Diabetes Research Final Project

Team 21 - Francis Chan, Evan Lum, Dylan Mochizuki, Gina Pak, Elise Pham, Khang Thai

2025-03-20

1. Introduction

Diabetes is a chronic disease characterized by high levels of glucose in the blood and linked to increased risk of damage to the eyes, kidney, nerves, and heart. The development of this condition is influenced by a variety of genetic, behavioral, and physiological factors. As a major public health concern, there is a growing effort to identify the factors contributing to its diagnosis. Among the key metrics used to assess diabetes risk are body mass index (BMI) and glycated hemoglobin (HbA1c) levels, both of which provide valuable insights into an individual's metabolic health. Furthermore, familial predisposition and age are well-documented risk factors that require deeper exploration to understand their roles in modifying diabetes risk.

This study seeks to research: Which factors interact and influence the likelihood of a diabetes diagnosis? We seek to gain valuable insights into the multifaceted elements contributing to diabetes, with potential implications for tailored prevention and intervention strategies.

2. Methods

Data Collection and Source

Our dataset used in this study consists of 9,538 medical records containing demographics, lifestyles, and biometric health factors related to diabetes diagnosis. The dataset is a secondary dataset and was likely collected from medical records, health surveys or electronic health systems. The data contains structured variables, including numerical values of glucose, BMI, HbA1c, etc., categorical variables for family history, medication use, and hypertension. We did not need to do any further cleaning with our dataset including removing NA values, filtering out impractical variables, or transforming any outliers.

Research Design

Our research aims to investigate the factors that contribute to diabetes diagnosis and examine how BMI, HBA1c, and family history interacts with diabetes risk. An observational study design will be used with a secondary dataset containing medical records of individuals. We will use statistical modeling to identify relationships between variables and diabetes diagnosis. The primary research questions include evaluating the combined effect of BMI and HbA1c on the likelihood of diabetes diagnosis and investigating whether family history changes the age-diabetes risk relationship. By doing so, we will be able to identify the significant factors and potential interaction effects that influence diabetes risk.

3. Analysis

We begin our project by running by first checking the VIFs of each predictor to check for multicollinearity. We find that both HbA1c and glucose have significantly high VIF scores.

This is the model without glucose.

```
##                            GVIF Df GVIF^(1/(2*Df))
## Age                    3.670190  1        1.915774
## Pregnancies            1.006458  1        1.003224
## BMI                    3.173633  1        1.781469
## BloodPressure          1.595918  1        1.263296
## HbA1c                  1.025278  1        1.012560
## LDL                    1.007540  1        1.003763
## HDL                    1.009917  1        1.004946
## Triglycerides          1.019409  1        1.009658
## WaistCircumference    74.626529  1        8.638665
## HipCircumference      47.807387  1        6.914289
## WHR                   67.792078  1        8.233594
## factor(FamilyHistory)  1.000001  1        1.000001
## factor(DietType)       1.025875  2        1.006407
## factor(Hypertension)   1.054709  1        1.026990
## factor(MedicationUse)  3.357799  1        1.832430
```

This is the model with glucose.

```
##                             GVIF Df GVIF^(1/(2*Df))
## Age                    13.155835  1        3.627097
## Pregnancies            21.132087  1        4.596965
## BMI                    10.939026  1        3.307420
## BloodPressure          25.265035  1        5.026434
## HbA1c                  21.108795  1        4.594431
## LDL                     6.697213  1        2.587897
## HDL                     8.717548  1        2.952549
## Triglycerides          19.742900  1        4.443298
## Glucose               324.742931  1       18.020625
## WaistCircumference   3278.810045  1       57.260895
## HipCircumference     1761.242794  1       41.967163
## WHR                  6124.710248  1       78.260528
## factor(FamilyHistory) 322.812777  1       17.966991
## factor(DietType)      209.754200  2        3.805640
## factor(Hypertension)    1.001004  1        1.000502
## factor(MedicationUse)  14.577590  1        3.818061
```

This implies that they are highly correlated and that it will be a challenge to assess which individual predictor is affecting the outcome. After doing some research we find that HbA1c is actually a measurement of the average glucose. This new definition explains the correlation between the two predictors. To account for this we decide to simply remove glucose from our model. We then run a logistic regression model to gain a better understanding of the rest of the predictors.

Next we use stepwise regression to find the best subset of predictors that will contribute the most in predicting the outcome. We came to the conclusion that the best model was with Age, BMI, HbA1c and Family History. Although every predictor in our reduced model was significant except for Family History, because it was included in the model we felt that it still played a role in predicting the outcome.

We then compare both the models using ANOVA that results in a non significant p-value. This means that the reduced model does not differ that much compared to the full model, implying that the reduced model works just as well as the full model and that the subset of variables is enough.

```
## Analysis of Deviance Table
##
## Model 1: Outcome ~ Age + BMI + HbA1c + factor(FamilyHistory)
## Model 2: Outcome ~ Age + Pregnancies + BMI + BloodPressure + HbA1c + LDL +
##     HDL + Triglycerides + WaistCircumference + HipCircumference +
```

```
##     WHR + factor(FamilyHistory) + factor(DietType) + factor(Hypertension) +
##     factor(MedicationUse)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      9533     1341.0
## 2      9521     1336.2 12    4.871   0.9622
```

We wanted to further validate our reduced model by calculating the VIF of the predictors where we saw low signs of multicollinearity.

```
##                  Age                HbA1c factor(FamilyHistory)
##             1.115542             1.002088              1.000001
##                  BMI
##             1.117578
```

Model Evaluation

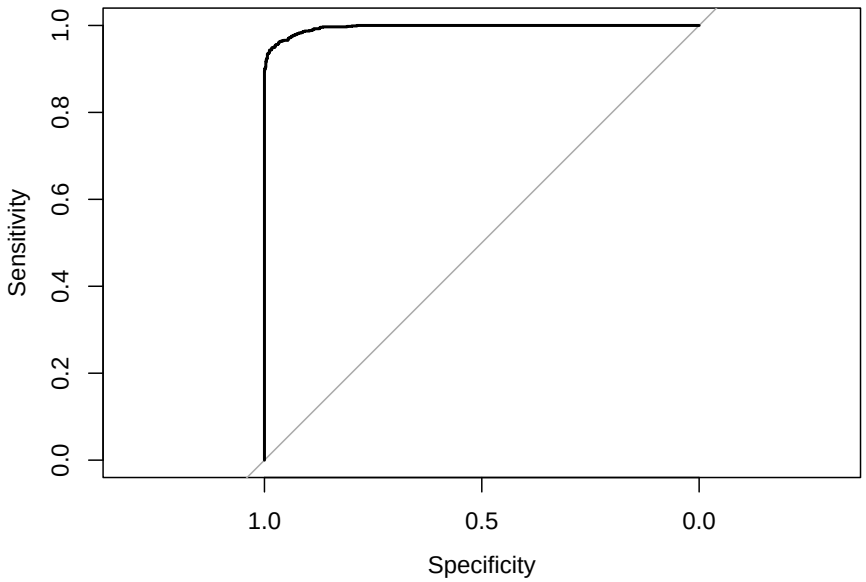
To assess the performance and reliability of our diabetes prediction model, we employed train-test split validation and 5-fold-cross-validation methods.

First, we performed a train-test split on the dataset by dividing it into 70% training data and 30% testing data. Using the glm() function in R, we trained a logistic regression model on the training data using our finalized model which includes BMI, HbA1c, and FamilyHistory as the predictors. We then applied the model to the testing data using the predict() function. We generated a confusion matrix to evaluate its performance. The confusion matrix revealed that the model achieved 97% accuracy, indicating that it correctly predicted and classified 97% of the cases in the test set. The high accuracy demonstrated the model's strong predictive capability in distinguishing between individuals with and without diabetes.

To further visualize the model's effectiveness, we plotted the Receiver Operating Characteristic (ROC) curve.

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```




```
## Area under the curve: 0.9959
```

The ROC curve above showed an area under the curve (AUC) of 0.99. An AUC close to 1.0 reflects excellent model performance and signifies that the model is highly effective in identifying both positive and negative

cases.

Next, to further validate the model and reduce the risk of overfitting, we performed 5-fold cross-validation. First, we divide the data into five equal subsets (folds). The model is then trained on four folds and tested on the remaining fold, and this process is repeated five times, which each fold serving as the testing set once. The performance metrics are average across all iterations. We chose k = 5 because it strikes a balance between bias and variance. Fewer folds could lead to a biased model while more folds could make the process computationally expensive.

The results of the 5-fold cross-validation shows that the model performed consistently well across all folds. The Root Mean Squared Error (RMSE) was approximately 0.146, indicating that the deviation between the predicted and the actual values is minimal, which reflects strong predictive accuracy. Meanwhile, the R-squared value of 0.906 suggests that the model explains approximately 90.6% of the variance in diabetes outcomes, demonstrating a strong fit. Additionally, the Mean Absolute Error (MAE) was 0.04, indicating that, on average the model's predictions deviated by only 0.04 units from the actual values, highlighting the precision of this predicting model.

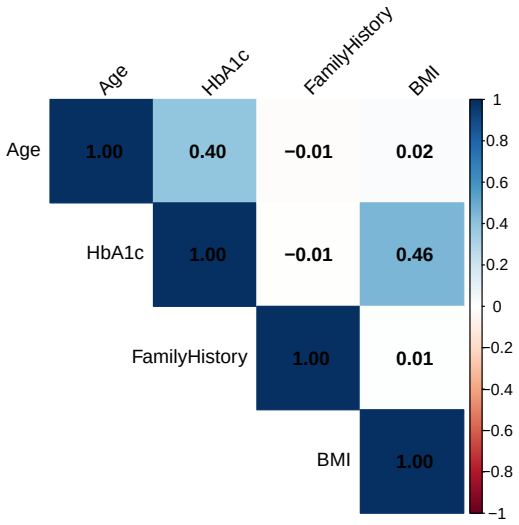
```
## Generalized Linear Model
##
## 9538 samples
##    4 predictor
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 7630, 7630, 7631, 7630, 7631
## Resampling results:
##
##   RMSE       Rsquared   MAE
##   0.1456907  0.9059009  0.04161462
```

Overall, the evaluation results from both train-test split validation and 5-fold cross-validation indicate that our model is highly reliable and accurate.

Further Exploration

In our logistic regression model, FamilyHistory variable was found to be statistically insignificant in predicting the outcome of diabetes. Despite the well-documented association between family history and diabetes risk, the lack of significance in our model suggests that, based on our dataset, FamilyHistory alone may not be a strong independent predictor when accounting for other variables such as BMI, HbA1c, and Age.

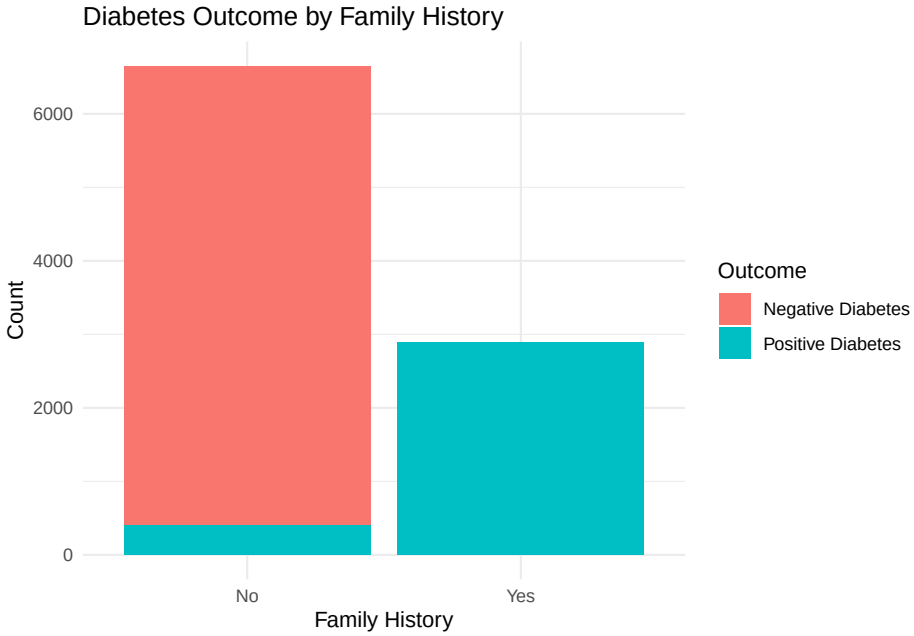
```
## corrplot 0.95 loaded
```



However, one of our key research questions is to determine whether family history modifies the relationship between age and diabetes risk. Given this objective, we felt it was important to further investigate this variable, even though it was not significant in the model. Family history is still widely recognized as a major risk factor diabetes, and we believe it remains clinically and practically relevant. To support this exploration, we incorporated visualizations to examine the relationship more closely.

4. Results

What proportion of diabetics have previous family history of the condition?



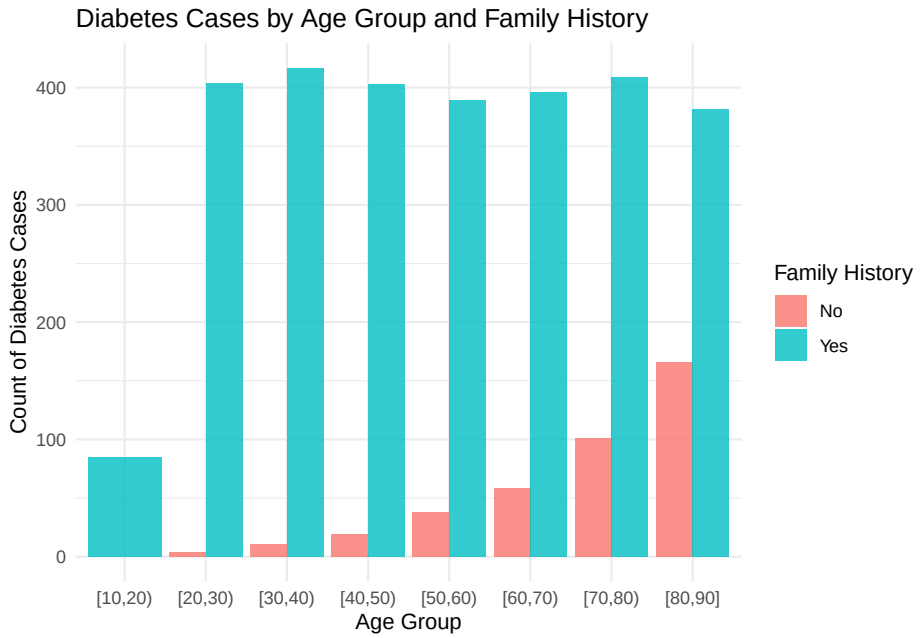
The bar chart above illustrates the distribution of diabetes outcomes based on family history. The chart displays two groups: individuals with no family history and those with a family history of diabetes. Within each group, the bars are divided by diabetes outcome, with red representing individuals without diabetes and blue representing those with diabetes.

From the visualization, it is evident that the no-family-history group is significantly larger. However, within

this group, the proportion of individuals diagnosed with diabetes is relatively small, as indicated by the thin blue section at the bottom of the bar. In contrast, the family-history group, despite being smaller in size, shows a higher proportion of diabetes case. The larger blue section in this group suggests that individuals with a family history are more likely to have diabetes.

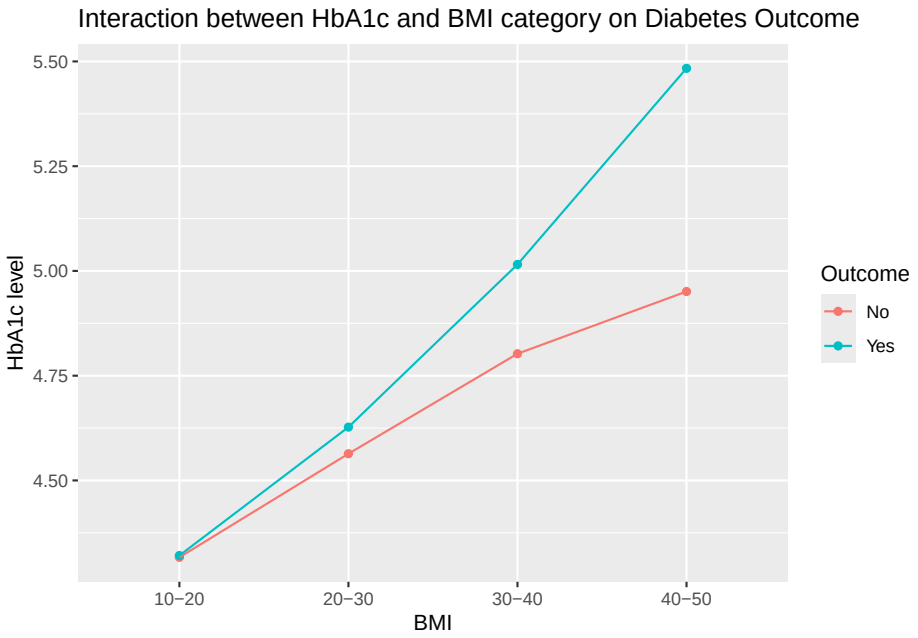
Even though family history was not statistically significant in our model, this visualization supports the notion that it may still play an important role in modifying diabetes risk, particularly when considered alongside age. The visual evidence suggests that individuals with a family history could be at greater risk, reinforcing the importance of further exploring potential interactions between family history, age, and other factors in future analyses.

Does family history modify the relationship between age and diabetes risk?



The bar chart above shows the distribution of diabetes individuals by age group, separated by family history. From the bar chart, we observed that the individuals with family history remain consistently high across all age groups. This indicates that individuals with a family history are at a higher risk of developing diabetes regardless of their age. In contrast, the individuals without family history show a distinct increase as age increases. While the number of diabetes cases without family history is relatively low in the younger age groups, it rises significantly in the older groups, particular from 50-60 onwards. This suggests that age becomes a stronger risk factor for diabetes among individuals without a family history.

Is there a combined effect of BMI and HBA1C on diabetes diagnosis?



The graph shows that on average individuals who do have diabetes usually have a higher BMI and HbA1c level. This suggests that those who have higher BMIs and higher HbA1c levels are more likely to be classified as having diabetes compared to those with lower BMI and lower HbA1c levels, which corresponds to what we see in real life. It is also worth noting that the discrepancy grows as BMI increases.

5. Limitations

Despite our rigorous methodology, this study has several limitations. One of the limitations is the possibility of missing confounders. Certain important risk factors, such as physical activity levels, dietary habits, and socioeconomic status, may not be fully accounted for in our dataset. Measurement errors also present a challenge, as inaccuracies in medical recordings could introduce inconsistencies. Variability in medical instruments can further affect the reliability of our results. Furthermore, while our stepwise regression approach identified the best subset of predictors, there is a risk of overfitting or model selection bias. The use of ANOVA helped validate the model choice, but it does not eliminate all potential biases.

6. Conclusion

In conclusion, our final model included the variables Age, BMI, HbA1c, and FamilyHistory as predictor variables of diabetes. We used model selection techniques such as stepwise regression, a comparison of full and reduced ANOVA, and an examination of VIF. We concluded that this final model included significant variables as well as variables we wanted to explore further with minimal collinearity. The variable Family History was the highest correlated variable with diabetes outcome, with a correlation coefficient of 0.85. However, it is important to note that our Family History variable includes selection bias amongst those who have a family history of diabetes. Every individual with a family history of diabetes in our study was positive for diabetes, leading to an inflated standard error, which may explain why the p-value for our Family History variable was so high. When we became aware of this large limitation in our dataset, we shifted our focus to individuals who did not have a family history of diabetes. Those who did not have a history of diabetes in their family had a 6% chance to test positive for diabetes. These cases did not occur until after the age of 20 and steadily increased up to the age of 90.

Our research aligns with real-world results. For example, older individuals have a higher risk of diabetes due

to increased insulin resistance, which leads to elevated glucose levels as the body struggles to produce enough hormones to regulate blood sugar. Having a high BMI produces a similar effect since visceral fat–commonly found around the abdomen–can produce hormones that interfere with insulin metabolism. Monitoring your glucose levels through HbA1c every one to three months provides a reliable measure of long-term blood sugar control while assessing diabetes management. Analyzing the relationships between features allows for early detection and prevention of diabetes by identifying high-risk groups. The factors that you have control over affecting diabetes may be limited, but it is also why you should take advantage to intervene early. By identifying such modifiable factors, strategies could be implemented to minimize risk.

In future investigation, we would want to explore more behind gender and diabetes complications such as glaucoma, heart disease, and kidney failure. Historically, diabetes is more commonly found in males whereas females face more serious risk of long-term health problems related to diabetes. This would also provide the opportunity to uncover the relationship between BMI and gender, helping to identify potential differences in diabetes risk factors. Understanding these gender-specific variations could improve targeted prevention and treatment strategies.