PM566 final project: Prediction model of Diabetes

Xiaoyu Zhu

Introduction

Diabetes is a chronic (long-lasting) health condition that affects how your body turns food into energy. Most of the food you eat is broken down into sugar (also called glucose) and released into your bloodstream. When your blood sugar goes up, it signals your pancreas to release insulin. The growing human and economic toll of diabetes has caused consternation worldwide. Not only is the number of people affected increasing at an alarming rate, but onset of the major forms of the disease occurs at ever younger ages. We now know that the reach of diabetes extends far beyond the classic acute metabolic and chronic vascular complications to increased risk of an ever-increasing array of conditions including Alzheimer disease, cancer, liver failure, bone fractures, depression, and hearing loss.In the U.S. one in three Medicare dollars is spent on care of people with diabetes, and the proportion of cardiovascular disease (CVD) risk attributable to diabetes is rising.

However, diabetes was being underestimated for its importance, and always being analysed as an effect rathan than results.

Data source

The data is from Harvard dataverse, the website is https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/JAW6AX&version=2.0 (https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/JAW6AX&version=2.0).

Examine variables

Lipotoxicity (1-17): lipotoxicity is a metabolic syndrome that results from the accumulation of lipid intermediates in non-adipose tissue, leading to cellular dysfunction and death.

Glucose (mmol/L): blood glucose level obtained by measuring plasma glucose concentration at 2 hours in an oral glucose tolerance test.

BloodPressure (mm Hg): the pressure of the blood in the circulatory system.

SkinThickness (mm): skin thickness is primarily determined by collagen content and is increased in insulin-dependent diabetes mellitus.

Insulin(μU/ml): insulin is an anabolic hormone that promotes glucose uptake.

BMI: body mass index (BMI) is a person’s weight in kilograms divided by the square of height in meters.

DiabetesPedigreeFunction (0:1 value generated from familial diabetes history/risk): diabetes pedigree function provides “a synthesis of the diabetes mellitus history in relatives and the genetic relationship of those relatives to the subject.” It generally provides scores of the likelihood of diabetes based on family history. The DPF uses information from parents, grandparents, siblings, aunts and uncles, and first cousins. It provides a measure of the expected genetic influence of affected and unaffected relatives on the subject’s eventual diabetes risk.

Age: age of the individual.

Outcome: diabetes test result (0 = Non-diabetic, 1 = Diabetic).

Key questions

Is there a significant difference in values of diabetes risk factors for those who have diagnosed with diabetes and those who are not? Is Diabetes Pedigree Function significantly associated with the onset of diabetes and other risk factors? Could these risk factors provide a reliable prediction of individual’s diabetes?

Methods

Analysis through assessing various plots, tables and graphs was performed to identify association between Diabetes Pedigree Function and diabetes test outcome, including the examination of the effect of lipotoxicity, glucose level, blood pressure, skin thickness, insulin level, BMI, and age. The data was cleaned by replacing extreme values to ’NA’s, shortening variable names, and creating new factor variable for better analysis. Skim function from skimr package was used to explore data. Dim, head and tail, summary, and table functions were used to check detailed observations. Age was stratified into four age groups(20-29, 30-39, 40-49, and 50+) for better understanding of relationship with diabetes. Outcome was binomial variable(0, 1) and it was transformed into factor variable (Non-diabetic and Diabetic).

Data Cleaning

- 0 values in all risk factors are speculated as an absence of the specific test outcome because there were significant difference in values between all 0s and the next minimum values of each variable. All of them are edited to “NA”s.
- Variable names are renamed into lowercase letters with shorter length.
- Created age groups (20-29, 30-39, 40-49, 50+) to compare proportion of the diabetes by age groups.

Show 10 entries

Search:

	lip	glu	bp	st	ins	bmi	dpf	age	outcome	age_group
1	6	148	72	35		33.6	0.627	50	1	50+
2	1	85	66	29		26.6	0.351	31	0	30-39
3	8	183	64			23.3	0.672	32	1	30-39
4	1	89	66	23	94	28.1	0.167	21	0	20-29
5		137	40	35	168	43.1	2.288	33	1	30-39
6	5	116	74			25.6	0.201	30	0	30-39
7	3	78	50	32	88	31	0.248	26	1	20-29
8	10	115				35.3	0.134	29	0	20-29
9	2	197	70	45	543	30.5	0.158	53	1	50+
10	8	125	96				0.232	54	1	50+

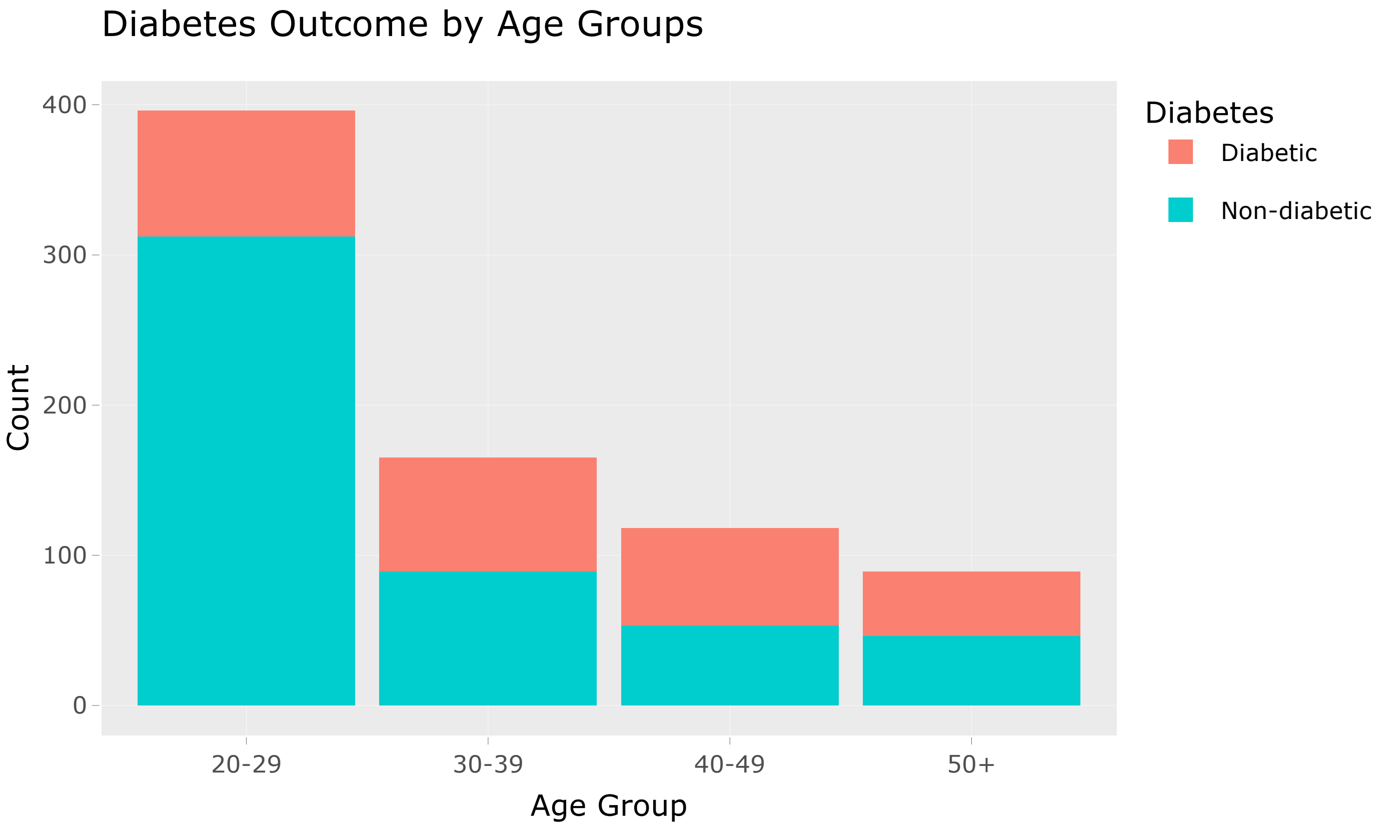
Showing 1 to 10 of 768 entries

Previous 1 2 3 4 5 … 77 Next

Results

Histogram of Diabetes outcome by age groups

- Age is a significant risk factor for diabetes. As age is a well-known confounder of most diseases, it could also play a role as a confounder when generating a prediction model. Through this histogram, I can confirm that age affects the onset of diabetes.



Mean Values of Each Risk Factors by Outcome Group

- All of the predictors are showing some differences in mean values by diabetes outcome, meaning that these predictors can be utilized for a prediction model.
- Diabetes test outcome: 0 (Non-diabetic), 1 (Diabetic).

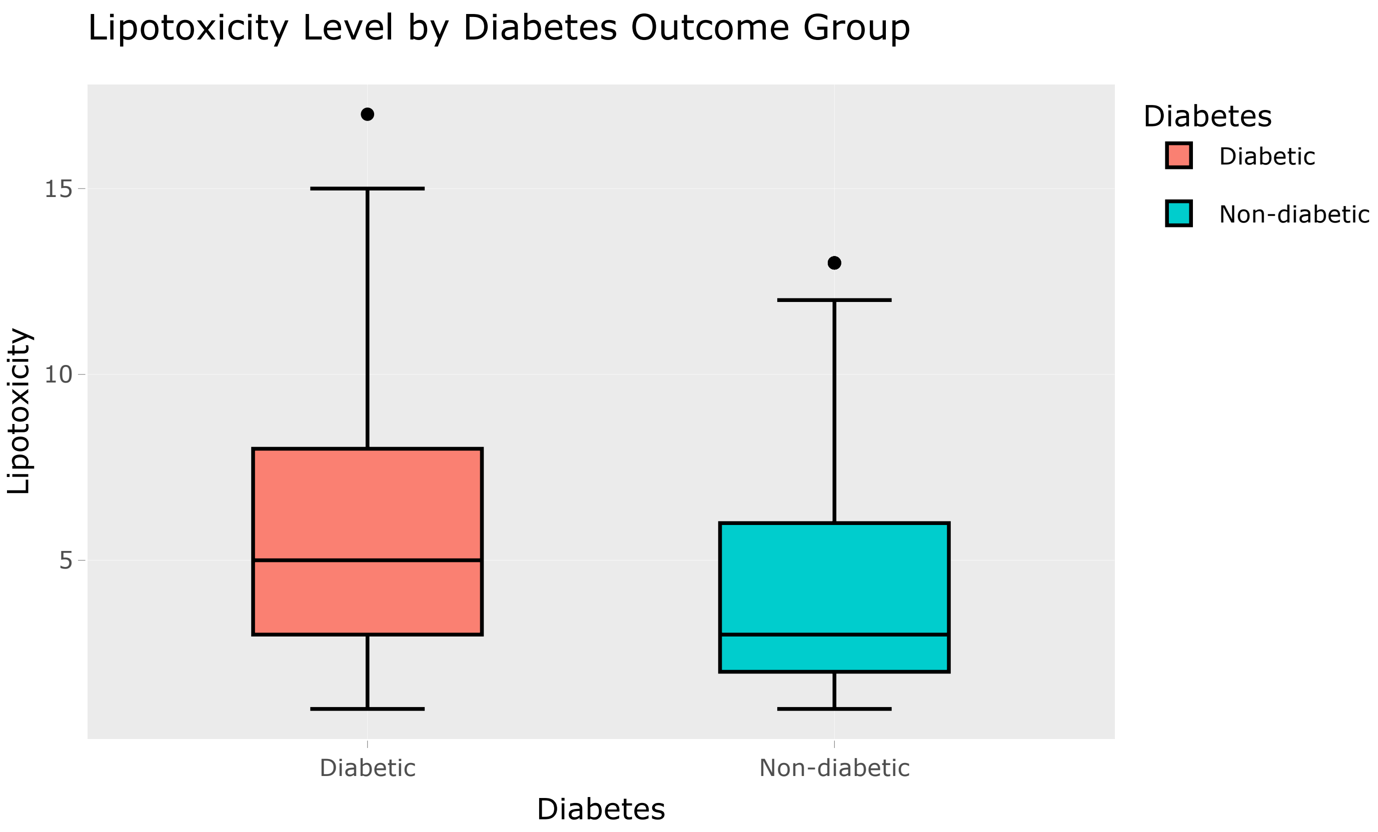
Mean of Each Risk Factors by Outcome Group

outcome	DiabetesPedigreeFunction	Lipotoxicity	Glucose	BloodPressur	SkinThickness	Insulin	BMI	Age
0	0.43	3.86	110.64	70.88	27.24	130.29	30.86	31.19
1	0.55	5.67	142.32	75.32	33.00	206.85	35.41	37.07

Box Plot of each risk factors by outcome

- Among the predictors, Glucose Level, Insulin level, and BMI show significant differences in mean values by outcome group.
- Insulin level is a direct indicator of discernment between type 1 and 2 diabetes. This predictor can be treated differently by the type of diabetes. In this dataset, it is presumed that there were more type 2 diabetes in the patient group as the result shows that mean insulin level is higher in the diabetic group.
- Although lipotoxicity is showing a meaningful gap between the diabetes and non-diabetes group, it will be excluded from further analysis because the data has a high NA proportion and the test for measuring lipotoxicity is not common.
- Skin thickning is a symptom detected from patients with insulin-dependent diabetes mellitus (IDDM). It means this data only applies to type 1 diabetes.
- Blood Pressure does not show significant difference by diabetes outcome.

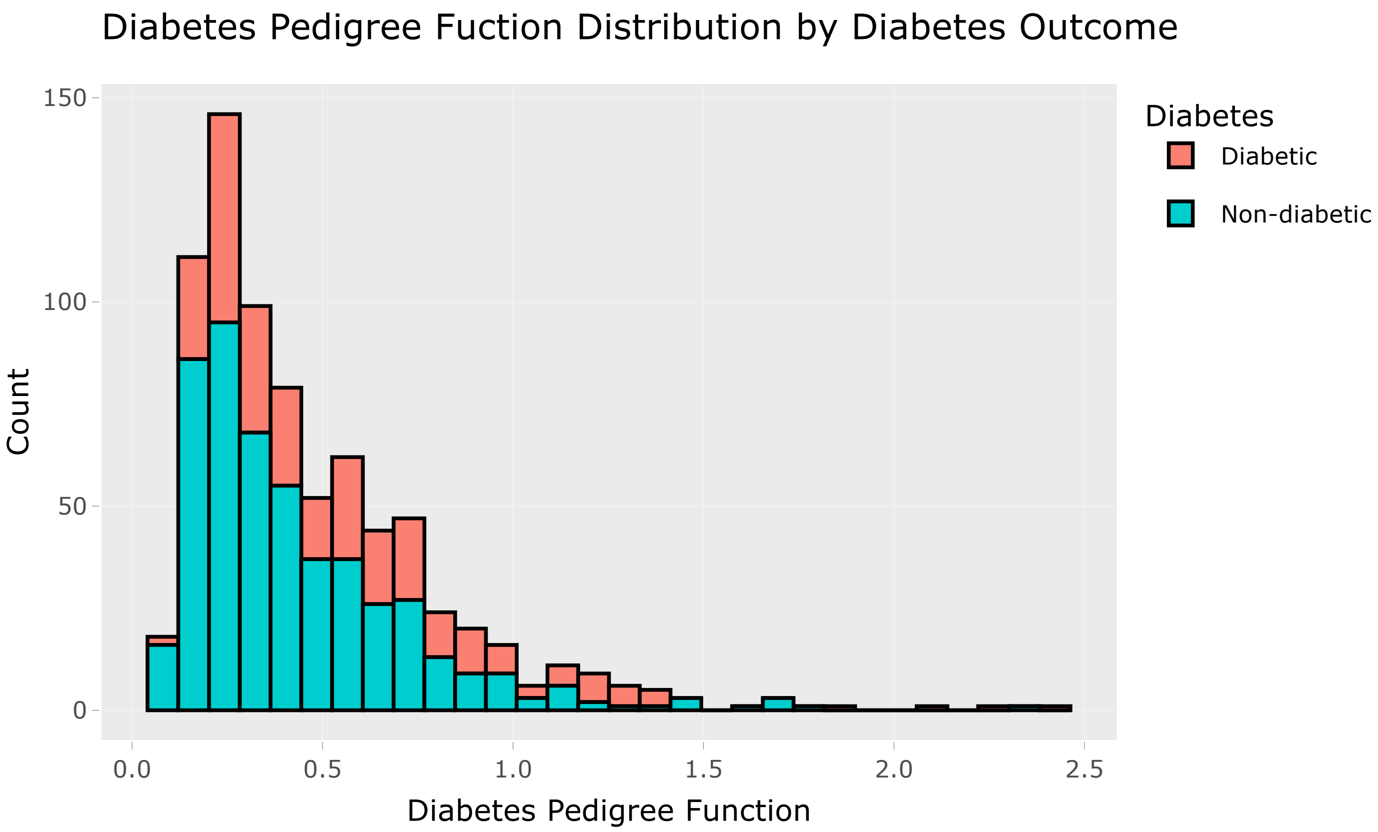
Lipotoxicity Glucose Level Blood Pressure Skin Thickness Insulin Level BMI



Plots of Diabetes Outcome by Diabetes Pedigree Function

- The histogram shows that Diabetes Pedigree Function follows Poisson distribution. Poisson regression could be used for further analysis.
- The proportion of diabetic outcome increases over Diabetes Pedigree Fucgion. From dpf 0.24 to 0.48 section, proportion of diabetic vs non-diabetic is approximately 1:2. In 0.56 to 0.72 section, the ratio is 2:3 and in 0.8 and over section, the ratio become close to 1:1.
- Boxplot also shows that there is a significant difference in Diabetes Pedigree Function value between diabetic and non-diabetic group. There is more research needed for the outliers without diabetes.

Histogram Boxplot

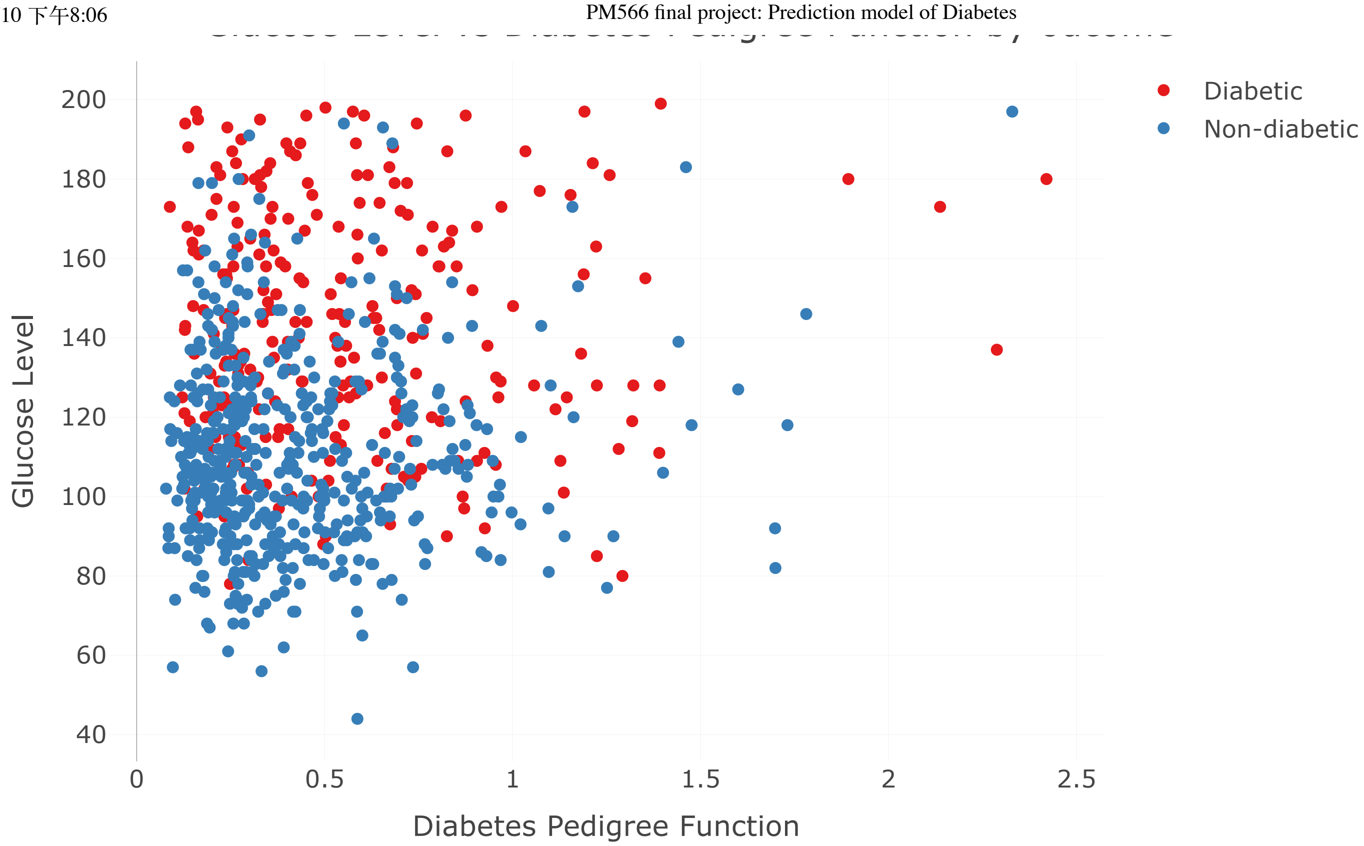


Scatter Plot Graph of Significant Risk Factors vs Diabetes Pedigree Function by Diabetes Outcome Group

- Glucose level and diabetes pedigree fuction interact well to distinguish diabetic section and non-diabetic section.
- Insulin level is not effective on marking off diabetic section in this scatter plot. However, it could be different when it only applies to the dataset of patients who entirely has one of two types of diabetes.
- BMI is also a good partner of diabetes pedigree function.

Glucose vs DPF Insulin vs DPF BMI vs DPF

Glucose Level vs Diabetes Pedigree Function by oucome



Prediction Model

- Based on th analysis above, prediction model was built including DPF, lipotoxicity, glucose level, insulin levle, BMI, and age as predictors. The final training model concluded with around 80% Accuracy, which indicates good prediction ability of the model. Along with the good accuracy value, around 60% Sensitivity, 90% Specificity, 70% Positive Prediction Value, and 80% Negative Prediction Value was obtained from the model.

```
## 
## 
##    Cell Contents
## |-------------------------|
## |                       N |
## |         N / Table Total |
## |-------------------------|
## 
## 
## Total Observations in Table:  768 
## 
## 
##           |         0 |         1 | 
##           |-----------|-----------|
##           |       500 |       268 | 
##           |     0.651 |     0.349 | 
##           |-----------|-----------|
## 
## 
## 
## 
```

```
## 
## Confusion Matrix and Statistics
## 
##           Reference
## Prediction   1   0
##          1  52  15
##          0  25 147
##                                          
##                 Total n : 239
##                Accuracy : 0.8326         
##                  95% CI : (0.7801, 0.8746)
##     No Information Rate : 0.6778         
##     P-Value [Acc > NIR] : 4.66e-08       
##                                          
##                   Kappa : 0.6033         
##  Mcnemar's Test P-Value : 0.1547         
##                                          
##             Sensitivity : 0.6753         
##             Specificity : 0.9074         
##          Pos Pred Value : 0.7761         
##          Neg Pred Value : 0.8547         
##              Prevalence : 0.3222         
##          Detection Rate : 0.2803         
##    Detection Prevalence : 0.2176         
##       Balanced Accuracy : 0.7914         
##          F-val Accuracy : 0.7222         
##      Matthews Cor.-Coef : 0.6063         
##                                          
##        'Positive' Class : 1              
```

Conclusion

Overall, most of the predictors displayed differences in mean values when it compares between diabetic and non-diabetic groups. The interesting predictor was diabetes pedigree function because, unlike other risk factors, DPF could be measured by relatives history and genetic data. This is the only predictor that can be obained by external sources other than individual’s biological test result. In the analysis, diabetes pedigree function showed its association with the onset of diabetes. Although it is hard to predict the risk of diabetes with diabetes pedigree fuction alone, the analysis showed a possibility of utilization of other risk factors combined with diabetes pedigree funtion in prediction model. In this study, glucose level and BMI provided evidence of significant association with onset of diabetes in conjunction of diabetes pedigree function. Insulin level and lipotoxicity also showed some association to be part of prediction model. Through this study, I found out that prediction model for diabetes could be built with relevant predictors such as diabetes pedigree function, glucose level, insulin level, lipotoxicity, BMI, and age. Further study will be needed with larger dataset for better accuracy. Overall, this study can be a good starting point of further research.