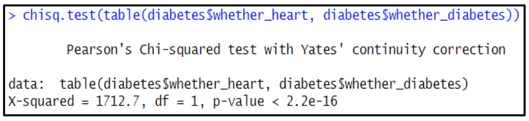


Graph 4-3

We can find that the proportion of individuals with heart disease among diabetic patients is higher than that among people without diabetes, which means that heart disease increases the likelihood of diabetes.

Correlation analysis

Before doing regressions, we have done Chi-Square test and plotted the coefficient heap map. Chi-Square test can be utilized to test whether two categorical variables are correlated, and the null hypothesis is they are not correlated. As shown in the result table 4-1, p values of the tests are less than 0.05, therefore we reject the null hypothesis and say the three categorical independent variables are correlated with diabetes.



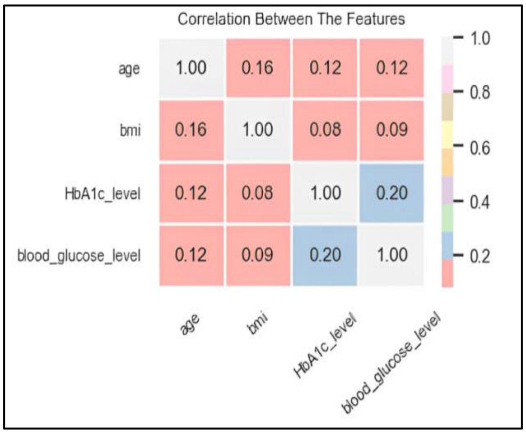
```
> chisq.test(table(diabetes$whether_heart, diabetes$whether_diabetes))

	Pearson's Chi-squared test with Yates' continuity correction

data:  table(diabetes$whether_heart, diabetes$whether_diabetes)
X-squared = 1712.7, df = 1, p-value < 2.2e-16
```

Table 5-1

It is shown in the graph 4-2 of heap map that the continuous independent variables are only slightly correlated, which means we can assume there is no high correlated relationship between continuous independent variables in our research.

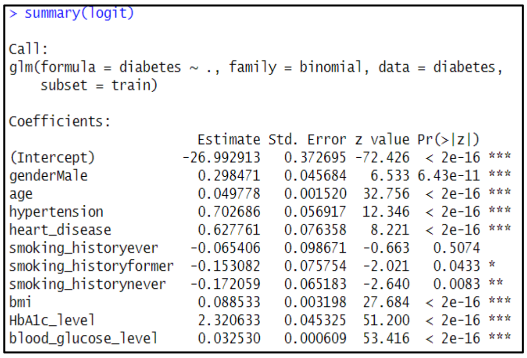


Graph 5-1

Logistic and Probit regression

Through the signs of coefficients in table 4-3 and 4-4, we can conclude that:

1. Nearly all the p values are statistically significant for they are less than 0.05.
2. The sign of coefficient of gender_male is positive, which indicates that males are more likely to develop diabetes than females.
3. The sign of coefficient that smoking history ‘never’ is negative, which means that people never smoke have lower likelihood of developing diabetes.
4. Coefficient of heart_disease is positive, which means that those who with heart disease more tend to get diabetes.

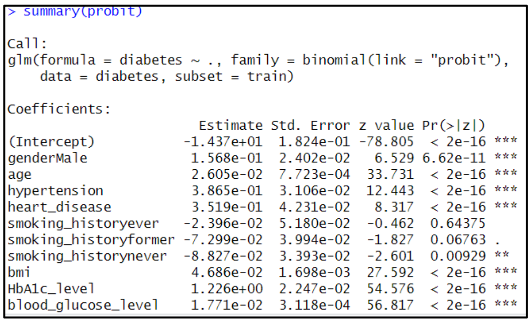


```
> summary(logit)

Call:
glm(formula = diabetes ~ ., family = binomial, data = diabetes,
    subset = train)

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)          -26.992913   0.372695 -72.426  < 2e-16 ***
genderMale             0.298471   0.045684   6.533 6.43e-11 ***
age                    0.049778   0.001520  32.756  < 2e-16 ***
hypertension           0.702686   0.056917  12.346  < 2e-16 ***
heart_disease          0.627761   0.076358   8.221  < 2e-16 ***
smoking_historyever   -0.065406   0.098671  -0.663   0.5074
smoking_historyformer -0.153082   0.075754  -2.021   0.0433 *
smoking_historynever  -0.172059   0.065183  -2.640   0.0083 **
bmi                    0.088533   0.003198  27.684  < 2e-16 ***
HbA1c_level            2.320633   0.045325  51.200  < 2e-16 ***
blood_glucose_level    0.032530   0.000609  53.416  < 2e-16 ***
```

Table 5-2



```
> summary(probit)

Call:
glm(formula = diabetes ~ ., family = binomial(link = "probit"),
    data = diabetes, subset = train)

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)           -1.437e+01  1.824e-01 -78.805  < 2e-16 ***
genderMale             1.568e-01  2.402e-02   6.529 6.62e-11 ***
age                    2.605e-02  7.723e-04  33.731  < 2e-16 ***
hypertension           3.865e-01  3.106e-02  12.443  < 2e-16 ***
heart_disease          3.519e-01  4.231e-02   8.317  < 2e-16 ***
smoking_historyever   -2.396e-02  5.180e-02  -0.462  0.64375
smoking_historyformer -7.299e-02  3.994e-02  -1.827  0.06763 .
smoking_historynever  -8.827e-02  3.393e-02  -2.601  0.00929 **
bmi                    4.686e-02  1.698e-03  27.592  < 2e-16 ***
HbA1c_level            1.226e+00  2.247e-02  54.576  < 2e-16 ***
blood_glucose_level    1.771e-02  3.118e-04  56.817  < 2e-16 ***
```

Table 5-3

Model Prediction Analysis

For model prediction, which is related to our fourth question, we use three metrics to measure the performance of different models in classification. They are accuracy, precision and recall: Accuracy is about overall correctness; Precision is about how many of the predicted positives are actually correct; Recall is about how many of the actual positives were successfully predicted.

To do the prediction, we cross-validated the model on 65 % of the test set and predicted the results with the remaining 35%.

The prediction we used are as followed:

(1) Linear Discriminant Analysis (LDA):

A class of approximations to the exchange –

correlation (XC) energy functional in density functional theory (DFT) that depend solely upon the value of the electronic density at each point in space.

(2) Quadratic Discriminant Analysis(QDA)

By calculating the distances between samples, the new sample is classified into the category of the K nearest training samples.

(3) Support vector machine

Its principle is to map the data into a high-dimensional space and find an optimal hyperplane that separates the data of different classes.

Hyperparameters: 'C': [0.1, 1, 10], 'kernel': ['linear', 'rbf', 'sigmoid'], 'gamma': [0.1, 1, 10]

(4) KneighborsClassifier

It classifies new samples into the category of the K nearest training samples based on the distance calculation between samples.

Hyperparameters: 'n_neighbors': [3, 5, 7, 10], 'weights': ['unoform', 'distance']

(5) Decision Tree

The principle of a decision tree is to recursively partition the data based on the features, in order to create a tree-like structure of decision nodes and leaf nodes.

Hyperparameters: criterion: ['gini', 'entropy']

(6) Random Forest Model

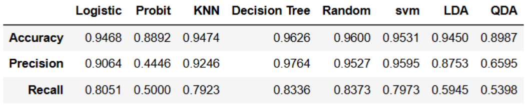
A bagging algorithm that constructs multiple decision trees during training and aggregates their predictions to improve accuracy and robustness of model.

Hyperparameters: criterion: ['gini', 'entropy']

The cross-validation accuracy of the random forest model is as high as 96.1% with a standard deviation of 0.002858. On the test set, the model achieves 96.03% accuracy, 95.45% precision, 83.73% recall and 88.46% F1 score.

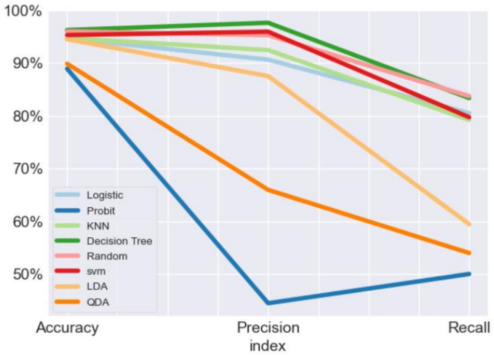
In the process of analyzing data using SVM, KNN, Decision Tree, Random Forest Model, grid search is first used to fit and find the optimal model through the training set, and then k=5 cross-validation is used to judge the goodness of fit of the model. Finally, the results of the test data are predicted, and then the confusion matrix and ROC curve are drawn and the accuracy, precision and recall of the prediction are calculated to judge the prediction ability of the model.

We used totally eight models for diabetes prediction analysis. The statistical indicators of all models can be plotted as shown in the figure below:



	Logistic	Probit	KNN	Decision Tree	Random	svm	LDA	QDA
Accuracy	0.9468	0.8892	0.9474	0.9626	0.9600	0.9531	0.9450	0.8987
Precision	0.9064	0.4446	0.9246	0.9764	0.9527	0.9595	0.8753	0.6595
Recall	0.8051	0.5000	0.7923	0.8336	0.8373	0.7973	0.5945	0.5398

Graph 6-1



Graph 6-2

By observing the three index, we can conclude decision Tree performed the best in predicting diabetes. Models like SVM and Random Forest, although more complex, did not perform as well. We thought this could be due to the limitations in the amount of data available and the high complexity of the models compared to the simplicity of the data features.

Summary of Research Questions

1. Men are more prone to diabetes than women.
2. Having a smoking history increases the probability of developing diabetes.
3. Heart disease increases the likelihood of developing diabetes.
4. On this dataset, the decision tree performs the best in predicting diabetes.

Outlook and Shortcomings

At the same time, we put forward two shortcomings of the research:

1. Some features in this dataset are not independent, as there may be a latent relationship between obesity and heart disease.
2. There may also be mediating effects between the independent variables and the dependent variable in this dataset. The future research can focus on these two aspects.

Appendix: task allocation table

Student Name	Student ID	Task
YAI Yuting	1155202878	Research Background，Questions，Report
LIU Xiaojia	1155203494	Dataset Description，Random Forest Model，PPT Slides
HONG Zhihui	1155208413	Regression Model， Data Visualization
YE Ruixin	1155206032	Regression Model， LDA and QDA Prediction Analysis
WANG Shuonan	1155201562	Model Construction，Regression Model，Prediction Analysis，Summary and Shortcomings