Group 7: Jiahe Huang, Jingjia Peng, Xinhe Wang

# 1 Introduction

Diabetes remains a significant social health challenge. In our project, we mainly focused on the following two aspects: Prevention and Early Detection of diabetes. To keep our conclusions relatively current, we used data from 2017 to March 2020 from the National Health and Nutrition Examination Survey [1]. To provide further details, our work focuses on two questions:

- Based on daily information such as lifestyle habits, how likely is a person to have diabetes?
- How to determine whether a person has diabetes based on basic urine physical examination data?

The first task focuses on the discovery and interpretation of lifestyle choices that can lower the risk of diabetes. The second task focuses mainly on feature engineering, which extracts important features from the dataset by developing prediction models for diabetes based on medical examination results.

# 2 Data

In this section, we first generally introduce the datasets we use; we also propose data preprocessing methods, which turn out to improve performance significantly. Considering the nature of NHANES datasets, each SEQN uniquely identifies an individual and remains consistent across all tables, which enables us to carry out complex data analysis using multiple datasets. The following sections discuss the datasets we used for the two problems.

## 2.1 Problem 1: Diabetes Prevention with Living Habits

### 2.1.1 Introduction to used datasets

For the prevention part, the supervisory learning task is based on the questionnaire Diabetes dataset and we divided people into two categories (healthy or with diabetes). We considered different predictors including age drinking habits, diet behavior, nephropathy background smoking habits, and weight status. The following table concludes the dataset we used for this problem.

| Name of Dataset | Description |
|---|---|
| P_DIQ.XPT | Dataset categorizing individuals as healthy or with diabetes |
| P_DEMO.XPT | Dataset including age and demographic variables |
| P_ALQ.XPT | Dataset on alcohol usage habits |
| P_DBQ.XPT | Dataset on diet behavior and nutrition |
| P_KIQ_U.XPT | Dataset on nephropathy background and kidney conditions |
| P_SMQ.XPT | Dataset on smoking habits and cigarette use |
| P_WHQ.XPT | Dataset on weight history and status |

Table 1: Description of Datasets for the first problem

### 2.1.2 Data Preprocessing methods

After merging the datasets based on "SEQN" (ID of participants) using inner join, we carry out research using two different approaches: the first one is to carefully select some variables that are supposed to be highly correlated with diabetes, and further do careful data preprocessing. The second method is to do minimal things on data preprocessing, which includes only dealing with NA values using Predictive Mean Matching. For the first method, the variables we selected are concluded in Table 2. The reason we carry out such data preprocessing is to mainly address the concern of connotations of values. For example, for the variable of drinking frequency, originally 0 represents never, 1 represents every day, and 2,3 follows by indicating frequencies in between. Even such values can be regarded as factors, listing them in a strictly monotonic order makes understanding and explanation easier.

| Variable Name | Meaning | Value | Description |
|---|---|---|---|
| DIQ010 | **Our Result**: If has diabetes | 0 | No |
| | | 1 | Yes |
| RIAGENDR | Gender | 0 | Female |
| | | 1 | Male |
| RIDAGEYR | Age | $\geq 80$ | Coded as 80 |
| ALQ121 | Drinking frequency | 1-11 | Higher value indicates higher frequency |
| DBQ700 | Is diet healthy | 1-5 | Lower value indicates healthier diet |
| KIQ022 | Is kidney weak | 0 | No |
| | | 1 | Yes |
| KIQ026 | If had kidney stone | 0 | No |
| | | 1 | Yes |
| SMQ020 | If smoked 100 cigarettes | 0 | No |
| | | 1 | Yes |
| WHQ030 | Weight | -1 | Underweight |
| | | 0 | Fit |
| | | 1 | Overweight |

Table 2: Table of Variable Encoding and Description

The following figure gives a basic visualization by considering the pairing relationship of the variables. We can notice from the plot that the results are correlated with the results.

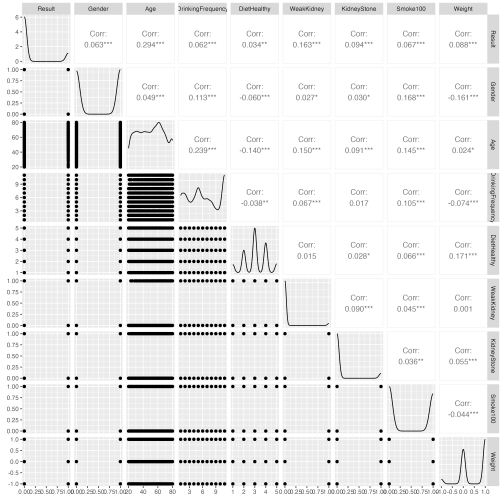


Figure 1: Visualized adjusted dataset.

## 2.2 Problem 2: Diabetes prediction through medical laboratory test

### 2.2.1 Introduction to used datasets

To determine whether a person has diabetes based on basic urine physical examination data, we merge multiple datasets containing urine physical examination data is combined as shown in Table 3. The response is generated based on the variable Fasting Glucose (mg/dL) in P_GLU.xpt. If this value is larger than 126, then one can be classified as having diabetes, and the response is 1, otherwise the response is 0.

| **File Name** | **Description** |
|---|---|
| P_GLU.XPT | Glucose Dataset |
| P_ALB_CR.XPT | Albumin & Creatinine - Urine |
| P_UTAS.XPT | Arsenic - Total - Urine |
| P_UAS.XPT | Arsenic - Speciated - Urine |
| P_UCM.XPT | Chromium - Urine |
| P_FR.XPT | Flame Retardants - Urine |
| P_UIO.XPT | Iodine - Urine |
| P_UHG.XPT | Mercury: Inorganic - Urine |
| P_UM.XPT | Metals - Urine |
| P_UNI.XPT | Nickel - Urine |
| P_PERNT.XPT | Perchlorate, Nitrate & Thiocyanate - Urine |
| P_UCFLOW.XPT | Urine Flow Rate |
| P_UVOC.XPT | Volatile Organic Compound (VOC) Metabolites - Urine |

Table 3: Description of Urine-Related Laboratory Result Datasets

### 2.2.2 Data preprocessing

It is worth noticing that, in many datasets included in this section, there is a variable named 'WTSAPRP', which indicates the weight of the observation. Since we are merging multiple datasets and the weights vary across tables, we discard this variable for simplicity, which means we consider that each observation shares the same weight.

Moreover, imputation is carefully considered in this problem. The columns containing only NA are dropped. Before a dataset is merged, the NAs are imputed by a 5-nearest neighbor. After the datasets are merged, as we are using outer join in this problem, the missing values are imputed with Predictive Mean Matching. As the dimension of data is still large after preprocessing, we are mostly concerned about the correlations between predictors and the response. We pick the first 9 predictors and plot their value against the response variable diabetes.

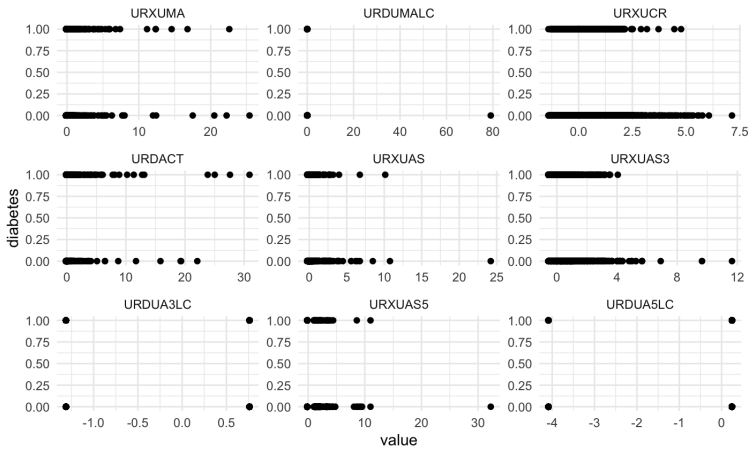


Figure 2: Plot between Predictors and the Diabetes Response Variable

We shall see that while some plot, such as "URDUMALC" seems to reveal that the response will always be 0 when this predictor is large, some plots don't reveal much information. For example, the plot of "URDUA3LC" and diabetes just shows four corners, and we could not notice any tendency just from the plot. Such a phenomenon calls for detailed analysis, as we will introduce below.

# 3 Methods

## 3.1 Diabetes prevention with living habits

To provide an accurate and explainable result, we use methods including LDA, QDA, and decision trees to address the problem. The reason why these three methods are as follows:

- LDAs and QDAs are suitable for classification problems that correlations of variables may not be investigated in detail before. As these methods explicitly consider correlations, they could provide into insights the data without complex attempts. Moreover,

empirically speaking, in this case, testing the hypothesis of equal covariance matrices across groups is not significantly cheaper than carrying out a QDA attempt. Therefore we choose to carry out both methods and compare the results.

- LDA and decision tree can generate **explainable** results. As our main goal is to provide insights into the influence of habits on the risk of diabetes, an explainable result is at the center of our research.

The general setup of all three methods is similar. We split the training data set by 80%. To make the model more stable, Cross-validation is introduced to prevent overfitting and tune parameters. We iterated over each value from 2 to 10 as the number of cross-validation folds and plotted corresponding training errors and testing errors of three model types in the same figure (Section 4.1). This is a classification problem so accuracy was chosen as the evaluation metric. Accuracy is defined as Equation 1:

$$Accuracy = \frac{\#correct\ testing\ prediction}{\#testing\ prediction} \tag{1}$$

Eventually, the approach with the highest testing accuracy thus the least testing error would be chosen as our final model.

To further examine the model we developed, which selects features in advance, we consider another approach using "method 2" mentioned in 2.1.2. After doing minimal data preprocessing, we directly employ LDA on the high dimensional data. We then sort the features by the magnitude of coefficients and extract the top 8 features. Using these features, we obtain another model, which can be used to compare with the feature selection approach.

## 3.2 Diabetes prediction through medical laboratory test

To carry out the prediction job based on pre-processed laboratory data, we use a logistic lasso, ridge regression, and random forests. The reasons are as follows

- These models generate **explainable** result. To make our study empirically easy to explain, we want to extract certain features so that people can directly reflect on their urine test results to get valuable results.

- These methods do feature selection effectively. By introducing $\ell_1$ or $\ell_2$ regularization terms, important features can be extracted from a large amount of features. For random forests, by considering the coefficients assigned to each feature, we could also extract the most influential features. **To effectively carry out this consideration, data are scaled to mean 0 and variance 1** so that different features will have an equal opportunity to contribute.

Now we introduce the setup of these methods. For all methods including logistic ridge/lasso regression and random forests, we use a $0.7 : 0.3$ train-test split of data. The three methods will use the same data for training and testing so that comparison across models can be performed fairly. The model's performance is evaluated also by accuracy, defined by Equation 1. For logistic ridge/lasso regression, cross-validation is used to select the best model.

# 4 Results

## 4.1 Diabetes self-prediction with living habits

Based on the methods in Section 3.1, we obtained the plot as Figure 3 by the feature selection method.

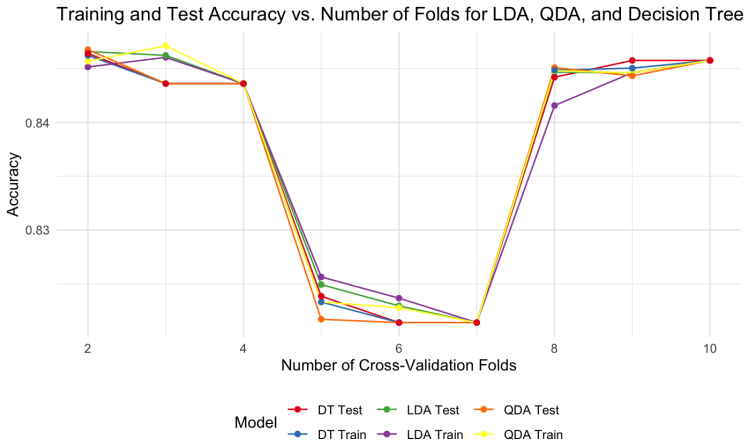


Figure 3: Training and testing accuracy vs. CV fold for LDA, QDA, and decision tree with all selected features.

We shall see that the three methods all have a good performance, and since QDA is not better than LDA, we focus on LDA for explanation. The LDA model with the best test accuracy occurs when the number of folds is 2, and the model can be concluded by

$$y = 0.342x_1 + 0.049x_2 - 0.020x_3 + 0.175x_4 + 1.744x_5 + 0.580x_6 + 0.126x_7 + 0.396x_8 - 3.779937 \tag{2}$$

From the model, we can see that "WeakKidney", "KidneyStone", and "Weight" are the most important features. We also spotted that the coefficient of $x_3$ (drinking frequency) is slightly negative. We could interpret it as of little significance, while the following source [2] could suggest that moderate consumption of alcohol **may** reduce the risk of diabetes.

One can self-predict the risk of having diabetes based on Equation 2. $y$ is typically between 0 and 1. The closer $y$ is to 1, the higher the risk it has. If one gets a high $y$, further medical laboratory tests are suggested.

We've also developed an LDA model that is based on minimum data preprocessing. The result is shown in Figure 4. The accuracy of this model is lower than the model with careful data pre-processing, which suggests our selection improves the model. The decision tree model shown in Figure 5 gives also an insight. The flexibility of the decision tree allows interpreting variables in another way: younger people may have a lower risk of having diabetes, while elder people may have a higher probability of diabetes if the kidney is not strong. By the nature of the decision tree, one can easily estimate the risk of diabetes by following the tree.

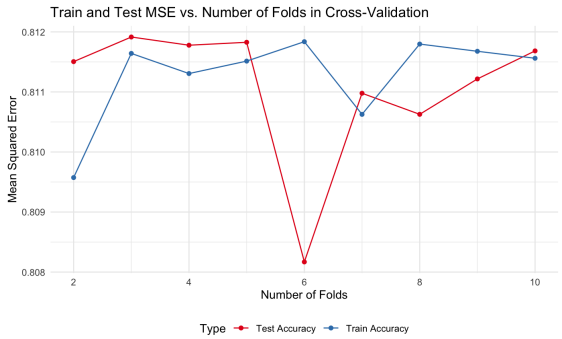


Figure 4: Minimal Data preprocessing LDA model

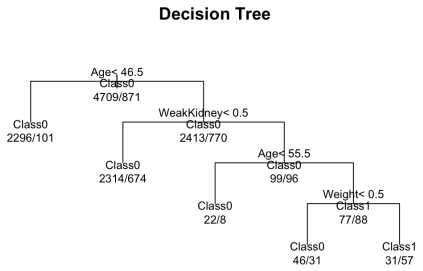


Figure 5: Decision Tree result

## 4.2 Diabetes prediction through medical laboratory test

We first provide the accuracies of the three models developed for this problem: the logistic lasso regression, the logistic ridge regression, and the random forest.

| Lasso | Ridge | Random Forest |
|---|---|---|
| 0.7574 | 0.7522 | 0.8672 |

Table 4: Problem 2 Model Accuracies

We shall see that the three methods have good accuracies, while the random forest methods significantly stand out. Now to further address the most important features, we extract the features with the largest magnitude of coefficients for all three models.

| Rank | Ridge | Lasso | Random Forest |
|---|---|---|---|
| 1 | URXPMM | **URDACT** | **URDACT** |
| 2 | URDUMMAL | URDUMMAL | URXUMA |
| 3 | **URDACT** | URXUDMA | URXUDMA |
| 4 | URXAAM | URDHEMLC | URXSCN |
| 5 | URXDPHP | URDCYALC | URXNO3 |

Table 5: Top 5 Features for Ridge, Lasso, and Random Forest Models

Although checking all these variables from a medical perspective goes beyond the discussion of this report, we could take a look at the top variable for the best model **URDACT**, we can conclude that it is important from a statistical perspective, and it is indeed, key data to detect diabetes according to National Health Service [3], which suggests that our method does extract the important features. To use the model developed, one can choose to run the model on their data to get a prediction, and they can also reflect on the most important features we've shown.

# 5 Conclusions

In this project, we apply LDA, QDA, and decision tree to address the concerns of prevention of diabetes based on daily information. We also apply logistic lasso/ridge regression and random forest to carry out feature engineering on urine data to select features that could indicate diabetes. Our answer to both questions not only provides models that could help statisticians analyze new data but also provides explainable results, especially features, that can direct people to prevent and detect diabetes.

We propose the following concerns that may cause our analysis to be misleading from both statistical and medical perspectives: Firstly, for both parts, we don't carry out detailed investigation on the meaning of variables. We cannot fully get rid of the risk of wrongly using value. Secondly, the assumptions of LDA are not checked in detail. Thirdly, the observation's weight is not considered, which means we carry on the research under the simplification that all observations contribute equally.

# 6 Contributions and Reproducibility

All group members equally contribute to the whole project.

- **Jiahe Huang:** She wrote codes for the first question and finished the corresponding parts in this report. She also designed the neural network structure in the Kaggle competition.
- **Jingjia Peng:** She wrote codes for the Kaggle competition and finished the Kaggle report. She also helps with the revision of reports.
- **Xinhe Wang:** He finalized codes for question 1 and wrote codes for question 2 of the open-ended report. He also revised question 1 of this report and wrote the remaining parts of this report.

All results shown in this report can be reproduced with a knit of Rmd files after the working directory is changed to yours. We provide two Rmd files in canvas submission stats415proj1part1.Rmd and stats415proj1part2.Rmd, which corresponds to results of the two problems. We also provide corresponding .html files for readers' direct reference. The second problem focuses on feature engineering, so the complete model is left for your reference in the HTML file because of its large number of parameters ( more than 100 for lasso).

# References


[1] CDC. "NHANES - National Health and Nutrition Examination Survey Homepage." *Centers for Disease Control and Prevention*, 2019, www.cdc.gov/nchs/nhanes/index.htm.

[2] Boston, 677 Huntington Avenue, and Ma 02115 +1495-1000. "Moderate Alcohol Intake May Decrease Men's Risk for Type 2 Diabetes." *News*, 15 Feb. 2011, www.hsph.harvard.edu/news/features/moderate-alcohol-intake-may-decrease-mens-risk-for-type-2-diabetes/.

[3] NHS, "ACR Test." https://www.nhs.uk/conditions/acr-test/.